(19) DANMARK

(10) **DK/EP 4289948 T3**

Patent- og Varemærkestyrelsen

(12) Oversættelse af europæisk patentskrift

(51) Int.Cl.: ***C 12 N 15/11 (2006.01)*** ***A 61 K 38/46 (2006.01)*** ***C 12 N 9/22 (2006.01)***
C 12 N 15/10 (2006.01) ***C 12 N 15/113 (2010.01)*** ***C 12 N 15/63 (2006.01)***
C 12 N 15/90 (2006.01)

(45) Oversættelsen bekendtgjort den: **2025-04-28**

(80) Dato for Den Europæiske Patentmyndigheds bekendtgørelse om meddelelse af patentet: **2025-02-26**

(86) Europæisk ansøgning nr.: **23187511.3**

(86) Europæisk indleveringsdag: **2013-03-15**

(87) Den europæiske ansøgnings publiceringsdag: **2023-12-13**

(30) Prioritet: **2012-05-25 US 201261652086 P** **2012-10-19 US 201261716256 P**
2013-01-28 US 201361757640 P **2013-02-15 US 201361765576 P**

(62) Stamansøgningsnr: **19157590.1**

(84) Designerede stater: **AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR**

(73) Patenthaver: **The Regents of the University of California, 1111 Franklin Street, 12th Floor, Oakland, CA 94607, USA**
Universität Wien, Universitätsring 1, 1010 Vienna, Østrig
Charpentier, Emmanuelle, Max Planck Unit for the Science of Pathogens , Virchowweg 12, 10117 Berlin, Tyskland

(72) Opfinder: **CHARPENTIER, Emmanuelle, , 10117 Berlin, Tyskland**
JINEK, Martin, , Berkeley, 94709, USA
DOUDNA CATE, James Harrison, , Berkeley, 94705, USA
LIM, Wendell, , San Francisco, 94118, USA
QI, Lei, , Albany, 94706, USA
CHYLINSKI, Krzysztof, , 1110 Vienna, Østrig
DOUDNA, Jennifer, , Berkeley, 94705, USA

(74) Fuldmægtig i Danmark: **Novagraaf Brevets, Bâtiment O2, 2 rue Sarah Bernhardt CS90017, F-92665 Asnières-sur-Seine cedex, Frankrig**

(54) Benævnelse: **FREMGANGSMÅDER OG SAMMENSÆTNINGER TIL RNA-DIRIGERET MÅL-DNA-MODIFICERING OG TIL RNA-DIRIGERET MODULERING AF TRANSKRIPTION**

(56) Fremdragne publikationer:
WO-A1-2010/021692
WO-A2-2008/108989
WO-A2-2011/072246
PAPWORTH M ET AL: "Designer zinc-finger proteins and their applications", GENE, ELSEVIER, AMSTERDAM, NL, vol. 366, no. 1, 17 January 2006 (2006-01-17), pages 27 - 38, XP024934269, ISSN: 0378-1119, [retrieved on 20060117], DOI: 10.1016/J.GENE.2005.09.011

Fortsættes ...

JEFFREY C MILLER ET AL: "A TALE nuclease architecture for efficient genome editing", NATURE BIOTECHNOLOGY, vol. 29, no. 2, 22 December 2010 (2010-12-22), New York, pages 143 - 148, XP055568321, ISSN: 1087-0156, DOI: 10.1038/nbt.1755
ELITZA DELTCHEVA ET AL: "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III", NATURE, vol. 471, no. 7340, 31 March 2011 (2011-03-31), pages 602 - 607, XP055308803, ISSN: 0028-0836, DOI: 10.1038/nature09886
MAKAROVA KIRA S ET AL: "Evolution and classification of the CRISPR-Cas systems", NATURE REVIEWS. MICROBIO, NATURE PUBLISHING GROUP, GB, vol. 9, no. 6, 9 May 2011 (2011-05-09), pages 467 - 477, XP009155547, ISSN: 1740-1526, DOI: 10.1038/NRMICRO2577
BLAKE WIEDENHEFT ET AL: "RNA-guided genetic silencing systems in bacteria and archaea", NATURE, vol. 482, no. 7385, 15 February 2012 (2012-02-15), pages 331 - 338, XP055116249, ISSN: 0028-0836, DOI: 10.1038/nature10886
M. JINEK ET AL: "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity", SCIENCE, vol. 337, no. 6096, 17 August 2012 (2012-08-17), pages 816 - 821, XP055299674, ISSN: 0036-8075, DOI: 10.1126/science.1225829
M. JINEK ET AL: "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity (Supplementary Material)", SCIENCE, vol. 337, no. 6096, 28 June 2012 (2012-06-28), US, pages - 821, XP055067747, ISSN: 0036-8075, DOI: 10.1126/science.1225829

DESCRIPTION

Description

BACKGROUND

[0001] About 60% of bacteria and 90% of archaea possess CRISPR (clustered regularly interspaced short palindromic repeats)/CRISPR-associated (Cas) system systems to confer resistance to foreign DNA elements. Type II CRISPR system from *Streptococcus pyogenes* involves only a single gene encoding the Cas9 protein and two RNAs - a mature CRISPR RNA (crRNA) and a partially complementary trans-acting RNA (tracrRNA) - which are necessary and sufficient for RNA-guided silencing of foreign DNAs. Deltcheva (Nature 471, 2011) relates to CRSPR RNA maturation by trans-encoded small RNA and host factor RNase III.

[0002] In recent years, engineered nuclease enzymes designed to target specific DNA sequences have attracted considerable attention as powerful tools for the genetic manipulation of cells and whole organisms, allowing targeted gene deletion, replacement and repair, as well as the insertion of exogenous sequences (transgenes) into the genome. Two major technologies for engineering site-specific DNA nucleases have emerged, both of which are based on the construction of chimeric endonuclease enzymes in which a sequence non-specific DNA endonuclease domain is fused to an engineered DNA binding domain. However, targeting each new genomic locus requires the design of a novel nuclease enzyme, making these approaches both time consuming and costly. In addition, both technologies suffer from limited precision, which can lead to unpredictable off-target effects. WO 2011/072246 relates to gene targeting with transcription activator-like nucleases.

[0003] The systematic interrogation of genomes and genetic reprogramming of cells involves targeting sets of genes for expression or repression. Currently the most common approach for targeting arbitrary genes for regulation is to use RNA interference (RNAi). This approach has limitations. For example, RNAi can exhibit significant off-target effects and toxicity.

[0004] There is need in the field for a technology that allows precise targeting of nuclease activity (or other protein activities) to distinct locations within a target DNA in a manner that does not require the design of a new protein for each new target sequence. In addition, there is a need in the art for methods of controlling gene expression with minimal off-target effects.

SUMMARY

[0005] Any reference to genetic modification of cells does not encompass the germline modification of human beings. The compositions of the invention do not comprise human germ cells, human embryos or human embryonic germ cells. The methods of the invention do not comprise the use of human embryos of human embryonic germ cells.

[0006] The invention provides a single-molecule DNA-targeting RNA which binds to a site-directed modifying polypeptide and targets said site-directed modifying polypeptide to a specific location within a target DNA, wherein said single-molecule DNA-targeting RNA comprises:

1. a) a DNA-targeting segment comprising a nucleotide sequence that is complementary to a sequence in said target DNA, and
2. b) a protein-binding segment that interacts with said site-directed modifying polypeptide which is a naturally-occurring Cas9 endonuclease, wherein the protein-binding segment comprises two complementary stretches of nucleotides that hybridize to form a double stranded RNA (dsRNA) duplex, and

wherein said single-molecule DNA-targeting RNA, together with said site-directed modifying polypeptide which is a naturally-occurring Cas9 endonuclease, provides for site-specific cleavage of said target DNA to generate a double-stranded break.

[0007] The invention provides a single-molecule DNA-targeting RNA comprising:

1. a) a DNA-targeting segment comprising a nucleotide sequence that is complementary to a target sequence in a target DNA, and
2. b) a protein-binding segment that interacts with a site-directed modifying polypeptide which is a naturally-occurring Cas9 endonuclease, wherein the protein-binding segment comprises two complementary stretches of nucleotides that hybridize to form a double stranded RNA (dsRNA) duplex,

wherein said nucleotide sequence that is complementary to a sequence in said target DNA is at least about 15 nucleotides.

[0008] The invention provides a DNA polynucleotide comprising a nucleotide sequence that encodes the DNA-targeting RNA of any one of the above embodiments.

[0009] The invention provides recombinant expression vector comprising the above DNA polynucleotide.

[0010] The invention provides a DNA-targeting RNA comprising:

1. a) a DNA-targeting segment comprising a nucleotide sequence that is complementary to a target sequence in a target DNA, and
2. b) a protein-binding segment that interacts with a naturally-occurring Cas9 polypeptide, wherein the protein-binding segment comprises two complementary stretches of nucleotides that hybridize to form a double stranded RNA (dsRNA) duplex,

wherein the DNA-targeting RNA comprises nucleic acid modifications selected from the group consisting of modified backbones and modified internucleoside linkages, nucleic acid mimetics, modified sugar moieties, base modifications and substitutions, and conjugates.

[0011] The invention provides a complex which is formed by

1. (i) a single-molecule DNA-targeting RNA of any one of the above embodiments or a DNA-targeting RNA of any one of the above embodiments, and
2. (ii) a site-directed modifying polypeptide which is a naturally-occurring Cas9 endonuclease.

[0012] The invention provides a composition comprising:

1. (i) a single-molecule DNA-targeting RNA of any one of the above embodiments or a DNA-targeting RNA of any one of the above embodiments, and
2. (ii) a site-directed modifying polypeptide which is a naturally-occurring Cas9 endonuclease.

[0013] The present disclosure provides a DNA-targeting RNA that comprises a targeting sequence and, together with a modifying polypeptide, provides for site-specific modification of a target DNA and/or a polypeptide associated with the target DNA. Compositions are also provided.

[0014] Features of the present disclosure include a DNA-targeting RNA comprising: (i) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (ii) a second segment that interacts with a site-directed modifying polypeptide. In some cases, the first segment comprises 8 nucleotides that have 100% complementarity to a sequence in the target DNA. In some cases, the second segment comprises a nucleotide sequence with at least 60% identity over a stretch of at least 8 contiguous nucleotides to any one of the nucleotide sequences set forth in SEQ ID NOs:431-682 (e.g., 431-562). In some cases, the second segment comprises a nucleotide sequence with at least 60% identity over a stretch of at least 8 contiguous nucleotides to any one of the nucleotide sequences set forth in SEQ ID NOs:563-682. In some cases, the site-directed modifying polypeptide comprises an amino acid sequence having at least about 75% amino acid sequence identity to amino acids 7-166 or 731-1003 of the Cas9/Csn1 amino acid sequence depicted in Figure 3, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-256 and 795-1346.

[0015] Features of the present disclosure include a DNA polynucleotide comprising a nucleotide sequence that encodes the DNA-targeting RNA. In some cases, a recombinant expression vector comprises the DNA polynucleotide. In some cases, the nucleotide sequence

encoding the DNA-targeting RNA is operably linked to a promoter. In some cases, the promoter is an inducible promoter. In some cases, the nucleotide sequence encoding the DNA-targeting RNA further comprises a multiple cloning site.

[0016] Features of the present disclosure include a recombinant expression vector comprising: (i) a nucleotide sequence encoding a DNA-targeting RNA, wherein the DNA-targeting RNA comprises: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) a nucleotide sequence encoding the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the DNA-targeting RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the DNA-targeting RNA.

[0017] Features of the present disclosure include a composition comprising: (i) a DNA-targeting RNA, the DNA-targeting RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide; and (ii) the site-directed modifying polypeptide, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the DNA-targeting RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the DNA-targeting RNA and wherein the site-directed modifying polypeptide is a naturally-occurring Cas9 endonuclease. In some cases, the first segment of the DNA-targeting RNA comprises 8 nucleotides that have at least 100% complementarity to a sequence in the target DNA. In some cases, the second segment of the DNA-targeting RNA comprises a nucleotide sequence with at least 60% identity over a stretch of at least 8 contiguous nucleotides to any one of the nucleotide sequences set forth in SEQ ID NOs:431-682 (e.g., SEQ ID NOs:563-682). In some cases, the second segment of the DNA-targeting RNA comprises a nucleotide sequence with at least 60% identity over a stretch of at least 8 contiguous nucleotides to any one of the nucleotide sequences set forth in SEQ ID NOs:431-562. The enzymatic activity modifies the target DNA. Tthe enzymatic activity is nuclease activity. In some cases, the DNA-targeting RNA is a double-molecule DNA-targeting RNA and the composition comprises both a targeter-RNA and an activator-RNA, the duplex-forming segments of which are complementary and hybridize to form the second segment of the DNA-targeting RNA. In some cases, the duplex-forming segment of the activator-RNA comprises a nucleotide sequence with at least 60% identity over a stretch of at least 8 contiguous nucleotides to any one of the nucleotide sequences set forth in SEQ ID NO:SEQ ID NOs:431-682.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018]

Figures 1A-B provide a schematic drawing of two exemplary subject DNA-targeting RNAs, each associated with a site-directed modifying polypeptide and with a target DNA.

Figure 2 depicts target DNA editing through double-stranded DNA breaks introduced using a Cas9/Csn1 site-directed modifying polypeptide and a DNA-targeting RNA.

Figures 3A-B depict the amino acid sequence of a Cas9/Csn1 protein from *Streptococcus pyogenes* (SEQ ID NO:8). Cas9 has domains homologous to both HNH and RuvC endonucleases. (**A**) Motifs 1-4 are overlined (**B**) Domains 1 and 2 are overlined.

Figures 4A-B depict the percent identity between the Cas9/Csn1 proteins from multiple species. (A) Sequence identity relative to *Streptococcus pyogenes.* For Example, Domain 1 is amino acids 7-166 and Domain 2 is amino acids 731-1003 of Cas9/Csn1 from *Streptococcus pyogenes* as depicted in Figure 3B. (B) Sequence identity relative to *Neisseria meningitidis.* For example, Domain 1 is amino acids 13-139 and Domain 2 is amino acids 475-750 of Cas9/Csn1 from *Neisseria meningitidis* (SEQ ID NO:79).

Figure 5 depicts a multiple sequence alignment of motifs 1-4 of Cas9/Csn1 proteins from various diverse species selected from the phylogenetic table in Figure 32 (see Figure 32, Figure 3A, and Table 1) (*Streptococcus pyogenes* (SEQ ID NO:8), *Legionella pneumophila* (SEQ ID NO: 17), *Gamma proteobacterium* (SEQ ID NO: 107), *Listeria innocua* (SEQ ID NO:3), *Lactobacillus gasseri* (SEQ ID NO: 152), *Eubacterium rectale* (SEQ ID NO:99), *Staphylococcus lugdunensis* (SEQ ID NO: 185), *Mycoplasma synoviae* (SEQ ID NO:22), *Mycoplasma mobile* (SEQ ID NO: 16), *Wolinella succinogenes* (SEQ ID NO: 10), *Flavobacterium columnare* (SEQ ID NO:235), *Fibrobacter succinogenes* (SEQ ID NO: 121), *Bacteroides fragilis* (SEQ ID NO:21), *Acidothermus cellulolyticus* (SEQ ID NO:42), and *Bifidobacterium dentium* (SEQ ID NO:131).

Figures 6A-B provide alignments of naturally occurring tracrRNA ("activator-RNA") sequences from various species (L. innocua (SEQ ID NO:268); S. pyogenes (SEQ ID NO:267); S. mutans (SEQ ID NO:269); S. thermophilus1 (SEQ ID NO:270); M. mobile (SEQ ID NO:274); N. meningitides (SEQ ID NO:272); P. multocida (SEQ ID NO:273); S. thermophilus2 (SEQ ID NO:271); and S. pyogenes (SEQ ID NO:267). (A) multiple sequence alignment of selected tracrRNA orthologues (AlignX, VectorNTI package, Invitrogen) associated with CRISPR/Cas loci of similar architecture and highly similar Cas9/Csn1 sequences. Black boxes represent shared nucleotides (**B**) multiple sequence alignment of selected tracrRNA orthologues (AlignX, VectorNTI package, Invitrogen) associated with CRISPR/Cas loci of different architecture and non-closely related Cas9/Csn1 sequences. Note the sequence similarity of N. meningitidis and P. multocida tracrRNA orthologues. Black boxes represent shared nucleotides. For more exemplary activator-RNA sequences, see SEQ ID NOs:431-562.

Figures 7A-B provide alignments of naturally occurring duplex-forming segments of crRNA ("targeter-RNA") sequences from various species (L. innocua (SEQ ID NO://); S. pyogenes (SEQ ID NO://); S. mutans (SEQ ID NO://); S. thermophilus1 (SEQ ID NO://); C. jejuni (SEQ ID NO://); S. pyogenes (SEQ ID NO://); F. novicida (SEQ ID NO://); M. mobile (SEQ ID NO://); N. meningitides (SEQ ID NO://); P. multocida (SEQ ID NO://); and S. thermophilus2 (SEQ ID NO://). (A) multiple sequence alignments of exemplary duplex-forming segment of targeter-RNA sequences (AlignX, VectorNTI package, Invitrogen) associated with the loci of similar architecture and highly similar Cas9/Csn1 sequences. (**B**) multiple sequence alignments of

exemplary duplex-forming segment of targeter-RNA sequences (AlignX, VectorNTI package, Invitrogen) associated with the loci of different architecture and diverse Cas9 sequences. Black boxes represent shared nucleotides. For more exemplary duplex-forming segments targeter-RNA sequences, see SEQ ID NOs:563-679.

Figure 8 provides a schematic of hybridization for naturally occurring duplex-forming segments of the crRNA ("targeter-RNA") with the duplex-forming segment of the corresponding tracrRNA orthologue ("activator-RNA"). Upper sequence, targeter-RNA; lower sequence, duplex-forming segment of the corresponding activator-RNA. The CRISPR loci belong to the Type II (Nmeni/CASS4) CRISPR/Cas system. Nomenclature is according to the CRISPR database (CRISPR DB). S. pyogenes (SEQ ID NO:// and //); S. mutans (SEQ ID NO:// and //); S. thermophilus1 (SEQ ID NO:// and //); S. thermophilus2 (SEQ ID NO:// and //); L. innocua (SEQ ID NO:// and //); T. denticola (SEQ ID NO:// and //); N. meningitides (SEQ ID NO:// and //); S. gordonii (SEQ ID NO:// and //); B. bifidum (SEQ ID NO:// and //); L. salivarius (SEQ ID NO:// and //); F. tularensis (SEQ ID NO:// and //); and L. pneumophila (SEQ ID NO:// and //). Note that some species contain each two Type II CRISPR loci. For more exemplary activator-RNA sequences, see SEQ ID NOs:431-562. For more exemplary duplex-forming segments targeter-RNA sequences, see SEQ ID NOs:563-679.

Figure 9 depicts example tracrRNA (activator-RNA) and crRNA (targeter-RNA) sequences from two species. A degree of interchangeability exists; for example, the *S.pyogenes* Cas9/Csn1 protein is functional with tracrRNA and crRNA derived from *L.innocua.* (|) denotes a canonical Watson-Crick base pair while (•) denotes a G-U wobble base pair. "Variable 20nt" or "20nt" represents the DNA-targeting segment that is complementary to a target DNA (this region can be up to about 100nt in length). Also shown is the design of single-molecule DNA-targeting RNA that incorporates features of the targeter-RNA and the activator-RNA. (Cas9/Csn1 protein sequences from a wide variety of species are depicted in Figure 3 and set forth as SEQ ID NOs:1-256 and 795-1346) *Streptococcus pyogenes:* top to bottom: (SEQ ID NO://, //, //); *Listeria innocua:* top to bottom: (SEQ ID NO://, //, //). The sequences provided are non-limiting examples and are meant to illustrate how single-molecule DNA-targeting RNAs and two-molecule DNA-targeting RNAs can be designed based on naturally existing sequences from a wide variety of species. Various examples of suitable sequences from a wide variety of species are set forth as follows (Cas9 protein: SEQ ID NOs: 1-259; tracrRNAs: SEQ ID NOs:431-562, or the complements thereof; crRNAs: SEQ ID NOs:563-679, or the complements thereof; and example single-molecule DNA-targeting RNAs: SEQ ID NOs:680-682).

Figures 10A-E show that Cas9 is a DNA endonuclease guided by two RNA molecules. Figure E (top to bottom, SEQ ID NOs: 278-280, and //).

Figures 11A-B demonstrate that Cas9 uses two nuclease domains to cleave the two strands in the target DNA.

Figures 12A-E illustrate that Cas9-catalyzed cleavage of target DNA requires an activating domain in tracrRNA and is governed by a seed sequence in the crRNA. Figure 12C (top to bottom, SEQ ID NO:278-280, and //); Figure 12D (top to bottom, SEQ ID NOs: 281-290); and

Figure 12E (top to bottom, SEQ ID NOs: 291-292, 283, 293-298).

Figures 13A-C show that a PAM is required to license target DNA cleavage by the Cas9-tracrRNA:crRNA complex.

Figures 14A-C illustrate that Cas9 can be programmed using a single engineered RNA molecule combining tracrRNA and crRNA features. Chimera A (SEQ ID NO:299); Chimera B (SEQ ID NO:300).

Figure 15 depicts the type II RNA-mediated CRISPR/Cas immune pathway.

Figures 16A-B depict purification of Cas9 nucleases.

Figures 17A-C show that Cas9 guided by dual-tracrRNA:crRNA cleaves protospacer plasmid and oligonucleotide DNA. Figure 17B (top to bottom, SEQ ID NOs: 301-303, and //); and Figure 17C (top to bottom, SEQ ID NO:304-306, and //).

Figures 18A-B show that Cas9 is a Mg2+-dependent endonuclease with 3'-5' exonuclease activity.

Figures 19A-C illustrate that dual-tracrRNA:crRNA directed Cas9 cleavage of target DNA is site specific. Figure 19C (top to bottom, SEQ ID NOs: 307-309, //, 337-339, and //).

Figures 20A-B show that dual-tracrRNA:crRNA directed Cas9 cleavage of target DNA is fast and efficient.

Figures 21A-B show that the HNH and RuvC-like domains of Cas9 direct cleavage of the complementary and noncomplementary DNA strand, respectively.

Figure 22 demonstrates that tracrRNA is required for target DNA recognition.

Figures 23A-B show that a minimal region of tracrRNA is capable of guiding dualtracrRNA: crRNA directed cleavage of target DNA.

Figures 24A-D demonstrate that dual-tracrRNA:crRNA guided target DNA cleavage by Cas9 can be species specific.

Figures 25A-C show that a seed sequence in the crRNA governs dual tracrRNA:crRNA directed cleavage of target DNA by Cas9. Figure 25A: target DNA probe 1 (SEQ ID NO:310); spacer 4 crRNA (1-42) (SEQ ID NO:311); tracrRNA (15-89) (SEQ ID NO://). Figure 25B left panel (SEQ ID NO:310).

Figures 26A-C demonstrate that the PAM sequence is essential for protospacer plasmid DNA cleavage by Cas9-tracrRNA:crRNA and for Cas9-mediated plasmid DNA interference in bacterial cells. Figure 26B (top to bottom, SEQ ID NOs:312-314); and Figure 26C (top to bottom, SEQ ID NO:315-320).

Figures 27A-C show that Cas9 guided by a single chimeric RNA mimicking dual tracrRNA:crRNA cleaves protospacer DNA. Figure 27C (top to bottom, SEQ ID NO:321-324).

Figures 28A-D depict de novo design of chimeric RNAs targeting the Green Fluorescent Protein (GFP) gene sequence. Figure 28B (top to bottom, SEQ ID NOs:325-326). Figure 28C: GFP1 target sequence (SEQ ID NO:327); GFP2 target sequence (SEQ ID NO:328); GFP3 target sequence (SEQ ID NO:329); GFP4 target sequence (SEQ ID NO:330); GFP5 target sequence (SEQ ID NO:331); GFP1 chimeric RNA (SEQ ID NO:332); GFP2 chimeric RNA (SEQ ID NO:333); GFP3 chimeric RNA (SEQ ID NO:334); GFP4 chimeric RNA (SEQ ID NO:335); GFP5 chimeric RNA (SEQ ID NO:336).

Figures 29A-E demonstrate that co-expression of Cas9 and guide RNA in human cells generates double-strand DNA breaks at the target locus. Figure 29C (top to bottom, SEQ ID NO:425-428).

Figures 30A-B demonstrate that cell lysates contain active Cas9:sgRNA and support site-specific DNA cleavage.

Figures 31A-B demonstrate that 3' extension of sgRNA constructs enhances site-specific NHEJ-mediated mutagenesis. Figure 31A (top to bottom, SEQ ID NO:428-430).

Figures 32A-B depict a phylogenetic tree of representative Cas9 sequences from various organisms (A) as well as Cas9 locus architectures for the main groups of the tree (B).

Figures 33A-E depict the architecture of type II CRISPR-Cas from selected bacterial species.

Figures 34A-B depict tracrRNA and pre-crRNA co-processing in selected type II CRISPR Cas systems. Figure 34A (top to bottom, SEQ ID NO://,//,//,//,//,//,//,//); Figure 34B (top to bottom, SEQ ID NO://,//,//,//).

Figure 35 depicts a sequence alignment of tracrRNA orthologues demonstrating the diversity of tracrRNA sequences.

Figures 36A-F depict the expression of bacterial tracrRNA orthologues and crRNAs revealed by deep RNA sequencing.

Figures 37A-O list all tracrRNA orthologues and mature crRNAs retrieved by sequencing for the bacterial species studied, including coordinates (region of interest) and corresponding cDNA sequences (5' to 3').

Figures 38 A-B present a table of bacterial species containing type II CRISPR-Cas loci characterized by the presence of the signature gene *cas9.* These sequences were used for phylogenetic analyses.

Figures 39A-B demonstrate that artificial sequences that share roughly 50% identity with naturally occurring a tracrRNAs and crRNAs can function with Cas9 to cleave target DNA as long as the structure of the protein-binding domain of the DNA-targeting RNA is conserved.

DEFINITIONS - PART I

[0019] The terms "polynucleotide" and "nucleic acid," used interchangeably herein, refer to a polymeric form of nucleotides of any length, either ribonucleotides or deoxyribonucleotides. Thus, this term includes, but is not limited to, single-, double-, or multi-stranded DNA or RNA, genomic DNA, cDNA, DNA-RNA hybrids, or a polymer comprising purine and pyrimidine bases or other natural, chemically or biochemically modified, non-natural, or derivatized nucleotide bases. "Oligonucleotide" generally refers to polynucleotides of between about 5 and about 100 nucleotides of single- or double-stranded DNA. However, for the purposes of this disclosure, there is no upper limit to the length of an oligonucleotide. Oligonucleotides are also known as "oligomers" or "oligos" and may be isolated from genes, or chemically synthesized by methods known in the art. The terms "polynucleotide" and "nucleic acid" should be understood to include, as applicable to the embodiments being described, single-stranded (such as sense or antisense) and double-stranded polynucleotides.

[0020] A "stem-loop structure" refers to a nucleic acid having a secondary structure that includes a region of nucleotides which are known or predicted to form a double strand (step portion) that is linked on one side by a region of predominantly single-stranded nucleotides (loop portion). The terms "hairpin" and "fold-back" structures are also used herein to refer to stem-loop structures. Such structures are well known in the art and these terms are used consistently with their known meanings in the art. As is known in the art, a stem-loop structure does not require exact base-pairing. Thus, the stem may include one or more base mismatches. Alternatively, the base-pairing may be exact, i.e. not include any mismatches.

[0021] By "hybridizable" or "complementary" or "substantially complementary" it is meant that a nucleic acid (e.g. RNA) comprises a sequence of nucleotides that enables it to non-covalently bind, i.e. form Watson-Crick base pairs and/or G/U base pairs, "anneal", or "hybridize," to another nucleic acid in a sequence-specific, antiparallel, manner (i.e., a nucleic acid specifically binds to a complementary nucleic acid) under the appropriate in vitro and/or in vivo conditions of temperature and solution ionic strength. As is known in the art, standard Watson-Crick base-pairing includes: adenine (A) pairing with thymidine (T), adenine (A) pairing with uracil (U), and guanine (G) pairing with cytosine (C) [DNA, RNA]. In addition, it is also known in the art that for hybridization between two RNA molecules (e.g., dsRNA), guanine (G) base pairs with uracil (U). For example, G/U base-pairing is partially responsible for the degeneracy (i.e., redundancy) of the genetic code in the context of tRNA anti-codon base-pairing with codons in mRNA. In the context of this disclosure, a guanine (G) of a protein-binding segment (dsRNA duplex) of a subject DNA-targeting RNA molecule is considered complementary to a uracil (U), and vice versa. As such, when a G/U base-pair can be made at a given nucleotide position a protein-binding segment (dsRNA duplex) of a subject DNA-targeting RNA molecule, the position is not considered to be non-complementary, but is instead considered to be complementary.

[0022] Hybridization and washing conditions are well known and exemplified in Sambrook, J., Fritsch, E. F. and Maniatis, T. Molecular Cloning: A Laboratory Manual, Second Edition, Cold

Spring Harbor Laboratory Press, Cold Spring Harbor (1989), particularly Chapter 11 and Table 11.1 therein; and Sambrook, J. and Russell, W., Molecular Cloning: A Laboratory Manual, Third Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (2001). The conditions of temperature and ionic strength determine the "stringency" of the hybridization.

[0023] Hybridization requires that the two nucleic acids contain complementary sequences, although mismatches between bases are possible. The conditions appropriate for hybridization between two nucleic acids depend on the length of the nucleic acids and the degree of complementation, variables well known in the art. The greater the degree of complementation between two nucleotide sequences, the greater the value of the melting temperature (Tm) for hybrids of nucleic acids having those sequences. For hybridizations between nucleic acids with short stretches of complementarity (e.g. complementarity over 35 or less, 30 or less, 25 or less, 22 or less, 20 or less, or 18 or less nucleotides) the position of mismatches becomes important (see Sambrook et al., supra, 11.7-11.8). Typically, the length for a hybridizable nucleic acid is at least about 10 nucleotides. Illustrative minimum lengths for a hybridizable nucleic acid are: at least about 15 nucleotides; at least about 20 nucleotides; at least about 22 nucleotides; at least about 25 nucleotides; and at least about 30 nucleotides). Furthermore, the skilled artisan will recognize that the temperature and wash solution salt concentration may be adjusted as necessary according to factors such as length of the region of complementation and the degree of complementation.

[0024] It is understood in the art that the sequence of polynucleotide need not be 100% complementary to that of its target nucleic acid to be specifically hybridizable or hybridizable. Moreover, a polynucleotide may hybridize over one or more segments such that intervening or adjacent segments are not involved in the hybridization event (e.g., a loop structure or hairpin structure). A polynucleotide can comprise at least 70%, at least 80%, at least 90%, at least 95%, at least 99%, or 100% sequence complementarity to a target region within the target nucleic acid sequence to which they are targeted. For example, an antisense nucleic acid in which 18 of 20 nucleotides of the antisense compound are complementary to a target region, and would therefore specifically hybridize, would represent 90 percent complementarity. In this example, the remaining noncomplementary nucleotides may be clustered or interspersed with complementary nucleotides and need not be contiguous to each other or to complementary nucleotides. Percent complementarity between particular stretches of nucleic acid sequences within nucleic acids can be determined routinely using BLAST programs (basic local alignment search tools) and PowerBLAST programs known in the art (Altschul et al., J. Mol. Biol., 1990, 215, 403-410; Zhang and Madden, Genome Res., 1997, 7, 649-656) or by using the Gap program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, Madison Wis.), using default settings, which uses the algorithm of Smith and Waterman (Adv. Appl. Math., 1981, 2, 482-489).

[0025] The terms "peptide," "polypeptide," and "protein" are used interchangeably herein, and refer to a polymeric form of amino acids of any length, which can include coded and non-coded amino acids, chemically or biochemically modified or derivatized amino acids, and polypeptides having modified peptide backbones.

[0026] "Binding" as used herein (e.g. with reference to an RNA-binding domain of a polypeptide) refers to a non-covalent interaction between macromolecules (e.g., between a protein and a nucleic acid). While in a state of non-covalent interaction, the macromolecules are said to be "associated" or "interacting" or "binding" (e.g., when a molecule X is said to interact with a molecule Y, it is meant the molecule X binds to molecule Y in a non-covalent manner). Not all components of a binding interaction need be sequence-specific (e.g., contacts with phosphate residues in a DNA backbone), but some portions of a binding interaction may be sequence-specific. Binding interactions are generally characterized by a dissociation constant (Kd) of less than 10^{-6} M, less than 10^{-7} M, less than 10^{-8} M, less than 10^{-9} M, less than 10^{-10} M, less than 10^{-11} M, less than 10^{-12} M, less than 10^{-11} M, less than 10^{-14} M, or less than 10^{-15} M. "Affinity" refers to the strength of binding, increased binding affinity being correlated with a lower Kd.

[0027] By "binding domain" it is meant a protein domain that is able to bind non-covalently to another molecule. A binding domain can bind to, for example, a DNA molecule (a DNA-binding protein), an RNA molecule (an RNA-binding protein) and/or a protein molecule (a protein-binding protein). In the case of a protein domain-binding protein, it can bind to itself (to form homodimers, homotrimers, etc.) and/or it can bind to one or more molecules of a different protein or proteins.

[0028] The term "conservative amino acid substitution" refers to the interchangeability in proteins of amino acid residues having similar side chains. For example, a group of amino acids having aliphatic side chains consists of glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains consists of serine and threonine; a group of amino acids having amide containing side chains consisting of asparagine and glutamine; a group of amino acids having aromatic side chains consists of phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains consists of lysine, arginine, and histidine; a group of amino acids having acidic side chains consists of glutamate and aspartate; and a group of amino acids having sulfur containing side chains consists of cysteine and methionine. Exemplary conservative amino acid substitution groups are: valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, and asparagine-glutamine.

[0029] A polynucleotide or polypeptide has a certain percent "sequence identity" to another polynucleotide or polypeptide, meaning that, when aligned, that percentage of bases or amino acids are the same, and in the same relative position, when comparing the two sequences. Sequence identity can be determined in a number of different manners. To determine sequence identity, sequences can be aligned using various methods and computer programs (e.g., BLAST, T-COFFEE, MUSCLE, MAFFT, etc.), available over the world wide web at sites including ncbi.nlm.nili.gov/BLAST, ebi.ac.uk/Tools/msa/tcoffee/, ebi.ac.uk/Tools/msa/muscle/, mafft.cbrc.jp/alignment/software/. See, e.g., Altschul et al. (1990), J. Mol. Bioi. 215:403-10.

[0030] A DNA sequence that "encodes" a particular RNA is a DNA nucleic acid sequence that is transcribed into RNA. A DNA polynucleotide may encode an RNA (mRNA) that is translated into protein, or a DNA polynucleotide may encode an RNA that is not translated into protein (e.g. tRNA, rRNA, or a DNA-targeting RNA; also called "non-coding" RNA or "ncRNA").

[0031] A "protein coding sequence" or a sequence that encodes a particular protein or polypeptide, is a nucleic acid sequence that is transcribed into mRNA (in the case of DNA) and is translated (in the case of mRNA) into a polypeptide in vitro or in vivo when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a start codon at the 5' terminus (N-terminus) and a translation stop nonsense codon at the 3' terminus (C-terminus). A coding sequence can include, but is not limited to, cDNA from prokaryotic or eukaryotic mRNA, genomic DNA sequences from prokaryotic or eukaryotic DNA, and synthetic nucleic acids. A transcription termination sequence will usually be located 3' to the coding sequence.

[0032] As used herein, a "promoter sequence" is a DNA regulatory region capable of binding RNA polymerase and initiating transcription of a downstream (3' direction) coding or non-coding sequence. For purposes of defining the present invention, the promoter sequence is bounded at its 3' terminus by the transcription initiation site and extends upstream (5' direction) to include the minimum number of bases or elements necessary to initiate transcription at levels detectable above background. Within the promoter sequence will be found a transcription initiation site, as well as protein binding domains responsible for the binding of RNA polymerase. Eukaryotic promoters will often, but not always, contain "TATA" boxes and "CAT" boxes. Various promoters, including inducible promoters, may be used to drive the various vectors of the present invention.

[0033] A promoter can be a constitutively active promoter (i.e., a promoter that is constitutively in an active/"ON" state), it may be an inducible promoter (i.e., a promoter whose state, active/"ON" or inactive/"OFF", is controlled by an external stimulus, e.g., the presence of a particular temperature, compound, or protein.), it may be a spatially restricted promoter (i.e., transcriptional control element, enhancer, etc.)(e.g., tissue specific promoter, cell type specific promoter, etc.), and it may be a temporally restricted promoter (i.e., the promoter is in the "ON" state or "OFF" state during specific stages of embryonic development or during specific stages of a biological process, e.g., hair follicle cycle in mice).

[0034] Suitable promoters can be derived from viruses and can therefore be referred to as viral promoters, or they can be derived from any organism, including prokaryotic or eukaryotic organisms. Suitable promoters can be used to drive expression by any RNA polymerase (e.g., pol I, pol II, pol III). Exemplary promoters include, but are not limited to the SV40 early promoter, mouse mammary tumor virus long terminal repeat (LTR) promoter; adenovirus major late promoter (Ad MLP); a herpes simplex virus (HSV) promoter, a cytomegalovirus (CMV) promoter such as the CMV immediate early promoter region (CMVIE), a rous sarcoma virus (RSV) promoter, a human U6 small nuclear promoter (U6) (Miyagishi et al. , Nature Biotechnology 20, 497 - 500 (2002)), an enhanced U6 promoter (e.g., Xia et al., Nucleic Acids

Res. 2003 Sep 1;31(17)), a human H1 promoter (H1), and the like.

[0035] Examples of inducible promoters include, but are not limited toT7 RNA polymerase promoter, T3 RNA polymerase promoter, Isopropyl-beta-D-thiogalactopyranoside (IPTG)-regulated promoter, lactose induced promoter, heat shock promoter, Tetracycline-regulated promoter, Steroid-regulated promoter, Metal-regulated promoter, estrogen receptor-regulated promoter, etc. Inducible promoters can therefore be regulated by molecules including, but not limited to, doxycycline; RNA polymerase, e.g., T7 RNA polymerase; an estrogen receptor; an estrogen receptor fusion; etc.

[0036] In some embodiments, the promoter is a spatially restricted promoter (i.e., cell type specific promoter, tissue specific promoter, etc.) such that in a multi-cellular organism, the promoter is active (i.e., "ON") in a subset of specific cells. Spatially restricted promoters may also be referred to as enhancers, transcriptional control elements, control sequences, etc. Any convenient spatially restricted promoter may be used and the choice of suitable promoter (e.g., a brain specific promoter, a promoter that drives expression in a subset of neurons, a promoter that drives expression in the germline, a promoter that drives expression in the lungs, a promoter that drives expression in muscles, a promoter that drives expression in islet cells of the pancreas, etc.) will depend on the organism. For example, various spatially restricted promoters are known for plants, flies, worms, mammals, mice, etc. Thus, a spatially restricted promoter can be used to regulate the expression of a nucleic acid encoding a subject site-directed modifying polypeptide in a wide variety of different tissues and cell types, depending on the organism. Some spatially restricted promoters are also temporally restricted such that the promoter is in the "ON" state or "OFF" state during specific stages of embryonic development or during specific stages of a biological process (e.g., hair follicle cycle in mice).

[0037] For illustration purposes, examples of spatially restricted promoters include, but are not limited to, neuron-specific promoters, adipocyte-specific promoters, cardiomyocyte-specific promoters, smooth muscle-specific promoters, photoreceptor-specific promoters, etc. Neuron-specific spatially restricted promoters include, but are not limited to, a neuron-specific enolase (NSE) promoter (see, e.g., EMBL HSENO2, X51956); an aromatic amino acid decarboxylase (AADC) promoter; a neurofilament promoter (see, e.g., GenBank HUMNFL, L04147); a synapsin promoter (see, e.g., GenBank HUMSYNIB, M55301); a thy-1 promoter (see, e.g., Chen et al. (1987) Cell 51:7-19; and Llewellyn, et al. (2010) Nat. Med. 16(10):1161-1166); a serotonin receptor promoter (see, e.g., GenBank S62283); a tyrosine hydroxylase promoter (TH) (see, e.g., Oh et al. (2009) Gene Ther 16:437; Sasaoka et al. (1992) Mol. Brain Res. 16:274; Boundy et al. (1998) J. Neurosci. 18:9989; and Kaneda et al. (1991) Neuron 6:583-594); a GnRH promoter (see, e.g., Radovick et al. (1991) Proc. Natl. Acad. Sci. USA 88:3402-3406); an L7 promoter (see, e.g., Oberdick et al. (1990) Science 248:223-226); a DNMT promoter (see, e.g., Bartge et al. (1988) Proc. Natl. Acad. Sci. USA 85:3648-3652); an enkephalin promoter (see, e.g., Comb et al. (1988) EMBO J. 17:3793-3805); a myelin basic protein (MBP) promoter; a Ca2+-calmodulin-dependent protein kinase II-alpha (CamKIIα) promoter (see, e.g., Mayford et al. (1996) Proc. Natl. Acad. Sci. USA 93:13250; and Casanova et al. (2001) Genesis 31:37); a CMV enhancer/platelet-derived growth factor-β promoter (see,

e.g., Liu et al. (2004) Gene Therapy 11:52-60); and the like.

[0038] Adipocyte-specific spatially restricted promoters include, but are not limited to aP2 gene promoter/enhancer, e.g., a region from -5.4 kb to +21 bp of a human aP2 gene (see, e.g., Tozzo et al. (1997) Endocrinol. 138:1604; Ross et al. (1990) Proc. Natl. Acad. Sci. USA 87:9590; and Pavjani et al. (2005) Nat. Med. 11:797); a glucose transporter-4 (GLUT4) promoter (see, e.g., Knight et al. (2003) Proc. Natl. Acad. Sci. USA 100:14725); a fatty acid translocase (FAT/CD36) promoter (see, e.g., Kuriki et al. (2002) Biol. Pharm. Bull. 25:1476; and Sato et al. (2002) J. Biol. Chem. 277:15703); a stearoyl-CoA desaturase-1 (SCD1) promoter (Tabor et al. (1999) J. Biol. Chem. 274:20603); a leptin promoter (see, e.g., Mason et al. (1998) Endocrinol. 139:1013; and Chen et al. (1999) Biochem. Biophys. Res. Comm. 262:187); an adiponectin promoter (see, e.g., Kita et al. (2005) Biochem. Biophys. Res. Comm. 331:484; and Chakrabarti (2010) Endocrinol. 151:2408); an adipsin promoter (see, e.g., Platt et al. (1989) Proc. Natl. Acad. Sci. USA 86:7490); a resistin promoter (see, e.g., Seo et al. (2003) Molec. Endocrinol. 17:1522); and the like.

[0039] Cardiomyocyte-specific spatially restricted promoters include, but are not limited to control sequences derived from the following genes: myosin light chain-2, α-myosin heavy chain, AE3, cardiac troponin C, cardiac actin, and the like. Franz et al. (1997) Cardiovasc. Res. 35:560-566; Robbins et al. (1995) Ann. N.Y. Acad. Sci. 752:492-505; Linn et al. (1995) Circ. Res. 76:584-591; Parmacek et al. (1994) Mol. Cell. Biol. 14:1870-1885; Hunter et al. (1993) Hypertension 22:608-617; and Sartorelli et al. (1992) Proc. Natl. Acad. Sci. USA 89:4047-4051.

[0040] Smooth muscle-specific spatially restricted promoters include, but are not limited to an SM22α promoter (see, e.g., Akyürek et al. (2000) Mol. Med. 6:983; and U.S. Patent No. 7,169,874); a smoothelin promoter (see, e.g., WO 2001/018048); an α-smooth muscle actin promoter; and the like. For example, a 0.4 kb region of the SM22α promoter, within which lie two CArG elements, has been shown to mediate vascular smooth muscle cell-specific expression (see, e.g., Kim, et al. (1997) Mol. Cell. Biol. 17, 2266-2278; Li, et al., (1996) J. Cell Biol. 132, 849-859; and Moessler, et al. (1996) Development 122, 2415-2425).

[0041] Photoreceptor-specific spatially restricted promoters include, but are not limited to, a rhodopsin promoter; a rhodopsin kinase promoter (Young et al. (2003) Ophthalmol. Vis. Sci. 44:4076); a beta phosphodiesterase gene promoter (Nicoud et al. (2007) J. Gene Med. 9:1015); a retinitis pigmentosa gene promoter (Nicoud et al. (2007) supra); an interphotoreceptor retinoid-binding protein (IRBP) gene enhancer (Nicoud et al. (2007) supra); an IRBP gene promoter (Yokoyama et al. (1992) Exp Eye Res. 55:225); and the like.

[0042] The terms "DNA regulatory sequences," "control elements," and "regulatory elements," used interchangeably herein, refer to transcriptional and translational control sequences, such as promoters, enhancers, polyadenylation signals, terminators, protein degradation signals, and the like, that provide for and/or regulate transcription of a non-coding sequence (e.g., DNA-targeting RNA) or a coding sequence (e.g., site-directed modifying polypeptide, or Cas9/Csn1 polypeptide) and/or regulate translation of an encoded polypeptide.

[0043] The term "naturally-occurring" or "unmodified" as used herein as applied to a nucleic acid, a polypeptide, a cell, or an organism, refers to a nucleic acid, polypeptide, cell, or organism that is found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by a human in the laboratory is naturally occurring.

[0044] The term "chimeric" as used herein as applied to a nucleic acid or polypeptide refers to two components that are defined by structures derived from different sources. For example, where "chimeric" is used in the context of a chimeric polypeptide (e.g., a chimeric Cas9/Csn1 protein), the chimeric polypeptide includes amino acid sequences that are derived from different polypeptides. A chimeric polypeptide may comprise either modified or naturally-occurring polypeptide sequences (e.g., a first amino acid sequence from a modified or unmodified Cas9/Csn1 protein; and a second amino acid sequence other than the Cas9/Csn1 protein). Similarly, "chimeric" in the context of a polynucleotide encoding a chimeric polypeptide includes nucleotide sequences derived from different coding regions (e.g., a first nucleotide sequence encoding a modified or unmodified Cas9/Csn1 protein; and a second nucleotide sequence encoding a polypeptide other than a Cas9/Csn1 protein).

[0045] The term "chimeric polypeptide" refers to a polypeptide which is made by the combination (i.e., "fusion") of two otherwise separated segments of amino sequence, usually through human intervention. A polypeptide that comprises a chimeric amino acid sequence is a chimeric polypeptide. Some chimeric polypeptides can be referred to as "fusion variants."

[0046] "Heterologous," as used herein, means a nucleotide or polypeptide sequence that is not found in the native nucleic acid or protein, respectively. For example, in a chimeric Cas9/Csn1 protein, the RNA-binding domain of a naturally-occurring bacterial Cas9/Csn1 polypeptide (or a variant thereof) may be fused to a heterologous polypeptide sequence (i.e. a polypeptide sequence from a protein other than Cas9/Csn1 or a polypeptide sequence from another organism). The heterologous polypeptide sequence may exhibit an activity (e.g., enzymatic activity) that will also be exhibited by the chimeric Cas9/Csn1 protein (e.g., methyltransferase activity, acetyltransferase activity, kinase activity, ubiquitinating activity, etc.). A heterologous nucleic acid sequence may be linked to a naturally-occurring nucleic acid sequence (or a variant thereof) (e.g., by genetic engineering) to generate a chimeric nucleotide sequence encoding a chimeric polypeptide. As another example, in a fusion variant Cas9 site-directed polypeptide, a variant Cas9 site-directed polypeptide may be fused to a heterologous polypeptide (i.e. a polypeptide other than Cas9), which exhibits an activity that will also be exhibited by the fusion variant Cas9 site-directed polypeptide. A heterologous nucleic acid sequence may be linked to a variant Cas9 site-directed polypeptide (e.g., by genetic engineering) to generate a nucleotide sequence encoding a fusion variant Cas9 site-directed polypeptide.

[0047] "Recombinant," as used herein, means that a particular nucleic acid (DNA or RNA) is the product of various combinations of cloning, restriction, polymerase chain reaction (PCR)

and/or ligation steps resulting in a construct having a structural coding or non-coding sequence distinguishable from endogenous nucleic acids found in natural systems. DNA sequences encoding polypeptides can be assembled from cDNA fragments or from a series of synthetic oligonucleotides, to provide a synthetic nucleic acid which is capable of being expressed from a recombinant transcriptional unit contained in a cell or in a cell-free transcription and translation system. Genomic DNA comprising the relevant sequences can also be used in the formation of a recombinant gene or transcriptional unit. Sequences of non-translated DNA may be present 5' or 3' from the open reading frame, where such sequences do not interfere with manipulation or expression of the coding regions, and may indeed act to modulate production of a desired product by various mechanisms (see "DNA regulatory sequences", below). Alternatively, DNA sequences encoding RNA (e.g., DNA-targeting RNA) that is not translated may also be considered recombinant. Thus, e.g., the term "recombinant" nucleic acid refers to one which is not naturally occurring, e.g., is made by the artificial combination of two otherwise separated segments of sequence through human intervention. This artificial combination is often accomplished by either chemical synthesis means, or by the artificial manipulation of isolated segments of nucleic acids, e.g., by genetic engineering techniques. Such is usually done to replace a codon with a codon encoding the same amino acid, a conservative amino acid, or a non-conservative amino acid. Alternatively, it is performed to join together nucleic acid segments of desired functions to generate a desired combination of functions. This artificial combination is often accomplished by either chemical synthesis means, or by the artificial manipulation of isolated segments of nucleic acids, e.g., by genetic engineering techniques. When a recombinant polynucleotide encodes a polypeptide, the sequence of the encoded polypeptide can be naturally occurring ("wild type") or can be a variant (e.g., a mutant) of the naturally occurring sequence. Thus, the term "recombinant" polypeptide does not necessarily refer to a polypeptide whose sequence does not naturally occur. Instead, a "recombinant" polypeptide is encoded by a recombinant DNA sequence, but the sequence of the polypeptide can be naturally occurring ("wild type") or non-naturally occurring (e.g., a variant, a mutant, etc.). Thus, a "recombinant" polypeptide is the result of human intervention, but may be a naturally occurring amino acid sequence.

[0048] A "vector" or "expression vector" is a replicon, such as plasmid, phage, virus, or cosmid, to which another DNA segment, i.e. an "insert", may be attached so as to bring about the replication of the attached segment in a cell.

[0049] An "expression cassette" comprises a DNA coding sequence operably linked to a promoter. "Operably linked" refers to a juxtaposition wherein the components so described are in a relationship permitting them to function in their intended manner. For instance, a promoter is operably linked to a coding sequence if the promoter affects its transcription or expression.

[0050] The terms "recombinant expression vector," or "DNA construct" are used interchangeably herein to refer to a DNA molecule comprising a vector and at least one insert. Recombinant expression vectors are usually generated for the purpose of expressing and/or propagating the insert(s), or for the construction of other recombinant nucleotide sequences. The insert(s) may or may not be operably linked to a promoter sequence and may or may not

be operably linked to DNA regulatory sequences.

[0051] A cell has been "genetically modified" or "transformed" or "transfected" by exogenous DNA, e.g. a recombinant expression vector, when such DNA has been introduced inside the cell. The presence of the exogenous DNA results in permanent or transient genetic change. The transforming DNA may or may not be integrated (covalently linked) into the genome of the cell. In prokaryotes, yeast, and mammalian cells for example, the transforming DNA may be maintained on an episomal element such as a plasmid. With respect to eukaryotic cells, a stably transformed cell is one in which the transforming DNA has become integrated into a chromosome so that it is inherited by daughter cells through chromosome replication. This stability is demonstrated by the ability of the eukaryotic cell to establish cell lines or clones that comprise a population of daughter cells containing the transforming DNA. A "clone" is a population of cells derived from a single cell or common ancestor by mitosis. A "cell line" is a clone of a primary cell that is capable of stable growth in vitro for many generations.

[0052] Suitable methods of genetic modification (also referred to as "transformation") include e.g., viral or bacteriophage infection, transfection, conjugation, protoplast fusion, lipofection, electroporation, calcium phosphate precipitation, polyethyleneimine (PEI)-mediated transfection, DEAE-dextran mediated transfection, liposome-mediated transfection, particle gun technology, calcium phosphate precipitation, direct micro injection, nanoparticle-mediated nucleic acid delivery (see, e.g., Panyam et., al Adv Drug Deliv Rev. 2012 Sep 13. pii: S0169-409X(12)00283-9. doi: 10.1016/j.addr.2012.09.023), and the like.

[0053] The choice of method of genetic modification is generally dependent on the type of cell being transformed and the circumstances under which the transformation is taking place (e.g., in vitro, ex vivo, or in vivo). A general discussion of these methods can be found in Ausubel, et al., Short Protocols in Molecular Biology, 3rd ed., Wiley & Sons, 1995.

[0054] A "target DNA" as used herein is a DNA polynucleotide that comprises a "target site" or "target sequence." The terms "target site" or "target sequence" or "target protospacer DNA" are used interchangeably herein to refer to a nucleic acid sequence present in a target DNA to which a DNA-targeting segment of a subject DNA-targeting RNA will bind (see Figure 1 and Figure 39), provided sufficient conditions for binding exist. For example, the target site (or target sequence) 5'-GAGCATATC-3' (SEQ ID NO://) within a target DNA is targeted by (or is bound by, or hybridizes with, or is complementary to) the RNA sequence 5'-GAUAUGCUC-3' (SEQ ID NO://). Suitable DNA/RNA binding conditions include physiological conditions normally present in a cell. Other suitable DNA/RNA binding conditions (e.g., conditions in a cell-free system) are known in the art; see, e.g., Sambrook, supra. The strand of the target DNA that is complementary to and hybridizes with the DNA-targeting RNA is referred to as the "complementary strand" and the strand of the target DNA that is complementary to the "complementary strand" (and is therefore not complementary to the DNA-targeting RNA) is referred to as the "noncomplementary strand" or "non-complementary strand" (see Figure 12).

[0055] By "site-directed modifying polypeptide" or "RNA-binding site-directed polypeptide" or

"RNA-binding site-directed modifying polypeptide" or "site-directed polypeptide" it is meant a polypeptide that binds RNA and is targeted to a specific DNA sequence. A site-directed modifying polypeptide as described herein is targeted to a specific DNA sequence by the RNA molecule to which it is bound. The RNA molecule comprises a sequence that is complementary to a target sequence within the target DNA, thus targeting the bound polypeptide to a specific location within the target DNA (the target sequence).

[0056] By "cleavage" it is meant the breakage of the covalent backbone of a DNA molecule. Cleavage can be initiated by a variety of methods including, but not limited to, enzymatic or chemical hydrolysis of a phosphodiester bond. Both single-stranded cleavage and double-stranded cleavage are possible, and double-stranded cleavage can occur as a result of two distinct single-stranded cleavage events. DNA cleavage can result in the production of either blunt ends or staggered ends. In certain embodiments, a complex comprising a DNA-targeting RNA and a site-directed modifying polypeptide is used for targeted double-stranded DNA cleavage.

[0057] "Nuclease" and "endonuclease" are used interchangeably herein to mean an enzyme which possesses catalytic activity for DNA cleavage.

[0058] By "cleavage domain" or "active domain" or "nuclease domain" of a nuclease it is meant the polypeptide sequence or domain within the nuclease which possesses the catalytic activity for DNA cleavage. A cleavage domain can be contained in a single polypeptide chain or cleavage activity can result from the association of two (or more) polypeptides. A single nuclease domain may consist of more than one isolated stretch of amino acids within a given polypeptide.

[0059] The RNA molecule that binds to the site-directed modifying polypeptide and targets the polypeptide to a specific location within the target DNA is referred to herein as the "DNA-targeting RNA" or "DNA-targeting RNA polynucleotide" (also referred to herein as a "guide RNA" or "gRNA"). A subject DNA-targeting RNA comprises two segments, a "DNA-targeting segment" and a "protein-binding segment." By "segment" it is meant a segment/section/region of a molecule, e.g., a contiguous stretch of nucleotides in an RNA. A segment can also mean a region/section of a complex such that a segment may comprise regions of more than one molecule. For example, in some cases the protein-binding segment (described below) of a DNA-targeting RNA is one RNA molecule and the protein-binding segment therefore comprises a region of that RNA molecule. In other cases, the protein-binding segment (described below) of a DNA-targeting RNA comprises two separate molecules that are hybridized along a region of complementarity. As an illustrative, non-limiting example, a protein-binding segment of a DNA-targeting RNA that comprises two separate molecules can comprise (i) base pairs 40-75 of a first RNA molecule that is 100 base pairs in length; and (ii) base pairs 10-25 of a second RNA molecule that is 50 base pairs in length. The definition of "segment," unless otherwise specifically defined in a particular context, is not limited to a specific number of total base pairs, is not limited to any particular number of base pairs from a given RNA molecule, is not limited to a particular number of separate molecules within a complex, and may include regions of

RNA molecules that are of any total length and may or may not include regions with complementarity to other molecules.

[0060] The DNA-targeting segment (or "DNA-targeting sequence") comprises a nucleotide sequence that is complementary to a specific sequence within a target DNA (the complementary strand of the target DNA). The protein-binding segment (or "protein-binding sequence") interacts with a site-directed modifying polypeptide. When the site-directed modifying polypeptide is a Cas9 or Cas9 related polypeptide (described in more detail below), site-specific cleavage of the target DNA occurs at locations determined by both (i) base-pairing complementarity between the DNA-targeting RNA and the target DNA; and (ii) a short motif (referred to as the protospacer adjacent motif (PAM)) in the target DNA.

[0061] The protein-binding segment of a subject DNA-targeting RNA comprises two complementary stretches of nucleotides that hybridize to one another to form a double stranded RNA duplex (dsRNA duplex).

[0062] In some embodiments, a subject nucleic acid (e.g., a DNA-targeting RNA, a nucleic acid comprising a nucleotide sequence encoding a DNA-targeting RNA; a nucleic acid encoding a site-directed polypeptide; etc.) comprises a modification or sequence that provides for an additional desirable feature (e.g., modified or regulated stability; subcellular targeting; tracking, e.g., a fluorescent label; a binding site for a protein or protein complex; etc.). Non-limiting examples include: a 5' cap (e.g., a 7-methylguanylate cap (m7G)); a 3' polyadenylated tail (i.e., a 3' poly(A) tail); a riboswitch sequence (e.g., to allow for regulated stability and/or regulated accessibility by proteins and/or protein complexes); a stability control sequence; a sequence that forms a dsRNA duplex (i.e., a hairpin)); a modification or sequence that targets the RNA to a subcellular location (e.g., nucleus, mitochondria, chloroplasts, and the like); a modification or sequence that provides for tracking (e.g., direct conjugation to a fluorescent molecule, conjugation to a moiety that facilitates fluorescent detection, a sequence that allows for fluorescent detection, etc.); a modification or sequence that provides a binding site for proteins (e.g., proteins that act on DNA, including transcriptional activators, transcriptional repressors, DNA methyltransferases, DNA demethylases, histone acetyltransferases, histone deacetylases, and the like); and combinations thereof.

[0063] In some embodiments, a DNA-targeting RNA comprises an additional segment at either the 5' or 3' end that provides for any of the features described above. For example, a suitable third segment can comprise a 5' cap (e.g., a 7-methylguanylate cap (m7G)); a 3' polyadenylated tail (i.e., a 3' poly(A) tail); a riboswitch sequence (e.g., to allow for regulated stability and/or regulated accessibility by proteins and protein complexes); a stability control sequence; a sequence that forms a dsRNA duplex (i.e., a hairpin)); a sequence that targets the RNA to a subcellular location (e.g., nucleus, mitochondria, chloroplasts, and the like); a modification or sequence that provides for tracking (e.g., direct conjugation to a fluorescent molecule, conjugation to a moiety that facilitates fluorescent detection, a sequence that allows for fluorescent detection, etc.); a modification or sequence that provides a binding site for proteins (e.g., proteins that act on DNA, including transcriptional activators, transcriptional

repressors, DNA methyltransferases, DNA demethylases, histone acetyltransferases, histone deacetylases, and the like); and combinations thereof.

[0064] A subject DNA-targeting RNA and a subject site-directed modifying polypeptide (i.e., site-directed polypeptide) form a complex (i.e., bind via non-covalent interactions). The DNA-targeting RNA provides target specificity to the complex by comprising a nucleotide sequence that is complementary to a sequence of a target DNA. The site-directed modifying polypeptide of the complex provides the site-specific activity. In other words, the site-directed modifying polypeptide is guided to a target DNA sequence (e.g. a target sequence in a chromosomal nucleic acid; a target sequence in an extrachromosomal nucleic acid, e.g. an episomal nucleic acid, a minicircle, etc.; a target sequence in a mitochondrial nucleic acid; a target sequence in a chloroplast nucleic acid; a target sequence in a plasmid; etc.) by virtue of its association with the protein-binding segment of the DNA-targeting RNA.

[0065] In some embodiments, a subject DNA-targeting RNA comprises two separate RNA molecules (RNA polynucleotides: an "activator-RNA" and a "targeter-RNA", see below) and is referred to herein as a "double-molecule DNA-targeting RNA" or a "two-molecule DNA-targeting RNA." In other embodiments, the subject DNA-targeting RNA is a single RNA molecule (single RNA polynucleotide) and is referred to herein as a "single-molecule DNA-targeting RNA," a "single-guide RNA," or an "sgRNA." The term "DNA-targeting RNA" or "gRNA" is inclusive, referring both to double-molecule DNA-targeting RNAs and to single-molecule DNA-targeting RNAs (i.e., sgRNAs).

[0066] An exemplary two-molecule DNA-targeting RNA comprises a crRNA-like ("CRISPR RNA" or "targeter-RNA" or "crRNA" or "crRNA repeat") molecule and a corresponding tracrRNA-like ("trans-acting CRISPR RNA" or "activator-RNA" or "tracrRNA") molecule. A crRNA-like molecule (targeter-RNA) comprises both the DNA-targeting segment (single stranded) of the DNA-targeting RNA and a stretch ("duplex-forming segment") of nucleotides that forms one half of the dsRNA duplex of the protein-binding segment of the DNA-targeting RNA. A corresponding tracrRNA-like molecule (activator-RNA) comprises a stretch of nucleotides (duplex-forming segment) that forms the other half of the dsRNA duplex of the protein-binding segment of the DNA-targeting RNA. In other words, a stretch of nucleotides of a crRNA-like molecule are complementary to and hybridize with a stretch of nucleotides of a tracrRNA-like molecule to form the dsRNA duplex of the protein-binding domain of the DNA-targeting RNA. As such, each crRNA-like molecule can be said to have a corresponding tracrRNA-like molecule. The crRNA-like molecule additionally provides the single stranded DNA-targeting segment. Thus, a crRNA-like and a tracrRNA-like molecule (as a corresponding pair) hybridize to form a DNA-targeting RNA. The exact sequence of a given crRNA or tracrRNA molecule is characteristic of the species in which the RNA molecules are found. Various crRNAs and tracrRNAs are depicted in corresponding complementary pairs in Figures 8. A subject double-molecule DNA-targeting RNA can comprise any corresponding crRNA and tracrRNA pair. A subject double-molecule DNA-targeting RNA can comprise any corresponding crRNA and tracrRNA pair.

[0067] The term "activator-RNA" is used herein to mean a tracrRNA-like molecule of a double-molecule DNA-targeting RNA. The term "targeter-RNA" is used herein to mean a crRNA-like molecule of a double-molecule DNA-targeting RNA. The term "duplex-forming segment" is used herein to mean the stretch of nucleotides of an activator-RNA or a targeter-RNA that contributes to the formation of the dsRNA duplex by hybridizing to a stretch of nucleotides of a corresponding activator-RNA or targeter-RNA molecule. In other words, an activator-RNA comprises a duplex-forming segment that is complementary to the duplex-forming segment of the corresponding targeter-RNA. As such, an activator-RNA comprises a duplex-forming segment while a targeter-RNA comprises both a duplex-forming segment and the DNA-targeting segment of the DNA-targeting RNA. Therefore, a subject double-molecule DNA-targeting RNA can be comprised of any corresponding activator-RNA and targeter-RNA pair.

[0068] The term "stem cell" is used herein to refer to a cell (e.g., plant stem cell, vertebrate stem cell) that has the ability both to self-renew and to generate a differentiated cell type (see Morrison et al. (1997) Cell 88:287-298). In the context of cell ontogeny, the adjective "differentiated", or "differentiating" is a relative term. A "differentiated cell" is a cell that has progressed further down the developmental pathway than the cell it is being compared with. Thus, pluripotent stem cells (described below) can differentiate into lineage-restricted progenitor cells (e.g., mesodermal stem cells), which in turn can differentiate into cells that are further restricted (e.g., neuron progenitors), which can differentiate into end-stage cells (i.e., terminally differentiated cells, e.g., neurons, cardiomyocytes, etc.), which play a characteristic role in a certain tissue type, and may or may not retain the capacity to proliferate further. Stem cells may be characterized by both the presence of specific markers (e.g., proteins, RNAs, etc.) and the absence of specific markers. Stem cells may also be identified by functional assays both in vitro and in vivo, particularly assays relating to the ability of stem cells to give rise to multiple differentiated progeny.

[0069] Stem cells of interest include pluripotent stem cells (PSCs). The term "pluripotent stem cell" or "PSC" is used herein to mean a stem cell capable of producing all cell types of the organism. Therefore, a PSC can give rise to cells of all germ layers of the organism (e.g., the endoderm, mesoderm, and ectoderm of a vertebrate). Pluripotent cells are capable of forming teratomas and of contributing to ectoderm, mesoderm, or endoderm tissues in a living organism. Pluripotent stem cells of plants are capable of giving rise to all cell types of the plant (e.g., cells of the root, stem, leaves, etc.).

[0070] PSCs of animals can be derived in a number of different ways. For example, embryonic stem cells (ESCs) are derived from the inner cell mass of an embryo whereas induced pluripotent stem cells (iPSCs) are derived from somatic cells (Takahashi et. al, Cell. 2007 Nov 30;131(5):861-72; Takahashi et. al, Nat Protoc. 2007;2(12):3081-9; Yu et. al, Science. 2007 Dec 21;318(5858):1917-20. Epub 2007 Nov 20). Because the term PSC refers to pluripotent stem cells regardless of their derivation, the term PSC encompasses the terms ESC and iPSC. PSCs may be in the form of an established cell line, they may be obtained directly from primary embryonic tissue, or they may be derived from a somatic cell. PSCs can be target cells of the methods described herein.

[0071] By "embryonic stem cell" (ESC) is meant a PSC that was isolated from an embryo, typically from the inner cell mass of the blastocyst. Stem cells of interest also include embryonic stem cells from other primates, such as Rhesus stem cells and marmoset stem cells. The stem cells may be obtained from any mammalian species, e.g. human, equine, bovine, porcine, canine, feline, rodent, e.g. mice, rats, hamster, primate, etc. (Thomson et al. (1998) Science 282:1145; Thomson et al. (1995) Proc. Natl. Acad. Sci USA 92:7844; Thomson et al. (1996) Biol. Reprod. 55:254; Shamblott et al., Proc. Natl. Acad. Sci. USA 95:13726, 1998). In culture, ESCs typically grow as flat colonies with large nucleo-cytoplasmic ratios, defined borders and prominent nucleoli. In addition, ESCs express SSEA-3, SSEA-4, TRA-1-60, TRA-1-81, and Alkaline Phosphatase, but not SSEA-1. Examples of methods of generating and characterizing ESCs may be found in, for example, US Patent No. 7,029,913, US Patent No. 5,843,780, and US Patent No. 6,200,806. Methods for proliferating hESCs in the undifferentiated form are described in WO 99/20741, WO 01/51616, and WO 03/020920.

[0072] By "embryonic germ stem cell" (EGSC) or "embryonic germ cell" or "EG cell" is meant a PSC that is derived from germ cells and/or germ cell progenitors, e.g. primordial germ cells, i.e. those that would become sperm and eggs. Embryonic germ cells (EG cells) are thought to have properties similar to embryonic stem cells as described above. Examples of methods of generating and characterizing EG cells may be found in, for example, US Patent No. 7,153,684; Matsui, Y., et al., (1992) Cell 70:841; Shamblott, M., et al. (2001) Proc. Natl. Acad. Sci. USA 98: 113; Shamblott, M., et al. (1998) Proc. Natl. Acad. Sci. USA, 95:13726; and Koshimizu, U., et al. (1996) Development, 122:1235.

[0073] By "induced pluripotent stem cell" or "iPSC" it is meant a PSC that is derived from a cell that is not a PSC (i.e., from a cell this is differentiated relative to a PSC). iPSCs can be derived from multiple different cell types, including terminally differentiated cells. iPSCs have an ES cell-like morphology, growing as flat colonies with large nucleo-cytoplasmic ratios, defined borders and prominent nuclei. In addition, iPSCs express one or more key pluripotency markers known by one of ordinary skill in the art, including but not limited to Alkaline Phosphatase, SSEA3, SSEA4, Sox2, Oct3/4, Nanog, TRA160, TRA181, TDGF 1, Dnmt3b, FoxD3, GDF3, Cyp26a1, TERT, and zfp42. Examples of methods of generating and characterizing iPSCs may be found in, for example, U.S. Patent Publication Nos. US20090047263, US20090068742, US20090191159, US20090227032, US20090246875, and US20090304646. Generally, to generate iPSCs, somatic cells are provided with reprogramming factors (e.g. Oct4, SOX2, KLF4, MYC, Nanog, Lin28, etc.) known in the art to reprogram the somatic cells to become pluripotent stem cells.

[0074] By "somatic cell" it is meant any cell in an organism that, in the absence of experimental manipulation, does not ordinarily give rise to all types of cells in an organism. In other words, somatic cells are cells that have differentiated sufficiently that they will not naturally generate cells of all three germ layers of the body, i.e. ectoderm, mesoderm and endoderm. For example, somatic cells would include both neurons and neural progenitors, the latter of which may be able to naturally give rise to all or some cell types of the central nervous system but

cannot give rise to cells of the mesoderm or endoderm lineages.

[0075] By "mitotic cell" it is meant a cell undergoing mitosis. Mitosis is the process by which a eukaryotic cell separates the chromosomes in its nucleus into two identical sets in two separate nuclei. It is generally followed immediately by cytokinesis, which divides the nuclei, cytoplasm, organelles and cell membrane into two cells containing roughly equal shares of these cellular components.

[0076] By "post-mitotic cell" it is meant a cell that has exited from mitosis, i.e., it is "quiescent", i.e. it is no longer undergoing divisions. This quiescent state may be temporary, i.e. reversible, or it may be permanent.

[0077] By "meiotic cell" it is meant a cell that is undergoing meiosis. Meiosis is the process by which a cell divides its nuclear material for the purpose of producing gametes or spores. Unlike mitosis, in meiosis, the chromosomes undergo a recombination step which shuffles genetic material between chromosomes. Additionally, the outcome of meiosis is four (genetically unique) haploid cells, as compared with the two (genetically identical) diploid cells produced from mitosis.

[0078] By "recombination" it is meant a process of exchange of genetic information between two polynucleotides. As used herein, "homology-directed repair (HDR)" refers to the specialized form DNA repair that takes place, for example, during repair of double-strand breaks in cells. This process requires nucleotide sequence homology, uses a "donor" molecule to template repair of a "target" molecule (i.e., the one that experienced the double-strand break), and leads to the transfer of genetic information from the donor to the target. Homology-directed repair may result in an alteration of the sequence of the target molecule (e.g., insertion, deletion, mutation), if the donor polynucleotide differs from the target molecule and part or all of the sequence of the donor polynucleotide is incorporated into the target DNA. In some embodiments, the donor polynucleotide, a portion of the donor polynucleotide, a copy of the donor polynucleotide, or a portion of a copy of the donor polynucleotide integrates into the target DNA.

[0079] By "non-homologous end joining (NHEJ)" it is meant the repair of double-strand breaks in DNA by direct ligation of the break ends to one another without the need for a homologous template (in contrast to homology-directed repair, which requires a homologous sequence to guide repair). NHEJ often results in the loss (deletion) of nucleotide sequence near the site of the double-strand break.

[0080] The terms "treatment", "treating" and the like are used herein to generally mean obtaining a desired pharmacologic and/or physiologic effect. The effect may be prophylactic in terms of completely or partially preventing a disease or symptom thereof and/or may be therapeutic in terms of a partial or complete cure for a disease and/or adverse effect attributable to the disease. "Treatment" as used herein covers any treatment of a disease or symptom in a mammal, and includes: (a) preventing the disease or symptom from occurring in

a subject which may be predisposed to acquiring the disease or symptom but has not yet been diagnosed as having it; (b) inhibiting the disease or symptom, i.e., arresting its development; or (c) relieving the disease, i.e., causing regression of the disease. The therapeutic agent may be administered before, during or after the onset of disease or injury. The treatment of ongoing disease, where the treatment stabilizes or reduces the undesirable clinical symptoms of the patient, is of particular interest. Such treatment is desirably performed prior to complete loss of function in the affected tissues. The non-claimed therapy will desirably be administered during the symptomatic stage of the disease, and in some cases after the symptomatic stage of the disease.

[0081] The terms "individual," "subject," "host," and "patient," are used interchangeably herein and refer to any mammalian subject for whom diagnosis, treatment, or therapy is desired, particularly humans.

[0082] General methods in molecular and cellular biochemistry can be found in such standard textbooks as Molecular Cloning: A Laboratory Manual, 3rd Ed. (Sambrook et al., HaRBor Laboratory Press 2001); Short Protocols in Molecular Biology, 4th Ed. (Ausubel et al. eds., John Wiley & Sons 1999); Protein Methods (Bollag et al., John Wiley & Sons 1996); Nonviral Vectors for Gene Therapy (Wagner et al. eds., Academic Press 1999); Viral Vectors (Kaplift & Loewy eds., Academic Press 1995); Immunology Methods Manual (I. Lefkovits ed., Academic Press 1997); and Cell and Tissue Culture: Laboratory Procedures in Biotechnology (Doyle & Griffiths, John Wiley & Sons 1998).

[0083] Before the present invention is further described, it is to be understood that this invention is not limited to particular embodiments described, as such may, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting, since the scope of the present invention will be limited only by the appended claims.

[0084] Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range, is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges, and are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the invention.

[0085] Certain ranges are presented herein with numerical values being preceded by the term "about." The term "about" is used herein to provide literal support for the exact number that it precedes, as well as a number that is near to or approximately the number that the term precedes. In determining whether a number is near to or approximately a specifically recited number, the near or approximating unrecited number may be a number which, in the context in which it is presented, provides the substantial equivalent of the specifically recited number.

[0086] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can also be used in the practice or testing of the present invention, the preferred methods and materials are now described.

[0087] The citation of any publication is for its disclosure prior to the filing date and should not be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention. Further, the dates of publication provided may be different from the actual publication dates which may need to be independently confirmed.

[0088] It is noted that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a polynucleotide" includes a plurality of such polynucleotides and reference to "the polypeptide" includes reference to one or more polypeptides and equivalents thereof known to those skilled in the art, and so forth. It is further noted that the claims may be drafted to exclude any optional element. As such, this statement is intended to serve as antecedent basis for use of such exclusive terminology as "solely," "only" and the like in connection with the recitation of claim elements, or use of a "negative" limitation.

[0089] It is appreciated that certain features of the invention, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of the invention, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable sub-combination. All combinations of the embodiments pertaining to the invention are specifically embraced by the present invention and are disclosed herein just as if each and every combination was individually and explicitly disclosed. In addition, all sub-combinations of the various embodiments and elements thereof are also specifically embraced by the present invention and are disclosed herein just as if each and every such sub-combination was individually and explicitly disclosed herein.

[0090] The publications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention. Further, the dates of publication provided may be different from the actual publication dates which may need to be independently confirmed.

DETAILED DESCRIPTION - PART I

[0091] The present disclosure provides a DNA-targeting RNA that comprises a targeting sequence and, together with a modifying polypeptide which is a naturally-occurring Cas9, provides for site-specific modification of a target DNA and/or a polypeptide associated with the

target DNA. Compositions comprising the DNA-targeting RNA are also provided.

NUCLEIC ACIDS

DNA-targeting RNA

[0092] The present disclosure provides a DNA-targeting RNA that directs the activities of an associated polypeptide (e.g., a site-directed modifying polypeptide) to a specific target sequence within a target DNA. A subject DNA-targeting RNA comprises: a first segment (also referred to herein as a "DNA-targeting segment" or a "DNA-targeting sequence") and a second segment (also referred to herein as a "protein-binding segment" or a "protein-binding sequence").

DNA-targeting segment of a DNA-targeting RNA

[0093] The DNA-targeting segment of a subject DNA-targeting RNA comprises a nucleotide sequence that is complementary to a sequence in a target DNA. In other words, the DNA-targeting segment of a subject DNA-targeting RNA interacts with a target DNA in a sequence-specific manner via hybridization (i.e., base pairing). As such, the nucleotide sequence of the DNA-targeting segment may vary and determines the location within the target DNA that the DNA-targeting RNA and the target DNA will interact. The DNA-targeting segment of a subject DNA-targeting RNA can be modified (e.g., by genetic engineering) to hybridize to any desired sequence within a target DNA.

[0094] The DNA-targeting segment can have a length of from about 12 nucleotides to about 100 nucleotides. For example, the DNA-targeting segment can have a length of from about 12 nucleotides (nt) to about 80 nt, from about 12 nt to about 50nt, from about 12 nt to about 40 nt, from about 12 nt to about 30 nt, from about 12 nt to about 25 nt, from about 12 nt to about 20 nt, or from about 12 nt to about 19 nt. For example, the DNA-targeting segment can have a length of from about 19 nt to about 20 nt, from about 19 nt to about 25 nt, from about 19 nt to about 30 nt, from about 19 nt to about 35 nt, from about 19 nt to about 40 nt, from about 19 nt to about 45 nt, from about 19 nt to about 50 nt, from about 19 nt to about 60 nt, from about 19 nt to about 70 nt, from about 19 nt to about 80 nt, from about 19 nt to about 90 nt, from about 19 nt to about 100 nt, from about 20 nt to about 25 nt, from about 20 nt to about 30 nt, from about 20 nt to about 35 nt, from about 20 nt to about 40 nt, from about 20 nt to about 45 nt, from about 20 nt to about 50 nt, from about 20 nt to about 60 nt, from about 20 nt to about 70 nt, from about 20 nt to about 80 nt, from about 20 nt to about 90 nt, or from about 20 nt to about 100 nt. The nucleotide sequence (the DNA-targeting sequence) of the DNA-targeting segment that is complementary to a nucleotide sequence (target sequence) of the target DNA can have a length at least about 12 nt. For example, the DNA-targeting sequence of the DNA-

targeting segment that is complementary to a target sequence of the target DNA can have a length at least about 12 nt, at least about 15 nt, at least about 18 nt, at least about 19 nt, at least about 20 nt, at least about 25 nt, at least about 30 nt, at least about 35 nt or at least about 40 nt. For example, the DNA-targeting sequence of the DNA-targeting segment that is complementary to a target sequence of the target DNA can have a length of from about 12 nucleotides (nt) to about 80 nt, from about 12 nt to about 50nt, from about 12 nt to about 45 nt, from about 12 nt to about 40 nt, from about 12 nt to about 35 nt, from about 12 nt to about 30 nt, from about 12 nt to about 25 nt, from about 12 nt to about 20 nt, from about 12 nt to about 19 nt, from about 19 nt to about 20 nt, from about 19 nt to about 25 nt, from about 19 nt to about 30 nt, from about 19 nt to about 35 nt, from about 19 nt to about 40 nt, from about 19 nt to about 45 nt, from about 19 nt to about 50 nt, from about 19 nt to about 60 nt, from about 20 nt to about 25 nt, from about 20 nt to about 30 nt, from about 20 nt to about 35 nt, from about 20 nt to about 40 nt, from about 20 nt to about 45 nt, from about 20 nt to about 50 nt, or from about 20 nt to about 60 nt. The nucleotide sequence (the DNA-targeting sequence) of the DNA-targeting segment that is complementary to a nucleotide sequence (target sequence) of the target DNA can have a length at least about 12 nt.

[0095] In some cases, the DNA-targeting sequence of the DNA-targeting segment that is complementary to a target sequence of the target DNA is 20 nucleotides in length. In some cases, the DNA-targeting sequence of the DNA-targeting segment that is complementary to a target sequence of the target DNA is 19 nucleotides in length.

[0096] The percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the target sequence of the target DNA can be at least 60% (e.g., at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, at least 97%, at least 98%, at least 99%, or 100%). In some cases, the percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the target sequence of the target DNA is 100% over the seven contiguous 5'-most nucleotides of the target sequence of the complementary strand of the target DNA. In some cases, the percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the target sequence of the target DNA is at least 60% over about 20 contiguous nucleotides. In some cases, the percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the target sequence of the target DNA is 100% over the fourteen contiguous 5'-most nucleotides of the target sequence of the complementary strand of the target DNA and as low as 0% over the remainder. In such a case, the DNA-targeting sequence can be considered to be 14 nucleotides in length (see Figure 12D-E). In some cases, the percent complementarity between the DNA-targeting sequence of the DNA-targeting segment and the target sequence of the target DNA is 100% over the seven contiguous 5'-most nucleotides of the target sequence of the complementary strand of the target DNA and as low as 0% over the remainder. In such a case, the DNA-targeting sequence can be considered to be 7 nucleotides in length.

Protein-binding segment of a DNA-targeting RNA

[0097] The protein-binding segment of a subject DNA-targeting RNA interacts with a site-directed modifying polypeptide. The subject DNA-targeting RNA guides the bound polypeptide to a specific nucleotide sequence within target DNA via the above mentioned DNA-targeting segment. The protein-binding segment of a subject DNA-targeting RNA comprises two stretches of nucleotides that are complementary to one another. The complementary nucleotides of the protein-binding segment hybridize to form a double stranded RNA duplex (dsRNA) (see Figures 1A and 1B).

[0098] A subject double-molecule DNA-targeting RNA comprises two separate RNA molecules. Each of the two RNA molecules of a subject double-molecule DNA-targeting RNA comprises a stretch of nucleotides that are complementary to one another such that the complementary nucleotides of the two RNA molecules hybridize to form the double stranded RNA duplex of the protein-binding segment (Figure 1A).

[0099] In some embodiments, the duplex-forming segment of the activator-RNA is at least about 60% identical to one of the activator-RNA (tracrRNA) molecules set forth in SEQ ID NOs:431-562, or a complement thereof, over a stretch of at least 8 contiguous nucleotides. For example, the duplex-forming segment of the activator-RNA (or the DNA encoding the duplex-forming segment of the activator-RNA) is at least about 60% identical, at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical, or 100 % identical, to one of the tracrRNA sequences set forth in SEQ ID NOs:431-562, or a complement thereof, over a stretch of at least 8 contiguous nucleotides.

[0100] In some embodiments, the duplex-forming segment of the targeter-RNA is at least about 60% identical to one of the targeter-RNA (crRNA) sequences set forth in SEQ ID NOs:563-679, or a complement thereof, over a stretch of at least 8 contiguous nucleotides. For example, the duplex-forming segment of the targeter-RNA (or the DNA encoding the duplex-forming segment of the targeter-RNA) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the crRNA sequences set forth in SEQ ID NOs:563-679, or a complement thereof, over a stretch of at least 8 contiguous nucleotides.

[0101] A two-molecule DNA-targeting RNA can be designed to allow for controlled (i.e., conditional) binding of a targeter-RNA with an activator-RNA. Because a two-molecule DNA-targeting RNA is not functional unless both the activator-RNA and the targeter-RNA are bound in a functional complex with dCas9, a two-molecule DNA-targeting RNA can be inducible (e.g., drug inducible) by rendering the binding between the activator-RNA and the targeter-RNA to be inducible. As one non-limiting example, RNA aptamers can be used to regulate (i.e., control) the binding of the activator-RNA with the targeter-RNA. Accordingly, the activator-RNA

and/or the targeter-RNA can comprise an RNA aptamer sequence.

[0102] RNA aptamers are known in the art and are generally a synthetic version of a riboswitch. The terms "RNA aptamer" and "riboswitch" are used interchangeably herein to encompass both synthetic and natural nucleic acid sequences that provide for inducible regulation of the structure (and therefore the availability of specific sequences) of the RNA molecule of which they are part. RNA aptamers usually comprise a sequence that folds into a particular structure (e.g., a hairpin), which specifically binds a particular drug (e.g., a small molecule). Binding of the drug causes a structural change in the folding of the RNA, which changes a feature of the nucleic acid of which the aptamer is a part. As non-limiting examples: (i) an activator-RNA with an aptamer may not be able to bind to the cognate targeter-RNA unless the aptamer is bound by the appropriate drug; (ii) a targeter-RNA with an aptamer may not be able to bind to the cognate activator-RNA unless the aptamer is bound by the appropriate drug; and (iii) a targeter-RNA and an activator-RNA, each comprising a different aptamer that binds a different drug, may not be able to bind to each other unless both drugs are present. As illustrated by these examples, a two-molecule DNA-targeting RNA can be designed to be inducible.

[0103] Examples of aptamers and riboswitches can be found, for example, in: Nakamura et al., Genes Cells. 2012 May;17(5):344-64; Vavalle et al., Future Cardiol. 2012 May;8(3):371-82; Citartan et al., Biosens Bioelectron. 2012 Apr 15;34(1): 1-11; and Liberman et al., Wiley Interdiscip Rev RNA. 2012 May-Jun;3(3):369-84.

[0104] Non-limiting examples of nucleotide sequences that can be included in a two-molecule DNA-targeting RNA include either of the sequences set forth in SEQ ID NOs:431-562, or complements thereof pairing with any sequences set forth in SEQ ID NOs:563-679, or complements thereof that can hybridize to form a protein binding segment.

[0105] A subject single-molecule DNA-targeting RNA comprises two stretches of nucleotides (a targeter-RNA and an activator-RNA) that are complementary to one another, are covalently linked by intervening nucleotides ("linkers" or "linker nucleotides"), and hybridize to form the double stranded RNA duplex (dsRNA duplex) of the protein-binding segment, thus resulting in a stem-loop structure (Figure 1B). The targeter-RNA and the activator-RNA can be covalently linked via the 3' end of the targeter-RNA and the 5' end of the activator-RNA. Alternatively, targeter-RNA and the activator-RNA can be covalently linked via the 5' end of the targeter-RNA and the 3' end of the activator-RNA.

[0106] The linker of a single-molecule DNA-targeting RNA can have a length of from about 3 nucleotides to about 100 nucleotides. For example, the linker can have a length of from about 3 nucleotides (nt) to about 90 nt, from about 3 nucleotides (nt) to about 80 nt, from about 3 nucleotides (nt) to about 70 nt, from about 3 nucleotides (nt) to about 60 nt, from about 3 nucleotides (nt) to about 50 nt, from about 3 nucleotides (nt) to about 40 nt, from about 3 nucleotides (nt) to about 30 nt, from about 3 nucleotides (nt) to about 20 nt or from about 3 nucleotides (nt) to about 10 nt. For example, the linker can have a length of from about 3 nt to

about 5 nt, from about 5 nt to about 10 nt, from about 10 nt to about 15 nt, from about 15 nt to about 20 nt, from about 20 nt to about 25 nt, from about 25 nt to about 30 nt, from about 30 nt to about 35 nt, from about 35 nt to about 40 nt, from about 40 nt to about 50 nt, from about 50 nt to about 60 nt, from about 60 nt to about 70 nt, from about 70 nt to about 80 nt, from about 80 nt to about 90 nt, or from about 90 nt to about 100 nt. In some embodiments, the linker of a single-molecule DNA-targeting RNA is 4 nt.

[0107] An exemplary single-molecule DNA-targeting RNA comprises two complementary stretches of nucleotides that hybridize to form a dsRNA duplex. In some embodiments, one of the two complementary stretches of nucleotides of the single-molecule DNA-targeting RNA (or the DNA encoding the stretch) is at least about 60% identical to one of the activator-RNA (tracrRNA) molecules set forth in SEQ ID Nos:431-562, or a complement thereof, over a stretch of at least 8 contiguous nucleotides. For example, one of the two complementary stretches of nucleotides of the single-molecule DNA-targeting RNA (or the DNA encoding the stretch) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the tracrRNA sequences set forth in SEQ ID Nos:431-562, or a complement thereof, over a stretch of at least 8 contiguous nucleotides.

[0108] In some embodiments, one of the two complementary stretches of nucleotides of the single-molecule DNA-targeting RNA (or the DNA encoding the stretch) is at least about 60% identical to one of the targeter-RNA (crRNA) sequences set forth in SEQ ID Nos:563-679, or a complement thereof, over a stretch of at least 8 contiguous nucleotides. For example, one of the two complementary stretches of nucleotides of the single-molecule DNA-targeting RNA (or the DNA encoding the stretch) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the crRNA sequences set forth in SEQ ID Nos:563-679, or a complement thereof, over a stretch of at least 8 contiguous nucleotides.

[0109] Appropriate naturally occurring cognate pairs of crRNAs and tracrRNAs can be routinely determined for SEQ ID Nos:431-679 by taking into account the species name and base-pairing (for the dsRNA duplex of the protein-binding domain) when determining appropriate cognate pairs (see Figure 8 as a non-limiting example).

[0110] With regard to both a subject single-molecule DNA-targeting RNA and to a subject double-molecule DNA-targeting RNA, **Figure 57** demonstrates that artificial sequences that share very little (roughly 50% identity) with naturally occurring a tracrRNAs and crRNAs can function with Cas9 to cleave target DNA as long as the structure of the protein-binding domain of the DNA-targeting RNA is conserved. Thus, RNA folding structure of a naturally occurring protein-binding domain of a DNA-targeting RNA can be taken into account in order to design artificial protein-binding domains (either two-molecule or single-molecule versions). As a non-limiting example, the functional artificial DNA-targeting RNA of Figure 57 was designed based

on the structure of the protein-binding segment of the naturally occurring DNA-targeting (e.g., including the same number of base pairs along the RNA duplex and including the same "bulge" region as present in the naturally occurring RNA). As structures can readily be produced by one of ordinary skill in the art for any naturally occurring crRNA:tracrRNA pair from any species (see SEQ ID NOs:431-679 for crRNA and tracrRNA sequences from a wide variety of species), an artificial DNA-targeting-RNA can be designed to mimic the natural structure for a given species when using the Cas9 (or a related Cas9, see Figure 32A) from that species. (see Figure 24D and related details in Example 1). Thus, a suitable DNA-targeting RNA can be an artificially designed RNA (non-naturally occurring) comprising a protein-binding domain that was designed to mimic the structure of a protein-binding domain of a naturally occurring DNA-targeting RNA. (see SEQ ID NOs:431-679, taking into account the species name when determining appropriate cognate pairs).

[0111] The protein-binding segment can have a length of from about 10 nucleotides to about 100 nucleotides. For example, the protein-binding segment can have a length of from about 15 nucleotides (nt) to about 80 nt, from about 15 nt to about 50 nt, from about 15 nt to about 40 nt, from about 15 nt to about 30 nt or from about 15 nt to about 25 nt.

[0112] Also with regard to both a subject single-molecule DNA-targeting RNA and to a subject double-molecule DNA-targeting RNA, the dsRNA duplex of the protein-binding segment can have a length from about 6 base pairs (bp) to about 50bp. For example, the dsRNA duplex of the protein-binding segment can have a length from about 6 bp to about 40 bp, from about 6 bp to about 30bp, from about 6 bp to about 25 bp, from about 6 bp to about 20 bp, from about 6 bp to about 15 bp, from about 8 bp to about 40 bp, from about 8 bp to about 30bp, from about 8 bp to about 25 bp, from about 8 bp to about 20 bp or from about 8 bp to about 15 bp. For example, the dsRNA duplex of the protein-binding segment can have a length from about from about 8 bp to about 10 bp, from about 10 bp to about 15 bp, from about 15 bp to about 18 bp, from about 18 bp to about 20 bp, from about 20 bp to about 25 bp, from about 25 bp to about 30 bp, from about 30 bp to about 35 bp, from about 35 bp to about 40 bp, or from about 40 bp to about 50 bp. In some embodiments, the dsRNA duplex of the protein-binding segment has a length of 36 base pairs. The percent complementarity between the nucleotide sequences that hybridize to form the dsRNA duplex of the protein-binding segment can be at least about 60%. For example, the percent complementarity between the nucleotide sequences that hybridize to form the dsRNA duplex of the protein-binding segment can be at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, or at least about 99% . In some cases, the percent complementarity between the nucleotide sequences that hybridize to form the dsRNA duplex of the protein-binding segment is 100%.

Site-directed modifying polypeptide

[0113] A subject DNA-targeting RNA and a site-directed modifying polypeptide which is a naturally-occurring Cas9 endonuclease form a complex. The DNA-targeting RNA provides

target specificity to the complex by comprising a nucleotide sequence that is complementary to a sequence of a target DNA (as noted above). The site-directed modifying polypeptide of the complex provides the site-specific activity. In other words, the site-directed modifying polypeptide is guided to a DNA sequence (e.g. a chromosomal sequence or an extrachromosomal sequence, e.g. an episomal sequence, a minicircle sequence, a mitochondrial sequence, a chloroplast sequence, etc.) by virtue of its association with at least the protein-binding segment of the DNA-targeting RNA (described above).

[0114] A site-directed modifying polypeptide (non-claimed as such) modifies target DNA (e.g., cleavage). A site-directed modifying polypeptide (non-claimed as such) is also referred to herein as a "site-directed polypeptide" or an "RNA binding site-directed modifying polypeptide."

[0115] The site-directed modifying polypeptide (non-claimed as such) is a naturally-occurring modifying polypeptide.

[0116] Exemplary naturally-occurring site-directed modifying polypeptides (non-claimed as such) are set forth in SEQ ID NOs: 1-255 as a non-limiting and non-exhaustive list of naturally occurring Cas9/Csn1 endonucleases. These naturally occurring polypeptides (non-claimed as such), as disclosed herein, bind a DNA-targeting RNA, are thereby directed to a specific sequence within a target DNA, and cleave the target DNA to generate a double strand break. A site-directed modifying polypeptide (non-claimed as such) comprises two portions, an RNA-binding portion and an activity portion. In some embodiments, a subject site-directed modifying polypeptide comprises: (i) an RNA-binding portion that interacts with a DNA-targeting RNA, wherein the DNA-targeting RNA comprises a nucleotide sequence that is complementary to a sequence in a target DNA; and (ii) an activity portion that exhibits site-directed enzymatic activity (activity for DNA cleavage), wherein the site of enzymatic activity is determined by the DNA-targeting RNA.

[0117] A subject site-directed modifying polypeptide (non-claimed as such) has enzymatic activity that modifies target DNA by nuclease activity.

Exemplary site-directed modifying polypeptides

[0118] In some cases, the site-directed modifying polypeptide (non-claimed as such) comprises an amino acid sequence having at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 99%, or 100%, amino acid sequence identity to amino acids 7-166 or 731-1003 of the Cas9/Csn1 amino acid sequence depicted in Figure 3, or to the corresponding portions in any of the amino acid sequences set forth as SEQ ID NOs: 1-256 and 795-1346.

Nucleic acid modifications

[0119] In some embodiments, a subject DNA-targeting RNA comprises one or more modifications, e.g., a base modification, a backbone modification, etc, to provide the nucleic acid with a new or enhanced feature (e.g., improved stability). As is known in the art, a nucleoside is a base-sugar combination. The base portion of the nucleoside is normally a heterocyclic base. The two most common classes of such heterocyclic bases are the purines and the pyrimidines. Nucleotides are nucleosides that further include a phosphate group covalently linked to the sugar portion of the nucleoside. For those nucleosides that include a pentofuranosyl sugar, the phosphate group can be linked to the 2', the 3', or the 5' hydroxyl moiety of the sugar. In forming oligonucleotides, the phosphate groups covalently link adjacent nucleosides to one another to form a linear polymeric compound. In turn, the respective ends of this linear polymeric compound can be further joined to form a circular compound, however, linear compounds are generally suitable. In addition, linear compounds may have internal nucleotide base complementarity and may therefore fold in a manner as to produce a fully or partially double-stranded compound. Within oligonucleotides, the phosphate groups are commonly referred to as forming the internucleoside backbone of the oligonucleotide. The normal linkage or backbone of RNA and DNA is a 3' to 5' phosphodiester linkage.

Modified backbones and modified internucleoside linkages

[0120] Examples of suitable DNA-targeting RNAs containing modifications include DNA-targeting RNAs containing modified backbones or non-natural internucleoside linkages. DNA-targeting RNAs having modified backbones include those that retain a phosphorus atom in the backbone and those that do not have a phosphorus atom in the backbone.

[0121] Suitable modified oligonucleotide backbones containing a phosphorus atom therein include, for example, phosphorothioates, chiral phosphorothioates, phosphorodithioates, phosphotriesters, aminoalkylphosphotriesters, methyl and other alkyl phosphonates including 3'-alkylene phosphonates, 5'-alkylene phosphonates and chiral phosphonates, phosphinates, phosphoramidates including 3'-amino phosphoramidate and aminoalkylphosphoramidates, phosphorodiamidates, thionophosphoramidates, thionoalkylphosphonates, thionoalkylphosphotriesters, selenophosphates and boranophosphates having normal 3'-5' linkages, 2'-5' linked analogs of these, and those having inverted polarity wherein one or more internucleotide linkages is a 3' to 3', 5' to 5' or 2' to 2' linkage. Suitable DNA-targeting RNAs having inverted polarity comprise a single 3' to 3' linkage at the 3'-most internucleotide linkage i.e. a single inverted nucleoside residue which may be a basic (the nucleobase is missing or has a hydroxyl group in place thereof). Various salts (such as, for example, potassium or sodium), mixed salts and free acid forms are also included.

[0122] In some embodiments, a subject DNA-targeting RNA comprises one or more phosphorothioate and/or heteroatom internucleoside linkages, in particular -CH_2-NH-O-CH_2-, -CH_2-N(CH_3)-O-CH_2- (known as a methylene (methylimino) or MMI backbone), -CH_2-O-N(CH_3)-CH_2-, -CH_2-N(CH_3)-N(CH_3)-CH_2- and -O-N(CH_3)-CH_2-CH_2- (wherein the native phosphodiester

internucleotide linkage is represented as -O-P(=O)(OH)-O-CH_2-). MMI type internucleoside linkages are disclosed in the above referenced U.S. Pat. No. 5,489,677. Suitable amide internucleoside linkages are disclosed in t U.S. Pat. No. 5,602,240.

[0123] Also suitable are DNA-targeting RNAs having morpholino backbone structures as described in, e.g., U.S. Pat. No. 5,034,506. For example, in some embodiments, a subject DNA-targeting RNAs comprises a 6-membered morpholino ring in place of a ribose ring. In some of these embodiments, a phosphorodiamidate or other non-phosphodiester internucleoside linkage replaces a phosphodiester linkage.

[0124] Suitable modified polynucleotide backbones that do not include a phosphorus atom therein have backbones that are formed by short chain alkyl or cycloalkyl internucleoside linkages, mixed heteroatom and alkyl or cycloalkyl internucleoside linkages, or one or more short chain heteroatomic or heterocyclic internucleoside linkages. These include those having morpholino linkages (formed in part from the sugar portion of a nucleoside); siloxane backbones; sulfide, sulfoxide and sulfone backbones; formacetyl and thioformacetyl backbones; methylene formacetyl and thioformacetyl backbones; riboacetyl backbones; alkene containing backbones; sulfamate backbones; methyleneimino and methylenehydrazino backbones; sulfonate and sulfonamide backbones; amide backbones; and others having mixed N, O, S and CH_2 component parts.

Mimetics

[0125] A subject DNA-targeting RNAs can be a nucleic acid mimetic. The term "mimetic" as it is applied to polynucleotides is intended to include polynucleotides wherein only the furanose ring or both the furanose ring and the internucleotide linkage are replaced with non-furanose groups, replacement of only the furanose ring is also referred to in the art as being a sugar surrogate. The heterocyclic base moiety or a modified heterocyclic base moiety is maintained for hybridization with an appropriate target nucleic acid. One such nucleic acid, a polynucleotide mimetic that has been shown to have excellent hybridization properties, is referred to as a peptide nucleic acid (PNA). In PNA, the sugar-backbone of a polynucleotide is replaced with an amide containing backbone, in particular an aminoethylglycine backbone. The nucleotides are retained and are bound directly or indirectly to aza nitrogen atoms of the amide portion of the backbone.

[0126] One polynucleotide mimetic that has been reported to have excellent hybridization properties is a peptide nucleic acid (PNA). The backbone in PNA compounds is two or more linked aminoethylglycine units which gives PNA an amide containing backbone. The heterocyclic base moieties are bound directly or indirectly to aza nitrogen atoms of the amide portion of the backbone. Representative U.S. patents that describe the preparation of PNA compounds include, but are not limited to: U.S. Pat. Nos. 5,539,082; 5,714,331; and 5,719,262.

[0127] Another class of polynucleotide mimetic that has been studied is based on linked morpholino units (morpholino nucleic acid) having heterocyclic bases attached to the morpholino ring. A number of linking groups have been reported that link the morpholino monomeric units in a morpholino nucleic acid. One class of linking groups has been selected to give a non-ionic oligomeric compound. The non-ionic morpholino-based oligomeric compounds are less likely to have undesired interactions with cellular proteins. Morpholino-based polynucleotides are non-ionic mimics of oligonucleotides which are less likely to form undesired interactions with cellular proteins (Dwaine A. Braasch and David R. Corey, Biochemistry, 2002, 41(14), 4503-4510). Morpholino-based polynucleotides are disclosed in U.S. Pat. No. 5,034,506. A variety of compounds within the morpholino class of polynucleotides have been prepared, having a variety of different linking groups joining the monomeric subunits.

[0128] A further class of polynucleotide mimetic is referred to as cyclohexenyl nucleic acids (CeNA). The furanose ring normally present in a DNA/RNA molecule is replaced with a cyclohexenyl ring. CeNA DMT protected phosphoramidite monomers have been prepared and used for oligomeric compound synthesis following classical phosphoramidite chemistry. Fully modified CeNA oligomeric compounds and oligonucleotides having specific positions modified with CeNA have been prepared and studied (see Wang et al., J. Am. Chem. Soc., 2000, 122, 8595-8602). In general the incorporation of CeNA monomers into a DNA chain increases its stability of a DNA/RNA hybrid. CeNA oligoadenylates formed complexes with RNA and DNA complements with similar stability to the native complexes. The study of incorporating CeNA structures into natural nucleic acid structures was shown by NMR and circular dichroism to proceed with easy conformational adaptation.

[0129] A further modification includes Locked Nucleic Acids (LNAs) in which the 2'-hydroxyl group is linked to the 4' carbon atom of the sugar ring thereby forming a 2'-C,4'-C-oxymethylene linkage thereby forming a bicyclic sugar moiety. The linkage can be a methylene (-CH_2-), group bridging the 2' oxygen atom and the 4' carbon atom wherein n is 1 or 2 (Singh et al., Chem. Commun., 1998, 4, 455-456). LNA and LNA analogs display very high duplex thermal stabilities with complementary DNA and RNA (Tm=+3 to +10° C), stability towards 3'-exonucleolytic degradation and good solubility properties. Potent and nontoxic antisense oligonucleotides containing LNAs have been described (Wahlestedt et al., Proc. Natl. Acad. Sci. U.S.A., 2000, 97, 5633-5638).

[0130] The synthesis and preparation of the LNA monomers adenine, cytosine, guanine, 5-methylcytosine, thymine and uracil, along with their oligomerization, and nucleic acid recognition properties have been described (Koshkin et al., Tetrahedron, 1998, 54, 3607-3630). LNAs and preparation thereof are also described in WO 98/39352 and WO 99/14226.

Modified sugar moieties

[0131] A subject DNA-targeting RNA for use can also include one or more substituted sugar

moieties. Suitable polynucleotides comprise a sugar substituent group selected from: OH; F; O-, S-, or N-alkyl; O-, S-, or N-alkenyl; O-, S- or N-alkynyl; or O-alkyl-O-alkyl, wherein the alkyl, alkenyl and alkynyl may be substituted or unsubstituted C.sub.1 to C_{10} alkyl or C_2 to C_{10} alkenyl and alkynyl. Particularly suitable are $O((CH_2)_nO)_mCH_3$, $O(CH_2)_nOCH_3$, $O(CH_2)_nNH_2$, $O(CH_2)_nCH_3$, $O(CH_2)_nONH_2$, and $O(CH_2)_nON((CH_2)_nCH_3)_2$, where n and m are from 1 to about 10. Other suitable polynucleotides comprise a sugar substituent group selected from: C_1 to C_{10} lower alkyl, substituted lower alkyl, alkenyl, alkynyl, alkaryl, aralkyl, O-alkaryl or O-aralkyl, SH, SCH_3, OCN, Cl, Br, CN, CF_3, OCF_3, $SOCH_3$, SO_2CH_3, ONO_2, NO_2, N_3, NH_2, heterocycloalkyl, heterocycloalkaryl, aminoalkylamino, polyalkylamino, substituted silyl, an RNA cleaving group, a reporter group, an intercalator, a group for improving the pharmacokinetic properties of an oligonucleotide, or a group for improving the pharmacodynamic properties of an oligonucleotide, and other substituents having similar properties. A suitable modification includes 2'-methoxyethoxy (2'-O-CH_2 CH_2OCH_3, also known as 2'-O-(2-methoxyethyl) or 2'-MOE) (Martin et al., Helv. Chim. Acta, 1995, 78, 486-504) i.e., an alkoxyalkoxy group. A further suitable modification includes 2'-dimethylaminooxyethoxy, i.e., a $O(CH_2)_2ON(CH_3)_2$ group, also known as 2'-DMAOE, as described in examples hereinbelow, and 2'-dimethylaminoethoxyethoxy (also known in the art as 2'-O-dimethyl-amino-ethoxy-ethyl or 2'-DMAEOE), i.e., 2'-O-CH_2-O-CH_2-$N(CH_3)_2$.

[0132] Other suitable sugar substituent groups include methoxy (-O-CH_3), aminopropoxy (--O CH_2 CH_2 CH_2NH_2), allyl (-CH_2-CH=CH_2), -O-allyl (--O-- CH_2-CH=CH_2) and fluoro (F). 2'-sugar substituent groups may be in the arabino (up) position or ribo (down) position. A suitable 2'-arabino modification is 2'-F. Similar modifications may also be made at other positions on the oligomeric compound, particularly the 3' position of the sugar on the 3' terminal nucleoside or in 2'-5' linked oligonucleotides and the 5' position of 5' terminal nucleotide. Oligomeric compounds may also have sugar mimetics such as cyclobutyl moieties in place of the pentofuranosyl sugar.

Base modifications and substitutions

[0133] A subject DNA-targeting RNA may also include nucleobase (often referred to in the art simply as "base") modifications or substitutions. As used herein, "unmodified" or "natural" nucleobases include the purine bases adenine (A) and guanine (G), and the pyrimidine bases thymine (T), cytosine (C) and uracil (U). Modified nucleobases include other synthetic and natural nucleobases such as 5-methylcytosine (5-me-C), 5-hydroxymethyl cytosine, xanthine, hypoxanthine, 2-aminoadenine, 6-methyl and other alkyl derivatives of adenine and guanine, 2-propyl and other alkyl derivatives of adenine and guanine, 2-thiouracil, 2-thiothymine and 2-thiocytosine, 5-halouracil and cytosine, 5-propynyl (-C=C-CH_3) uracil and cytosine and other alkynyl derivatives of pyrimidine bases, 6-azo uracil, cytosine and thymine, 5-uracil (pseudouracil), 4-thiouracil, 8-halo, 8-amino, 8-thiol, 8-thioalkyl, 8-hydroxyl and other 8-substituted adenines and guanines, 5-halo particularly 5-bromo, 5-trifluoromethyl and other 5-

substituted uracils and cytosines, 7-methylguanine and 7-methyladenine, 2-F-adenine, 2-aminoadenine, 8-azaguanine and 8-azaadenine, 7-deazaguanine and 7-deazaadenine and 3-deazaguanine and 3-deazaadenine. Further modified nucleobases include tricyclic pyrimidines such as phenoxazine cytidine(1H-pyrimido(5,4-b)(1,4)benzoxazin-2(3H)-one), phenothiazine cytidine (1H-pyrimido(5,4-b)(1,4)benzothiazin-2(3H)-one), G-clamps such as a substituted phenoxazine cytidine (e.g. 9-(2-aminoethoxy)-H-pyrimido(5,4-(b) (1,4)benzoxazin-2(3H)-one), carbazole cytidine (2H-pyrimido(4,5-b)indol-2-one), pyridoindole cytidine (H-pyrido(3',2':4,5)pyrrolo(2,3-d)pyrimidin-2-one).

[0134] Heterocyclic base moieties may also include those in which the purine or pyrimidine base is replaced with other heterocycles, for example 7-deaza-adenine, 7-deazaguanosine, 2-aminopyridine and 2-pyridone. Further nucleobases include those disclosed in U.S. Pat. No. 3,687,808, those disclosed in The Concise Encyclopedia Of Polymer Science And Engineering, pages 858-859, Kroschwitz, J. I., ed. John Wiley & Sons, 1990, those disclosed by Englisch et al., Angewandte Chemie, International Edition, 1991, 30, 613, and those disclosed by Sanghvi, Y. S., Chapter 15, Antisense Research and Applications, pages 289-302, Crooke, S. T. and Lebleu, B., ed., CRC Press, 1993. Certain of these nucleobases are useful for increasing the binding affinity of an oligomeric compound. These include 5-substituted pyrimidines, 6-azapyrimidines and N-2, N-6 and O-6 substituted purines, including 2-aminopropyladenine, 5-propynyluracil and 5-propynylcytosine. 5-methylcytosine substitutions have been shown to increase nucleic acid duplex stability by 0.6-1.2° C. (Sanghvi et al., eds., Antisense Research and Applications, CRC Press, Boca Raton, 1993, pp. 276-278) and are suitable base substitutions, e.g., when combined with 2'-O-methoxyethyl sugar modifications.

Conjugates

[0135] Another possible modification of a subject DNA-targeting RNA involves chemically linking to the polynucleotide one or more moieties or conjugates which enhance the activity, cellular distribution or cellular uptake of the oligonucleotide. These moieties or conjugates can include conjugate groups covalently bound to functional groups such as primary or secondary hydroxyl groups. Conjugate groups include, but are not limited to, intercalators, reporter molecules, polyamines, polyamides, polyethylene glycols, polyethers, groups that enhance the pharmacodynamic properties of oligomers, and groups that enhance the pharmacokinetic properties of oligomers. Suitable conjugate groups include, but are not limited to, cholesterols, lipids, phospholipids, biotin, phenazine, folate, phenanthridine, anthraquinone, acridine, fluoresceins, rhodamines, coumarins, and dyes. Groups that enhance the pharmacodynamic properties include groups that improve uptake, enhance resistance to degradation, and/or strengthen sequence-specific hybridization with the target nucleic acid. Groups that enhance the pharmacokinetic properties include groups that improve uptake, distribution, metabolism or excretion of a subject nucleic acid.

[0136] Conjugate moieties include but are not limited to lipid moieties such as a cholesterol moiety (Letsinger et al., Proc. Natl. Acad. Sci. USA, 1989, 86, 6553-6556), cholic acid

(Manoharan et al., Bioorg. Med. Chem. Let., 1994, 4, 1053-1060), a thioether, e.g., hexyl-S-tritylthiol (Manoharan et al., Ann. N.Y. Acad. Sci., 1992, 660, 306-309; Manoharan et al., Bioorg. Med. Chem. Let., 1993, 3, 2765-2770), a thiocholesterol (Oberhauser et al., Nucl. Acids Res., 1992, 20, 533-538), an aliphatic chain, e.g., dodecandiol or undecyl residues (Saison-Behmoaras et al., EMBO J., 1991, 10, 1111-1118; Kabanov et al., FEBS Lett., 1990, 259, 327-330; Svinarchuk et al., Biochimie, 1993, 75, 49-54), a phospholipid, e.g., di-hexadecyl-rac-glycerol or triethylammonium 1,2-di-0-hexadecyl-rac-glycero-3-H-phosphonate (Manoharan et al., Tetrahedron Lett., 1995, 36, 3651-3654; Shea et al., Nucl. Acids Res., 1990, 18, 3777-3783), a polyamine or a polyethylene glycol chain (Manoharan et al., Nucleosides & Nucleotides, 1995, 14, 969-973), or adamantane acetic acid (Manoharan et al., Tetrahedron Lett., 1995, 36, 3651-3654), a palmityl moiety (Mishra et al., Biochim. Biophys. Acta, 1995, 1264, 229-237), or an octadecylamine or hexylamino-carbonyl-oxycholesterol moiety (Crooke et al., J. Pharmacol. Exp. Ther., 1996, 277, 923-937.\

[0137] A conjugate may include a "Protein Transduction Domain" or PTD (also known as a CPP - cell penetrating peptide), which may refer to a polypeptide, polynucleotide, carbohydrate, or organic or inorganic compound that facilitates traversing a lipid bilayer, micelle, cell membrane, organelle membrane, or vesicle membrane. A PTD attached to another molecule, which can range from a small polar molecule to a large macromolecule and/or a nanoparticle, facilitates the molecule traversing a membrane, for example going from extracellular space to intracellular space, or cytosol to within an organelle. In some embodiments, a PTD is covalently linked to a nucleic acid (e.g., a DNA-targeting RNA, a polynucleotide encoding a DNA-targeting RNA, etc.). Exemplary PTDs include but are not limited to a minimal undecapeptide protein transduction domain (corresponding to residues 47-57 of HIV-1 TAT comprising YGRKKRRQRRR; SEQ ID NO:264); a polyarginine sequence comprising a number of arginines sufficient to direct entry into a cell (e.g., 3, 4, 5, 6, 7, 8, 9, 10, or 10-50 arginines); a VP22 domain (Zender et al. (2002) Cancer Gene Ther. 9(6):489-96); an Drosophila Antennapedia protein transduction domain (Noguchi et al. (2003) Diabetes 52(7):1732-1737); a truncated human calcitonin peptide (Trehin et al. (2004) Pharm. Research 21:1248-1256); polylysine (Wender et al. (2000) Proc. Natl. Acad. Sci. USA 97:13003-13008); RRQRRTSKLMKR (SEQ ID NO:265); Transportan GWTLNSAGYLLGKINLKALAALAKKIL (SEQ ID NO:266); KALAWEAKLAKALAKALAKHLAKALAKALKCEA (SEQ ID NO:267); and RQIKIWFQNRRMKWKK (SEQ ID NO:268). Exemplary PTDs include but are not limited to, YGRKKRRQRRR (SEQ ID NO:264), RKKRRQRRR (SEQ ID NO:269); an arginine homopolymer of from 3 arginine residues to 50 arginine residues; Exemplary PTD domain amino acid sequences include, but are not limited to, any of the following: YGRKKRRQRRR (SEQ ID NO:264); RKKRRQRR (SEQ ID NO:270); YARAAARQARA (SEQ ID NO:271); THRLPRRRRRR (SEQ ID NO:272); and GGRRARRRRRR (SEQ ID NO:273). In some embodiments, the PTD is an activatable CPP (ACPP) (Aguilera et al. (2009) Integr Biol (Camb) June; 1(5-6): 371-381). ACPPs comprise a polycationic CPP (e.g., Arg9 or "R9") connected via a cleavable linker to a matching polyanion (e.g., Glu9 or "E9"), which reduces the net charge to nearly zero and thereby inhibits adhesion and uptake into cells. Upon cleavage of the linker, the polyanion is released, locally unmasking the polyarginine and its inherent adhesiveness, thus "activating" the ACPP to traverse the membrane.

Exemplary DNA-targeting RNAs

[0138] In some embodiments, a suitable DNA-targeting RNA comprises two separate RNA polynucleotide molecules. The first of the two separate RNA polynucleotide molecules (the activator-RNA) comprises a nucleotide sequence having at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100% nucleotide sequence identity over a stretch of at least 8 contiguous nucleotides to any one of the nucleotide sequences set forth in SEQ ID NOs:431-562, or complements thereof. The second of the two separate RNA polynucleotide molecules (the targeter-RNA) comprises a nucleotide sequence having at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100% nucleotide sequence identity over a stretch of at least 8 contiguous nucleotides to any one of the nucleotide sequences set forth in SEQ ID NOs:563-679, or complements thereof.

[0139] In some embodiments, a suitable DNA-targeting RNA is a single RNA polynucleotide and comprises a first nucleotide sequence having at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100% nucleotide sequence identity over a stretch of at least 8 contiguous nucleotides to any one of the nucleotide sequences set forth in SEQ ID NOs:431-562 and a second nucleotide sequence having at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or 100% nucleotide sequence identity over a stretch of at least 8 contiguous nucleotides to any one of the nucleotide sequences set forth in SEQ ID NOs: 463-679.

[0140] In some embodiments, the DNA-targeting RNA is a double-molecule DNA-targeting RNA and the targeter-RNA comprises the sequence 5'GUUUUAGAGCUA-3' (SEQ ID NO:679) linked at its 5' end to a stretch of nucleotides that are complementary to a target DNA. In some embodiments, the DNA-targeting RNA is a double-molecule DNA-targeting RNA and the activator-RNA comprises the sequence 5' UAGCAAGUUAAAAUAAGGCUAGUCCG-3' (SEQ ID NO://).

[0141] In some embodiments, the DNA-targeting RNA is a single-molecule DNA-targeting RNA and comprises the sequence 5'-GUUUUAGAGCUA-linker-UAGCAAGUUAAAAUAAGGCUAGUCCG-3' linked at its 5' end to a stretch of nucleotides that are complementary to a target DNA (where "linker" denotes any a linker nucleotide sequence that can comprise any nucleotide sequence) (SEQ ID NO://). Other exemplary single-molecule DNA-targeting RNAs include those set forth in SEQ ID NOs: 680-682.

Nucleic acids encoding a subject DNA-targeting RNA

[0142] The present disclosure provides a nucleic acid comprising a nucleotide sequence encoding a subject single-molecule DNA-targeting RNA. In some embodiments, a subject single-molecule DNA-targeting RNA-encoding nucleic acid is an expression vector, e.g., a recombinant expression vector.

[0143] The recombinant expression vector may be a viral construct, e.g., a recombinant adeno-associated virus construct (see, e.g., U.S. Patent No. 7,078,387), a recombinant adenoviral construct, a recombinant lentiviral construct, a recombinant retroviral construct, etc.

[0144] Suitable expression vectors include, but are not limited to, viral vectors (e.g. viral vectors based on vaccinia virus; poliovirus; adenovirus (see, e.g., Li et al., Invest Opthalmol Vis Sci 35:2543 2549, 1994; Borras et al., Gene Ther 6:515 524, 1999; Li and Davidson, PNAS 92:7700 7704, 1995; Sakamoto et al., H Gene Ther 5:1088 1097, 1999; WO 94/12649, WO 93/03769; WO 93/19191; WO 94/28938; WO 95/11984 and WO 95/00655); adeno-associated virus (see, e.g., Ali et al., Hum Gene Ther 9:81 86, 1998, Flannery et al., PNAS 94:6916 6921, 1997; Bennett et al., Invest Opthalmol Vis Sci 38:2857 2863, 1997; Jomary et al., Gene Ther 4:683 690, 1997, Rolling et al., Hum Gene Ther 10:641 648, 1999; Ali et al., Hum Mol Genet 5:591 594, 1996; Srivastava in WO 93/09239, Samulski et al., J. Vir. (1989) 63:3822-3828; Mendelson et al., Virol. (1988) 166:154-165; and Flotte et al., PNAS (1993) 90:10613-10617); SV40; herpes simplex virus; human immunodeficiency virus (see, e.g., Miyoshi et al., PNAS 94:10319 23, 1997; Takahashi et al., J Virol 73:7812 7816, 1999); a retroviral vector (e.g., Murine Leukemia Virus, spleen necrosis virus, and vectors derived from retroviruses such as Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, a lentivirus, human immunodeficiency virus, myeloproliferative sarcoma virus, and mammary tumor virus); and the like.

[0145] Numerous suitable expression vectors are known to those of skill in the art, and many are commercially available. The following vectors are provided by way of example; for eukaryotic host cells: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, and pSVLSV40 (Pharmacia). However, any other vector may be used so long as it is compatible with the host cell.

[0146] Depending on the host/vector system utilized, any of a number of suitable transcription and translation control elements, including constitutive and inducible promoters, transcription enhancer elements, transcription terminators, etc. may be used in the expression vector (see e.g., Bitter et al. (1987) Methods in Enzymology, 153:516-544).

[0147] In some embodiments, a nucleotide sequence encoding a single-molecule DNA-targeting RNA is operably linked to a control element, e.g., a transcriptional control element, such as a promoter. The transcriptional control element may be functional in either a eukaryotic cell, e.g., a mammalian cell; or a prokaryotic cell (e.g., bacterial or archaeal cell). In some embodiments, a nucleotide sequence encoding a single-molecule DNA-targeting RNA is

operably linked to multiple control elements that allow expression of the nucleotide sequence encoding a single-molecule DNA-targeting RNA in both prokaryotic and eukaryotic cells.

[0148] Non-limiting examples of suitable eukaryotic promoters (promoters functional in a eukaryotic cell) include those from cytomegalovirus (CMV) immediate early, herpes simplex virus (HSV) thymidine kinase, early and late SV40, long terminal repeats (LTRs) from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art. The expression vector may also contain a ribosome binding site for translation initiation and a transcription terminator. The expression vector may also include appropriate sequences for amplifying expression. The expression vector may also include nucleotide sequences encoding protein tags (e.g., 6xHis tag, hemagglutinin tag, green fluorescent protein, etc.) that are fused to the site-directed modifying polypeptide, thus resulting in a chimeric polypeptide.

[0149] In some embodiments, a nucleotide sequence encoding a single-molecule DNA-targeting RNA is operably linked to an inducible promoter. In some embodiments, a nucleotide sequence encoding a single-molecule DNA-targeting RNA is operably linked to a constitutive promoter.

[0150] Methods of introducing a nucleic acid into a host cell are known in the art, and any known method can be used to introduce a nucleic acid (e.g., an expression construct) into a cell. Suitable methods include e.g., viral or bacteriophage infection, transfection, conjugation, protoplast fusion, lipofection, electroporation, calcium phosphate precipitation, polyethyleneimine (PEI)-mediated transfection, DEAE-dextran mediated transfection, liposome-mediated transfection, particle gun technology, calcium phosphate precipitation, direct micro injection, nanoparticle-mediated nucleic acid delivery (see, e.g., Panyam et., al Adv Drug Deliv Rev. 2012 Sep 13. pii: S0169-409X(12)00283-9. doi: 10.1016/j.addr.2012.09.023), and the like.

[0151] As discussed above, a subject DNA-targeting RNA and a naturally-occurring Cas9 polypeptide form a complex. The DNA-targeting RNA provides target specificity to the complex by comprising a nucleotide sequence that is complementary to a sequence of a target DNA. The naturally-occurring Cas9 polypeptide of the complex provides the site-specific activity. A subject complex modifies a target DNA, leading to DNA cleavage, The target DNA may be, for example, chromosomal DNA in cells *in vitro,* etc.

[0152] In some cases, the site-directed modifying polypeptide exhibits nuclease activity that cleaves target DNA at a target DNA sequence defined by the region of complementarity between the DNA-targeting RNA and the target DNA. The site-directed modifying polypeptide is a naturally-occurring Cas9 polypeptide, and site-specific cleavage of the target DNA occurs at locations determined by both (i) base-pairing complementarity between the DNA-targeting RNA and the target DNA; and (ii) a short motif [referred to as the protospacer adjacent motif (PAM)] in the target DNA. In some embodiments (e.g., when Cas9 from S. pyogenes, or a closely related Cas9, is used (see SEQ ID NOs: 1-256 and 795-1346)), the PAM sequence of

the non-complementary strand is 5'-XGG-3', where X is any DNA nucleotide and X is immediately 3' of the target sequence of the non-complementary strand of the target DNA (see Figure 10). As such, the PAM sequence of the complementary strand is 5'-CCY-3', where Y is any DNA nucleotide and Y is immediately 5' of the target sequence of the complementary strand of the target DNA (see Figure 10 where the PAM of the non-complementary strand is 5'-GGG-3' and the PAM of the complementary strand is 5'-CCC-3'). In some such embodiments, X and Y can be complementary and the X-Y base pair can be any basepair (e.g., X=C and Y=G; X=G and Y=C; X=A and Y=T, X=T and Y=A).

[0153] In some cases, different Cas9 proteins (i.e., Cas9 proteins from various species) may be advantageous to use in the various provided methods in order to capitalize on various enzymatic characteristics of the different Cas9 proteins (e.g., for different PAM sequence preferences; for increased or decreased enzymatic activity; for an increased or decreased level of cellular toxicity; to change the balance between NHEJ, homology-directed repair, single strand breaks, double strand breaks, etc.). Cas9 proteins from various species (see SEQ ID NOs: 1-256 and 795-1346) may require different PAM sequences in the target DNA. Thus, for a particular Cas9 protein of choice, the PAM sequence requirement may be different than the 5'-XGG-3' sequence described above.

[0154] Many Cas9 orthologues from a wide variety of species have been identified herein and the proteins share only a few identical amino acids. All identified Cas9 orthologs have the same domain architecture with a central HNH endonuclease domain and a split RuvC/RNaseH domain (See Figures 3A, 3B, Figure 5, and Table 1). Cas9 proteins share 4 key motifs with a conserved architecture. Motifs 1, 2, and 4 are RuvC like motifs while motif 3 is an HNH-motif.

[0155] The nuclease activity cleaves target DNA to produce double strand breaks. These breaks are then repaired by the cell in one of two ways: non-homologous end joining, and homology-directed repair (Figure 2). In non-homologous end joining (NHEJ), the double-strand breaks are repaired by direct ligation of the break ends to one another. As such, no new nucleic acid material is inserted into the site, although some nucleic acid material may be lost, resulting in a deletion. In homology-directed repair, a donor polynucleotide with homology to the cleaved target DNA sequence is used as a template for repair of the cleaved target DNA sequence, resulting in the transfer of genetic information from the donor polynucleotide to the target DNA. As such, new nucleic acid material may be inserted/copied into the site. In some cases, a target DNA is contacted with a subject donor polynucleotide. The modifications of the target DNA due to NHEJ and/or homology-directed repair lead to, for example, gene correction, gene replacement, gene tagging, transgene insertion, nucleotide deletion, gene disruption, gene mutation, etc.

[0156] Accordingly, cleavage of DNA by a naturally-occurring Cas9 polypeptide may be used to delete nucleic acid material from a target DNA sequence (e.g., to disrupt a gene that makes cells susceptible to infection (e.g. the CCR5 or CXCR4 gene, which makes T cells susceptible to HIV infection), to remove disease-causing trinucleotide repeat sequences in neurons, to create gene knockouts and mutations as disease models in research, etc.) by cleaving the

target DNA sequence and allowing the cell to repair the sequence in the absence of an exogenously provided donor polynucleotide.

[0157] In some embodiments, a subject naturally-occurring Cas9 polypeptide can be codon-optimized. This type of optimization is known in the art and entails the mutation of foreign-derived DNA to mimic the codon preferences of the intended host organism or cell while encoding the same protein. Thus, the codons are changed, but the encoded protein remains unchanged. For example, if the intended target cell was a human cell, a human codon-optimized Cas9 (or variant, e.g., enzymatically inactive variant) would be a suitable site-directed modifying polypeptide (see SEQ ID NO:256 for an example). Any suitable site-directed modifying polypeptide (e.g., any naturally-occurring Cas9 such as any of the sequences set forth in SEQ ID NOs: 1-256 and 795-1346) can be codon optimized. As another non-limiting example, if the intended host cell were a mouse cell, than a mouse codon-optimized Cas9 (or variant, e.g., enzymatically inactive variant) would be a suitable site-directed modifying polypeptide. While codon optimization is not required, it is acceptable and may be preferable in certain cases.

Nucleic acids encoding a subject single-molecule DNA-targeting RNA

[0158] In some embodiments, a subject method involves contacting a target DNA or introducing into a cell (or a population of cells) one or more nucleic acids comprising nucleotide sequences encoding a single-molecule DNA-targeting RNA. Suitable nucleic acids comprising nucleotide sequences encoding a single-molecule DNA-targeting RNA include expression vectors, where an expression vector comprising a nucleotide sequence encoding a single-molecule DNA-targeting RNA is a "recombinant expression vector."

[0159] In some embodiments, the recombinant expression vector is a viral construct, e.g., a recombinant adeno-associated virus construct (see, e.g., U.S. Patent No. 7,078,387), a recombinant adenoviral construct, a recombinant lentiviral construct, etc.

[0160] Suitable expression vectors include, but are not limited to, viral vectors (e.g. viral vectors based on vaccinia virus; poliovirus; adenovirus (see, e.g., Li et al., Invest Opthalmol Vis Sci 35:2543 2549, 1994; Borras et al., Gene Ther 6:515 524, 1999; Li and Davidson, PNAS 92:7700 7704, 1995; Sakamoto et al., H Gene Ther 5:1088 1097, 1999; WO 94/12649, WO 93/03769; WO 93/19191; WO 94/28938; WO 95/11984 and WO 95/00655); adeno-associated virus (see, e.g., Ali et al., Hum Gene Ther 9:81 86, 1998, Flannery et al., PNAS 94:6916 6921, 1997; Bennett et al., Invest Opthalmol Vis Sci 38:2857 2863, 1997; Jomary et al., Gene Ther 4:683 690, 1997, Rolling et al., Hum Gene Ther 10:641 648, 1999; Ali et al., Hum Mol Genet 5:591 594, 1996; Srivastava in WO 93/09239, Samulski et al., J. Vir. (1989) 63:3822-3828; Mendelson et al., Virol. (1988) 166:154-165; and Flotte et al., PNAS (1993) 90:10613-10617); SV40; herpes simplex virus; human immunodeficiency virus (see, e.g., Miyoshi et al., PNAS 94:10319 23, 1997; Takahashi et al., J Virol 73:7812 7816, 1999); a retroviral vector (e.g., Murine Leukemia Virus, spleen necrosis virus, and vectors derived from retroviruses such as

Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, a lentivirus, human immunodeficiency virus, myeloproliferative sarcoma virus, and mammary tumor virus); and the like.

[0161] Numerous suitable expression vectors are known to those of skill in the art, and many are commercially available. The following vectors are provided by way of example; for eukaryotic host cells: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, and pSVLSV40 (Pharmacia). However, any other vector may be used so long as it is compatible with the host cell.

[0162] In some embodiments, a nucleotide sequence encoding a single-molecule DNA-targeting RNA is operably linked to a control element, e.g., a transcriptional control element, such as a promoter. The transcriptional control element may be functional in either a eukaryotic cell, e.g., a mammalian cell, or a prokaryotic cell (e.g., bacterial or archaeal cell). In some embodiments, a nucleotide sequence encoding a single-molecule DNA-targeting RNA is operably linked to multiple control elements that allow expression of the nucleotide sequence encoding a DNA-targeting RNA and/or a site-directed modifying polypeptide in both prokaryotic and eukaryotic cells.

[0163] Depending on the host/vector system utilized, any of a number of suitable transcription and translation control elements, including constitutive and inducible promoters, transcription enhancer elements, transcription terminators, etc. may be used in the expression vector (e.g., U6 promoter, H1 promoter, etc.; see above) (see e.g., Bitter et al. (1987) Methods in Enzymology, 153:516-544).

[0164] Nucleotides encoding a single-molecule DNA-targeting RNA (introduced as DNA) may be provided to the cells using well-developed transfection techniques; see, e.g. Angel and Yanik (2010) PLoS ONE 5(7): e11756, and the commercially available TransMessenger® reagents from Qiagen, Stemfect™ RNA Transfection Kit from Stemgent, and TransIT®-mRNA Transfection Kit from Mirus Bio LLC. See also Beumer et al. (2008) Efficient gene targeting in Drosophila by direct embryo injection with zinc-finger nucleases. PNAS 105(50):19821-19826. Alternatively, nucleic acids encoding a single-molecule DNA-targeting RNA may be provided on DNA vectors. Many vectors, e.g. plasmids, cosmids, minicircles, phage, viruses, etc., useful for transferring nucleic acids into target cells are available. The vectors comprising the nucleic acid(s) may be maintained episomally, e.g. as plasmids, minicircle DNAs, viruses such cytomegalovirus, adenovirus, etc., or they may be integrated into the target cell genome, through homologous recombination or random integration, e.g. retrovirus-derived vectors such as MMLV, HIV-1, ALV, etc.

[0165] Vectors may be provided directly to the subject cells. In other words, the cells are contacted with vectors comprising the nucleic acid encoding single-molecule DNA-targeting RNA such that the vectors are taken up by the cells. Methods for contacting cells with nucleic acid vectors that are plasmids, including electroporation, calcium chloride transfection, microinjection, and lipofection are well known in the art. For viral vector delivery, the cells are

contacted with viral particles comprising the nucleic acid encoding a single-molecule DNA-targeting RNA. Retroviruses, for example, lentiviruses, are particularly suitable to the method of the invention. Commonly used retroviral vectors are "defective", i.e. unable to produce viral proteins required for productive infection. Rather, replication of the vector requires growth in a packaging cell line. To generate viral particles comprising nucleic acids of interest, the retroviral nucleic acids comprising the nucleic acid are packaged into viral capsids by a packaging cell line. Different packaging cell lines provide a different envelope protein (ecotropic, amphotropic or xenotropic) to be incorporated into the capsid, this envelope protein determining the specificity of the viral particle for the cells (ecotropic for murine and rat; amphotropic for most mammalian cell types including human, dog and mouse; and xenotropic for most mammalian cell types except murine cells). The appropriate packaging cell line may be used to ensure that the cells are targeted by the packaged viral particles. Methods of introducing the retroviral vectors comprising the nucleic acid encoding the reprogramming factors into packaging cell lines and of collecting the viral particles that are generated by the packaging lines are well known in the art. Nucleic acids can also introduced by direct micro-injection (e.g., injection of RNA into a zebrafish embryo).

[0166] Vectors used for providing the nucleic acids encoding single-molecule DNA-targeting RNA to the subject cells will typically comprise suitable promoters for driving the expression, that is, transcriptional activation, of the nucleic acid of interest. In other words, the nucleic acid of interest will be operably linked to a promoter. This may include ubiquitously acting promoters, for example, the CMV-β-actin promoter, or inducible promoters, such as promoters that are active in particular cell populations or that respond to the presence of drugs such as tetracycline. By transcriptional activation, it is intended that transcription will be increased above basal levels in the target cell by at least about 10 fold, by at least about 100 fold, more usually by at least about 1000 fold. In addition, vectors used for providing a single-molecule DNA-targeting RNA to the subject cells may include nucleic acid sequences that encode for selectable markers in the target cells, so as to identify cells that have taken up the single-molecule DNA-targeting RNA.

[0167] A subject DNA-targeting RNA may instead be used to contact DNA or introduced into cells as RNA. Methods of introducing RNA into cells are known in the art and may include, for example, direct injection, transfection, or any other method used for the introduction of DNA.

[0168] Also included in the subject invention are DNA-targeting RNAs that have been modified using ordinary molecular biological techniques and synthetic chemistry so as , to change the target sequence specificity, to optimize solubility properties, or to render them more suitable as a therapeutic agent.

[0169] The naturally-occurring Cas9 endonuclease may also be isolated and purified in accordance with conventional methods of recombinant synthesis. A lysate may be prepared of the expression host and the lysate purified using HPLC, exclusion chromatography, gel electrophoresis, affinity chromatography, or other purification technique. For the most part, the compositions which are used will comprise at least 20% by weight of the desired product, more

usually at least about 75% by weight, preferably at least about 95% by weight, and for therapeutic purposes, usually at least about 99.5% by weight, in relation to contaminants related to the method of preparation of the product and its purification. Usually, the percentages will be based upon total protein.

[0170] To induce DNA cleavage and recombination, or any desired modification to a target DNA the DNA-targeting RNA and/or the naturally-occurring Cas9 endonuclease and/or the donor polynucleotide, whether they be introduced as nucleic acids or polypeptides, are provided to the cells for about 30 minutes to about 24 hours, e.g., 1 hour, 1.5 hours, 2 hours, 2.5 hours, 3 hours, 3.5 hours 4 hours, 5 hours, 6 hours, 7 hours, 8 hours, 12 hours, 16 hours, 18 hours, 20 hours, or any other period from about 30 minutes to about 24 hours, which may be repeated with a frequency of about every day to about every 4 days, e.g., every 1.5 days, every 2 days, every 3 days, or any other frequency from about every day to about every four days. The agent(s) may be provided to the cells one or more times, e.g. one time, twice, three times, or more than three times, and the cells allowed to incubate with the agent(s) for some amount of time following each contacting event e.g. 16-24 hours, after which time the media is replaced with fresh media and the cells are cultured further.

[0171] In cases in which two or more different targeting complexes are provided to the cell (e.g., two different DNA-targeting RNAs that are complementary to different sequences within the same or different target DNA), the complexes may be provided simultaneously (e.g. as two polypeptides and/or nucleic acids), or delivered simultaneously. Alternatively, they may be provided consecutively, e.g. the targeting complex being provided first, followed by the second targeting complex, etc. or vice versa.

[0172] Typically, an effective amount of the DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide is provided to the target DNA or cells to induce cleavage. An effective amount of the DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide is the amount to induce a 2--fold increase or more in the amount of target modification observed between two homologous sequences relative to a negative control, e.g. a cell contacted with an empty vector or irrelevant polypeptide. That is to say, an effective amount or dose of the DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide will induce a 2-fold increase, a 3-fold increase, a 4-fold increase or more in the amount of target modification observed at a target DNA region, in some instances a 5-fold increase, a 6-fold increase or more, sometimes a 7-fold or 8-fold increase or more in the amount of recombination observed, e.g. an increase of 10-fold, 50-fold, or 100-fold or more, in some instances, an increase of 200-fold, 500-fold, 700-fold, or 1000-fold or more, e.g. a 5000-fold, or 10,000-fold increase in the amount of recombination observed. The amount of target modification may be measured by any convenient method. For example, a silent reporter construct comprising complementary sequence to the targeting segment (targeting sequence) of the DNA-targeting RNA flanked by repeat sequences that, when recombined, will reconstitute a nucleic acid encoding an active reporter may be cotransfected into the cells, and the amount of reporter protein assessed after contact with the DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or

donor polynucleotide, e.g. 2 hours, 4 hours, 8 hours, 12 hours, 24 hours, 36 hours, 48 hours, 72 hours or more after contact with the DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide. As another, more sensitivity assay, for example, the extent of recombination at a genomic DNA region of interest comprising target DNA sequences may be assessed by PCR or Southern hybridization of the region after contact with a DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide, e.g. 2 hours, 4 hours, 8 hours, 12 hours, 24 hours, 36 hours, 48 hours, 72 hours or more after contact with the DNA-targeting RNA and/or site-directed modifying polypeptide and/or donor polynucleotide.

[0173] Contacting the cells with a DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide may occur in any culture media and under any culture conditions that promote the survival of the cells. For example, cells may be suspended in any appropriate nutrient medium that is convenient, such as Iscove's modified DMEM or RPMI 1640, supplemented with fetal calf serum or heat inactivated goat serum (about 5-10%), L-glutamine, a thiol, particularly 2-mercaptoethanol, and antibiotics, e.g. penicillin and streptomycin. The culture may contain growth factors to which the cells are responsive. Growth factors, as defined herein, are molecules capable of promoting survival, growth and/or differentiation of cells, either in culture or in the intact tissue, through specific effects on a transmembrane receptor. Growth factors include polypeptides and non-polypeptide factors. Conditions that promote the survival of cells are typically permissive of nonhomologous end joining and homology-directed repair.

[0174] In applications in which it is desirable to insert a polynucleotide sequence into a target DNA sequence, a polynucleotide comprising a donor sequence to be inserted is also provided to the cell. By a "donor sequence" or "donor polynucleotide" it is meant a nucleic acid sequence to be inserted at the cleavage site induced by a site-directed modifying polypeptide. The donor polynucleotide will contain sufficient homology to a genomic sequence at the cleavage site, e.g. 70%, 80%, 85%, 90%, 95%, or 100% homology with the nucleotide sequences flanking the cleavage site, e.g. within about 50 bases or less of the cleavage site, e.g. within about 30 bases, within about 15 bases, within about 10 bases, within about 5 bases, or immediately flanking the cleavage site, to support homology-directed repair between it and the genomic sequence to which it bears homology. Approximately 25, 50, 100, or 200 nucleotides, or more than 200 nucleotides, of sequence homology between a donor and a genomic sequence (or any integral value between 10 and 200 nucleotides, or more) will support homology-directed repair. Donor sequences can be of any length, e.g. 10 nucleotides or more, 50 nucleotides or more, 100 nucleotides or more, 250 nucleotides or more, 500 nucleotides or more, 1000 nucleotides or more, 5000 nucleotides or more, etc.

[0175] The donor sequence is typically not identical to the genomic sequence that it replaces. Rather, the donor sequence may contain at least one or more single base changes, insertions, deletions, inversions or rearrangements with respect to the genomic sequence, so long as sufficient homology is present to support homology-directed repair. In some embodiments, the donor sequence comprises a non-homologous sequence flanked by two regions of homology,

such that homology-directed repair between the target DNA region and the two flanking sequences results in insertion of the non-homologous sequence at the target region. Donor sequences may also comprise a vector backbone containing sequences that are not homologous to the DNA region of interest and that are not intended for insertion into the DNA region of interest. Generally, the homologous region(s) of a donor sequence will have at least 50% sequence identity to a genomic sequence with which recombination is desired. In certain embodiments, 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% sequence identity is present. Any value between 1% and 100% sequence identity can be present, depending upon the length of the donor polynucleotide.

[0176] The donor sequence may comprise certain sequence differences as compared to the genomic sequence, e.g. restriction sites, nucleotide polymorphisms, selectable markers (e.g., drug resistance genes, fluorescent proteins, enzymes etc.), etc., which may be used to assess for successful insertion of the donor sequence at the cleavage site or in some cases may be used for other purposes (e.g., to signify expression at the targeted genomic locus). In some cases, if located in a coding region, such nucleotide sequence differences will not change the amino acid sequence, or will make silent amino acid changes (i.e., changes which do not affect the structure or function of the protein). Alternatively, these sequences differences may include flanking recombination sequences such as FLPs, loxP sequences, or the like, that can be activated at a later time for removal of the marker sequence.

[0177] The donor sequence may be provided to the cell as single-stranded DNA, single-stranded RNA, double-stranded DNA, or double-stranded RNA. It may be introduced into a cell in linear or circular form. If introduced in linear form, the ends of the donor sequence may be protected (e.g., from exonucleolytic degradation) by methods known to those of skill in the art. For example, one or more dideoxynucleotide residues are added to the 3' terminus of a linear molecule and/or self-complementary oligonucleotides are ligated to one or both ends. See, for example, Chang et al. (1987) Proc. Natl. Acad Sci USA 84:4959-4963; Nehls et al. (1996) Science 272:886-889. Additional methods for protecting exogenous polynucleotides from degradation include, but are not limited to, addition of terminal amino group(s) and the use of modified internucleotide linkages such as, for example, phosphorothioates, phosphoramidates, and O-methyl ribose or deoxyribose residues. As an alternative to protecting the termini of a linear donor sequence, additional lengths of sequence may be included outside of the regions of homology that can be degraded without impacting recombination. A donor sequence can be introduced into a cell as part of a vector molecule having additional sequences such as, for example, replication origins, promoters and genes encoding antibiotic resistance. Moreover, donor sequences can be introduced as naked nucleic acid, as nucleic acid complexed with an agent such as a liposome or poloxamer, or can be delivered by viruses (e.g., adenovirus, AAV), as described above for nucleic acids encoding a DNA-targeting RNA and/or site-directed modifying polypeptide and/or donor polynucleotide.

[0178] Pharmaceutical preparations are compositions that include one or more a DNA-targeting RNA and/or site-directed modifying polypeptide and/or donor polynucleotide present in a pharmaceutically acceptable vehicle. "Pharmaceutically acceptable vehicles" may be

vehicles approved by a regulatory agency of the Federal or a state government or listed in the U.S. Pharmacopeia or other generally recognized pharmacopeia for use in mammals, such as humans. The term "vehicle" refers to a diluent, adjuvant, excipient, or carrier with which a compound of the invention is formulated for administration to a mammal. Such pharmaceutical vehicles can be lipids, e.g. liposomes, e.g. liposome dendrimers; liquids, such as water and oils, including those of petroleum, animal, vegetable or synthetic origin, such as peanut oil, soybean oil, mineral oil, sesame oil and the like, saline; gum acacia, gelatin, starch paste, talc, keratin, colloidal silica, urea, and the like. In addition, auxiliary, stabilizing, thickening, lubricating and coloring agents may be used. Pharmaceutical compositions may be formulated into preparations in solid, semisolid, liquid or gaseous forms, such as tablets, capsules, powders, granules, ointments, solutions, suppositories, injections, inhalants, gels, microspheres, and aerosols. As such, administration of the a DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide can be achieved in various ways, including oral, buccal, rectal, parenteral, intraperitoneal, intradermal, transdermal, intratracheal, intraocular, etc., administration. The active agent may be systemic after administration or may be localized by the use of regional administration, intramural administration, or use of an implant that acts to retain the active dose at the site of implantation. The active agent may be formulated for immediate activity or it may be formulated for sustained release.

[0179] For some conditions, particularly central nervous system conditions, it may be necessary to formulate agents to cross the blood-brain barrier (BBB). One strategy for drug delivery through the blood-brain barrier (BBB) entails disruption of the BBB, either by osmotic means such as mannitol or leukotrienes, or biochemically by the use of vasoactive substances such as bradykinin. The potential for using BBB opening to target specific agents to brain tumors is also an option. A BBB disrupting agent can be co-administered with the therapeutic compositions of the invention when the compositions are administered by intravascular injection. Other strategies to go through the BBB may entail the use of endogenous transport systems, including Caveolin-1 mediated transcytosis, carrier-mediated transporters such as glucose and amino acid carriers, receptor-mediated transcytosis for insulin or transferrin, and active efflux transporters such as p-glycoprotein. Active transport moieties may also be conjugated to the therapeutic compounds for use in the invention to facilitate transport across the endothelial wall of the blood vessel. Alternatively, drug delivery of therapeutics agents behind the BBB may be by local delivery, for example by intrathecal delivery, e.g. through an Ommaya reservoir (see e.g. US Patent Nos. 5,222,982 and 5385582); by bolus injection, e.g. by a syringe, e.g. intravitreally or intracranially; by continuous infusion, e.g. by cannulation, e.g. with convection (see e.g. US Application No. 20070254842); or by implanting a device upon which the agent has been reversibly affixed (see e.g. US Application Nos. 20080081064 and 20090196903).

[0180] Typically, an effective amount of a DNA-targeting RNA and/or site-directed modifying polypeptide and/or donor polynucleotide are provided. As discussed above with regard to non-claimed ex vivo methods, an effective amount or effective dose of a DNA-targeting RNA and/or site-directed modifying polypeptide and/or donor polynucleotide in vivo is the amount to induce

a 2 fold increase or more in the amount of recombination observed between two homologous sequences relative to a negative control, e.g. a cell contacted with an empty vector or irrelevant polypeptide. The amount of recombination may be measured by any convenient method, e.g. as described above and known in the art. The calculation of the effective amount or effective dose of a DNA-targeting RNA and/or site-directed modifying polypeptide and/or donor polynucleotide to be administered is within the skill of one of ordinary skill in the art, and will be routine to those persons skilled in the art. The final amount to be administered will be dependent upon the route of administration and upon the nature of the disorder or condition that is to be treated.

[0181] The effective amount given to a particular patient will depend on a variety of factors, several of which will differ from patient to patient. A competent clinician will be able to determine an effective amount of a therapeutic agent to administer to a patient to halt or reverse the progression the disease condition as required. Utilizing LD50 animal data, and other information available for the agent, a clinician can determine the maximum safe dose for an individual, depending on the route of administration. For instance, an intravenously administered dose may be more than an intrathecally administered dose, given the greater body of fluid into which the therapeutic composition is being administered. Similarly, compositions which are rapidly cleared from the body may be administered at higher doses, or in repeated doses, in order to maintain a therapeutic concentration. Utilizing ordinary skill, the competent clinician will be able to optimize the dosage of a particular therapeutic in the course of routine clinical trials.

[0182] For inclusion in a medicament, a DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide may be obtained from a suitable commercial source. As a general proposition, the total pharmaceutically effective amount of the a DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide administered parenterally per dose will be in a range that can be measured by a dose response curve.

[0183] Therapies based on a DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotides, i.e. preparations of a DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide to be used for therapeutic administration, must be sterile. Sterility is readily accomplished by filtration through sterile filtration membranes (e.g., 0.2 μm membranes). Therapeutic compositions generally are placed into a container having a sterile access port, for example, an intravenous solution bag or vial having a stopper pierceable by a hypodermic injection needle. The therapies based on a DNA-targeting RNA and/or naturally-occurring Cas9 endonuclease and/or donor polynucleotide may be stored in unit or multi-dose containers, for example, sealed ampules or vials, as an aqueous solution or as a lyophilized formulation for reconstitution. As an example of a lyophilized formulation, 10-mL vials are filled with 5 ml of sterile-filtered 1% (w/v) aqueous solution of compound, and the resulting mixture is lyophilized. The infusion solution is prepared by reconstituting the lyophilized compound using bacteriostatic Water-for-Injection.

[0184] Pharmaceutical compositions can include, depending on the formulation desired, pharmaceutically-acceptable, non-toxic carriers of diluents, which are defined as vehicles commonly used to formulate pharmaceutical compositions for animal or human administration. The diluent is selected so as not to affect the biological activity of the combination. Examples of such diluents are distilled water, buffered water, physiological saline, PBS, Ringer's solution, dextrose solution, and Hank's solution. In addition, the pharmaceutical composition or formulation can include other carriers, adjuvants, or non-toxic, nontherapeutic, nonimmunogenic stabilizers, excipients and the like. The compositions can also include additional substances to approximate physiological conditions, such as pH adjusting and buffering agents, toxicity adjusting agents, wetting agents and detergents.

[0185] The composition can also include any of a variety of stabilizing agents, such as an antioxidant for example. When the pharmaceutical composition includes a polypeptide, the polypeptide can be complexed with various well-known compounds that enhance the in vivo stability of the polypeptide, or otherwise enhance its pharmacological properties (e.g., increase the half-life of the polypeptide, reduce its toxicity, enhance solubility or uptake). Examples of such modifications or complexing agents include sulfate, gluconate, citrate and phosphate. The nucleic acids or polypeptides of a composition can also be complexed with molecules that enhance their in vivo attributes. Such molecules include, for example, carbohydrates, polyamines, amino acids, other peptides, ions (e.g., sodium, potassium, calcium, magnesium, manganese), and lipids.

[0186] Further guidance regarding formulations that are suitable for various types of administration can be found in Remington's Pharmaceutical Sciences, Mace Publishing Company, Philadelphia, Pa., 17th ed. (1985). For a brief review of methods for drug delivery, see, Langer, Science 249:1527-1533 (1990).

[0187] The pharmaceutical compositions can be administered for prophylactic and/or therapeutic treatments. Toxicity and therapeutic efficacy of the active ingredient can be determined according to standard pharmaceutical procedures in cell cultures and/or experimental animals, including, for example, determining the LD50 (the dose lethal to 50% of the population) and the ED50 (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD50/ED50. Therapies that exhibit large therapeutic indices are preferred.

[0188] The data obtained from cell culture and/or animal studies can be used in formulating a range of dosages for humans. The dosage of the active ingredient typically lines within a range of circulating concentrations that include the ED50 with low toxicity. The dosage can vary within this range depending upon the dosage form employed and the route of administration utilized.

[0189] The components used to formulate the pharmaceutical compositions are preferably of high purity and are substantially free of potentially harmful contaminants (e.g., at least National Food (NF) grade, generally at least analytical grade, and more typically at least pharmaceutical

grade). Moreover, compositions intended for in vivo use are usually sterile. To the extent that a given compound must be synthesized prior to use, the resulting product is typically substantially free of any potentially toxic agents, particularly any endotoxins, which may be present during the synthesis or purification process. Compositions for parental administration are also sterile, substantially isotonic and made under GMP conditions.

[0190] The effective amount of a therapeutic composition to be given to a particular patient will depend on a variety of factors, several of which will differ from patient to patient. A competent clinician will be able to determine an effective amount of a therapeutic agent to administer to a patient to halt or reverse the progression the disease condition as required. Utilizing LD50 animal data, and other information available for the agent, a clinician can determine the maximum safe dose for an individual, depending on the route of administration. For instance, an intravenously administered dose may be more than an intrathecally administered dose, given the greater body of fluid into which the therapeutic composition is being administered. Similarly, compositions which are rapidly cleared from the body may be administered at higher doses, or in repeated doses, in order to maintain a therapeutic concentration. Utilizing ordinary skill, the competent clinician will be able to optimize the dosage of a particular therapeutic in the course of routine clinical trials.

COMPOSITIONS

[0191] The present invention provides a composition comprising a subject DNA-targeting RNA and a naturally-occurring Cas9 endonuclease. A subject composition is useful for carrying out a non-claimed method of the present disclosure, e.g., a non-claimed method for site-specific modification of a target DNA; etc.

Compositions comprising a DNA-targeting RNA and a site-directed modifying polypeptide

[0192] The present invention provides a composition comprising: (i) a DNA-targeting RNA or a DNA polynucleotide encoding the same; and ii) a naturally-occurring Cas9 endonuclease. In some cases

[0193] The present invention provides a composition comprising: (i) a DNA-targeting RNA, as described above, the DNA-targeting RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide which is a naturally-occurring Cas9 endonuclease; and (ii) the site-directed modifying polypeptide, or a polynucleotide encoding the same, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the DNA-targeting RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the DNA-

targeting RNA.

[0194] In some instances, a subject composition comprises: a composition comprising: (i) a subject DNA-targeting RNA, the DNA-targeting RNA comprising: (a) a first segment comprising a nucleotide sequence that is complementary to a sequence in a target DNA; and (b) a second segment that interacts with a site-directed modifying polypeptide which is a naturally-occurring Cas9 endonuclease; and (ii) the site-directed modifying polypeptide, the site-directed modifying polypeptide comprising: (a) an RNA-binding portion that interacts with the DNA-targeting RNA; and (b) an activity portion that exhibits site-directed enzymatic activity, wherein the site of enzymatic activity is determined by the DNA-targeting RNA.

[0195] In some embodiments, a subject composition includes both RNA molecules of a double-molecule DNA-targeting RNA. As such, in some embodiments, a subject composition includes an activator-RNA that comprises a duplex-forming segment that is complementary to the duplex-forming segment of a targeter-RNA (see Figure 1A). The duplex-forming segments of the activator-RNA and the targeter-RNA hybridize to form the dsRNA duplex of the protein-binding segment of the DNA-targeting RNA. The targeter-RNA further provides the DNA-targeting segment (single stranded) of the DNA-targeting RNA and therefore targets the DNA-targeting RNA to a specific sequence within the target DNA. As one non-limiting example, the duplex-forming segment of the activator-RNA comprises a nucleotide sequence that has at least about 70%, at least about 80%, at least about 90%, at least about 95%, at least about 98%, or 100% identity with the sequence 5'-UAGCAAGUUAAAAU-3' (SEQ ID NO:562). As another non-limiting example, the duplex-forming segment of the targeter-RNA comprises a nucleotide sequence that has at least about 70%, at least about 80%, at least about 90%, at least about 95%, at least about 98%, or 100% identity with the sequence 5'-GUUUUAGAGCUA-3' (SEQ ID NO:679).

[0196] A subject composition can comprise, in addition to i) a subject DNA-targeting RNA; and ii) a which is a naturally-occurring Cas9 endonuclease, one or more of: a salt, e.g., NaCl, $MgCl_2$, KCl, $MgSO_4$, etc.; a buffering agent, e.g., a Tris buffer, HEPES, MES, MES sodium salt, MOPS, TAPS, etc.; a solubilizing agent; a detergent, e.g., a non-ionic detergent such as Tween-20, etc.; a protease inhibitor; a reducing agent (e.g., dithiothreitol); and the like.

[0197] In some cases, the components of the composition are individually pure, e.g., each of the components is at least about 75%, at least about 80%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or at least 99%, pure. In some cases, the individual components of a subject composition are pure before being added to the composition.

[0198] For example, in some embodiments, a site-directed modifying polypeptide present in a subject composition is pure, e.g., at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 98%, at least about 99%, or more than 99% pure, where "% purity" means that the site-directed modifying polypeptide is the recited percent free from other proteins (e.g., proteins other than the site-directed modifying

polypeptide), other macromolecules, or contaminants that may be present during the production of the site-directed modifying polypeptide.

DEFINITIONS - PART II

[0199] The term "naturally-occurring" or "unmodified" as used herein as applied to a nucleic acid, a polypeptide, a cell, or an organism, refers to a nucleic acid, polypeptide, cell, or organism that is found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by a human in the laboratory is naturally occurring.

[0200] "Heterologous," as used herein, means a nucleotide or polypeptide sequence that is not found in the native nucleic acid or protein, respectively. For example, in a fusion variant Cas9 site-directed polypeptide, a variant Cas9 site-directed polypeptide may be fused to a heterologous polypeptide (i.e. a polypeptide other than Cas9). The heterologous polypeptide may exhibit an activity (e.g., enzymatic activity) that will also be exhibited by the fusion variant Cas9 site-directed polypeptide. A heterologous nucleic acid sequence may be linked to a variant Cas9 site-directed polypeptide (e.g., by genetic engineering) to generate a nucleotide sequence encoding a fusion variant Cas9 site-directed polypeptide.

[0201] The term "chimeric polypeptide" refers to a polypeptide which is not naturally occurring, e.g., is made by the artificial combination of two otherwise separated segments of amino sequence through human intervention. Thus, a chimeric polypeptide is also the result of human intervention. Thus, a polypeptide that comprises a chimeric amino acid sequence is a chimeric polypeptide.

[0202] By "site-directed polypeptide" or "RNA-binding site-directed polypeptide" or "RNA-binding site-directed polypeptide" it is meant a polypeptide that binds RNA and is targeted to a specific DNA sequence. A site-directed polypeptide as described herein is targeted to a specific DNA sequence by the RNA molecule to which it is bound. The RNA molecule comprises a sequence that is complementary to a target sequence within the target DNA, thus targeting the bound polypeptide to a specific location within the target DNA (the target sequence).

[0203] In some embodiments, a subject nucleic acid (e.g., a DNA-targeting RNA, a nucleic acid comprising a nucleotide sequence encoding a single.molecule DNA-targeting RNA; etc.) comprises a modification or sequence that provides for an additional desirable feature (e.g., modified or regulated stability; subcellular targeting; tracking, e.g., a fluorescent label; a binding site for a protein or protein complex; etc.). Non-limiting examples include: a 5' cap (e.g., a 7-methylguanylate cap (m^7G)); a 3' polyadenylated tail (i.e., a 3' poly(A) tail); a riboswitch sequence (e.g., to allow for regulated stability and/or regulated accessibility by proteins and/or protein complexes); a modification or sequence that targets the RNA to a subcellular location (e.g., nucleus, mitochondria, chloroplasts, and the like); a modification or

sequence that provides for tracking (e.g., direct conjugation to a fluorescent molecule, conjugation to a moiety that facilitates fluorescent detection, a sequence that allows for fluorescent detection, etc.); a modification or sequence that provides a binding site for proteins (e.g., proteins that act on DNA, including transcriptional activators, transcriptional repressors, DNA methyltransferases, DNA demethylases, histone acetyltransferases, histone deacetylases, and the like); and combinations thereof.

[0204] In some embodiments, a DNA-targeting RNA comprises an additional segment at either the 5' or 3' end that provides for any of the features described above. For example, a suitable third segment can comprise a 5' cap (e.g., a 7-methylguanylate cap (m^7G)); a 3' polyadenylated tail (i.e., a 3' poly(A) tail); a riboswitch sequence (e.g., to allow for regulated stability and/or regulated accessibility by proteins and protein complexes); a sequence that targets the RNA to a subcellular location (e.g., nucleus, mitochondria, chloroplasts, and the like); a modification or sequence that provides for tracking (e.g., direct conjugation to a fluorescent molecule, conjugation to a moiety that facilitates fluorescent detection, a sequence that allows for fluorescent detection, etc.); a modification or sequence that provides a binding site for proteins (e.g., proteins that act on DNA, including transcriptional activators, transcriptional repressors, DNA methyltransferases, DNA demethylases, histone acetyltransferases, histone deacetylases, and the like); and combinations thereof.

[0205] A subject DNA-targeting RNA and a subject site-directed polypeptide which is a naturally-occurring Cas9 endonuclease form a complex (i.e., bind via non-covalent interactions). The DNA-targeting RNA provides target specificity to the complex by comprising a nucleotide sequence that is complementary to a sequence of a target DNA. The site-directed polypeptide of the complex provides the site-specific activity. In other words, the site-directed polypeptide is guided to a target DNA sequence (e.g. a target sequence in a chromosomal nucleic acid; a target sequence in an extrachromosomal nucleic acid, e.g. an episomal nucleic acid, a minicircle, etc.; a target sequence in a mitochondrial nucleic acid; a target sequence in a chloroplast nucleic acid; a target sequence in a plasmid; etc.) by virtue of its association with the protein-binding segment of the DNA-targeting RNA.

[0206] In some embodiments, a subject DNA-targeting RNA comprises two separate RNA molecules (RNA polynucleotides) and is referred to herein as a "double-molecule DNA-targeting RNA" or a "two-molecule DNA-targeting RNA." In other embodiments, a subject DNA-targeting RNA is a single RNA molecule (single RNA polynucleotide) and is referred to herein as a "single-molecule DNA-targeting RNA.". If not otherwise specified, the term "DNA-targeting RNA" is inclusive, referring to both single-molecule DNA-targeting RNAs and double-molecule DNA-targeting RNAs.

[0207] A subject two-molecule DNA-targeting RNA comprises two separate RNA molecules (a "targeter-RNA" and an "activator-RNA"). Each of the two RNA molecules of a subject two-molecule DNA-targeting RNA comprises a stretch of nucleotides that are complementary to one another such that the complementary nucleotides of the two RNA molecules hybridize to form the double stranded RNA duplex of the protein-binding segment.

[0208] A subject single-molecule DNA-targeting RNA comprises two stretches of nucleotides (a targeter-RNA and an activator-RNA) that are complementary to one another, are covalently linked by intervening nucleotides ("linkers" or "linker nucleotides"), and hybridize to form the double stranded RNA duplex (dsRNA duplex) of the protein-binding segment, thus resulting in a stem-loop structure. The targeter-RNA and the activator-RNA can be covalently linked via the 3' end of the targeter-RNA and the 5' end of the activator-RNA. Alternatively, targeter-RNA and the activator-RNA can be covalently linked via the 5' end of the targeter-RNA and the 3' end of the activator-RNA.

[0209] An exemplary two-molecule DNA-targeting RNA comprises a crRNA-like ("CRISPR RNA" or "targeter-RNA" or "crRNA" or "crRNA repeat") molecule and a corresponding tracrRNA-like ("trans-acting CRISPR RNA" or "activator-RNA" or "tracrRNA") molecule. A crRNA-like molecule (targeter-RNA) comprises both the DNA-targeting segment (single stranded) of the DNA-targeting RNA and a stretch ("duplex-forming segment") of nucleotides that forms one half of the dsRNA duplex of the protein-binding segment of the DNA-targeting RNA. A corresponding tracrRNA-like molecule (activator-RNA) comprises a stretch of nucleotides (duplex-forming segment) that forms the other half of the dsRNA duplex of the protein-binding segment of the DNA-targeting RNA. In other words, a stretch of nucleotides of a crRNA-like molecule are complementary to and hybridize with a stretch of nucleotides of a tracrRNA-like molecule to form the dsRNA duplex of the protein-binding domain of the DNA-targeting RNA. As such, each crRNA-like molecule can be said to have a corresponding tracrRNA-like molecule. The crRNA-like molecule additionally provides the single stranded DNA-targeting segment. Thus, a crRNA-like and a tracrRNA-like molecule (as a corresponding pair) hybridize to form a DNA-targeting RNA. The exact sequence of a given crRNA or tracrRNA molecule is characteristic of the species in which the RNA molecules are found.

[0210] The term "activator-RNA" is used herein to mean a tracrRNA-like molecule of a double-molecule DNA-targeting RNA. The term "targeter-RNA" is used herein to mean a crRNA-like molecule of a double-molecule DNA-targeting RNA. The term "duplex-forming segment" is used herein to mean the stretch of nucleotides of an activator-RNA or a targeter-RNA that contributes to the formation of the dsRNA duplex by hybridizing to a stretch of nucleotides of a corresponding activator-RNA or targeter-RNA molecule. In other words, an activator-RNA comprises a duplex-forming segment that is complementary to the duplex-forming segment of the corresponding targeter-RNA. As such, an activator-RNA comprises a duplex-forming segment while a targeter-RNA comprises both a duplex-forming segment and the DNA-targeting segment of the DNA-targeting RNA. Therefore, a subject double-molecule DNA-targeting RNA can be comprised of any corresponding activator-RNA and targeter-RNA pair.

[0211] A two-molecule DNA-targeting RNA can be designed to allow for controlled (i.e., conditional) binding of a targeter-RNA with an activator-RNA. Because a two-molecule DNA-targeting RNA is not functional unless both the activator-RNA and the targeter-RNA are bound in a functional complex with Cas9, a two-molecule DNA-targeting RNA can be inducible (e.g., drug inducible) by rendering the binding between the activator-RNA and the targeter-RNA to

be inducible. As one non-limiting example, RNA aptamers can be used to regulate (i.e., control) the binding of the activator-RNA with the targeter-RNA. Accordingly, the activator-RNA and/or the targeter-RNA can comprise an RNA aptamer sequence.

[0212] RNA aptamers are known in the art and are generally a synthetic version of a riboswitch. The terms "RNA aptamer" and "riboswitch" are used interchangeably herein to encompass both synthetic and natural nucleic acid sequences that provide for inducible regulation of the structure (and therefore the availability of specific sequences) of the RNA molecule of which they are part. RNA aptamers usually comprise a sequence that folds into a particular structure (e.g., a hairpin), which specifically binds a particular drug (e.g., a small molecule). Binding of the drug causes a structural change in the folding of the RNA, which changes a feature of the nucleic acid of which the aptamer is a part. As non-limiting examples: (i) an activator-RNA with an aptamer may not be able to bind to the cognate targeter-RNA unless the aptamer is bound by the appropriate drug; (ii) a targeter-RNA with an aptamer may not be able to bind to the cognate activator-RNA unless the aptamer is bound by the appropriate drug; and (iii) a targeter-RNA and an activator-RNA, each comprising a different aptamer that binds a different drug, may not be able to bind to each other unless both drugs are present. As illustrated by these examples, a two-molecule DNA-targeting RNA can be designed to be inducible.

[0213] Examples of aptamers and riboswitches can be found, for example, in: Nakamura et al., Genes Cells. 2012 May;17(5):344-64; Vavalle et al., Future Cardiol. 2012 May;8(3):371-82; Citartan et al., Biosens Bioelectron. 2012 Apr 15;34(1):1-11; and Liberman et al., Wiley Interdiscip Rev RNA. 2012 May-Jun;3(3):369-84.

[0214] Non-limiting examples of nucleotide sequences that can be included in a two-molecule DNA-targeting RNA include targeter RNAs (e.g., SEQ ID NOs:566-567) that can pair with the duplex forming segment of any one of the activator RNAs set forth in SEQ ID NOs:671-678.

[0215] An exemplary single-molecule DNA-targeting RNA comprises two complementary stretches of nucleotides that hybridize to form a dsRNA duplex. In some embodiments, one of the two complementary stretches of nucleotides of the single-molecule DNA-targeting RNA (or the DNA encoding the stretch) is at least about 60% identical to one of the activator-RNA (tracrRNA) sequences set forth in SEQ ID NOs:431-562 over a stretch of at least 8 contiguous nucleotides. For example, one of the two complementary stretches of nucleotides of the single-molecule DNA-targeting RNA (or the DNA encoding the stretch) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the tracrRNA sequences set forth in SEQ ID NOs:431-562over a stretch of at least 8 contiguous nucleotides.

[0216] In some embodiments, one of the two complementary stretches of nucleotides of the single-molecule DNA-targeting RNA (or the DNA encoding the stretch) is at least about 60% identical to one of the targeter-RNA (crRNA) sequences set forth in SEQ ID NOs:563-679 over

a stretch of at least 8 contiguous nucleotides. For example, one of the two complementary stretches of nucleotides of the single-molecule DNA-targeting RNA (or the DNA encoding the stretch) is at least about 65% identical, at least about 70% identical, at least about 75% identical, at least about 80% identical, at least about 85% identical, at least about 90% identical, at least about 95% identical, at least about 98% identical, at least about 99% identical or 100 % identical to one of the crRNA sequences set forth in SEQ ID NOs:563-679 over a stretch of at least 8 contiguous nucleotides.

[0217] Definitions provided in "Definitions - Part I" are also applicable to the instant section; see "Definitions - Part I" for additional clarification of terms.

[0218] Before the present invention is further described, it is to be understood that this invention is not limited to particular embodiments described, as such may, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting, since the scope of the present invention will be limited only by the appended claims.

[0219] Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range, is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges, and are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the invention.

[0220] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can also be used in the practice or testing of the present invention, the preferred methods and materials are now described.

[0221] It must be noted that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "an enzymatically inactive Cas9 polypeptide" includes a plurality of such polypeptides and reference to "the target nucleic acid" includes reference to one or more target nucleic acids and equivalents thereof known to those skilled in the art, and so forth. It is further noted that the claims may be drafted to exclude any optional element. As such, this statement is intended to serve as antecedent basis for use of such exclusive terminology as "solely," "only" and the like in connection with the recitation of claim elements, or use of a "negative" limitation.

[0222] It is appreciated that certain features of the invention, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single

embodiment. Conversely, various features of the invention, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable sub-combination. All combinations of the embodiments pertaining to the invention are specifically embraced by the present invention and are disclosed herein just as if each and every combination was individually and explicitly disclosed. In addition, all sub-combinations of the various embodiments and elements thereof are also specifically embraced by the present invention and are disclosed herein just as if each and every such sub-combination was individually and explicitly disclosed herein.

[0223] The publications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention. Further, the dates of publication provided may be different from the actual publication dates which may need to be independently confirmed.

EXAMPLES

[0224] The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the present invention, and are not intended to limit the scope of what the inventors regard as their invention nor are they intended to represent that the experiments below are all or the only experiments performed. Efforts have been made to ensure accuracy with respect to numbers used (e.g. amounts, temperature, etc.) but some experimental errors and deviations should be accounted for. Unless indicated otherwise, parts are parts by weight, molecular weight is weight average molecular weight, temperature is in degrees Celsius, and pressure is at or near atmospheric. Standard abbreviations may be used, e.g., bp, base pair(s); kb, kilobase(s); pl, picoliter(s); s or sec, second(s); min, minute(s); h or hr, hour(s); aa, amino acid(s); kb, kilobase(s); bp, base pair(s); nt, nucleotide(s); i.m., intramuscular(ly); i.p., intraperitoneal(ly); s.c., subcutaneous(ly); and the like.

[0225] Example 1: Use of Cas9 to generate modifications in target DNA.

MATERIALS AND METHODS

Bacterial strains and culture conditions

[0226] *Streptococcus pyogenes,* cultured in THY medium (Todd Hewitt Broth (THB, Bacto, Becton Dickinson) supplemented with 0.2% yeast extract (Oxoid)) or on TSA (trypticase soy agar, BBL, Becton Dickinson) supplemented with 3% sheep blood, was incubated at 37°C in an atmosphere supplemented with 5% CO_2 without shaking. Escherichia coli, cultured in Luria-

Bertani (LB) medium and agar, was incubated at 37°C with shaking. When required, suitable antibiotics were added to the medium at the following final concentrations: ampicillin, 100 µg/ml for E. *coli;* chloramphenicol, 33 µg/ml for *Escherichia coli;* kanamycin, 25 µg/ml for E. *coli* and 300 µg/ml for S. *pyogenes.* Bacterial cell growth was monitored periodically by measuring the optical density of culture aliquots at 620 nm using a microplate reader (SLT Spectra Reader).

Transformation of bacterial cells

[0227] Plasmid DNA transformation into E. *coli* cells was performed according to a standard heat shock protocol. Transformation of S. *pyogenes* was performed as previously described with some modifications. The transformation assay performed to monitor in vivo CRISPR/Cas activity on plasmid maintenance was essentially carried out as described previously. Briefly, electrocompetent cells of S. *pyogenes* were equalized to the same cell density and electroporated with 500 ng of plasmid DNA. Every transformation was plated two to three times and the experiment was performed three times independently with different batches of competent cells for statistical analysis. Transformation efficiencies were calculated as CFU (colony forming units) per µg of DNA. Control transformations were performed with sterile water and backbone vector pEC85.

DNA manipulations

[0228] DNA manipulations including DNA preparation, amplification, digestion, ligation, purification, agarose gel electrophoresis were performed according to standard techniques with minor modifications. Protospacer plasmids for the in vitro cleavage and S. *pyogenes* transformation assays were constructed as described previously (4). Additional pUC19-based protospacer plasmids for in vitro cleavage assays were generated by ligating annealed oligonucleotides between digested EcoRI and BamHI sites in pUC19. The GFP gene-containing plasmid has been described previously (41). Kits (Qiagen) were used for DNA purification and plasmid preparation. Plasmid mutagenesis was performed using QuikChange® II XL kit (Stratagene) or QuikChange site-directed mutagenesis kit (Agilent). VBC-Biotech Services, Sigma-Aldrich and Integrated DNA Technologies supplied the synthetic oligonucleotides and RNAs.

Oligonucleotides for *in vitro* transcription templates

Templates for *in vitro* transcribed CRISPR Type II-A tracrRNA and crRNAs of *S. pyogenes* (for tracrRNA - PCR on chr. DNA SF370; for crRNA - annealing of two oligonucleotides)

T7-tracrRNA (75nt)

[0229]

OLEC1521 (F 5' tracrRNA): SEQ ID NO:340

OLEC1522 (R 3' tracrRNA): SEQ ID NO:341

T7-crRNA (template)

[0230]

OLEC2176 (F crRNA-sp1): SEQ ID NO:342

OLEC2178 (R crRNA-sp1): SEQ ID NO:343

OLEC2177 (F crRNA-sp2): SEQ ID NO:344

OLEC2179 (R crRNA-sp2): SEQ ID NO:345

Templates for in vitro transcribed N. meningitidis tracrRNA and engineered crRNA-sp2 (for tracrRNA - PCR on chr. DNA Z2491; for crRNA - annealing of two oligonucleotides)

T7-tracrRNA

[0231]

OLEC2205 (F predicted 5'): SEQ ID NO:346

OLEC2206 (R predicted 3'): SEQ ID NO:347

T7-crRNA (template)

[0232]

OLEC2209 (F sp2(*speM*) + *N.m.* repeat): SEQ ID NO:348

OLEC2214 (R sp2(*speM*) + *N.m.* repeat): SEQ ID NO:349

Templates for in vitro transcribed L. innocua tracrRNA and engineered crRNA-sp2 (for tracrRNA - PCR on chr. DNA Clip11262; for crRNA - annealing of two oligonucleotides)

T7-tracrRNA

[0233]

OLEC2203 (F predicted 5'): SEQ ID NO:350

OLEC2204 (R predicted 3'): SEQ ID NO:351

T7-crRNA (template)

[0234]

OLEC2207 (F sp2(*speM*) + *L.in.* repeat): SEQ ID NO:352

OLEC2212 (R sp2(*speM*) + *L.in.* repeat): SEQ ID NO:353

Oligonucleotides for constructing plasmids with protospacer for in vitro and in vivo studies

Plasmids for *speM* (spacer 2 (CRISPR Type II-A, SF370; protospacer prophage ø8232.3 from MGAS8232) analysis *in vitro* and in *S. pyogenes* (template: chr. DNA MGAS8232 or plasmids containing *speM* fragments)

pEC287

[0235]

OLEC1555 (F *speM*): SEQ ID NO:354

OLEC1556 (R *speM*): SEQ ID NO:355

pEC488

[0236]

OLEC2145 (F *speM*): SEQ ID NO:356

OLEC2146 (R *speM*): SEQ ID NO:357

pEC370

[0237]

OLEC1593 (F pEC488 protospacer 2 A22G): SEQ ID NO:358

OLEC1594 (R pEC488 protospacer 2 A22G): SEQ ID NO:359

pEC371

[0238]

OLEC1595 (F pEC488 protospacer 2 T10C): SEQ ID NO:360

OLEC1596 (R pEC488 protospacer 2 T10C): SEQ ID NO:361

pEC372

[0239]

OLEC2185 (F pEC488 protospacer 2 T7A): SEQ ID NO:362

OLEC2186 (R pEC488 protospacer 2 T7A): SEQ ID NO:363

pEC373

[0240]

OLEC2187 (F pEC488 protospacer 2 A6T): SEQ ID NO:364

OLEC2188 (R pEC488 protospacer 2 A6T): SEQ ID NO:365

pEC374

[0241]

OLEC2235 (F pEC488 protospacer 2 A5T): SEQ ID NO:366

OLEC2236 (R pEC488 protospacer 2 A5T): SEQ ID NO:367

pEC375

[0242]

OLEC2233 (F pEC488 protospacer 2 A4T): SEQ ID NO:368

OLEC2234 (R pEC488 protospacer 2 A4T): SEQ ID NO:369

pEC376

[0243]

OLEC2189 (F pEC488 protospacer 2 A3T): SEQ ID NO:370

OLEC2190 (R pEC488 protospacer 2 A3T): SEQ ID NO:371

pEC377

[0244]

OLEC2191 (F pEC488 protospacer 2 PAM G1C): SEQ ID NO:372

OLEC2192 (R pEC488 protospacer 2 PAM G1C): SEQ ID NO:373

pEC378

[0245]

OLEC2237 (F pEC488 protospacer 2 PAM GG1, 2CC): SEQ ID NO:374

OLEC2238 (R pEC488 protospacer 2 PAM GG1, 2CC): SEQ ID NO:375

Plasmids for SPy_0700 (spacer 1 (CRISPR Type II-A, SF370; protospacer prophage ø370.1 from SF370) analysis *in vitro* and in *S. pyogenes* (template: chr. DNA SF370 or plasmids containing SPy_0700 fragments)

pEC489

[0246]

OLEC2106 (F Spy_0700): SEQ ID NO:376

OLEC2107 (R Spy_0700): SEQ ID NO:377

pEC573

[0247]

OLEC2941 (F PAM TG1, 2GG): SEQ ID NO:378

OLEC2942 (R PAM TG1, 2GG): SEQ ID NO:379

Oligonucleotides for verification of plasmid constructs and cutting sites by sequencing analysis

ColE1 (pEC85)

[0248] oliRN228 (R sequencing): SEQ ID NO:380

speM (pEC287)

[0249]

OLEC1557 (F sequencing): SEQ ID NO:381

OLEC1556 (R sequencing): SEQ ID NO:382

repDEG-pAMbeta1 (pEC85)

[0250] OLEC787 (F sequencing): SEQ ID NO:383

Oligonucleotides for *in vitro* cleavage assays

crRNA

[0251]

Spacer 1 crRNA (1-42): SEQ ID NO:384

Spacer 2 crRNA (1-42): SEQ ID NO:385

Spacer 4 crRNA (1-42): SEQ ID NO:386

Spacer 2 crRNA (1-36): SEQ ID NO:387

Spacer 2 crRNA (1-32): SEQ ID NO:388

Spacer 2 crRNA (11-42): SEQ ID NO:389

tracrRNA

[0252]

(4-89): SEQ ID NO:390

(15-89): SEQ ID NO:391

(23-89): SEQ ID NO:392

(15-53): SEQ ID NO:393

(15-44): SEQ ID NO:394

(15-36): SEQ ID NO:395

(23-53): SEQ ID NO:396

(23-48): SEQ ID NO:397

(23-44): SEQ ID NO:398

(1-26): SEQ ID NO:399

chimeric RNAs

[0253]

Spacer 1 - chimera A: SEQ ID NO:400

Spacer 1 - chimera B: SEQ ID NO:401

Spacer 2 - chimera A: SEQ ID NO:402

Spacer 2 - chimera B: SEQ ID NO:403

Spacer 4 - chimera A: SEQ ID NO:404

Spacer 4 - chimera B: SEQ ID NO:405

GFP1: SEQ ID NO:406

GFP2: SEQ ID NO:407

GFP3: SEQ ID NO:408

GFP4: SEQ ID NO:409

GFP5: SEQ ID NO:410

DNA oligonucleotides as substrates for cleavage assays (protospacer in bold, PAM underlined)

[0254]

protospacer 1 - complementary - WT: SEQ ID NO:411

protospacer 1 - noncomplementary - WT: SEQ ID NO:412

protospacer 2 - complementary - WT: SEQ ID NO:413

protospacer 2 - noncomplementary - WT: SEQ ID NO:414

protospacer 4 - complementary - WT: SEQ ID NO:415

protospacer 4 - noncomplementary - WT: SEQ ID NO:416

protospacer 2 - complementary - PAM1: SEQ ID NO:417

protospacer 2 - noncomplementary - PAM1: SEQ ID NO:418

protospacer 2 - complementary - PAM2: SEQ ID NO:419

protospacer 2 - noncomplementary - PAM2: SEQ ID NO:420

protospacer 4 - complementary - PAM1: SEQ ID NO:421

protospacer 4 - noncomplementary - PAM1: SEQ ID NO:422

protospacer 4 - complementary - PAM2: SEQ ID NO:423

protospacer 4 - noncomplementary - PAM2: SEQ ID NO:424

In vitro transcription and purification of RNA

[0255] RNA was in vitro transcribed using T7 Flash in vitro Transcription Kit (Epicentre, Illumina company) and PCR-generated DNA templates carrying a T7 promoter sequence. RNA was gel-purified and quality-checked prior to use. The primers used for the preparation of RNA templates from S. pyogenes SF370, Listeria innocua Clip 11262 and Neisseria meningitidis A Z2491 are described above.

Protein purification

[0256] The sequence encoding Cas9 (residues 1-1368) was PCRamplified from the genomic DNA of S. pyogenes SF370 and inserted into a custom pET-based expression vector using ligation-independent cloning (LIC). The resulting fusion construct contained an N-terminal hexahistidine-maltose binding protein (His6- MBP) tag, followed by a peptide sequence containing a tobacco etch virus (TEV) protease cleavage site. The protein was expressed in E. coli strain BL21 Rosetta 2 (DE3) (EMD Biosciences), grown in 2xTY medium at 18°C for 16 h following induction with 0.2 mM IPTG. The protein was purified by a combination of affinity, ion exchange and size exclusion chromatographic steps. Briefly, cells were lysed in 20 mM Tris pH 8.0, 500 mM NaCl, 1 mM TCEP (supplemented with protease inhibitor cocktail (Roche)) in a homogenizer (Avestin). Clarified lysate was bound in batch to Ni-NTA agarose (Qiagen). The resin was washed extensively with 20 mM Tris pH 8.0, 500 mM NaCl and the bound protein was eluted in 20 mM Tris pH 8.0, 250 mM NaCl, 10% glycerol. The His6-MBP affinity tag was removed by cleavage with TEV protease, while the protein was dialyzed overnight against 20 mM HEPES pH 7.5, 150 mM KCl, 1 mM TCEP, 10% glycerol. The cleaved Cas9 protein was separated from the fusion tag by purification on a 5 ml SP Sepharose HiTrap column (GE Life Sciences), eluting with a linear gradient of 100 mM - 1 M KCl. The protein was further purified by size exclusion chromatography on a Superdex 200 16/60 column in 20 mM HEPES pH 7.5, 150 mM KCl and 1 mM TCEP. Eluted protein was concentrated to ~8 mg/ml, flash-frozen in liquid nitrogen and stored at -80°C. Cas9 D10A, H840A and D10A/H840A point mutants were generated using the QuikChange site-directed mutagenesis kit (Agilent) and confirmed by DNA sequencing. The proteins were purified following the same procedure as for the wildtype Cas9 protein.

[0257] Cas9 orthologs from Streptococcus thermophilus (LMD-9,YP_820832.1), L. innocua (Clip11262, NP_472073.1), Campylobacter jejuni (subsp. jejuni NCTC 11168, YP _002344900.1) and N. meningitidis (Z2491, YP_002342100.1) were expressed in BL21 Rosetta (DE3) pLysS cells (Novagen) as His6-MBP (N. meningitidis and C. jejuni), His6-Thioredoxin (L. innocua) and His6-GST (S. thermophilus) fusion proteins, and purified essentially as for S. pyogenes Cas9 with the following modifications. Due to large amounts of co-purifying nucleic acids, all four Cas9 proteins were purified by an additional heparin sepharose step prior to gel filtration, eluting the bound protein with a linear gradient of 100 mM - 2 M KCl. This successfully removed nucleic acid contamination from the C. jejuni, N. meningitidis and L. innocua proteins, but failed to remove co-purifying nucleic acids from the S. thermophilus Cas9 preparation. All proteins were concentrated to 1-8 mg/ml in 20 mM HEPES

pH 7.5, 150 mM KCl and 1 mM TCEP, flash-frozen in liquid N2 and stored at -80°C.

Plasmid DNA cleavage assay

[0258] Synthetic or in vitro-transcribed tracrRNA and crRNA were pre-annealed prior to the reaction by heating to 95°C and slowly cooling down to room temperature. Native or restriction digest-linearized plasmid DNA (300 ng (~8 nM)) was incubated for 60 min at 37°C with purified Cas9 protein (50-500 nM) and tracrRNA:crRNA duplex (50-500 nM, 1: 1) in a Cas9 plasmid cleavage buffer (20 mM HEPES pH 7.5, 150 mM KCl, 0.5 mM DTT, 0.1 mM EDTA) with or without 10 mM $MgCl_2$. The reactions were stopped with 5X DNA loading buffer containing 250 mM EDTA, resolved by 0.8 or 1% agarose gel electrophoresis and visualized by ethidium bromide staining. For the Cas9 mutant cleavage assays, the reactions were stopped with 5X SDS loading buffer (30% glycerol, 1.2% SDS, 250 mM EDTA) prior to loading on the agarose gel.

Metal-dependent cleavage assay

[0259] Protospacer 2 plasmid DNA (5 nM) was incubated for 1 h at 37°C with Cas9 (50 nM) pre-incubated with 50 nM tracrRNA:crRNA-sp2 in cleavage buffer (20 mM HEPES pH 7.5, 150 mM KCl, 0.5 mM DTT, 0.1 mM EDTA) supplemented with 1, 5 or 10 mM $MgCl_2$, 1 or 10 mM of $MnCl_2$, $CaCl_2$, $ZnCl_2$, $CoCl_2$, $NiSO_4$ or $CuSO_4$. The reaction was stopped by adding 5X SDS loading buffer (30% glycerol, 1.2% SDS, 250 mM EDTA), resolved by 1% agarose gel electrophoresis and visualized by ethidium bromide staining.

Single-turnover assay

[0260] Cas9 (25 nM) was pre-incubated 15 min at 37°C in cleavage buffer (20 mM HEPES pH 7.5, 150 mM KCl, 10 mM $MgCl_2$, 0.5 mM DTT, 0.1 mM EDTA) with duplexed tracrRNA:crRNA-sp2 (25 nM, 1: 1) or both RNAs (25 nM) not preannealed and the reaction was started by adding protospacer 2 plasmid DNA (5 nM). The reaction mix was incubated at 37°C. At defined time intervals, samples were withdrawn from the reaction, 5X SDS loading buffer (30% glycerol, 1.2% SDS, 250 mM EDTA) was added to stop the reaction and the cleavage was monitored by 1% agarose gel electrophoresis and ethidium bromide staining. The same was done for the single turnover kinetics without pre-incubation of Cas9 and RNA, where protospacer 2 plasmid DNA (5 nM) was mixed in cleavage buffer with duplex tracrRNA: crRNA-sp2 (25 nM) or both RNAs (25 nM) not pre-annealed, and the reaction was started by addition of Cas9 (25 nM). Percentage of cleavage was analyzed by densitometry and the average of three independent experiments was plotted against time. The data were fit by nonlinear regression analysis and the cleavage rates (k_{obs} [min^{-1}]) were calculated.

Multiple-turnover assay

[0261] Cas9 (1 nM) was pre-incubated for 15 min at 37°C in cleavage buffer (20 mM HEPES pH 7.5, 150 mM KCl, 10 mM $MgCl_2$, 0.5 mM DTT, 0.1 mM EDTA) with pre-annealed tracrRNA:crRNA-sp2 (1 nM, 1: 1). The reaction was started by addition of protospacer 2 plasmid DNA (5 nM). At defined time intervals, samples were withdrawn and the reaction was stopped by adding 5X SDS loading buffer (30% glycerol, 1.2% SDS, 250 mM EDTA). The cleavage reaction was resolved by 1% agarose gel electrophoresis, stained with ethidium bromide and the percentage of cleavage was analyzed by densitometry. The results of four independent experiments were plotted against time (min).

Oligonucleotide DNA cleavage assay

[0262] DNA oligonucleotides (10 pmol) were radiolabeled by incubating with 5 units T4 polynucleotide kinase (New England Biolabs) and ~3-6 pmol (~20-40 mCi) [γ-32P]-ATP (Promega) in 1X T4 polynucleotide kinase reaction buffer at 37°C for 30 min, in a 50 µL reaction. After heat inactivation (65°C for 20 min), reactions were purified through an Illustra MicroSpin G-25 column (GE Healthcare) to remove unincorporated label. Duplex substrates (100 nM) were generated by annealing labeled oligonucleotides with equimolar amounts of unlabeled complementary oligonucleotide at 95°C for 3 min, followed by slow cooling to room temperature. For cleavage assays, tracrRNA and crRNA were annealed by heating to 95°C for 30 s, followed by slow cooling to room temperature. Cas9 (500 nM final concentration) was pre-incubated with the annealed tracrRNA:crRNA duplex (500 nM) in cleavage assay buffer (20 mM HEPES pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 5% glycerol) in a total volume of 9 µl. Reactions were initiated by the addition of 1 µl target DNA (10 nM) and incubated for 1 h at 37°C. Reactions were quenched by the addition of 20 µl of loading dye (5 mM EDTA, 0.025% SDS, 5% glycerol in formamide) and heated to 95°C for 5 min. Cleavage products were resolved on 12% denaturing polyacrylamide gels containing 7 M urea and visualized by phosphorimaging (Storm, GE Life Sciences). Cleavage assays testing PAM requirements (Figure 13B) were carried out using DNA duplex substrates that had been pre-annealed and purified on an 8% native acrylamide gel, and subsequently radiolabeled at both 5' ends. The reactions were set-up and analyzed as above.

Electrophoretic mobility shift assays

[0263] Target DNA duplexes were formed by mixing of each strand (10 nmol) in deionized water, heating to 95°C for 3 min and slow cooling to room temperature. All DNAs were purified on 8% native gels containing 1X TBE. DNA bands were visualized by UV shadowing, excised, and eluted by soaking gel pieces in DEPC-treated H_2O. Eluted DNA was ethanol precipitated

and dissolved in DEPC-treated H_2O. DNA samples were 5' end labeled with [γ-32P]-ATP using T4 polynucleotide kinase (New England Biolabs) for 30 min at 37°C. PNK was heat denatured at 65°C for 20 min, and unincorporated radiolabel was removed using an Illustra MicroSpin G-25 column (GE Healthcare). Binding assays were performed in buffer containing 20 mM HEPES pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT and 10% glycerol in a total volume of 10 µl. Cas9 D10A/H840A double mutant was programmed with equimolar amounts of pre-annealed tracrRNA:crRNA duplex and titrated from 100 pM to 1 µM. Radiolabeled DNA was added to a final concentration of 20 pM. Samples were incubated for 1 h at 37°C and resolved at 4°C on an 8% native polyacrylamide gel containing 1X TBE and 5 mM $MgCl_2$. Gels were dried and DNA visualized by phosphorimaging.

In silico analysis of DNA and protein sequences

[0264] Vector NTI package (Invitrogen) was used for DNA sequence analysis (Vector NTI) and comparative sequence analysis of proteins (AlignX).

In silico modeling of RNA structure and co-folding

[0265] In silico predictions were performed using the Vienna RNA package algorithms (42, 43). RNA secondary structures and co-folding models were predicted with RNAfold and RNAcofold, respectively and visualized with VARNA (44).

RESULTS

[0266] Bacteria and archaea have evolved RNA mediated adaptive defense systems called clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) that protect organisms from invading viruses and plasmids (1-3). We show that in a subset of these systems, the mature crRNA that is base-paired to trans-activating crRNA (tracrRNA) forms a two-RNA structure that directs the CRISPR-associated protein Cas9 to introduce double-stranded (ds) breaks in target DNA. At sites complementary to the crRNA-guide sequence, the Cas9 HNH nuclease domain cleaves the complementary strand, whereas the Cas9 RuvC-like domain cleaves the noncomplementary strand. The dual-tracrRNA:crRNA, when engineered as a single RNA chimera, also directs sequence-specific Cas9 dsDNA cleavage. These studies reveal a family of endonucleases that use dual-RNAs for site-specific DNA cleavage and highlight the ability to exploit the system for RNA-programmable genome editing.

[0267] CRISPR/Cas defense systems rely on small RNAs for sequence-specific detection and silencing of foreign nucleic acids. CRISPR/Cas systems are composed of cas genes organized in operon(s) and CRISPR array(s) consisting of genome-targeting sequences (called spacers)

interspersed with identical repeats (1-3). CRISPR/Cas-mediated immunity occurs in three steps. In the adaptive phase, bacteria and archaea harboring one or more CRISPR loci respond to viral or plasmid challenge by integrating short fragments of foreign sequence (protospacers) into the host chromosome at the proximal end of the CRISPR array (1-3). In the expression and interference phases, transcription of the repeat spacer element into precursor CRISPR RNA (pre-crRNA) molecules followed by enzymatic cleavage yields the short crRNAs that can pair with complementary protospacer sequences of invading viral or plasmid targets (4-11). Target recognition by crRNAs directs the silencing of the foreign sequences by means of Cas proteins that function in complex with the crRNAs (10, 12-20).

[0268] There are three types of CRISPR/Cas systems (21-23). The type I and III systems share some overarching features: specialized Cas endonucleases process the pre-crRNAs, and oncemature, each crRNA assembles into a large multi-Cas protein complex capable of recognizing and cleaving nucleic acids complementary to the crRNA. In contrast, type II systems process precrRNAs by a different mechanism in which a trans-activating crRNA (tracrRNA) complementary to the repeat sequences in pre-crRNA triggers processing by the double-stranded (ds) RNAspecific ribonuclease RNase III in the presence of the Cas9 (formerly Csn1) protein (Figure 15) (4, 24). Cas9 is thought to be the sole protein responsible for crRNA-guided silencing of foreign DNA (25-27).

[0269] We show that in type II systems, Cas9 proteins constitute a family of enzymes that require a base-paired structure formed between the activating tracrRNA and the targeting crRNA to cleave target dsDNA. Site-specific cleavage occurs at locations determined by both base-pairing complementarity between the crRNA and the target protospacer DNA and a short motif [referred to as the protospacer adjacent motif (PAM)] juxtaposed to the complementary region in the target DNA. Our study further demonstrates that the Cas9 endonuclease family can be programmed with single RNA molecules to cleave specific DNA sites, thereby facilitating the development of a simple and versatile RNA-directed system to generate dsDNA breaks for genome targeting and editing.

Cas9 is a DNA endonuclease guided by two RNAs

[0270] Cas9, the hallmark protein of type II systems, has been hypothesized to be involved in both crRNA maturation and crRNA-guided DNA interference (Figure 15) (4, 25-27). Cas9 is involved in crRNA maturation (4), but its direct participation in target DNA destruction has not been investigated. To test whether and how Cas9 might be capable of target DNA cleavage, we used an overexpression system to purify Cas9 protein derived from the pathogen Streptococcus pyogenes (Figure 16, see supplementary materials and methods) and tested its ability to cleave a plasmid DNA or an oligonucleotide duplex bearing a protospacer sequence complementary to a mature crRNA, and a bona fide PAM. We found that mature crRNA alone was incapable of directing Cas9-catalyzed plasmid DNA cleavage (Figure 10A and Figure 17A). However, addition of tracrRNA, which can pair with the repeat sequence of crRNA and is essential to crRNA maturation in this system, triggered Cas9 to cleave plasmid DNA (Figure

10A and Figure 17A). The cleavage reaction required both magnesium and the presence of a crRNA sequence complementary to the DNA; a crRNA capable of tracrRNA base pairing but containing a noncognate target DNA-binding sequence did not support Cas9-catalyzed plasmid cleavage (Figure 10A; Figure 17A, compare crRNA-sp2 to crRNA-sp1; and Figure 18A). We obtained similar results with a short linear dsDNA substrate (Figure 10B and Figure 17, B and C). Thus, the trans-activating tracrRNA is a small noncoding RNA with two critical functions: triggering pre-crRNA processing by the enzyme RNase III (4) and subsequently activating crRNA-guided DNA cleavage by Cas9.

[0271] Cleavage of both plasmid and short linear dsDNA by tracrRNA:crRNA-guided Cas9 is site specific (Figure 10, C to E, and Figure 19, A and B). Plasmid DNA cleavage produced blunt ends at a position three base pairs upstream of the PAM sequence (Figure 10, C and E, and Figure 19, A and C) (26). Similarly, within short dsDNA duplexes, the DNA strand that is complementary to the target-binding sequence in the crRNA (the complementary strand) is cleaved at a site three base pairs upstream of the PAM(Figure 10, D and E, and Figure 19, B and C). The noncomplementary DNA strand is cleaved at one or more sites within three to eight base pairs upstream of the PAM. Further investigation revealed that the noncomplementary strand is first cleaved endonucleolytically and subsequently trimmed by a 3'-5' exonuclease activity (Figure 18B). The cleavage rates by Cas9 under single-turnover conditions ranged from 0.3 to 1 min-1, comparable to those of restriction endonucleases (Figure 20A), whereas incubation of wildtype (WT) Cas9-tracrRNA:crRNA complex with a fivefold molar excess of substrate DNA provided evidence that the dual-RNA-guided Cas9 is a multiple-turnover enzyme (Figure 20B). In contrast to the CRISPR type I Cascade complex (18), Cas9 cleaves both linearized and supercoiled plasmids (Figures 10A and 11A). Therefore, an invading plasmid can, in principle, be cleaved multiple times by Cas9 proteins programmed with different crRNAs.

[0272] Figure 10 (A) Cas9 was programmed with a 42-nucleotide crRNA-sp2 (crRNA containing a spacer 2 sequence) in the presence or absence of 75-nucleotide tracrRNA. The complex was added to circular or XhoI-linearized plasmid DNA bearing a sequence complementary to spacer 2 and a functional PAM. crRNA-sp1, specificity control; M, DNA marker; kbp, kilo-base pair. See Figure 17A. **(B)** Cas9 was programmed with crRNA-sp2 and tracrRNA (nucleotides 4 to 89). The complex was incubated with double- or single-stranded DNAs harboring a sequence complementary to spacer 2 and a functional PAM (4). The complementary or noncomplementary strands of the DNA were 5'-radiolabeled and annealed with a nonlabeled partner strand. nt, nucleotides. See Figure 17, B and C. **(C)** Sequencing analysis of cleavage products from Figure 10A. Termination of primer extension in the sequencing reaction indicates the position of the cleavage site. The 3' terminal A overhang (asterisks) is an artifact of the sequencing reaction. See Figure 19, A and C. **(D)** The cleavage products from Figure 10B were analyzed alongside 5' end-labeled size markers derived from the complementary and noncomplementary strands of the target DNA duplex. M, marker; P, cleavage product. See Figure 19, B and C. **(E)** Schematic representation of tracrRNA, crRNA-sp2, and protospacer 2 DNA sequences. Regions of crRNA complementarity to tracrRNA (overline) and the protospacer DNA (underline) are represented. The PAM sequence is

labeled; cleavage sites mapped in (C) and (D) are represented by white-filled arrows (C), a black-filled arrow [(D), complementary strand], and a black bar [(D), noncomplementary strand].

[0273] Figure 15 depicts the type II RNA-mediated CRISPR/Cas immune pathway. The expression and interference steps are represented in the drawing. The type II CRISPR/Cas loci are composed of an operon of four genes encoding the proteins Cas9, Cas1, Cas2 and Csn2, a CRISPR array consisting of a leader sequence followed by identical repeats (black rectangles) interspersed with unique genome-targeting spacers (diamonds) and a sequence encoding the trans-activating tracrRNA. Represented here is the type II CRISPR/Cas locus of S. pyogenes SF370 (Accession number NC_002737) (4). Experimentally confirmed promoters and transcriptional terminator in this locus are indicated (4). The CRISPR array is transcribed as a precursor CRISPR RNA (pre-crRNA) molecule that undergoes a maturation process specific to the type II systems (4). In S. pyogenes SF370, tracrRNA is transcribed as two primary transcripts of 171 and 89 nt in length that have complementarity to each repeat of the pre-crRNA. The first processing event involves pairing of tracrRNA to pre-crRNA, forming a duplex RNA that is recognized and cleaved by the housekeeping endoribonuclease RNase III in the presence of the Cas9 protein. RNase III-mediated cleavage of the duplex RNA generates a 75-nt processed tracrRNA and a 66-nt intermediate crRNAs consisting of a central region containing a sequence of one spacer, flanked by portions of the repeat sequence. A second processing event, mediated by unknown ribonuclease(s), leads to the formation of mature crRNAs of 39 to 42 nt in length consisting of 5'-terminal spacer-derived guide sequence and repeat-derived 3'-terminal sequence. Following the first and second processing events, mature tracrRNA remains paired to the mature crRNAs and bound to the Cas9 protein. In this ternary complex, the dual tracrRNA:crRNA structure acts as guide RNA that directs the endonuclease Cas9 to the cognate target DNA. Target recognition by the Cas9-tracrRNA:crRNA complex is initiated by scanning the invading DNA molecule for homology between the protospacer sequence in the target DNA and the spacer-derived sequence in the crRNA. In addition to the DNA protospacer-crRNA spacer complementarity, DNA targeting requires the presence of a short motif (NGG, where N can be any nucleotide) adjacent to the protospacer (protospacer adjacent motif - PAM). Following pairing between the dual-RNA and the protospacer sequence, an R-loop is formed and Cas9 subsequently introduces a double-stranded break (DSB) in the DNA. Cleavage of target DNA by Cas9 requires two catalytic domains in the protein. At a specific site relative to the PAM, the HNH domain cleaves the complementary strand of the DNA while the RuvC-like domain cleaves the noncomplementary strand.

[0274] Figure 16 (A) *S. pyogenes* Cas9 was expressed in *E. coli* as a fusion protein containing an N-terminal His6-MBP tag and purified by a combination of affinity, ion exchange and size exclusion chromatographic steps. The affinity tag was removed by TEV protease cleavage following the affinity purification step. Shown is a chromatogram of the final size exclusion chromatography step on a Superdex 200 (16/60) column. Cas9 elutes as a single monomeric peak devoid of contaminating nucleic acids, as judged by the ratio of absorbances at 280 and 260 nm. Inset; eluted fractions were resolved by SDS-PAGE on a 10% polyacrylamide gel and

stained with SimplyBlue Safe Stain (Invitrogen). **(B)** SDS-PAGE analysis of purified Cas9 orthologs. Cas9 orthologs were purified as described in Supplementary Materials and Methods. 2.5 µg of each purified Cas9 were analyzed on a 4-20% gradient polyacrylamide gel and stained with SimplyBlue Safe Stain.

[0275] Figure 17 (also see Figure 10). The protospacer 1 sequence originates from *S. pyogenes* SF370 (M1) SPy_0700, target of *S. pyogenes* SF370 crRNAsp1 (4). Here, the protospacer 1 sequence was manipulated by changing the PAM from a nonfunctional sequence (TTG) to a functional one (TGG). The protospacer 4 sequence originates from *S. pyogenes* MGAS10750 (M4) MGAS10750_Spy1285, target of *S. pyogenes* SF370 crRNA-sp4 (4). (A) Protospacer 1 plasmid DNA cleavage guided by cognate tracrRNA:crRNA duplexes. The cleavage products were resolved by agarose gel electrophoresis and visualized by ethidium bromide staining. M, DNA marker; fragment sizes in base pairs are indicated. **(B)** Protospacer 1 oligonucleotide DNA cleavage guided by cognate tracrRNA:crRNA-sp1 duplex. The cleavage products were resolved by denaturing polyacrylamide gel electrophoresis and visualized by phosphorimaging. Fragment sizes in nucleotides are indicated. **(C)** Protospacer 4 oligonucleotide DNA cleavage guided by cognate tracrRNA:crRNA-sp4 duplex. The cleavage products were resolved by denaturing polyacrylamide gel electrophoresis and visualized by phosphorimaging. Fragment sizes in nucleotides are indicated. **(A, B, C)** Experiments in (A) were performed as in Figure 10A; in (B) and in (C) as in Figure 10B. **(B, C)** A schematic of the tracrRNA:crRNA target DNA interaction is shown below. The regions of crRNA complementarity to tracrRNA and the protospacer DNA are overlined and underlined, respectively. The PAM sequence is labeled.

[0276] Figure 18 (also see Figure 10). **(A)** Protospacer 2 plasmid DNA was incubated with Cas9 complexed with tracrRNA:crRNA-sp2 in the presence of different concentrations of Mg^{2+}, Mn^{2+}, Ca^{2+}, Zn^{2+}, Co^{2+}, Ni^{2+} or Cu^{2+}. The cleavage products were resolved by agarose gel electrophoresis and visualized by ethidium bromide staining. Plasmid forms are indicated. **(B)** A protospacer 4 oligonucleotide DNA duplex containing a PAM motif was annealed and gel-purified prior to radiolabeling at both 5' ends. The duplex (10 nM final concentration) was incubated with Cas9 programmed with tracrRNA (nucleotides 23-89) and crRNAsp4 (500 nM final concentration, 1:1). At indicated time points (min), 10 µl aliquots of the cleavage reaction were quenched with formamide buffer containing 0.025% SDS and 5 mM EDTA, and analyzed by denaturing polyacrylamide gel electrophoresis as in Figure 10B. Sizes in nucleotides are indicated.

[0277] Figure 19 (A) Mapping of protospacer 1 plasmid DNA cleavage. Cleavage products from Figure 17A were analyzed by sequencing as in Figure 10C. Note that the 3' terminal A overhang (asterisk) is an artifact of the sequencing reaction. **(B)** Mapping of protospacer 4 oligonucleotide DNA cleavage. Cleavage products from Figure 17C were analyzed by denaturing polyacrylamide gel electrophoresis alongside 5' endlabeled oligonucleotide size markers derived from the complementary and noncomplementary strands of the protospacer 4 duplex DNA. M, marker; P, cleavage product. Lanes 1-2: complementary strand. Lanes 3-4: non-complementary strand. Fragment sizes in nucleotides are indicated. **(C)** Schematic

representations of tracrRNA, crRNA-sp1 and protospacer 1 DNA sequences (top) and tracrRNA, crRNAsp4 and protospacer 4 DNA sequences (bottom). tracrRNA:crRNA forms a dual-RNA structure directed to complementary protospacer DNA through crRNA-protospacer DNA pairing. The regions of crRNA complementary to tracrRNA and the protospacer DNA are overlined and underlined, respectively. The cleavage sites in the complementary and noncomplementary DNA strands mapped in (A) (top) and (B) (bottom) are represented with arrows (A and B, complementary strand) and a black bar (B, noncomplementary strand) above the sequences, respectively.

[0278] Figure 20 (A) Single turnover kinetics of Cas9 under different RNA pre-annealing and protein-RNA pre-incubation conditions. Protospacer 2 plasmid DNA was incubated with either Cas9 pre-incubated with pre-annealed tracrRNA:crRNA-sp2 (◦), Cas9 not pre-incubated with pre-annealed tracrRNA:crRNA-sp2 (•), Cas9 pre-incubated with not pre-annealed tracrRNA and crRNA-sp2 (□) or Cas9 not pre-incubated with not pre-annealed RNAs (■). The cleavage activity was monitored in a time-dependent manner and analyzed by agarose gel electrophoresis followed by ethidium bromide staining. The average percentage of cleavage from three independent experiments is plotted against the time (min) and fitted with a nonlinear regression. The calculated cleavage rates (k_{obs}) are shown in the table. The results suggest that the binding of Cas9 to the RNAs is not rate-limiting under the conditions tested. Plasmid forms are indicated. The obtained k_{obs} values are comparable to those of restriction endonucleases which are typically of the order of 1-10 per min (45-47). **(B)** Cas9 is a multiple turnover endonuclease. Cas9 loaded with duplexed tracrRNA:crRNA-sp2 (1 nM, 1:1:1 - indicated with gray line on the graph) was incubated with a 5-fold excess of native protospacer 2 plasmid DNA. Cleavage was monitored by withdrawing samples from the reaction at defined time intervals (0 to 120 min) followed by agarose gel electrophoresis analysis (top) and determination of cleavage product amount (nM) (bottom). Standard deviations of three independent experiments are indicated. In the time interval investigated, 1 nM Cas9 was able to cleave ~2.5 nM plasmid DNA. **Each Cas9 nuclease domain cleaves one DNA strand**

[0279] Cas9 contains domains homologous to both HNH and RuvC endonucleases (Figure 11A and Figure 3) (21-23, 27, 28). We designed and purified Cas9 variants containing inactivating point mutations in the catalytic residues of either the HNH or RuvC-like domains (Figure 11A and Figure 3) (23, 27). Incubation of these variant Cas9 proteins with native plasmid DNA showed that dual-RNA-guided mutant Cas9 proteins yielded nicked open circular plasmids, whereas the WT Cas9 protein-tracrRNA:crRNA complex produced a linear DNA product (Figures 10A and 11A and figures 17A and 25A). This result indicates that the Cas9 HNH and RuvC-like domains each cleave one plasmid DNA strand. To determine which strand of the target DNA is cleaved by each Cas9 catalytic domain, we incubated the mutant Cas9-tracrRNA:crRNA complexes with short dsDNA substrates in which either the complementary or noncomplementary strand was radiolabeled at its 5' end. The resulting cleavage products indicated that the Cas9 HNH domain cleaves the complementary DNA strand, whereas the Cas9 RuvC-like domain cleaves the noncomplementary DNA strand (Figure 11B and Figure 21B).

[0280] Figure 11(A) (Top) Schematic representation of Cas9 domain structure showing the positions of domain mutations. D10A, Asp10→Ala10; H840A; His840-Ala840. Complexes of WT or nuclease mutant Cas9 proteins with tracrRNA: crRNA-sp2 were assayed for endonuclease activity as in Figure 10A. **(B)** Complexes of WT Cas9 or nuclease domain mutants with tracrRNA and crRNA-sp2 were tested for activity as in Figure 10B.

[0281] Figure 3 The amino-acid sequence of Cas9 from *S. pyogenes* (SEQ ID NO:8) is represented. Cas9/Csn1 proteins from various diverse species have 2 domains that include motifs homologous to both HNH and RuvC endonucleases. **(A)** Motifs 1-4 (motif numbers are marked on left side of sequence) are shown for *S. pyogenes* Cas9/Csn1. The three predicted RuvC-like motifs (1, 2, 4) and the predicted HNH motif (3) are overlined. Residues Asp10 and His840, which were substituted by Ala in this study are highlighted by an asterisk above the sequence. Underlined residues are highly conserved among Cas9 proteins from different species. Mutations in underlined residues are likely to have functional consequences on Cas9 activity. Note that in the present study coupling of the two nuclease-like activities is experimentally demonstrated (Figure 11 and Figure 21). **(B)** Domains 1 (amino acids 7-166) and 2 (amino acids 731-1003), which include motifs 1-4, are depicted for *S. pyogenes* Cas9/Csn1. Refer to **Table 1**and **Figure 5** for additional information.

[0282] Figure 21 Protospacer DNA cleavage by cognate tracrRNA:crRNA-directed Cas9 mutants containing mutations in the HNH or RuvC-like domain. **(A)** Protospacer 1 plasmid DNA cleavage. The experiment was performed as in Figure 11A. Plasmid DNA conformations and sizes in base pairs are indicated. **(B)** Protospacer 4 oligonucleotide DNA cleavage. The experiment was performed as in Figure 11B. Sizes in nucleotides are indicated.

Dual-RNA requirements for target DNA binding and cleavage

[0283] tracrRNA might be required for target DNA binding and/or to stimulate the nuclease activity of Cas9 downstream of target recognition. To distinguish between these possibilities, we used an electrophoretic mobility shift assay to monitor target DNA binding by catalytically inactive Cas9 in the presence or absence of crRNA and/or tracrRNA. Addition of tracrRNA substantially enhanced target DNA binding by Cas9, whereas we observed little specific DNA binding with Cas9 alone or Cas9- crRNA (Figure 22). This indicates that tracrRNA is required for target DNA recognition, possibly by properly orienting the crRNA for interaction with the complementary strand of target DNA. The predicted tracrRNA:crRNA secondary structure includes base pairing between the 22 nucleotides at the 3' terminus of the crRNA and a segment near the 5' end of the mature tracrRNA (Figure 10E). This interaction creates a structure in which the 5'-terminal 20 nucleotides of the crRNA, which vary in sequence in different crRNAs, are available for target DNA binding. The bulk of the tracrRNA downstream of the crRNA base pairing region is free to form additional RNA structure(s) and/or to interact with Cas9 or the target DNA site. To determine whether the entire length of the tracrRNA is necessary for site specific Cas9-catalyzed DNA cleavage, we tested Cas9-tracrRNA:crRNA complexes reconstituted using full-length mature (42-nucleotide) crRNA and various truncated

forms of tracrRNA lacking sequences at their 5' or 3' ends. These complexes were tested for cleavage using a short target dsDNA. A substantially truncated version of the tracrRNA retaining nucleotides 23 to 48 of the native sequence was capable of supporting robust dual-RNA-guided Cas9-catalyzed DNA cleavage (Figure 12, A and C, and Figure 23, A and B). Truncation of the crRNA from either end showed that Cas9-catalyzed cleavage in the presence of tracrRNA could be triggered with crRNAs missing the 3'-terminal 10 nucleotides (Figure 12, B and C). In contrast, a 10-nucleotide deletion from the 5' end of crRNA abolished DNA cleavage by Cas9 (Figure 12B). We also analyzed Cas9 orthologs from various bacterial species for their ability to support S. *pyogenes* tracrRNA:crRNA-guided DNA cleavage. In contrast to closely related *S. pyogenes* Cas9 orthologs, more distantly related orthologs were not functional in the cleavage reaction (Figure 24). Similarly, S. pyogenes Cas9 guided by tracrRNA:crRNA duplexes originating from more distant systems was unable to cleave DNA efficiently (Figure 24). Species specificity of dual-RNA-guided cleavage of DNA indicates coevolution of Cas9, tracrRNA, and the crRNA repeat, as well as the existence of a still unknown structure and/or sequence in the dual- RNA that is critical for the formation of the ternary complex with specific Cas9 orthologs.

[0284] To investigate the protospacer sequence requirements for type II CRISPR/Cas immunity in bacterial cells, we analyzed a series of protospacer-containing plasmid DNAs harboring single-nucleotide mutations for their maintenance following transformation in S. pyogenes and their ability to be cleaved by Cas9 in vitro. In contrast to point mutations introduced at the 5' end of the protospacer, mutations in the region close to the PAM and the Cas9 cleavage sites were not tolerated in vivo and resulted in decreased plasmid cleavage efficiency in vitro (Figure 12D). Our results are in agreement with a previous report of protospacer escape mutants selected in the type II CRISPR system from S. thermophilus in vivo (27, 29). Furthermore, the plasmid maintenance and cleavage results hint at the existence of a "seed" region located at the 3' end of the protospacer sequence that is crucial for the interaction with crRNA and subsequent cleavage by Cas9. In support of this notion, Cas9 enhanced complementary DNA strand hybridization to the crRNA; this enhancement was the strongest in the 3'-terminal region of the crRNA targeting sequence (Figure 25A-C). Corroborating this finding, a contiguous stretch of at least 13 base pairs between the crRNA and the target DNA site proximal to the PAM is required for efficient target cleavage, whereas up to six contiguous mismatches in the 5'-terminal region of the protospacer are tolerated (Figure 12E). These findings are reminiscent of the previously observed seed-sequence requirements for target nucleic acid recognition in Argonaute proteins (30, 31) and the Cascade and Csy CRISPR complexes (13, 14).

[0285] Figure 12 (A) Cas9-tracrRNA: crRNA complexes were reconstituted using 42-nucleotide crRNA-sp2 and truncated tracrRNA constructs and were assayed for cleavage activity as in Figure 10B. **(B)** Cas9 programmed with full-length tracrRNA and crRNA-sp2 truncations was assayed for activity as in (A). **(C)** Minimal regions of tracrRNA and crRNA capable of guiding Cas9-mediated DNA cleavage (shaded region). **(D)** Plasmids containing WT or mutant protospacer 2 sequences with indicated point mutations were cleaved in vitro by programmed Cas9 as in Figure 10A and used for transformation assays of WT or pre-crRNA-

deficient S. pyogenes. The transformation efficiency was calculated as colony-forming units (CFU) per microgram of plasmid DNA. Error bars represent SDs for three biological replicates. **(E)** Plasmids containing WT and mutant protospacer 2 inserts with varying extent of crRNA-target DNA mismatches (bottom) were cleaved in vitro by programmed Cas9 (top). The cleavage reactions were further digested with XmnI. The 1880- and 800-bp fragments are Cas9-generated cleavage products. M, DNA marker.

[0286] Figure 22 Electrophoretic mobility shift assays were performed using protospacer 4 target DNA duplex and Cas9 (containing nuclease domain inactivating mutations D10A and H840) alone or in the presence of crRNA-sp4, tracrRNA (75nt), or both. The target DNA duplex was radiolabeled at both 5' ends. Cas9 (D10/H840A) and complexes were titrated from 1 nM to 1 μM. Binding was analyzed by 8% native polyacrylamide gel electrophoresis and visualized by phosphorimaging. Note that Cas9 alone binds target DNA with moderate affinity. This binding is unaffected by the addition of crRNA, suggesting that this represents sequence nonspecific interaction with the dsDNA. Furthermore, this interaction can be outcompeted by tracrRNA alone in the absence of crRNA. In the presence of both crRNA and tracrRNA, target DNA binding is substantially enhanced and yields a species with distinct electrophoretic mobility, indicative of specific target DNA recognition.

[0287] Figure 23 A fragment of tracrRNA encompassing a part of the crRNA paired region and a portion of the downstream region is sufficient to direct cleavage of protospacer oligonucleotide DNA by Cas9. (A) Protospacer 1 oligonucleotide DNA cleavage and (B) Protospacer 4 oligonucleotide DNA cleavage by Cas9 guided with a mature cognate crRNA and various tracrRNA fragments. **(A, B)** Sizes in nucleotides are indicated.

[0288] Figure 24 Like Cas9 from *S. pyogenes,* the closely related Cas9 orthologs from the Gram-positive bacteria *L. innocua* and *S. thermophilus* cleave protospacer DNA when targeted by tracrRNA:crRNA from *S. pyogenes.* However, under the same conditions, DNA cleavage by the less closely related Cas9 orthologs from the Gram negative bacteria *C. jejuni* and *N. meningitidis* is not observed. Spy, *S. pyogenes* SF370 (Accession Number NC_002737); Sth, *S. thermophilus* LMD-9 (STER_1477 Cas9 ortholog; Accession Number NC_008532); Lin, *L. innocua* Clip11262 (Accession Number NC_003212); Cje, *C. jejuni* NCTC 11168 (Accession Number NC_002163); Nme, *N. meningitidis* A Z2491 (Accession Number NC_003116). **(A)** Cleavage of protospacer plasmid DNA. Protospacer 2 plasmid DNA (300 ng) was subjected to cleavage by different Cas9 orthologs (500 nM) guided by hybrid tracrRNA:crRNA-sp2 duplexes (500 nM, 1: 1) from different species. To design the RNA duplexes, we predicted tracrRNA sequences from *L. innocua* and *N. meningitidis* based on previously published Northern blot data *(4).* The dual-hybrid RNA duplexes consist of species specific tracrRNA and a heterologous crRNA. The heterologous crRNA sequence was engineered to contain S. *pyogenes* DNA-targeting sp2 sequence at the 5' end fused to *L. innocua* or *N. meningitidis* tracrRNA-binding repeat sequence at the 3' end. Cas9 orthologs from S. *thermophilus* and *L. innocua,* but not from *N. meningitidis* or *C. jejuni,* can be guided by *S. pyogenes* tracrRNA:crRNA-sp2 to cleave protospacer 2 plasmid DNA, albeit with slightly decreased efficiency. Similarly, the hybrid *L. innocua* tracrRNA:crRNA-sp2 can guide S. *pyogenes* Cas9 to

cleave the target DNA with high efficiency, whereas the hybrid *N. meningitidis* tracrRNA:crRNA-sp2 triggers only slight DNA cleavage activity by *S. pyogenes* Cas9. As controls, *N. meningitidis* and *L. innocua* Cas9 orthologs cleave protospacer 2 plasmid DNA when guided by the cognate hybrid tracrRNA:crRNA-sp2. Note that as mentioned above, the tracrRNA sequence of *N. meningitidis* is predicted only and has not yet been confirmed by RNA sequencing. Therefore, the low efficiency of cleavage could be the result of either low activity of the Cas9 orthologs or the use of a nonoptimally designed tracrRNA sequence. **(B)** Cleavage of protospacer oligonucleotide DNA. 5'-end radioactively labeled complementary strand oligonucleotide (10 nM) pre-annealed with unlabeled noncomplementary strand oligonucleotide (protospacer 1) (10 nM) (left) or 5'-end radioactively labeled noncomplementary strand oligonucleotide (10 nM) pre-annealed with unlabeled complementary strand oligonucleotide (10 nM) (right) (protospacer 1) was subjected to cleavage by various Cas9 orthologs (500 nM) guided by tracrRNA:crRNA-sp1 duplex from S. *pyogenes* (500 nM, 1: 1). Cas9 orthologs from S. *thermophilus* and *L. innocua,* but not from *N. meningitidis* or *C. jejuni* can be guided by S. *pyogenes* cognate dual-RNA to cleave the protospacer oligonucleotide DNA, albeit with decreased efficiency. Note that the cleavage site on the complementary DNA strand is identical for all three orthologs. Cleavage of the noncomplementary strand occurs at distinct positions. **(C)** Amino acid sequence identity of Cas9 orthologs. *S. pyogenes, S. thermophilus* and *L. innocua* Cas9 orthologs share high percentage of amino acid identity. In contrast, the *C. jejuni* and *N. meningitidis* Cas9 proteins differ in sequence and length (~300-400 amino acids shorter). **(D)** Co-foldings of engineered species-specific heterologous crRNA sequences with the corresponding tracrRNA orthologs from S. *pyogenes* (experimentally confirmed, *(4)), L. innocua* (predicted) or *N. meningitidis* (predicted). tracrRNAs; crRNA spacer 2 fragments; and crRNA repeat fragments are traced and labeled. *L. innocua* and *S. pyogenes* hybrid tracrRNA:crRNA-sp2 duplexes share very similar structural characteristics, albeit distinct from the *N. meningitidis* hybrid tracrRNA:crRNA. Together with the cleavage data described above in (A) and (B), the co-folding predictions would indicate that the species-specificity cleavage of target DNA by Cas9-tracrRNA:crRNA is dictated by a still unknown structural feature in the tracrRNA:crRNA duplex that is recognized specifically by a cognate Cas9 ortholog. It was predicted that the species-specificity of cleavage observed in (A) and (B) occurs at the level of binding of Cas9 to dual-tracrRNA:crRNA. Dual-RNA guided Cas9 cleavage of target DNA can be species specific. Depending on the degree of diversity/evolution among Cas9 proteins and tracrRNA:crRNA duplexes, Cas9 and dual- RNA orthologs are partially interchangeable.

[0289] Figure 25 A series of 8-nucleotide DNA probes complementary to regions in the crRNA encompassing the DNA-targeting region and tracrRNA-binding region were analyzed for their ability to hybridize to the crRNA in the context of a tracrRNA:crRNA duplex and the Cas9-tracrRNA:crRNA ternary complex. **(A)** Schematic representation of the sequences of DNA probes used in the assay and their binding sites in crRNA-sp4. **(B-C)** Electrophoretic mobility shift assays of target DNA probes with tracrRNA:crRNA-sp4 or Cas9-tracrRNA:crRNA-sp4. The tracrRNA(15-89) construct was used in the experiment. Binding of the duplexes or complexes to target oligonucleotide DNAs was analyzed on a 16% native polyacrylamide gel and visualized by phosphorimaging.

A short sequence motif dictates R-loop formation

[0290] In multiple CRISPR/Cas systems, recognition of self versus nonself has been shown to involve a short sequence motif that is preserved in the foreign genome, referred to as the PAM(27, 29, 32-34). PAM motifs are only a few base pairs in length, and their precise sequence and position vary according to the CRISPR/Cas system type (32). In the S. pyogenes type II system, the PAM conforms to an NGG consensus sequence, containing two G: C base pairs that occur one base pair downstream of the crRNA binding sequence, within the target DNA (4). Transformation assays demonstrated that the GG motif is essential for protospacer plasmid DNA elimination by CRISPR/Cas in bacterial cells (Figure 26A), consistent with previous observations in S. thermophilus (27). The motif is also essential for in vitro protospacer plasmid cleavage by tracrRNA:crRNA-guided Cas9 (Figure 26B). To determine the role of the PAM in target DNA cleavage by the Cas9-tracrRNA: crRNA complex, we tested a series of dsDNA duplexes containing mutations in the PAM sequence on the complementary or noncomplementary strands, or both (Figure 13A). Cleavage assays using these substrates showed that Cas9- catalyzed DNA cleavage was particularly sensitive to mutations in the PAM sequence on the noncomplementary strand of the DNA, in contrast to complementary strand PAM recognition by type I CRISPR/Cas systems (18, 34). Cleavage of target single-stranded DNAs was unaffected by mutations of the PAM motif. This observation suggests that the PAM motif is required only in the context of target dsDNA and may thus be required to license duplex unwinding, strand invasion, and the formation of an R-loop structure. When we used a different crRNA-target DNA pair (crRNA-sp4 and protospacer 4 DNA), selected due to the presence of a canonical PAM not present in the protospacer 2 target DNA, we found that both G nucleotides of the PAM were required for efficient Cas9-catalyzed DNA cleavage (Figure 13B and Figure 26C). To determine whether the PAM plays a direct role in recruiting the Cas9-tracrRNA:crRNA complex to the correct target DNA site, we analyzed binding affinities of the complex for target DNA sequences by native gel mobility shift assays (Figure 13C). Mutation of either G in the PAM sequence substantially reduced the affinity of Cas9-tracrRNA: crRNA for the target DNA. This finding illustrates a role for the PAM sequence in target DNA binding by Cas9.

[0291] Figure 13 (A) Dual RNA-programmed Cas9 was tested for activity as in Figure 10B. WT and mutant PAM sequences in target DNAs are indicated with lines. **(B)** Protospacer 4 target DNA duplexes (labeled at both 5' ends) containing WT and mutant PAM motifs were incubated with Cas9 programmed with tracrRNA:crRNA-sp4 (nucleotides 23 to 89). At the indicated time points (in minutes), aliquots of the cleavage reaction were taken and analyzed as in Figure 10B. (C) Electrophoretic mobility shift assays were performed using RNA-programmed Cas9 (D10A/H840A) and protospacer 4 target DNA duplexes [same as in (B)] containing WT and mutated PAM motifs. The Cas9 (D10A/H840A)-RNA complex was titrated from 100 pM to 1 mM.

[0292] Figure 26 (A) Mutations of the PAM sequence in protospacer 2 plasmid DNA abolish

interference of plasmid maintenance by the Type II CRISPR/Cas system in bacterial cells. Wild-type protospacer 2 plasmids with a functional or mutated PAM were transformed into wild-type (strain SF370, also named EC904) and pre-crRNA-deficient mutant (EC1479) *S. pyogenes* as in Figure 12D. PAM mutations are not tolerated by the Type II CRISPR/Cas system in vivo. The mean values and standard deviations of three biological replicates are shown. **(B)** Mutations of the PAM sequence in protospacer plasmid DNA abolishes cleavage by Cas9-tracrRNA:crRNA. Wild type protospacer 2 plasmid with a functional or mutated PAM were subjected to Cas9 cleavage as in Figure 10A. The PAM mutant plasmids are not cleaved by the Cas9-tracrRNA:crRNA complex. (C) Mutations of the canonical PAM sequence abolish interference of plasmid maintenance by the Type II CRISPR/Cas system in bacterial cells. Wild-type protospacer 4 plasmids with a functional or mutated PAM were cleaved with Cas9 programmed with tracrRNA and crRNA-sp2. The cleavage reactions were carried out in the presence of the XmnI restriction endonuclease to visualize the Cas9 cleavage products as two fragments (~1880 and ~800 bp). Fragment sizes in base pairs are indicated.

Cas9 can be programmed with a single chimeric RNA

[0293] Examination of the likely secondary structure of the tracrRNA:crRNA duplex (Figures 10E and 12C) suggested the possibility that the features required for site-specific Cas9-catalyzed DNA cleavage could be captured in a single chimeric RNA. Although the tracrRNA:crRNA target-selection mechanism works efficiently in nature, the possibility of a single RNA-guided Cas9 is appealing due to its potential utility for programmed DNA cleavage and genome editing (Figure 1A-B). We designed two versions of a chimeric RNA containing a target recognition sequence at the 5' end followed by a hairpin structure retaining the base-pairing interactions that occur between the tracrRNA and the crRNA (Figure 14A). This single transcript effectively fuses the 3' end of crRNA to the 5' end of tracrRNA, thereby mimicking the dual-RNA structure required to guide site-specific DNA cleavage by Cas9. In cleavage assays using plasmid DNA, we observed that the longer chimeric RNA was able to guide Cas9-catalyzed DNA cleavage in a manner similar to that observed for the truncated tracrRNA:crRNA duplex (Figure 14A and Figure 27, A and C). The shorter chimeric RNA did not work efficiently in this assay, confirming that nucleotides that are 5 to 12 positions beyond the tracrRNA:crRNA base-pairing interaction are important for efficient Cas9 binding and/or target recognition. We obtained similar results in cleavage assays using short dsDNA as a substrate, further indicating that the position of the cleavage site in target DNA is identical to that observed using the dual tracrRNA:crRNA as a guide (Figure 14B and Figure 27, B and C). Finally, to establish whether the design of chimeric RNA might be universally applicable, we engineered five different chimeric guide RNAs to target a portion of the gene encoding the green-fluorescent protein (GFP) (Figure 28, A to C) and tested their efficacy against a plasmid carrying the GFP coding sequence in vitro. In all five cases, Cas9 programmed with these chimeric RNAs efficiently cleaved the plasmid at the correct target site (Figure 14C and Figure 28D), indicating that rational design of chimeric RNAs is robust and could, in principle, enable targeting of any DNA sequence of interest with few constraints beyond the presence of a GG dinucleotide adjacent to the targeted sequence.

[0294] Figure 1 A DNA-targeting RNA comprises a single stranded "DNA-targeting segment" and a "protein-binding segment," which comprises a stretch of double stranded RNA. (A) A DNA-targeting RNA can comprise two separate RNA molecules (referred to as a "double-molecule" or "two-molecule" DNA-targeting RNA). A double-molecule DNA-targeting RNA comprises a "targeter-RNA" and an "activator-RNA." (B) A DNA-targeting RNA can comprise a single RNA molecule (referred to as a "single-molecule" DNA-targeting RNA). A single-molecule DNA-targeting RNA comprises "linker nucleotides."

[0295] Figure 14 (A) A plasmid harboring protospacer 4 target sequence and a WT PAM was subjected to cleavage by Cas9 programmed with tracrRNA(4-89):crRNA-sp4 duplex or in vitro-transcribed chimeric RNAs constructed by joining the 3' end of crRNA to the 5' end of tracrRNA with a GAAA tetraloop. Cleavage reactions were analyzed by restriction mapping with XmnI. Sequences of chimeric RNAs A and B are shown with DNA-targeting (underline), crRNA repeat-derived sequences (overlined), and tracrRNA-derived (dashed underlined) sequences. **(B)** Protospacer 4 DNA duplex cleavage reactions were performed as in Figure 10B. **(C)** Five chimeric RNAs designed to target the GFP gene were used to program Cas9 to cleave a GFP gene-containing plasmid. Plasmid cleavage reactions were performed as in Figure 12E, except that the plasmid DNA was restriction mapped with AvrII after Cas9 cleavage.

[0296] Figure 27 (A) A single chimeric RNA guides Cas9-catalyzed cleavage of cognate protospacer plasmid DNA (protospacer 1 and protospacer 2). The cleavage reactions were carried out in the presence of the XmnI restriction endonuclease to visualize the Cas9 cleavage products as two fragments (~1880 and ~800 bp). Fragment sizes in base pairs are indicated. **(B)** A single chimeric RNA guides Cas9-catalyzed cleavage of cognate protospacer oligonucleotide DNA (protospacer 1 and protospacer 2). Fragment sizes in nucleotides are indicated. **(C)** Schematic representations of the chimeric RNAs used in the experiment. Sequences of chimeric RNAs A and B are shown with the 5' protospacer DNA-targeting sequence of crRNA (underlined), the tracrRNA-binding sequence of crRNA (overlined) and tracrRNA-derived sequence (dashed underlined).

[0297] Figure 28 (A) Schematic representation of the GFP expression plasmid pCFJ127. The targeted portion of the GFP open reading frame is indicated with a black arrowhead. **(B)** Close-up of the sequence of the targeted region. Sequences targeted by the chimeric RNAs are shown with gray bars. PAM dinucleotides are boxed. A unique SalI restriction site is located 60 bp upstream of the target locus. **(C)** Left: Target DNA sequences are shown together with their adjacent PAM motifs. Right: Sequences of the chimeric guide RNAs. **(D)** pCFJ127 was cleaved by Cas9 programmed with chimeric RNAs GFP1-5, as indicated. The plasmid was additionally digested with SalI and the reactions were analyzed by electrophoresis on a 3% agarose gel and visualized by staining with SYBR Safe.

Conclusions

[0298] A DNA interference mechanism was identified, involving a dual-RNA structure that directs a Cas9 endonuclease to introduce site-specific double-stranded breaks in target DNA. The tracrRNA:crRNA-guided Cas9 protein makes use of distinct endonuclease domains (HNH and RuvC-like domains) to cleave the two strands in the target DNA. Target recognition by Cas9 requires both a seed sequence in the crRNA and a GG dinucleotide-containing PAM sequence adjacent to the crRNA-binding region in the DNA target. We further show that the Cas9 endonuclease can be programmed with guide RNA engineered as a single transcript to target and cleave any dsDNA sequence of interest. The system is efficient, versatile, and programmable by changing the DNA target-binding sequence in the guide chimeric RNA. Zinc-finger nucleases and transcription-activator-like effector nucleases have attracted considerable interest as artificial enzymes engineered to manipulate genomes (35-38). This represents alternative methodology based on RNA-programmed Cas9 that facilitates gene-targeting and genome-editing applications.

References Cited

[0299]

1. 1. B. Wiedenheft, S. H. Sternberg, J. A. Doudna, Nature 482, 331 (2012).
2. 2. D. Bhaya, M. Davison, R. Barrangou, Annu. Rev. Genet. 45, 273 (2011).
3. 3. M. P. Terns, R. M. Terns, Curr. Opin. Microbiol. 14, 321 (2011).
4. 4. E. Deltcheva et al., Nature 471, 602 (2011).
5. 5. J. Carte, R. Wang, H. Li, R. M. Terns, M. P. Terns, Genes Dev. 22, 3489 (2008).
6. 6. R. E. Haurwitz, M. Jinek, B. Wiedenheft, K. Zhou, J. A. Doudna, Science 329, 1355 (2010).
7. 7. R. Wang, G. Preamplume, M. P. Terns, R. M. Terns, H. Li, Structure 19, 257 (2011).
8. 8. E. M. Gesner, M. J. Schellenberg, E. L. Garside, M. M. George, A. M. Macmillan, Nat. Struct. Mol. Biol. 18, 688 (2011).
9. 9. A. Hatoum-Aslan, I. Maniv, L. A. Marraffini, Proc. Natl. Acad. Sci. U.S.A. 108, 21218 (2011).
10. 10. S. J. J. Brouns et al., Science 321, 960 (2008).
11. 11. D. G. Sashital, M. Jinek, J. A. Doudna, Nat. Struct. Mol. Biol. 18, 680 (2011).
12. 12. N. G. Lintner et al., J. Biol. Chem. 286, 21643 (2011).
13. 13. E. Semenova et al., Proc. Natl. Acad. Sci. U.S.A. 108, 10098 (2011).
14. 14. B. Wiedenheft et al., Proc. Natl. Acad. Sci. U.S.A. 108, 10092 (2011).
15. 15. B. Wiedenheft et al., Nature 477, 486 (2011).
16. 16. C. R. Hale et al., Cell 139, 945 (2009).
17. 17. J. A. L. Howard, S. Delmas, I. Ivančić-Baće, E. L. Bolt, Biochem. J. 439, 85 (2011).
18. 18. E. R. Westra et al., Mol. Cell 46, 595 (2012).
19. 19. C. R. Hale et al., Mol. Cell 45, 292 (2012).
20. 20. J. Zhang et al., Mol. Cell 45, 303 (2012).
21. 21. K. S. Makarova et al., Nat. Rev. Microbiol. 9, 467 (2011).
22. 22. K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, E. V. Koonin, Biol. Direct 1,

7 (2006).
23. 23. K. S. Makarova, L. Aravind, Y. I. Wolf, E. V. Koonin, Biol. Direct 6, 38 (2011).
24. 24. S. Gottesman, Nature 471, 588 (2011).
25. 25. R. Barrangou et al., Science 315, 1709 (2007).
26. 26. J. E. Garneau et al., Nature 468, 67 (2010).
27. 27. R. Sapranauskas et al., Nucleic Acids Res. 39, 9275 (2011).
28. 28. G. K. Taylor, D. F. Heiter, S. Pietrokovski, B. L. Stoddard, Nucleic Acids Res. 39, 9705 (2011).
29. 29. H. Deveau et al., J. Bacteriol. 190, 1390 (2008).
30. 30. B. P. Lewis, C. B. Burge, D. P. Bartel, Cell 120, 15 (2005).
31. 31. G. Hutvagner, M. J. Simard, Nat. Rev. Mol. Cell Biol. 9, 22 (2008).
32. 32. F. J. M. Mojica, C. Díez-Villaseñor, J. Garcia-Martinez, C. Almendros, Microbiology 155, 733 (2009).
33. 33. L. A. Marraffini, E. J. Sontheimer, Nature 463, 568 (2010).
34. 34. D. G. Sashital, B. Wiedenheft, J. A. Doudna, Mol. Cell 46, 606 (2012).
35. 35. M. Christian et al., Genetics 186, 757 (2010).
36. 36. J. C. Miller et al., Nat. Biotechnol. 29, 143 (2011).
37. 37. F. D. Umov, E. J. Rebar, M. C. Holmes, H. S. Zhang, P. D. Gregory, Nat. Rev. Genet. 11, 636 (2010).
38. 38. D. Carroll, Gene Ther. 15, 1463 (2008).
39. 39. J. Sambrook, E. F. Fritsch, T. Maniatis, Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1989).
40. 40. M. G. Caparon, J. R. Scott, Genetic manipulation of pathogenic streptococci. Methods Enzymol. 204, 556 (1991). doi:10.1016/0076-6879(91)04028-M Medline
41. 41. C. Frøkjær-Jensen et al., Single-copy insertion of transgenes in Caenorhabditis elegans. Nat. Genet. 40, 1375 (2008). doi:10.1038/ng.248 Medline
42. 42. R. B. Denman, Using RNAFOLD to predict the activity of small catalytic RNAs. Biotechniques 15, 1090 (1993). Medline
43. 43. I. L. Hofacker, P. F. Stadler, Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics 22, 1172 (2006). doi:10.1093/bioinformatics/btl023 Medline
44. 44. K. Darty, A. Denise, Y. Ponty, VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 25, 1974 (2009). doi:10.1093/bioinformatics/btp250 Medline

Example 2: RNA-programmed genome editing in human cells

[0300] Data provided below demonstrate that Cas9 can be expressed and localized to the nucleus of human cells, and that it assembles with single-guide RNA ("sgRNA"; encompassing the features required for both Cas9 binding and DNA target site recognition) in a human cell. These complexes can generate double stranded breaks and stimulate non-homologous end

joining (NHEJ) repair in genomic DNA at a site complementary to the sgRNA sequence, an activity that requires both Cas9 and the sgRNA. Extension of the RNA sequence at its 3' end enhances DNA targeting activity in living cells. Further, experiments using extracts from transfected cells show that sgRNA assembly into Cas9 is the limiting factor for Cas9-mediated DNA cleavage. These results demonstrate that RNA-programmed genome editing works in living cells and *in vivo.*

MATERIALS AND METHODS

Plasmid design and construction

[0301] The sequence encoding *Streptococcus pyogenes* Cas9 (residues 1-1368) fused to an HA epitope (amino acid sequence DAYPYDVPDYASL (SEQ ID NO:274)), a nuclear localization signal (amino acid sequence PKKKRKVEDPKKKRKVD (SEQ ID NO:275)) was codon optimized for human expression and synthesized by GeneArt. The DNA sequence is SEQ ID NO:276 and the protein sequence is SEQ ID NO:277. Ligation-independent cloning (LIC) was used to insert this sequence into a pcDNA3.1-derived GFP and mCherry LIC vectors (vectors 6D and 6B, respectively, obtained from the UC Berkeley MacroLab), resulting in a Cas9-HA-NLS-GFP and Cas9-HA-NLS-mCherry fusions expressed under the control of the CMV promoter. Guide sgRNAs were expressed using expression vector pSilencer 2.1-U6 puro (Life Technologies) and pSuper (Oligoengine). RNA expression constructs were generated by annealing complementary oligonucleotides to form the RNA-coding DNA sequence and ligating the annealed DNA fragment between the BamHI and HindIII sites in pSilencer 2.1-U6 puro and BglII and HindIII sites in pSuper.

Cell culture conditions and DNA transfections

[0302] HEK293T cells were maintained in Dulbecco's modified eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS) in a 37°C humidified incubator with 5% CO_2. Cells were transiently transfected with plasmid DNA using either X-tremeGENE DNA Transfection Reagent (Roche) or Turbofect Transfection Reagent (Thermo Scientific) with recommended protocols. Briefly, HEK293T cells were transfected at 60-80% confluency in 6-well plates using 0.5 µg of the Cas9 expression plasmid and 2.0 µg of the RNA expression plasmid. The transfection efficiencies were estimated to be 30-50% for Tubofect (Figures 29E and 37A-B) and 80-90% for X-tremegene (Figure 31B), based on the fraction of GFP-positive cells observed by fluorescence microscopy. 48 hours post transfection, cells were washed with phosphate buffered saline (PBS) and lysed by applying 250 µl lysis buffer (20 mM Hepes pH 7.5, 100 mM potassium chloride (KCl), 5 mM magnesium chloride ($MgCl_2$), 1 mM dithiothreitol (DTT), 5% glycerol, 0.1% Triton X-100, supplemented with Roche Protease Inhibitor cocktail) and then rocked for 10 min at 4°C. The resulting cell lysate was divided into aliquots for further

analysis. Genomic DNA was isolated from 200 µl cell lysate using the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's protocol.

Western blot analysis of Cas9 expression

[0303] HEK293T , transfected with the Cas9-HA-NLS-GFP expression plasmid, were harvested and lysed 48 hours post transfection as above. 5 ul of lysate were eletrophoresed on a 10% SDS polyacrylamide gel, blotter onto a PVDF membrane and probed with HRP-conjugated anti-HA antibody (Sigma, 1: 1000 dilution in 1x PBS).

Surveyor assay

[0304] The Surveyor assay was performed as previously described [10,12,13]. Briefly, the human clathrin light chain A (CLTA) locus was PCR amplified from 200 ng of genomic DNA using a high fidelity polymerase, Herculase II Fusion DNA Polymerase (Agilent Technologies) and forward primer 5'-GCAGCAGAAGAAGCCTTTGT-3' (SEQ ID NO://) and reverse primer 5'-TTCCTCCTCTCCCTCCTCTC-3' (SEQ ID NO://). 300 ng of the 360 bp amplicon was then denatured by heating to 95°C and slowly reannealed using a heat block to randomly rehybridize wild type and mutant DNA strands. Samples were then incubated with Cel-1 nuclease (Surveyor Kit, Transgenomic) for 1 hour at 42°C. Cel-1 recognizes and cleaves DNA helices containing mismatches (wild type:mutant hybridization). Cel-1 nuclease digestion products were separated on a 10% acrylamide gel and visualized by staining with SYBR Safe (Life Technologies). Quantification of cleavage bands was performed using ImageLab software (Bio-Rad). The percent cleavage was determined by dividing the average intensity of cleavage products (160-200 bps) by the sum of the intensities of the uncleaved PCR product (360 bp) and the cleavage product.

***In vitro* transcription**

[0305] Guide RNA was *in vitro* transcribed using recombinant T7 RNA polymerase and a DNA template generated by annealing complementary synthetic oligonucleotides as previously described [14]. RNAs were purified by electrophoresis on 7M urea denaturing acrylamide gel, ethanol precipitated, and dissolved in DEPC-treated water.

Northern blot analysis

[0306] RNA was purified from HEK293T cells using the mirVana small-RNA isolation kit (Ambion). For each sample, 800 ng of RNA were separated on a 10% urea-PAGE gel after denaturation for 10 min at 70°C in RNA loading buffer (0.5X TBE (pH7.5), 0.5 mg/ml

bromophenol blue, 0.5 mg xylene cyanol and 47% formamide). After electrophoresis at 10W in 0.5X TBE buffer until the bromophenol blue dye reached the bottom of the gel, samples were electroblotted onto a Nytran membrane at 20 volts for 1.5 hours in 0.5X TBE. The transferred RNAs were crosslinked onto the Nytran membrane in UV-Crosslinker (Strategene) and were pre-hybridized at 45°C for 3 hours in a buffer containing 40% formamide, 5X SSC, 3X Dernhardt's (0.1% each of ficoll, polyvinylpyrollidone, and BSA) and 200 µg/ml Salmon sperm DNA. The pre-hybridized membranes were incubated overnight in the prehybridization buffer supplemented with 5'-^{32}P-labeled antisense DNA oligo probe at 1 million cpm/ml. After several washes in SSC buffer (final wash in 0.2X SCC), the membranes were imaged phosphorimaging.

***In vitro* cleavage assay**

[0307] Cell lysates were prepared as described above and incubated with CLTA-RFP donor plasmid [10]. Cleavage reactions were carried out in a total volume of 20 µl and contained 10 µl lysate, 2 µl of 5x cleavage buffer (100 mM HEPES pH 7.5, 500 mM KCl, 25 mM $MgCl_2$, 5 mM DTT, 25% glycerol) and 300 ng plasmid. Where indicated, reactions were supplemented with 10 pmol of *in vitro* transcribed CLTA1 sgRNA. Reactions were incubated at 37°C for one hour and subsequently digested with 10 U of XhoI (NEB) for an additional 30 min at 37°C. The reactions were stopped by the addition of Proteinase K (Thermo Scientific) and incubated at 37°C for 15 min. Cleavage products were analyzed by electrophoresis on a 1% agarose gel and stained with SYBR Safe. The presence of ~2230 and ~3100 bp fragments is indicative of Cas9-mediated cleavage.

RESULTS

[0308] To test whether Cas9 could be programmed to cleave genomic DNA in living cells, Cas9 was co-expressed together with an sgRNA designed to target the human clathrin light chain (CLTA) gene. The CLTA genomic locus has previously been targeted and edited using ZFNs [10]. We first tested the expression of a human-codon-optimized version of the *Streptococcus pyogenes* Cas9 protein and sgRNA in human HEK293T cells. The 160 kDa Cas9 protein was expressed as a fusion protein bearing an HA epitope, a nuclear localization signal (NLS), and green fluorescent protein (GFP) attached to the C-terminus of Cas9 (Figure 29A). Analysis of cells transfected with a vector encoding the GFP-fused Cas9 revealed abundant Cas9 expression and nuclear localization (Figure 29B). Western blotting confirmed that the Cas9 protein is expressed largely intact in extracts from these cells (Figure 29A). To program Cas9, we expressed sgRNA bearing a 5'-terminal 20-nucleotide sequence complementary to the target DNA sequence, and a 42-nucleotide 3'-terminal stem loop structure required for Cas9 binding (Figure 29C). This 3'-terminal sequence corresponds to the minimal stem-loop structure that has previously been used to program Cas9 *in vitro* [8]. The expression of this sgRNA was driven by the human U6 (RNA polymerase III) promoter [11]. Northern blotting

analysis of RNA extracted from cells transfected with the U6 promoter-driven sgRNA plasmid expression vector showed that the sgRNA is indeed expressed, and that their stability is enhanced by the presence of Cas9 (Figure 29D).

[0309] Figure 29 demonstrates that co-expression of Cas9 and guide RNA in human cells generates double-strand DNA breaks at the target locus. **(A)** Top; schematic diagram of the Cas9-HA-NLS-GFP expression construct. Bottom; lysate from HEK293T cells transfected with the Cas9 expression plasmid was analyzed by Western blotting using an anti-HA antibody. **(B)** Fluorescence microscopy of HEK293T cells expressing Cas9-HA-NLS-GFP. **(C)** Design of a single-guide RNA (sgRNA, i.e., a single-molecule DNA-targeting RNA) targeting the human CLTA locus. Top; schematic diagram of the sgRNA target site in exon 7 of the human CLTA gene. The target sequence that hybridizes to the guide segment of CLTA1 sgRNA is indicated by "CLTA1 sgRNA." The GG di-nucleotide protospacer adjacent motif (PAM) is marked by an arrow. Black lines denote the DNA binding regions of the control ZFN protein. The translation stop codon of the CLTA open reading frame is marked with a dotted line for reference. Middle; schematic diagram of the sgRNA expression construct. The RNA is expressed under the control of the U6 Pol III promoter and a poly(T) tract that serves as a Pol III transcriptional terminator signal. Bottom; sgRNA-guided cleavage of target DNA by Cas9. The sgRNA consists of a 20-nt 5'-terminal guide segment followed by a 42-nt stem-loop structure required for Cas9 binding. Cas9-mediated cleavage of the two target DNA strands occurs upon unwinding of the target DNA and formation of a duplex between the guide segment of the sgRNA and the target DNA. This is dependent on the presence of a PAM motif (appropriate for the Cas9 being used, e.g., GG dinucleotide, see Example 1 above) downstream of the target sequence in the target DNA. Note that the target sequence is inverted relative to the upper diagram. **(D)** Northern blot analysis of sgRNA expression in HEK239T cells. **(E)** Surveyor nuclease assay of genomic DNA isolated from HEK293T cells expressing Cas9 and/or CLTA sgRNA. A ZFN construct previously used to target the CLTA locus [10] was used as a positive control for detecting DSB-induced DNA repair by non-homologous end joining.

[0310] Next we investigated whether site-specific DSBs are generated in HEK293T cells transfected with Cas9-HA-NLS-mCherry and the CLTA1 sgRNA. To do this, we probed for minor insertions and deletions in the locus resulting from imperfect repair by DSB-induced NHEJ using the Surveyor nuclease assay [12]. The region of genomic DNA targeted by Cas9:sgRNA is amplified by PCR and the resulting products are denatured and reannealed. The rehybridized PCR products are incubated with the mismatch recognition endonuclease Cel-1 and resolved on an acrylamide gel to identify Cel-1 cleavage bands. As DNA repair by NHEJ is typically induced by a DSB, a positive signal in the Surveyor assay indicates that genomic DNA cleavage has occurred. Using this assay, we detected cleavage of the CLTA locus at a position targeted by the CLTA1 sgRNA (Figure 29E). A pair of ZFNs that target a neighboring site in the CLTA locus provided a positive control in these experiments [10].

[0311] To determine if either Cas9 or sgRNA expression is a limiting factor in the observed genome editing reactions, lysates prepared from the transfected cells were incubated with plasmid DNA harboring a fragment of the CLTA gene targeted by the CLTA1 sgRNA. Plasmid

DNA cleavage was not observed upon incubation with lysate prepared from cells transfected with the Cas9-HA-NLS-GFP expression vector alone, consistent with the Surveyor assay results. However, robust plasmid cleavage was detected when the lysate was supplemented with *in vitro* transcribed CLTA1 sgRNA (Figure 30A). Furthermore, lysate prepared from cells transfected with both Cas9 and sgRNA expression vectors supported plasmid cleavage, while lysates from cells transfected with the sgRNA-encoding vector alone did not (Figure 30A). These results suggest that a limiting factor for Cas9 function in human cells could be assembly with the sgRNA. We tested this possibility directly by analyzing plasmid cleavage in lysates from cells transfected as before in the presence and absence of added exogenous sgRNA. Notably, when exogenous sgRNA was added to lysate from cells transfected with both the Cas9 and sgRNA expression vectors, a substantial increase in DNA cleavage activity was observed (Figure 30B). This result indicates that the limiting factor for Cas9 function in HEK293T cells is the expression of the sgRNA or its loading into Cas9.

[0312] Figure 30 demonstrates that cell lysates contain active Cas9:sgRNA and support site-specific DNA cleavage. **(A)** Lysates from cells transfected with the plasmid(s) indicated at left were incubated with plasmid DNA containing a PAM and the target sequence complementary to the CLTA1 sgRNA; where indicated, the reaction was supplemented with 10 pmol of in vitro transcribed CLTA1 sgRNA; secondary cleavage with XhoI generated fragments of ~2230 and ~3100 bp fragments indicative of Cas9-mediated cleavage. A control reaction using lysate from cells transfected with a ZFN expression construct shows fragments of slightly different size reflecting the offset of the ZFN target site relative to the CLTA1 target site. **(B)** Lysates from cells transfected with Cas9-GFP expression plasmid and, where indicated, the CLTA1 sgRNA expression plasmid, were incubated with target plasmid DNA as in (A) in the absence or presence of in vitro-transcribed CLTA1 sgRNA.

[0313] As a means of enhancing the Cas9:sgRNA assembly *in living cells,* we next tested the effect of extending the presumed Cas9-binding region of the guide RNA. Two new versions of the CLTA1 sgRNA were designed to include an additional six or twelve base pairs in the helix that mimics the base-pairing interactions between the crRNA and tracrRNA (Figure 31A). Additionally, the 3'-end of the guide RNA was extended by five nucleotides based on the native sequence of the S. *pyogenes* tracrRNA [9]. Vectors encoding these 3' extended sgRNAs under the control of either the U6 or H1 Pol III promoters were transfected into cells along with the Cas9-HA-NLS-GFP expression vector and site-specific genome cleavage was tested using the Surveyor assay (Figure 31B). The results confirmed that cleavage required both Cas9 and the CLTA1 sgRNA, but did not occur when either Cas9 or the sgRNA were expressed alone. Furthermore, we observed substantially increased frequencies of NHEJ, as detected by Cel-1 nuclease cleavage, while the frequency of NHEJ mutagenesis obtained with the control ZFN pair was largely unchanged.

[0314] Figure 31 demonstrates that 3' extension of sgRNA constructs enhances site-specific NHEJ-mediated mutagenesis. **(A)** The construct for CLTA1 sgRNA expression (top) was designed to generate transcripts containing the original Cas9-binding sequence (v1.0), or dsRNA duplexes extended by 4 base pairs (v2.1) or 10 base pairs (v2.2). **(B)** Surveyor

nuclease assay of genomic DNA isolated from HEK293T cells expressing Cas9 and/or CLTA sgRNA v1.0, v2.1 or v2.2. A ZFN construct previously used to target the CLTA locus [10] was used as a positive control for detecting DSB-induced DNA repair by non-homologous end joining.

[0315] The results thus provide the framework for implementing Cas9 as a facile molecular tool for diverse genome editing applications. A powerful feature of this system is the potential to program Cas9 with multiple sgRNAs in the same cell, either to increase the efficiency of targeting at a single locus, or as a means of targeting several loci simultaneously. Such strategies would find broad application in genome-wide experiments and large-scale research efforts such as the development of multigenic disease models.

Example 3: The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems

[0316] We searched for all putative type II CRISPR-Cas loci currently existing in publicly available bacterial genomes by screening for sequences homologous to Cas9, the hallmark protein of the type II system. We constructed a phylogenetic tree from a multiple sequence alignment of the identified Cas9 orthologues. The CRISPR repeat length and gene organization of cas operons of the associated type II systems were analyzed in the different Cas9 subclusters. A subclassification of type II loci was proposed and further divided into subgroups based on the selection of 75 representative Cas9 orthologues. We then predicted tracrRNA sequences mainly by retrieving CRISPR repeat sequences and screening for anti-repeats within or in the vicinity of the cas genes and CRISPR arrays of selected type II loci. Comparative analysis of sequences and predicted structures of chosen tracrRNA orthologues was performed. Finally, we determined the expression and processing profiles of tracrRNAs and crRNAs from five bacterial species.

MATERIALS AND METHODS

Bacterial strains and culture conditions

[0317] The following media were used to grow bacteria on plates: TSA (trypticase soy agar, Trypticase™ Soy Agar (TSA II) BD BBL, Becton Dickinson) supplemented with 3% sheep blood for S. mutans (UA159), and BHI (brain heart infusion, BD Bacto™ Brain Heart Infusion, Becton Dickinson) agar for L. innocua (Clip11262). When cultivated in liquid cultures, THY medium (Todd Hewitt Broth (THB, Bacto, Becton Dickinson) supplemented with 0.2% yeast extract (Servabacter®) was used for S. mutans, BHI broth for L. innocua, BHI liquid medium containing 1% vitamin-mix VX (Difco, Becton Dickinson) for N. meningitidis (A Z2491), MH (Mueller Hinton Broth, Oxoid) Broth including 1% vitamin-mix VX for C. jejuni (NCTC 11168; ATCC 700819) and

TSB (Tryptic Soy Broth, BD BBL™ Trypticase™ Soy Broth) for F. novicida (U112). S. mutans was incubated at 37°C, 5% CO2 without shaking. Strains of L. innocua, N. meningitidis and F. novicida were grown aerobically at 37°C with shaking. C. jejuni was grown at 37°C in microaerophilic conditions using campygen (Oxoid) atmosphere. Bacterial cell growth was followed by measuring the optical density of cultures at 620 nm (OD_{620} nm) at regular time intervals using a microplate reader (BioTek PowerWave™).

Sequencing of bacterial small RNA libraries.

[0318] C. jejuni NCTC 11168 (ATCC 700819), F. novicida U112, L. innocua Clip11262, N. meningitidis A Z2491 and S. mutans UA159 were cultivated until mid-logarithmic growth phase and total RNA was extracted with TRIzol (Sigma-Aldrich). 10 µg of total RNA from each strain were treated with TURBO™ DNase (Ambion) to remove any residual genomic DNA. Ribosomal RNAs were removed by using the Ribo-Zero™ rRNA Removal Kits® for Gram-positive or Gram-negative bacteria (Epicentre) according to the manufacturer's instructions. Following purification with the RNA Clean & Concentrator™-5 kit (Zymo Research), the libraries were prepared using ScriptMiner™ Small RNA-Seq Library Preparation Kit (Multiplex, Illumina® compatible) following the manufacturer's instructions. RNAs were treated with the Tobacco Acid Pyrophosphatase (TAP) (Epicentre). Columns from RNA Clean & Concentrator™-5 (Zymo Research) were used for subsequent RNA purification and the Phusion® High-Fidelity DNA Polymerase (New England Biolabs) was used for PCR amplification. Specific userdefined barcodes were added to each library (RNA-Seq Barcode Primers (Illumina®- compatible) Epicentre) and the samples were sequenced at the Next Generation Sequencing (CSF NGS Unit; on the web at "csf." followed by "ac.at") facilty of the Vienna Biocenter, Vienna, Austria (Illumina single end sequencing).

Analysis of tracrRNA and crRNA sequencing data

[0319] The RNA sequencing reads were split up using the illumina2bam tool and trimmed by (i) removal of Illumina adapter sequences (cutadapt 1.0) and (ii) removal of 15 nt at the 3' end to improve the quality of reads. After removal of reads shorter than 15 nt, the cDNA reads were aligned to their respective genome using Bowtie by allowing 2 mismatches: C. jejuni (GenBank: NC_002163), F. novicida (GenBank: NC_008601), N. meningitidis (GenBank: NC_003116), L. innocua (GenBank: NC_003212) and S. mutans (GenBank: NC_004350). Coverage of the reads was calculated at each nucleotide position separately for both DNA strands using BEDTools-Version-2.15.0. A normalized wiggle file containing coverage in read per million (rpm) was created and visualized using the Integrative Genomics Viewer (IGV) tool ("www." followed by "broadinstitute.org/igv/") (Figure 36). Using SAMTools flagstat[80] the proportion of mapped reads was calculated on a total of mapped 9914184 reads for C. jejuni, 48205 reads

for F. novicida, 13110087 reads for N. meningitidis, 161865 reads L. innocua and 1542239 reads for S. mutans. A file containing the number of reads starting (5') and ending (3') at each single nucleotide position was created and visualized in IGV. For each tracrRNA orthologue and crRNA, the total number of reads retrieved was calculated using SAMtools.

Cas9 sequence analysis, multiple sequence alignment and guide tree construction

[0320] Position-Specific Iterated (PSI)-BLAST program was used to retrieve homologues of the Cas9 family in the NCBI non redundant database. Sequences shorter than 800 amino acids were discarded. The BLASTClust program set up with a length coverage cutoff of 0.8 and a score coverage threshold (bit score divided by alignment length) of 0.8 was used to cluster the remaining sequences (Figure 38). This procedure produced 78 clusters (48 of those were represented by one sequence only). One (or rarely a few representatives) were selected from each cluster and multiple alignment for these sequences was constructed using the MUSCLE program with default parameters, followed by a manual correction on the basis of local alignments obtained using PSI-BLAST and HHpred programs. A few more sequences were unalignable and also excluded from the final alignments. The confidently aligned blocks with 272 informative positions were used for maximum likelihood tree reconstruction using the FastTree program with the default parameters: JTT evolutionary model, discrete gamma model with 20 rate categories. The same program was used to calculate the bootstrap values.

[0321] Figure 38 depicts sequences that were grouped according to the BLASTclust clustering program. Only sequences longer than 800 amino acids were selected for the BLASTclust analysis (see Materials and Methods). Representative strains harboring *cas9* orthologue genes were used. Some sequences did not cluster, but were verified as Cas9 sequences due to the presence of conserved motifs and/or other *cas* genes in their immediate vicinity.

Analysis of CRISPR-Cas loci

[0322] The CRISPR repeat sequences were retrieved from the CRISPRdb database or predicted using the CRISPRFinder tool (Grissa I et al., BMC Bioinformatics 2007; 8:172; Grissa I et al., Nucleic Acids Res 2007). The cas genes were identified using the BLASTp algorithm and/or verified with the KEGG database (on the web at "www." followed by kegg.jp/).

In silico prediction and analysis of tracrRNA orthologues

[0323] The putative antirepeats were identified using the Vector NTI® software (Invitrogen) by screening for additional, degenerated repeat sequences that did not belong to the repeat-spacer array on both strands of the respective genomes allowing up to 15 mismatches. The transcriptional promoters and rho-independent terminators were predicted using the BDGP

Neural Network Promoter Prediction program ("www." followed by fruitfly.org/seq_tools/promoter.html) and the TransTennHP software, respectively. The multiple sequence alignments were performed using the MUSCLE program with default parameters. The alignments were analyzed for the presence of conserved structure motifs using the RNAalifold algorithm of the Vienna RNA package 2.0.

RESULTS

Type II CRISPR-Cas systems are widespread in bacteria.

[0324] In addition to the tracrRNA-encoding DNA and the repeat-spacer array, type II CRISPR-Cas loci are typically composed of three to four cas genes organized in an operon (Figure 32A-B). Cas9 is the signature protein characteristic for type II and is involved in the steps of expression and interference. Cas1 and Cas2 are core proteins that are shared by all CRISPR-Cas systems and are implicated in spacer acquisition. Csn2 and Cas4 are present in only a subset of type II systems and were suggested to play a role in adaptation. To retrieve a maximum number of type II CRISPR-Cas loci, containing tracrRNA, we first screened publicly available genomes for sequences homologous to already annotated Cas9 proteins. 235 Cas9 orthologues were identified in 203 bacterial species. A set of 75 diverse sequences representative of all retrieved Cas9 orthologues were selected for further analysis (Figure 32, Figure 38, and Materials and Methods).

[0325] Figure 32 depicts (A) a phylogenetic tree of representative Cas9 sequences from various organisms as well as (B) representative Cas9 locus architecture. Bootstrap values calculated for each node are indicated. Same color branches represent selected subclusters of similar Cas9 orthologues. CRISPR repeat length in nucleotides, average Cas9 protein size in amino acids (aa) and consensus locus architecture are shown for every subcluster. *-gi|116628213 **-gi|116627542 †- gi|34557790 ‡- gi|34557932. Type II-A is characterized by *cas9- csx12, cas1, cas2, cas4.* Type II-B is characterized by *cas9, cas1, cas2* followed by a *csn2* variant. Type II-C is characterized by a conserved *cas9, cas1, cas2* operon (See also Figure 38).

[0326] Next, we performed a multiple sequence alignment of the selected Cas9 orthologues. The comparative analysis revealed high diversities in amino acid composition and protein size. The Cas9 orthologues share only a few identical amino acids and all retrieved sequences have the same domain architecture with a central HNH endonuclease domain and splitted RuvC/RNaseH domain. The lengths of Cas9 proteins range from 984 (Campylobacter jejuni) to 1629 (Francisella novicida) amino acids with typical sizes of ~1100 or ~1400 amino acids. Due to the high diversity of Cas9 sequences, especially in the length of the inter-domain regions, we selected only well-aligned, informative positions of the prepared alignment to reconstruct a phylogenetic tree of the analyzed sequences (Figure 32 and Materials and Methods). Cas9

orthologues grouped into three major, monophyletic clusters with some outlier sequences. The observed topology of the Cas9 tree is well in agreement with the current classification of type II loci, with previously defined type II-A and type II-B forming separate, monophyletic clusters. To further characterize the clusters, we examined in detail the cas operon compositions and CRISPR repeat sequences of all listed strains.

Cas9 subclustering reflects diversity in type II CRISPR-Cas loci architecture

[0327] A deeper analysis of selected type II loci revealed that the clustering of Cas9 orthologue sequences correlates with the diversity in CRISPR repeat length. For most of the type II CRISPR-Cas systems, the repeat length is 36 nucleotides (nt) with some variations for two of the Cas9 tree subclusters. In the type II-A cluster (Figure 32) that comprises loci encoding the long Cas9 orthologue, previously named Csx12, the CRISPR repeats are 37 nt long. The small subcluster composed of sequences from bacteria belonging to the Bacteroidetes phylum (Figure 32) is characterized by unusually long CRISPR repeats, up to 48 nt in size. Furthermore, we noticed that the subclustering of Cas9 sequences correlates with distinct cas operon architectures, as depicted in Figure 32. The third major cluster (Figure 32) and the outlier loci (Figure 32), consist mainly of the minimum operon composed of the cas9, cas 1 and cas2 genes, with an exception of some incomplete loci that are discussed later. All other loci of the two first major clusters are associated with a fourth gene, mainly cas4, specific to type II-A or csn2- like, specific to type II-B (Figure 32). We identified genes encoding shorter variants of the Csn2 protein, Csn2a, within loci similar to type II-B S. pyogenes CRISPR01 and S. thermophilus CRISPR3 (Figure 32). The longer variant of Csn2, Csn2b, was found associated with loci similar to type II-B S. thermophilus CRISPR1 (Figure 32). Interestingly, we identified additional putative cas genes encoding proteins with no obvious sequence similarity to previously described Csn2 variants. One of those uncharacterized proteins is exclusively associated with type II-B loci of Mycoplasma species (Figure 32 and Figure 33). Two others were found encoded in type II-B loci of Staphylococcus species (Figure 33). In all cases the cas operon architecture diversity is thus consistent with the subclustering of Cas9 sequences. These characteristics together with the general topology of the Cas9 tree divided into three major, distinct, monophyletic clusters, led us to propose a new, further division of the type II CRISPR-Cas system into three subtypes. Type II-A is associated with Csx12- like Cas9 and Cas4, type II-B is associated with Csn2-like and type II-C only contains the minimal set of the cas9, cas 1 and cas2 genes, as depicted in Figure 32.

[0328] **Figure 33** depicts the architecture of type II CRISPR-Cas from selected bacterial species. The vertical bars group the loci that code for Cas9 orthologues belonging to the same tree subcluster (compare with **Figure 32**). Horizontal black bar, leader sequence; black rectangles and diamonds, repeat-spacer array. Predicted anti-repeats are represented by arrows indicating the direction of putative tracrRNA orthologue transcription. Note that for the loci that were not verified experimentally, the CRISPR repeat-spacer array is considered here to be transcribed from the same strand as the *cas* operon. The transcription direction of the putative tracrRNA orthologue is indicated accordingly.

***In silico* predictions of novel tracrRNA orthologues**

[0329] Type II loci selected earlier based on the 75 representative Cas9 orthologues were screened for the presence of putative tracrRNA orthologues. Our previous analysis performed on a restricted number of tracrRNA sequences revealed that neither the sequences of tracrRNAs nor their localization within the CRISPR-Cas loci seemed to be conserved. However, as mentioned above, tracrRNAs are also characterized by an anti-repeat sequence capable of base-pairing with each of the pre-crRNA repeats to form tracrRNA:precrRNA repeat duplexes that are cleaved by RNase III in the presence of Cas9. To predict novel tracrRNAs, we took advantage of this characteristic and used the following workflow: (i) screen for potential anti-repeats (sequence base-pairing with CRISPR repeats) within the CRISPR-Cas loci, (ii) select anti-repeats located in the intergenic regions, (iii) validate CRISPR anti-repeat:repeat base-pairing, and (iv) predict promoters and Rho-independent transcriptional terminators associated to the identified tracrRNAs.

[0330] To screen for putative anti-repeats, we retrieved repeat sequences from the CRISPRdb database or, when the information was not available, we predicted the repeat sequences using the CRISPRfinder software. In our previous study, we showed experimentally that the transcription direction of the repeat-spacer array compared to that of the cas operon varied among loci. Here RNA sequencing analysis confirmed this observation. In some of the analyzed loci, namely in F. novicida, N. meningitidis and C. jejuni, the repeat-spacer array is transcribed in the opposite direction of the cas operon (see paragraph 'Deep RNA sequencing validates expression of novel tracrRNA orthologues' and Figures 33 and 34) while in S. pyogenes, S. mutans, S. thermophilus and L. innocua, the array and the cas operon are transcribed in the same direction. These are the only type II repeat-spacer array expression data available to date. To predict the transcription direction of other repeat-spacer arrays, we considered the previous observation according to which the last repeats of the arrays are usually mutated. This remark is in agreement with the current spacer acquisition model, in which typically the first repeat of the array is duplicated upon insertion of a spacer sequence during the adaptation phase. For 37 repeat spacer arrays, we were able to identify the mutated repeat at the putative end of the arrays. We observed that the predicted orientation of transcription for the N. meningitidis and C. jejuni repeat-spacer array would be opposite to the orientation determined experimentally (RNA sequencing and Northern blot analysis). As the predicted orientation is not consistent within the clusters and as in most of the cases we could detect potential promoters on both ends of the arrays, we considered transcription of the repeat-spacer arrays to be in the same direction as transcription of the cas operon, if not validated otherwise.

[0331] Figure 34 depicts tracrRNA and pre-crRNA co-processing in selected type II CRISPR Cas systems. CRISPR loci architectures with verified positions and directions of tracrRNA and pre-crRNA transcription are shown. Top sequences, pre-crRNA repeats; bottom sequences, tracrRNA sequences base-pairing with crRNA repeats. Putative RNA processing sites as

revealed by RNA sequencing are indicated with arrowheads. For each locus, arrowhead sizes represent relative amounts of the retrieved 5' and 3' ends (see also **Figure 37**).

[0332] Figure 37 lists all tracrRNA orthologues and mature crRNAs retrieved by sequencing for the bacterial species studied, including coordinates (region of interest) and corresponding cDNA sequences (5' to 3'). The arrows represent the transcriptional direction (strand). Number of cDNA reads (calculated using SAMtools), coverage numbers (percentage of mapped reads) and predominant ends associated with each transcript are indicated. Numbers of reads starting or stopping at each nucleotide position around the 5' and 3' ends of each transcript are displayed. The sizes of each crRNA mature forms are indicated. The number allocated to each crRNA species corresponds to the spacer sequence position in the pre-crRNA, according to the CRISPRdb. The number allocated to each tracrRNA species corresponds to different forms of the same transcript.

[0333] We then screened the selected CRISPR-Cas loci including sequences located 1 kb upstream and downstream on both strands for possible repeat sequences that did not belong to the repeat-spacer array, allowing up to 15 mismatches. On average, we found one to three degenerated repeat sequences per locus that would correspond to anti-repeats of tracrRNA orthologues and selected the sequences located within the intergenic regions. The putative anti-repeats were found in four typical localizations: upstream of the cas9 gene, in the region between cas9 and cas1, and upstream or downstream of the repeat-spacer array (Figure 33). For every retrieved sequence, we validated the extent of base-pairing formed between the repeat and anti-repeat (Figure 44) by predicting the possible RNA:RNA interaction and focusing especially on candidates with longer and perfect complementarity region forming an optimal double-stranded structure for RNase III processing. To predict promoters and transcriptional terminators flanking the anti-repeat, we set the putative transcription start and termination sites to be included within a region located maximally 200 nt upstream and 100 nt downstream of the anti-repeat sequence, respectively, based on our previous observations[26]. As mentioned above, experimental information on the transcriptional direction of most repeat-spacer arrays of type II systems is lacking. The in silico promoter prediction algorithms often give false positive results and point to putative promoters that would lead to the transcription of repeat-spacer arrays from both strands. In some cases we could not predict transcriptional terminators, even though the tracrRNA orthologue expression could be validated experimentally, as exemplified by the C. jejuni locus (see paragraph 'Deep RNA sequencing validates expression of novel tracrRNA orthologues'). We suggest to consider promoter and transcriptional terminator predictions only as a supportive, but not essential, step of the guideline described above.

[0334] Figure 44 depicts predicted pre-crRNA repeat:tracrRNA anti-repeat basepairing in selected bacterial species. [b]The CRISPR loci belong to the type II (Nmeni/CASS4) CRISPR-Cas system. Nomenclature is according to the CRISPR database (CRISPRdb). Note that S. thermophilus LMD-9 and W. succinogenes contain two type II loci. [c]Upper sequence, pre-crRNA repeat consensus sequence (5' to 3'); lower sequence, tracrRNA homologue sequence

annealing to the repeat (anti-repeat; 3' to 5'). Note that the repeat sequence given is based on the assumption that the CRISPR repeat-spacer array is transcribed from the same strand as the cas operon. For the sequences that were validated experimentally in this study, RNA sequencing data were taken into account to determine the base-pairing. See Figure 33. [d]Two possible anti-repeats were identified in the F. tularensis subsp. novicida, W. succinogenes and gamma proteobacterium HTCC5015 type II-A loci. Upper sequence pairing, anti-repeat within the putative leader sequence; lower sequence pairing, anti-repeat downstream of the repeat spacer array. See Figure 33. [e]Two possible anti-repeats were identified in the S. wadsworthensis type II-A locus. Upper sequence pairing, anti-repeat; lower sequence pairing, anti-repeat within the putative leader sequence See Figure 33. [f]Two possible anti-repeats were identified in the L. gasseri type II-B locus. Upper sequence pairing, anti-repeat upstream of cas9; lower sequence pairing, anti-repeat between the cas9 and cas1 genes. See Figure 33. [g]Two possible anti-repeats were identified in the C. jejuni type II-C loci. Upper sequence pairing, anti-repeat upstream of cas9; lower sequence pairing, anti-repeat downstream of the repeat-spacer array. See Figure 33. [h]Two possible anti-repeats were identified in the R. rubrum type II-C locus. Upper sequence pairing, antirepeat downstream of the repeat-spacer array; lower sequence pairing, anti-repeat upstream of cas1. See Figure 33.

A plethora of tracrRNA orthologues

[0335] We predicted putative tracrRNA orthologues for 56 of the 75 loci selected earlier. The results of predictions are depicted in Figure 33. As already mentioned, the direction of tracrRNA transcription indicated in this figure is hypothetical and based on the indicated direction of repeat-spacer array transcription. As previously stated, sequences encoding putative tracrRNA orthologues were identified upstream, within and downstream of the cas operon, as well as downstream of the repeat spacer arrays, including the putative leader sequences, commonly found in type II-A loci (Figure 33). However, we observed that anti-repeats of similar localization within CRISPR-Cas loci can be transcribed in different directions (as observed when comparing e.g. Lactobacillus rhamnosus and Eubacterium rectale or Mycoplasma mobile and S. pyogenes or N. meningitidis) (Figure 33). Notably, loci grouped within a same subcluster of the Cas9 guide tree share a common architecture with respect to the position of the tracrRNA-encoding gene. We identified anti-repeats around the repeat-spacer array in type II-A loci, and mostly upstream of the cas9 gene in types II-B and II-C with several notable exceptions for the putative tracrRNA located between cas9 and cas1 in three distinct subclusters of type II-B.

Some type II CRISPR-Cas loci have defective repeat-spacer arrays and/or tracrRNA orthologues

[0336] For six type II loci (Fusobacterium nucleatum, Aminomonas paucivorans, Helicobacter

mustelae, Azospirillum sp., Prevotella ruminicola and Akkermansia muciniphila), we identified potential anti-repeats with weak base-pairing to the repeat sequence or located within the open reading frames. Notably, in these loci, a weak anti-repeat within the open reading frame of the gene encoding a putative ATPase in A. paucivorans, a strong anti-repeat within the first 100 nt of the cas9 gene in Azospirillum sp. B510 and a strong anti-repeat overlapping with both cas9 and cas1 in A. muciniphila were identified (Figure 33). For twelve additional loci (Peptoniphilus duerdenii, Coprococcus catus, Acidaminococcus intestini, Catenibacterium mitsuokai, Staphylococcus pseudintermedius, Ilyobacter polytropus, Elusimicrobium minutum, Bacteroides fragilis, Acidothermus cellulolyticus, Corynebacterium diphteriae, Bifidobacterium longum and Bifidobacterium dentium), we could not detect any putative anti-repeat. There is no available information on pre-crRNA expression and processing in these CRISPR-Cas loci. Thus, the functionality of type II systems in the absence of a clearly defined tracrRNA orthologue remains to be addressed. For seven analyzed loci we could not identify any repeat spacer array (Parasutterella excrementihominis, Bacillus cereus, Ruminococcus albus, Rhodopseudomonas palustris, Nitrobacter hamburgensis, Bradyrhizobium sp. and Prevotella micans) (Figure 33) and in three of those (Bradyrhizobium sp. BTAi1, N. hamburgensis and B. cereus) we detected cas9 as a single gene with no other cas genes in the vicinity. For these three loci, we failed to predict any small RNA sequence upstream or downstream of the cas9 gene. In the case of R. albus and P. excrementihominis, the genomic contig containing cas9 is too short to allow prediction of the repeat spacer array.

Deep RNA sequencing validates expression of novel tracrRNA orthologues

[0337] To verify the in silico tracrRNA predictions and determine tracrRNA:pre-crRNA coprocessing patterns, RNAs from selected Gram-positive (S. mutans and L. innocua) and Gram-negative (N. meningitidis, C. jejuni and F. novicida) bacteria were analyzed by deep sequencing. Sequences of tracrRNA orthologues and processed crRNAs were retrieved (Figure 36 and Figure 37). Consistent with previously published differential tracrRNA sequencing data in S. pyogenes[26], tracrRNA orthologues were highly represented in the libraries, ranging from 0.08 to 6.2% of total mapped reads. Processed tracrRNAs were also more abundant than primary transcripts, ranging from 66% to more than 95% of the total amount of tracrRNA reads (Figure 36 and Figure 37).

[0338] Figure 36 depicts the expression of bacterial tracrRNA orthologues and crRNAs revealed by deep RNA sequencing. Expression profiles of tracrRNA orthologues and crRNAs of selected bacterial strains are represented along the corresponding genomes by bar charts (Images captured from the Integrative Genomics Viewer (IGV) tool). Campylobacter jejuni (GenBank: NC_002163), Francisella novicida (GenBank: NC_008601), Neisseria meningitidis (GenBank: NC_003116), Listeria innocua (GenBank: NC _003212) and Streptococcus mutans (GenBank: NC_004350). Genomic coordinates are given. [a]Sequence coverage calculated using BEDTools-Version-2.15.0 (Scale given in reads per million). [b]Distribution of reads starting (5') and ending (3') at each nucleotide position are indicated (Scale given in numbers of

reads). Upper panels correspond to transcripts from the positive strand and lower panels correspond to transcripts from the negative strand. The negative coverage values and peaks presented below the axes indicate transcription from the negative strand of the genome. Predominant 5'- and 3'-ends of the reads are plotted for all RNAs. Note that given the low quality of L. innocua cDNA library, the reads are shortened for crRNAs, and an accumulation of the reads at the 3' end of tracrRNA is observed, presumably due to RNA degradation.

[0339] To assess the 5' ends of tracrRNA primary transcripts, we analyzed the abundance of all 5' end reads of tracrRNA and retrieved the most prominent reads upstream or in the vicinity of the 5' end of the predicted anti-repeat sequence. The 5' ends of tracrRNA orthologues were further confirmed using the promoter prediction algorithm. The identified 5' ends of tracrRNAs from S. mutans, L. innocua and N. meningitidis correlated with both in silico predictions and Northern blot analysis of tracrRNA expression[26]. The most prominent 5' end of C. jejuni tracrRNA was identified in the middle of the anti-repeat sequence. Five nucleotides upstream, an additional putative 5' end correlating with the in silico prediction and providing longer sequence of interaction with the CRISPR repeat sequence was detected. We retrieved relatively low amount of reads from the F. novicida library that corresponded almost exclusively to processed transcripts. Analysis of the very small amount of reads of primary transcripts provided a 5' end that corresponded to the strong in silico promoter predictions. Northern blot probing of F. novicida tracrRNA further confirmed the validity of the predictions showing the low abundance of transcripts of around 90 nt in length. The results are listed in **Table 2.** For all examined species, except N. meningitidis, primary tracrRNA transcripts were identified as single small RNA species of 75 to 100 nt in length. In the case of N. meningitidis, we found a predominant primary tracrRNA form of ~110 nt and a putative longer transcript of ~170 nt represented by a very low amount of reads and detected previously as a weak band by Northern blot analysis.

Table 2. Selected tracrRNA orthologues

Strains[a]	**Transcript**	**5'-end"**			**3'-end[c]**	**Length (nt)**
		RNA-seq		**Predicted**		
		First read	**Most prominent**			
S. pyogenes SF370	*primary*	-	854 546	-	854 376	171
	primary	-	854 464	-		89
	processed	-	854 450	-		~75
C. jejuni NCTC 11168	*primary*	1 455 497	1 455 502	1 455 497	1 455 570	~75
	processed	-	1 455 509	-		~60
L. innocua Clip11262	*primary*	2 774 774	2 774 774	2 774 773	2 774 863	~90
	processed	-	2 774 788	-		~75
S. mutans UA159	*primary*	1 335 040	1 335 040	1 355 039	1 335 141	~100

Strains[a]	Transcript	5'-end"			3'-end[c]	Length (nt)
		RNA-seq		Predicted		
		First read	Most prominent			
	processed	-	1 335 054	-		~85
			1 335 062			~80
N. meningitidis A Z2491	*primary*	614 158	614 162	614 154	614 333	~175
	primary	614 223	614 225	614 223		~110
	processed	-	614 240	-		~90
F. novicida U112	*primary*	817 144	-	817 145	817 065	~80
				817 154		
	processed	-	817 138	-		~75
			817 128			~65
S. thermophilus LMD-9	*primary*	-	-	1 384 330	1 384 425	~95
	primary	-	-	646 654	646 762	~110
P. multocida Pm70	*primary*	-	-	1 327 287	1 327 396	~110
M. mobile 163K	*primary*	-	-	49 470	49 361	~110

"tracrRNA orthologues of S. *thermophilus, P. multocida* and *M. mobile* were predicted *in silico.* [b]RNA-seq, revealed by RNA sequencing (Table S3); first read, first 5'-end position retrieved by sequencing; most prominent, abundant 5'-end according to RNA-seq data; predicted, *in silico* prediction of transcription start site; underlined, 5'-end chosen for the primary tracrRNA to be aligned.

[c]Estimated 3' end according to RNA-seq data and transcriptional terminator prediction.

tracrRNA and pre-crRNA co-processing sites lie in the anti-repeat:repeat region.

[0340] We examined the processed tracrRNA transcripts by analyzing abundant tracrRNA 5' ends within the predicted anti-repeat sequence and abundant mature crRNA 3' ends (Figure 34 and 45). In all species, we identified the prominent 5' ends of tracrRNA orthologues that could result from co-processing of the tracrRNA:pre-crRNA repeat duplexes by RNase III. We also identified the processed 5'-ends of crRNAs that most probably result from a second maturation event by putative trimming, consistently with previous observations. Noteworthy, in the closely related RNA pairs of S. pyogenes, S. mutans and L. innocua, we observed the

same processing site around the G:C basepair in the middle of the anti-repeat sequence. In both S. mutans and L. innocua, we detected additional prominent tracrRNA 5' ends and crRNA 3' ends that could suggest further trimming of the tracrRNA:crRNA duplex, with 3'-end of crRNA being shortened additionally to the already mentioned 5'-end trimming, following the RNase III-catalyzed first processing event. Similarly, in C. jejuni we found only a small amount of crRNA 3' ends that would fit to the RNase III processing patterns and retrieved the corresponding 5' ends of processed tracrRNA. Thus, the putative trimming of tracrRNA:crRNA duplexes after initial cleavage by RNase III would result in a shorter repeat-derived part in mature crRNAs, producing shorter tracrRNA:crRNA duplexes stabilized by a triple G:C base-pairing for interaction with the endonuclease Cas9 and subsequent cleavage of target DNAs. The N. meningitidis RNA duplex seems to be processed at two primary sites further to the 3' end of the CRISPR repeat, resulting in a long repeat-derived part in mature crRNA and stable RNA:RNA interaction despite the central bulge within the duplex. Interestingly, the tracrRNA:pre-crRNA duplex of F. novicida seems to be cleaved within the region of low complementarity and some of the retrieved abundant 5' ends of tracrRNA suggest its further trimming without concomitant trimming of crRNA. Differences in primary transcript sizes and in the location of processing sites result in various lengths of processed tracrRNAs ranging from ~65 to 85 nt. The coordinates and sizes of the prominent processed tracrRNA transcripts are shown in Table 2 and Figure 37. The observed processing patterns of tracrRNA and crRNA are well in agreement with the previously proposed model of two maturation events. The putative further trimming of some of the tracrRNA 5'-ends and crRNA 3'-ends could stem from the second maturation event or alternatively, be an artifact of the cDNA library preparation or RNA sequencing. The nature of these processings remains to be investigated further.

Sequences of tracrRNA orthologues are highly diverse

[0341] Sequences similarities of selected tracrRNA orthologues were also determined. We performed multiple sequence alignments of primary tracrRNA transcripts of S. pyogenes (89 nt form only), S. mutans, L. innocua and N. meningitidis (110 nt form only), S. thermophilus, P. multocida and M. mobile (**Table 2, Figure 35**). We observed high diversity in tracrRNA sequences but significant conservation of sequences from closely related CRISPR-Cas loci. tracrRNAs from L. innocua, S. pyogenes, S. mutans and S. thermophiles share on average 77% identity and tracrRNAs from N. meningitidis and P. multocida share 82% identity according to pairwise alignments. The average identity of the analyzed tracrRNA sequences is 56%, comparable to the identity of random RNA sequences. This observation further confirms that the prediction of tracrRNA orthologues based on sequence similarity can be performed only in the case of closely related loci. We also sought for possible tracrRNA structure conservation but could not find any significant similarity except one co-variation and conserved transcriptional terminator structure (**Figure 35**).

[0342] Figure 35 depicts sequence diversity of tracrRNA orthologues. tracrRNA sequence multiple alignment. S. thermophilus and S. thermophilus2, tracrRNA associated with SEQ ID NO:41 and SEQ ID NO:40 Cas9 orthologues, accordingly. Black, highly conserved; dark grey,

conserved; light grey, weakly conserved. Predicted consensus structure is depicted on the top of the alignment. Arrows indicate the nucleotide covariations. S. pyogenes SF370, S. mutans UA159, L. innocua Clip11262, C. jejuni NCTC 11168, F. novicida U112 and N. meningitidis A Z2491 tracrRNAs were validated by RNA sequencing and Northern blot analysis. S. thermophiles LMD-9 tracrRNA was validated by Northern blot analysis. P. multocida Pm70 tracrRNA was predicted from high similarity of the CRISPR-Cas locus with that of N. meningitidis A Z2491. M. mobile 163K tracrRNA was predicted in silico from strong predictions of transcriptional promoter and terminator.

Example 4: Cas9 can use artificial guide RNAs, not existing in nature, to perform target DNA cleavage

[0343] An artificial crRNA and an artificial tracrRNA were designed based on the protein-binding segment of naturally occurring transcripts of S. pyogenes crRNA and tracrRNAs, modified to mimic the asymmetric bulge within natural S. pyogenes crRNA:tracrRNA duplex (see the bulge in the protein-binding domain of both the artificial (top) and natural (bottom) RNA molecules depicted in Figure 39A). The artificial tracrRNA sequence shares less than 50% identity with the natural tracrRNA. The predicted secondary structure of the crRNA:tracrRNA protein-binding duplex is the same for both RNA pairs, but the predicted structure of the rest of the RNAs is much different.

[0344] Figure 39 demonstrates that artificial sequences that share very little (roughly 50% identity) with naturally occurring a tracrRNAs and crRNAs can function with Cas9 to cleave target DNA as long as the structure of the protein-binding domain of the DNA-targeting RNA is conserved. (**A**) Co-folding of *S. pyogenes* tracrRNA and crRNA and artificial tracrRNA and crRNA. (**B**) Combinations of *S. pyogenes* Cas9 and tracrRNA:crRNA orthologs were used to perform plasmid DNA cleavage assays .Spy - *S. pyogenes,* Lin - *L. innocua,* Nme - N. *meningitidis,* Pmu - *P. multocida. S. pyogenes* Cas9 can be guided by some, but not all tracrRNA:crRNA orthologs naturally occurring in selected bacterial species. Notably, *S. pyogenes* Cas9 can be guided by the artificial tracrRNA:crRNA pair, which was designed based on the structure of the protein-binding segment of the naturally occurring DNA-targeting RNA using sequence completely unrelated to the CRISPR system.

[0345] The artificial "tracrRNA" (activator RNA) used was 5'-GUUUUCCCUUUUCAAAGAAAUCUCCUGGGCACCUAUCUUCUUAGGUGCCCUCCCU UGUUUAAACCUGACCAGUUAACCGGCUGGUUAGGUUUUU-3' (SEQ ID NO: 1347). The artificial "crRNA" (targeter RNA) used was: 5'- GAGAUUUAUGAAAAGGGAAAAC-3' (SEQ ID NO: 1348).

Example 5: Generation of non-human Transgenic Organisms

[0346] A transgenic mouse expressing Cas9 (either unmodified, modified to have reduced enzymatic activity, modified as a fusion protein for any of the purposes outline above) is generated using a convenient method known to one of ordinary skill in the art (e.g., (i) gene knock-in at a targeted locus (e.g., ROSA 26) of a mouse embryonic stem cell (ES cell) followed by blastocyst injection and the generation of chimeric mice; (ii) injection of a randomly integrating transgene into the pronucleus of a fertilized mouse oocyte followed by ovum implantation into a pseudopregnant female; etc.). The Cas9 protein is under the control of a promoter that expresses at least in embryonic stem cells, and may be additionally under temporal or tissue-specific control (e.g., drug inducible, controlled by a Cre/Lox based promoter system, etc.). Once a line of transgenic Cas9 expressing mice is generated, embryonic stem cells are isolated and cultured and in some cases ES cells are frozen for future use. Because the isolated ES cells express Cas9 (and in some cases the expression is under temporal control (e.g., drug inducible), new knock-out or knock-in cells (and therefore mice) are rapidly generated at any desired locus in the genome by introducing an appropriately designed DNA-targeting RNA that targets the Cas9 to a particular locus of choice. Such a system, and many variations thereof, is used to generate new genetically modified organisms at any locus of choice. When modified Cas9 is used to modulate transcription and/or modify DNA and/or modify polypeptides associated with DNA, the ES cells themselves (or any differentiated cells derived from the ES cells (e.g., an entire mouse, a differentiated cell line, etc.) are used to study to properties of any gene of choice (or any expression product of choice, or any genomic locus of choice) simply by introducing an appropriate DNA-targeting RNA into a desired Cas9 expressing cell.

REFERENCES CITED IN THE DESCRIPTION

Cited references

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- WO2011072246A [0002]
- US7169874B [0040]
- WO2001018048A [0040]
- US7029913B [0071]

- US5843780A [0071]
- US6200806B [0071]
- WO9920741A [0071]
- WO0151616A [0071]
- WO03020920A [0071]
- US7153684B [0072]
- US20090047263A [0073]
- US20090068742A [0073]
- US20090191159A [0073]
- US20090227032A [0073]
- US20090246875A [0073]
- US20090304646A [0073]
- US5489677A [0122]
- USNO5602240A [0122]
- US5034506A [0123] [0127]
- US5539082A [0126]
- US5714331A [0126]
- US5719262A [0126]
- WO9839352A [0130]
- WO9914226A [0130]
- US3687808A [0134]
- US7078387B [0143] [0159]
- WO9412649A [0144] [0160]
- WO9303769A [0144] [0160]
- WO9319191A [0144] [0160]
- WO9428938A [0144] [0160]
- WO9511984A [0144] [0160]
- WO9500655A [0144] [0160]
- WO9309239A [0144] [0160]
- US5222982A [0179]
- US5385582A [0179]
- US20070254842A [0179]
- US20080081064A [0179]
- US20090196903A [0179]

Non-patent literature cited in the description

- **SAMBROOK, J.FRITSCH, E. FMANIATIS, T**Molecular Cloning: A Laboratory ManualCold Spring Harbor Laboratory Press, Cold Spring Harbor19890000 [0022]
- **SAMBROOK, J.RUSSELL, W.**Molecular Cloning: A Laboratory ManualCold Spring

Harbor Laboratory Press, Cold Spring Harbor20010000 [0022]
- **ALTSCHUL et al.**J. Mol. Biol., 1990, vol. 215, 403-410 [0024]
- **ZHANGMADDEN**Genome Res., 1997, vol. 7, 649-656 [0024]
- **SMITHWATERMAN**Adv. Appl. Math., 1981, vol. 2, 482-489 [0024]
- **ALTSCHUL et al.**J. Mol. Bioi., 1990, vol. 215, 403-10 [0029]
- **MIYAGISHI et al.**Nature Biotechnology, 2002, vol. 20, 497-500 [0034]
- **XIA et al.**Nucleic Acids Res., 2003, vol. 31, 17 [0034]
- **CHEN et al.**Cell, 1987, vol. 51, 7-19 [0037]
- **LLEWELLYN et al.**Nat. Med., 2010, vol. 16, 101161-1166 [0037]
- GenBankS62283 [0037]
- **OH et al.**Gene Ther, 2009, vol. 16, 437- [0037]
- **SASAOKA et al.**Mol. Brain Res, 1992, vol. 16, 274- [0037]
- **BOUNDY et al.**J. Neurosci., 1998, vol. 18, 9989- [0037]
- **KANEDA et al.**Neuron, 1991, vol. 6, 583-594 [0037]
- **RADOVICK et al.**Proc. Natl. Acad. Sci. USA, 1991, vol. 88, 3402-3406 [0037]
- **OBERDICK et al.**Science, 1990, vol. 248, 223-226 [0037]
- **BARTGE et al.**Proc. Natl. Acad. Sci. USA, vol. 85, 3648-3652 [0037]
- **COMB et al.**EMBO J., 1988, vol. 17, 3793-3805 [0037]
- **MAYFORD et al.**Proc. Natl. Acad. Sci. USA, 1996, vol. 93, 13250- [0037]
- **CASANOVA et al.**Genesis, 2001, vol. 31, 37- [0037]
- **LIU et al.**Gene Therapy, 2004, vol. 11, 52-60 [0037]
- **TOZZO et al.**Endocrinol, 1997, vol. 138, 1604- [0038]
- **ROSS et al.**Proc. Natl. Acad. Sci. USA, 1990, vol. 87, 9590- [0038]
- **PAVJANI et al.**Nat. Med., 2005, vol. 11, 797- [0038]
- **KNIGHT et al.**Proc. Natl. Acad. Sci. USA, 2003, vol. 100, 14725- [0038]
- **KURIKI et al.**Biol. Pharm. Bull., 2002, vol. 25, 1476- [0038]
- **SATO et al.**J. Biol. Chem., 2002, vol. 277, 15703- [0038]
- **TABOR et al.**J. Biol. Chem., 1999, vol. 274, 20603- [0038]
- **MASON et al.**Endocrinol, 1998, vol. 139, 1013- [0038]
- **CHEN et al.**Biochem. Biophys. Res. Comm., 1999, vol. 262, 187- [0038]
- **KITA et al.**Biochem. Biophys. Res. Comm., 2005, vol. 331, 484- [0038]
- **CHAKRABARTI**Endocrinol, 2010, vol. 151, 2408- [0038]
- **PLATT et al.**Proc. Natl. Acad. Sci. USA, vol. 86, 7490- [0038]
- **SEO et al.**Molec. Endocrinol., 2003, vol. 17, 1522- [0038]
- **FRANZ et al.**Cardiovasc. Res., 1997, vol. 35, 560-566 [0039]
- **ROBBINS et al.**Ann. N.Y. Acad. Sci., 1995, vol. 752, 492-505 [0039]
- **LINN et al.**Circ. Res., 1995, vol. 76, 584-591 [0039]
- **PARMACEK et al.**Mol. Cell. Biol., 1994, vol. 14, 1870-1885 [0039]
- **HUNTER et al.**Hypertension, 1993, vol. 22, 608-617 [0039]
- **SARTORELLI et al.**Proc. Natl. Acad. Sci. USA, 1992, vol. 89, 4047-4051 [0039]
- **AKYÜREK et al.**Mol. Med., 2000, vol. 6, 983- [0040]
- **KIM et al.**Mol. Cell. Biol., 1997, vol. 17, 2266-2278 [0040]
- **LI et al.**J. Cell Biol., 1996, vol. 132, 849-859 [0040]
- **MOESSLER et al.**Development, 1996, vol. 122, 2415-2425 [0040]

- **YOUNG et al.**Ophthalmol. Vis. Sci., 2003, vol. 44, 4076- [0041]
- **NICOUD et al.**J. Gene Med, 2007, vol. 9, 1015- [0041]
- **YOKOYAMA et al.**Exp Eye Res., 1992, vol. 55, 225- [0041]
- **PANYAM**Adv Drug Deliv Rev, 2012, 1200283-9 [0052]
- **AUSUBEL et al.**Short Protocols in Molecular BiologyWiley & Sons19950000 [0053]
- **MORRISON et al.**Cell, 1997, vol. 88, 287-298 [0068]
- **TAKAHASHI**Cell, 2007, vol. 131, 5861-72 [0070]
- **TAKAHASHI**Nat Protoc., 2007, vol. 2, 123081-9 [0070]
- **YU**Science, 2007, vol. 318, 58581917-20 [0070]
- **THOMSON et al.**Science, 1998, vol. 282, 1145- [0071]
- **THOMSON et al.**Proc. Natl. Acad. Sci USA, 1995, vol. 92, 7844- [0071]
- **THOMSON et al.**Biol. Reprod., 1996, vol. 55, 254- [0071]
- **SHAMBLOTT et al.**Proc. Natl. Acad. Sci. USA, 1998, vol. 95, 13726- [0071]
- **MATSUI, Y. et al.**Cell, 1992, vol. 70, 841- [0072]
- **SHAMBLOTT, M. et al.**Proc. Natl. Acad. Sci. USA, 2001, vol. 98, 113- [0072]
- **SHAMBLOTT, M. et al.**Proc. Natl. Acad. Sci. USA, 1998, vol. 95, 13726- [0072]
- **KOSHIMIZU, U. et al.**Development, 1996, vol. 122, 1235- [0072]
- **SAMBROOK et al.**Molecular Cloning: A Laboratory ManualHaRBor Laboratory Press20010000 [0082]
- Short Protocols in Molecular BiologyJohn Wiley & Sons19990000 [0082]
- **BOLLAG et al.**Protein MethodsJohn Wiley & Sons19960000 [0082]
- Vectors for Gene TherapyAcademic Press19990000 [0082]
- Viral VectorsAcademic Press19950000 [0082]
- Immunology Methods ManualAcademic Press19970000 [0082]
- **DOYLEGRIFFITHS**Cell and Tissue Culture: Laboratory Procedures in BiotechnologyJohn Wiley & Sons19980000 [0082]
- **NAKAMURA et al.**Genes Cells., 2012, vol. 17, 5344-64 [0103] [0213]
- **VAVALLE et al.**Future Cardiol., 2012, vol. 8, 3371-82 [0103] [0213]
- **CITARTAN et al.**Biosens Bioelectron., 2012, vol. 34, 11-11 [0103] [0213]
- **LIBERMAN et al.**Wiley Interdiscip Rev RNA, 2012, vol. 3, 3369-84 [0103] [0213]
- **DWAINE A. BRAASCHDAVID R. COREY**Biochemistry, 2002, vol. 41, 144503-4510 [0127]
- **WANG et al.**J. Am. Chem. Soc., 2000, vol. 122, 8595-8602 [0128]
- **SINGH et al.**Chem. Commun., 1998, vol. 4, 455-456 [0129]
- **WAHLESTEDT et al.**Proc. Natl. Acad. Sci. U.S.A., 2000, vol. 97, 5633-5638 [0129]
- **KOSHKIN et al.**Tetrahedron, 1998, vol. 54, 3607-3630 [0130]
- **MARTIN et al.**Helv. Chim. Acta, 1995, vol. 78, 486-504 [0131]
- The Concise Encyclopedia Of Polymer Science And EngineeringJohn Wiley & Sons19900000858-859 [0134]
- **ENGLISCH et al.**Angewandte Chemie19910000vol. 30, 613- [0134]
- **SANGHVI, Y. S.**Antisense Research and ApplicationsCRC Press19930000289-302 [0134]
- Antisense Research and ApplicationsCRC Press19930000276-278 [0134]
- **LETSINGER et al.**Proc. Natl. Acad. Sci. USA, 1989, vol. 86, 6553-6556 [0136]

- **MANOHARAN et al.**Bioorg. Med. Chem. Let., 1994, vol. 4, 1053-1060 [0136]
- **MANOHARAN et al.**Ann. N.Y. Acad. Sci., 1992, vol. 660, 306-309 [0136]
- **MANOHARAN et al.**Bioorg. Med. Chem. Let., 1993, vol. 3, 2765-2770 [0136]
- **OBERHAUSER et al.**Nucl. Acids Res., 1992, vol. 20, 533-538 [0136]
- **SAISON-BEHMOARAS et al.**EMBO J., 1991, vol. 10, 1111-1118 [0136]
- **KABANOV et al.**FEBS Lett., 1990, vol. 259, 327-330 [0136]
- **SVINARCHUK et al.**Biochimie, 1993, vol. 75, 49-54 [0136]
- **MANOHARAN et al.**Tetrahedron Lett., 1995, vol. 36, 3651-3654 [0136] [0136]
- **SHEA et al.**Nucl. Acids Res., 1990, vol. 18, 3777-3783 [0136]
- **MANOHARAN et al.**Nucleosides & Nucleotides, 1995, vol. 14, 969-973 [0136]
- **MISHRA et al.**Biochim. Biophys. Acta, 1995, vol. 1264, 229-237 [0136]
- **CROOKE et al.**J. Pharmacol. Exp. Ther., 1996, vol. 277, 923-937 [0136]
- **ZENDER et al.**Cancer Gene Ther., 2002, vol. 9, 6489-96 [0137]
- **NOGUCHI et al.**Diabetes, 2003, vol. 52, 71732-1737 [0137]
- **TREHIN et al.**Pharm. Research, 2004, vol. 21, 1248-1256 [0137]
- **WENDER et al.**Proc. Natl. Acad. Sci. USA, 2000, vol. 97, 13003-13008 [0137]
- **AGUILERA et al.**Integr Biol (Camb), 2009, vol. 1, 5-6371-381 [0137]
- **LI et al.**Invest Opthalmol Vis Sci, 1994, vol. 35, 2543-2549 [0144] [0160]
- **; BORRAS et al.**Gene Ther, 1999, vol. 6, 515524- [0144]
- **LIDAVIDSON**PNAS, 1995, vol. 92, 7700-7704 [0144]
- **SAKAMOTO et al.**H Gene Ther, 1999, vol. 5, 1088-1097 [0144]
- **ALI et al.**Hum Gene Ther, 1998, vol. 9, 81-86 [0144] [0160]
- **FLANNERY et al.**PNAS, 1997, vol. 94, 6916-6921 [0144]
- **BENNETT et al.**Invest Opthalmol Vis Sci, 1997, vol. 38, 2857-2863 [0144] [0160]
- **JOMARY et al.**Gene Ther, 1997, vol. 4, 683-690 [0144] [0160]
- **ROLLING et al.**Hum Gene Ther, 1999, vol. 10, 641-648 [0144]
- **ALI et al.**Hum Mol Genet, 1996, vol. 5, 591-594 [0144]
- **SAMULSKI et al.**J. Vir., 1989, vol. 63, 3822-3828 [0144] [0160]
- **MENDELSON et al.**Virol., 1988, vol. 166, 154-165 [0144]
- **FLOTTE et al.**PNAS, 1993, vol. 90, 10613-10617 [0144] [0160]
- **MIYOSHI et al.**PNAS, 1997, vol. 94, 10319-23 [0144] [0160]
- **TAKAHASHI et al.**J Virol, 1999, vol. 73, 7812-7816 [0144]
- **BITTER et al.**Methods in Enzymology, 1987, vol. 153, 516-544 [0146] [0163]
- **PANYAM**Adv Drug Deliv Rev, 2012, 0169-409 [0150]
- **BORRAS et al.**Gene Ther, 1999, vol. 6, 515-524 [0160]
- **LIDAVIDSON**PNAS, 1995, vol. 92, 7700-7704 [0160]
- **SAKAMOTO et al.**H Gene Ther, 1999, vol. 5, 1088-1097 [0160]
- **FLANNERY et al.**PNAS, 1997, vol. 94, [0160]
- **ROLLING et al.**Hum Gene Ther, 1999, vol. 10, 641-648 [0160]
- **ALI et al.**Hum Mol Genet, 1996, vol. 5, 591-594 [0160]
- **MENDELSON et al.**Virol, 1988, vol. 166, 154-165 [0160]
- **TAKAHASHI et al.**J Virol, 1999, vol. 73, 7812-7816 [0160]
- **YANIK**PLoS ONE, 2010, vol. 5, 7e11756- [0164]
- **BEUMER et al.**Efficient gene targeting in Drosophila by direct embryo injection with zinc-

finger nucleasesPNAS, 2008, vol. 105, 5019821-19826 [0164]

- **CHANG et al.**Proc. Natl. Acad Sci USA, 1987, vol. 84, 4959-4963 [0177]
- **NEHLS et al.**Science, 1996, vol. 272, 886-889 [0177]
- Remington's Pharmaceutical SciencesMace Publishing Company19850000 [0186]
- **LANGER**Science, 1990, vol. 249, 1527-1533 [0186]
- **B. WIEDENHEFTS. H. STERNBERGJ. A. DOUDNA**Nature, 2012, vol. 482, 331- [0299]
- **D. BHAYAM. DAVISONR. BARRANGOU**Annu. Rev. Genet., 2011, vol. 45, 273- [0299]
- **M. P. TERNSR. M. TERNS**Curr. Opin. Microbiol., 2011, vol. 14, 321- [0299]
- **E. DELTCHEVA et al.**Nature, 2011, vol. 471, 602- [0299]
- **J. CARTER. WANGH. LIR. M. TERNSM. P. TERNS**Genes Dev, 2008, vol. 22, 3489- [0299]
- **R. E. HAURWITZM. JINEKB. WIEDENHEFTK. ZHOUJ. A. DOUDNA**Science, 2010, vol. 329, 1355- [0299]
- **R. WANGG. PREAMPLUMEM. P. TERNSR. M. TERNSH. LI**Structure, 2011, vol. 19, 257- [0299]
- **E. M. GESNERM. J. SCHELLENBERGE. L. GARSIDEM. M. GEORGEA. M. MACMILLAN**Nat. Struct. Mol. Biol., 2011, vol. 18, 688- [0299]
- **A. HATOUM-ASLANI. MANIVL. A. MARRAFFINI**Proc. Natl. Acad. Sci. U.S.A., 2011, vol. 108, 21218- [0299]
- **S. J. J. BROUNS et al.**Science, 2008, vol. 321, 960- [0299]
- **D. G. SASHITALM. JINEKJ. A. DOUDNA**Nat. Struct. Mol. Biol., 2011, vol. 18, 680- [0299]
- **N. G. LINTNER et al.**J. Biol. Chem., 2011, vol. 286, 21643- [0299]
- **E. SEMENOVA et al.**Proc. Natl. Acad. Sci. U.S.A., 2011, vol. 108, 10098- [0299]
- **B. WIEDENHEFT et al.**Proc. Natl. Acad. Sci. U.S.A., 2011, vol. 108, 10092- [0299]
- **B. WIEDENHEFT et al.**Nature, 2011, vol. 477, 486- [0299]
- **C. R. HALE et al.**Cell, 2009, vol. 139, 945- [0299]
- **J. A. L. HOWARDS. DELMASI. IVANE. L. BOLT**Biochem. J., 2011, vol. 439, 85- [0299]
- **E. R. WESTRA et al.**Mol. Cell, 2012, vol. 46, 595- [0299]
- **C. R. HALE et al.**Mol. Cell, 2012, vol. 45, 292- [0299]
- **J. ZHANG et al.**Mol. Cell, 2012, vol. 45, 303- [0299]
- **K. S. MAKAROVA et al.**Nat. Rev. Microbiol., 2011, vol. 9, 467- [0299]
- **K. S. MAKAROVAN. V. GRISHINS. A. SHABALINAY. I. WOLFE. V. KOONIN**Biol. Direct, 2006, vol. 1, 7- [0299]
- **K. S. MAKAROVAL. ARAVINDY. I. WOLFE. V. KOONIN**Biol. Direct, 2011, vol. 6, 38- [0299]
- **S. GOTTESMAN**Nature, 2011, vol. 471, 588- [0299]
- **R. BARRANGOU et al.**Science, 2007, vol. 315, 1709- [0299]
- **J. E. GARNEAU et al.**Nature, 2010, vol. 468, 67- [0299]
- **R. SAPRANAUSKAS et al.**Nucleic Acids Res., 2011, vol. 39, 9275- [0299]
- **G. K. TAYLORD. F. HEITERS. PIETROKOVSKIB. L. STODDARD**Nucleic Acids Res., 2011, vol. 39, 9705- [0299]
- **H. DEVEAU et al.**J. Bacteriol., 2008, vol. 190, 1390- [0299]
- **B. P. LEWISC. B. BURGED. P. BARTEL**Cell, 2005, vol. 120, 15- [0299]

- **G. HUTVAGNERM. J. SIMARD**Nat. Rev. Mol. Cell Biol, 2008, vol. 9, 22- [0299]
- **F. J. M. MOJICAC. DÍEZ-VILLASEÑORJ. GARCIA-MARTINEZC. ALMENDROS**Microbiology, 2009, vol. 155, 733- [0299]
- **L. A. MARRAFFINIE. J. SONTHEIMER**Nature, 2010, vol. 463, 568 [0299]
- **D. G. SASHITALB. WIEDENHEFTJ. A. DOUDNA**Mol. Cell, 2012, vol. 46, 606- [0299]
- **M. CHRISTIAN et al.**Genetics, 2010, vol. 186, 757- [0299]
- **J. C. MILLER et al.**Nat. Biotechnol., 2011, vol. 29, 143- [0299]
- **F. D. UMOVE. J. REBARM. C. HOLMESH. S. ZHANGP. D. GREGORY**Nat. Rev. Genet., 2010, vol. 11, 636- [0299]
- **D. CARROLL**Gene Ther., 2008, vol. 15, 1463- [0299]
- **J. SAMBROOKE. F. FRITSCHT. MANIATIS**Molecular Cloning: A Laboratory ManualCold Spring Harbor Laboratory Press, Cold Spring Harbor19890000 [0299]
- **M. G. CAPARONJ. R. SCOTT**Genetic manipulation of pathogenic streptococci.Methods Enzymol., 1991, vol. 204, 556- [0299]
- **C. FRØKJÆR-JENSEN et al.**Single-copy insertion of transgenes in Caenorhabditis elegansNat. Genet., 2008, vol. 40, 1375- [0299]
- **R. B. DENMAN**Using RNAFOLD to predict the activity of small catalytic RNAsBiotechniques, 1993, vol. 15, 1090- [0299]
- **I. L. HOFACKERP. F. STADLER**Memory efficient folding algorithms for circular RNA secondary structuresBioinformatics, 2006, vol. 22, 1172- [0299]
- **K. DARTYA. DENISEY. PONTY**VARNA: Interactive drawing and editing of the RNA secondary structureBioinformatics, 2009, vol. 25, 1974- [0299]
- **GRISSA I et al.**BMC Bioinformatics, 2007, vol. 8, 172- [0322]
- **GRISSA I et al.**Nucleic Acids Res, 2007, [0322]

PATENTKRAV

1. DNA-målrettet enkeltmolekyle-RNA, der binder sig til et site-rettet modificerende polypeptid og målretter det site-rettede modificerende polypeptid mod en specifik placering inde i et mål-DNA, hvor det DNA-målrettede enkeltmolekyle-RNA omfatter:

(a) et DNA-målrettet segment omfattende en nukleotidsekvens, der er komplementær til en sekvens i mål-DNA'et, og

(b) et proteinbindende segment, der interagerer med det site-rettede modificerende polypeptid, som er en naturligt forekommende Cas9-endonuklease, hvor det proteinbindende segment omfatter to komplementære nukleotidstykker, der hybridiserer til dannelse af et dobbeltstrenget RNA (dsRNA)-dupleks, og

hvor det DNA-målrettede enkeltmolekyle-RNA, sammen med det site-rettede modificerende polypeptid, som er en naturligt forekommende Cas9-endonuklease, tilvejebringer site-specifik spaltning af mål-DNA'et, så der dannes et dobbeltstrenget brud.

2. DNA-målrettet enkeltmolekyle-RNA ifølge krav 1, hvor nukleotidsekvensen af det DNA-målrettede segment, som er komplementært til en sekvens i mål-DNA'et, er større end 15 nukleotider.

3. DNA-målrettet enkeltmolekyle-RNA, der binder sig til et site-rettet modificerende polypeptid og målretter det site-rettede modificerende polypeptid mod en specifik placering inde i et mål-DNA, hvor det DNA-målrettede enkeltmolekyle-RNA omfatter:

(a) et DNA-målrettet segment omfattende en nukleotidsekvens, der er komplementær til en målsekvens i et mål-DNA, og

(b) et proteinbindende segment, der interagerer med det site-rettede modificerende polypeptid, som er en naturligt forekommende Cas9-endonuklease, hvor det proteinbindende segment omfatter to komplementære nukleotidstykker, der hybridiserer til dannelse af et dobbeltstrenget RNA (dsRNA)-dupleks,

hvor nukleotidsekvensen, der er komplementær til en sekvens i mål-DNA'et, er større end 15 nukleotider.

4. DNA-målrettet enkeltmolekyle-RNA ifølge krav 3,

hvor det DNA-målrettede enkeltmolekyle-RNA, sammen med et site-rettet modificerende polypeptid, som er en naturligt forekommende Cas9-endonuklease, tilvejebringer site-specifik spaltning af mål-DNA'et, så der dannes et dobbeltstrenget brud.

5. DNA-målrettet enkeltmolekyle-RNA ifølge et hvilket som helst af kravene 1 og 4, hvor det er det DNA-målrettede enkeltmolekyle-RNA, der tilvejebringer specificitet til et kompleks dannet af det DNA-målrettede RNA og den naturligt forekommende Cas9-endonuklease for mål-DNA'et ved hjælp af den sekvens af det DNA-målrettede segment, der er komplementær til sekvensen i mål-DNA'et, og hvor det er den naturligt forekommende Cas9-endonuklease af komplekset, der tilvejebringer nukleaseaktivitet til komplekset, hvilken nukleaseaktivitet spalter mål-DNA'et ved en mål-DNA-sekvens, der er defineret ved komplementaritetsregionen mellem det DNA-målrettede RNA og mål-DNA'et.

6. DNA-målrettet enkeltmolekyle-RNA ifølge et hvilket som helst af kravene 1-5, hvor de to komplementære nukleotidstykker er kovalent bundet ved hjælp af mellemliggende nukleotider.

7. DNA-målrettet enkeltmolekyle-RNA ifølge et hvilket som helst af kravene 1-6, hvor det første af de to komplementære stykker af det proteinbindende segment omfatter et crRNA-gentagelsesstykke, og hvor det andet af de to komplementære stykker af det proteinbindende segment omfatter et tracrRNA-stykke.

8. DNA-målrettet enkeltmolekyle-RNA ifølge krav 7, hvor crRNA-gentagelsesstykket og tracrRNA-stykket er fra *S. pyogenes.*

9. DNA-målrettet enkeltmolekyle-RNA ifølge et hvilket som helst af kravene 1-8, hvor den naturligt forekommende Cas9-endonuklease er fra *S. pyogenes.*

10. DNA-målrettet enkeltmolekyle-RNA ifølge et hvilket som helst af kravene 1-9, hvor det DNA-målrettede enkeltmolekyle-RNA omfatter nukleinsyremodifikationer

udvalgt fra gruppen bestående af modificerede backbones og modificerede internukleosidbindinger, nukleinsyremimetika, modificerede sukkermolekyldele, basemodifikationer og -substitutioner og konjugater.

11. DNA-polynukleotid omfattende en nukleotidsekvens, der koder for det DNA-målrettede RNA ifølge et hvilket som helst af kravene 1 til 9.

12. Rekombinant ekspressionsvektor omfattende DNA-polynukleotidet ifølge krav 11.

13. Rekombinant ekspressionsvektor ifølge krav 12, hvor nukleotidsekvensen, der koder for det DNA-målrettede RNA, er funktionsmæssigt bundet til en promoter.

14. Rekombinant ekspressionsvektor ifølge krav 13, hvor promoteren er en inducerbar promoter.

15. DNA-målrettet RNA, der binder sig til et a site-rettet modificerende polypeptid og målretter det site-rettede modificerende polypeptid mod en specifik placering inde i et mål-DNA, hvor det DNA-målrettede RNA omfatter:

(a) et DNA-målrettet segment omfattende en nukleotidsekvens, der er komplementær til en målsekvens i et mål-DNA, og

(b) et proteinbindende segment, der interagerer med det site-rettede modificerende polypeptid, som er en naturligt forekommende Cas9-endonuklease, hvor det proteinbindende segment omfatter to komplementære nukleotidstykker, der hybridiserer til dannelse af et dobbeltstrenget RNA (dsRNA)-dupleks,

hvor det DNA-målrettede RNA omfatter nukleinsyremodifikationer udvalgt fra gruppen bestående af modificerede backbones og modificerede internukleosidbindinger, nukleinsyremimetika, modificerede sukkermolekyldele, basemodifikationer og -substitutioner og konjugater.

16. DNA-målrettet RNA ifølge krav 15, hvor det DNA-målrettede RNA er et DNA-målrettet dobbeltmolekyle-RNA.

17. DNA-målrettet RNA ifølge krav 15, hvor det DNA-målrettede RNA er et DNA-målrettet enkeltmolekyle-RNA.

18. DNA-målrettet enkeltmolekyle-RNA ifølge krav 17, hvor de to komplementære nukleotidstykker er kovalent bundet ved hjælp af mellemliggende nukleotider.

19. DNA-målrettet RNA ifølge et hvilket som helst af kravene 15-18,

hvor det DNA-målrettede RNA sammen med et site-rettet modificerende polypeptid, som er en naturligt forekommende Cas9-endonuklease, tilvejebringer site-specifik spaltning af mål-DNA'et, så der dannes et dobbeltstrenget brud.

20. DNA-målrettet RNA ifølge krav 18,

hvor det er det DNA-målrettede RNA, der tilvejebringer specificitet til et kompleks dannet af det DNA-målrettede RNA og den naturligt forekommende Cas9-endonuklease for mål-DNA'et ved hjælp af den sekvens af det DNA-målrettede segment, der er komplementær til sekvensen i mål-DNA'et, og

hvor det er den naturligt forekommende Cas9-endonuklease af komplekset, der tilvejebringer nukleaseaktivitet til komplekset, hvilken nukleaseaktivitet spalter mål-DNA'et ved en mål-DNA-sekvens, der er defineret ved komplementaritetsregionen mellem det DNA-målrettede RNA og mål-DNA'et.

21. DNA-målrettet RNA ifølge et hvilket som helst af kravene 15-20, hvor det første af de to komplementære stykker af det proteinbindende segment omfatter et crRNA-gentagelsesstykke, og hvor det andet af de to komplementære stykker af det proteinbindende segment omfatter et tracrRNA-stykke.

22. DNA-målrettet RNA ifølge krav 21, hvor crRNA-gentagelsesstykket og tracrRNA-stykket er fra *S. pyogenes.*

23. DNA-målrettet RNA ifølge et hvilket som helst af kravene 15-22, hvor den naturligt forekommende Cas9-endonuklease er fra *S. pyogenes.*

24. Kompleks, der er dannet af

(i) et DNA-målrettet enkeltmolekyle-RNA ifølge et hvilket som helst af kravene 1-10 eller et DNA-målrettet RNA ifølge et hvilket som helst af kravene 15-23 og

(ii) et site-rettet modificerende polypeptid, som er en naturligt forekommende Cas9-endonuklease.

25. Sammensætning, der omfatter:

(i) et DNA-målrettet enkeltmolekyle-RNA ifølge et hvilket som helst af kravene 1-10 eller et DNA-målrettet RNA ifølge et hvilket som helst af kravene 15-23 og

(ii) et site-rettet modificerende polypeptid, som er en naturligt forekommende Cas9-endonuklease.

DRAWINGS

Drawing

FIGURE 1

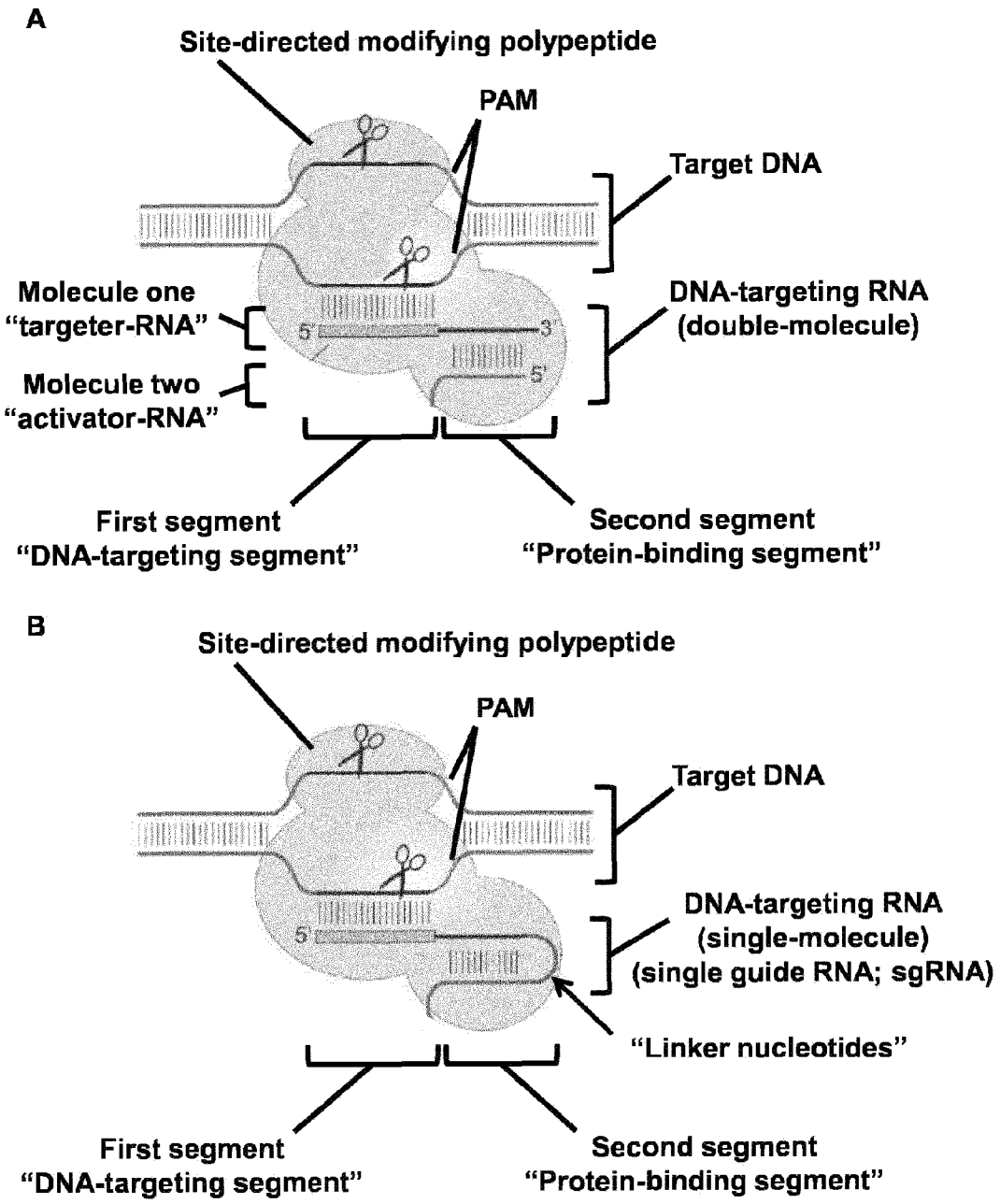


FIGURE 2

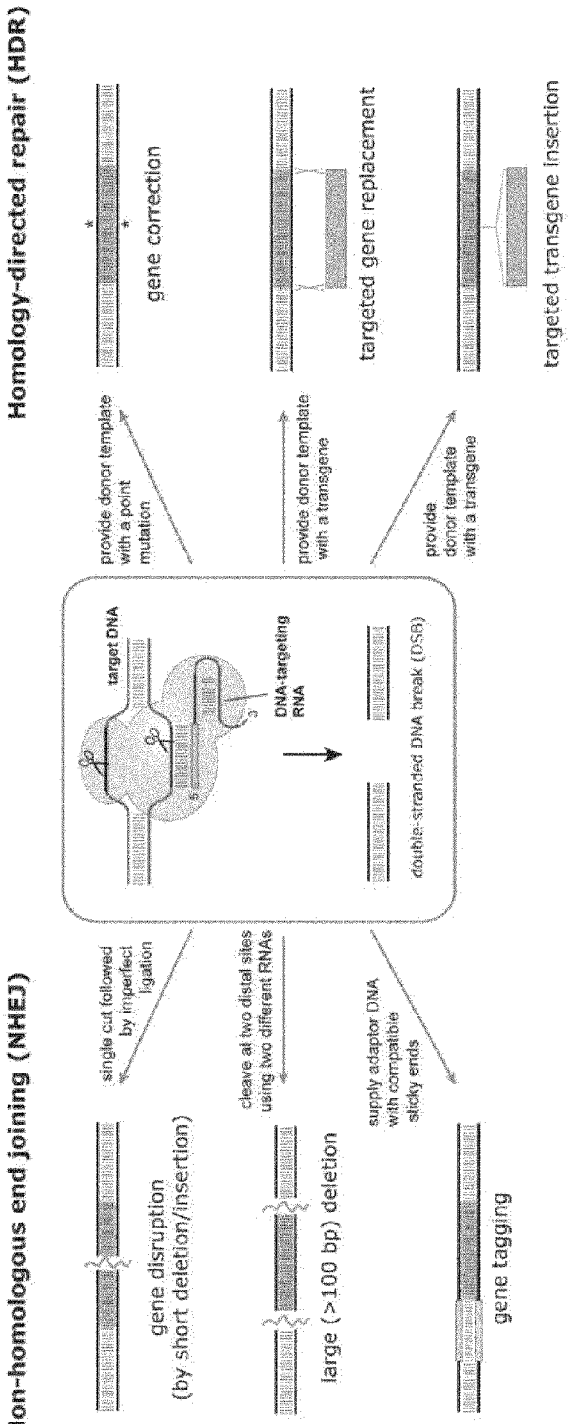

Homology-directed repair (HDR)
Non-homologous end joining (NHEJ)
target DNA
DNA-targeting RNA
double-stranded DNA break (DSB)
single cut followed by imperfect ligation
cleave at two distal sites using two different RNAs
supply adaptor DNA with compatible sticky ends
provide donor template with a point mutation
provide donor template with a transgene
provide donor template with a transgene
gene disruption (by short deletion/insertion)
large (>100 bp) deletion
gene tagging
gene correction
targeted gene replacement
targeted transgene insertion

FIGURE 3A

Cas9/Csn1 Streptococcus pyogenes

motifs

1 MDKKYSIGLDIGTNSVGWAVITDDYKVPSKKLKGLGNTDRHGIKKNLIGALL
FDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEE
SFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLADSTDKVDLRLIYL
ALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASRVDA
KAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAED
AKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDATLLSDILRVNSEITK
APLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGG
ASQEEFYKFIKPILEKMDGTEELLAKLNREDLLRKQRTFDNGSIPYQIHLGEL
HAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEE
TITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYN
ELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIEC
FDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFED
REMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTI
LDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGS
2 PAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERM
KRIEEGIKELGSQILKEYPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRL
3 SDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYW
RQLLNAKLITQRKFDNLTKAERGGLSELDKVGFIKRQLVETRQITKHVAQILD
4 SRMNTKYDENDKLIREVRVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAY
LNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYS
NIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQV
NIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAYSVL
VVAKVEKGKSKKLKSVKELLGITIMERSSFEKDPIDFLEAKGYKEVRKDLIIKL
PKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSP
EDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPI
REQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLIHQSITGLY
ETRIDLSQLGGD

FIGURE 3B

Cas9/Csn1 Streptococcus pyogenes

Domains

1 MDKKYS**IGLDIGTNSVGWAVITDDYKVPSKKLKGLGNTDRHGIKKNLIGAL**
LFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLE
ESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLADSTDKVDLRLI
YLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASRV
DAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLA
EDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDATLLSDILRVNSEI
TKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYID
GGASQEEFYKFIKPILEKMDGTEELLAKLNREDLLRKQRTFDNGSIPYQIHL
GELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRK
SEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTV
YNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKI
ECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLF
EDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSG
KTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLA
2 GS**PAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSR**
ERMKRIEEGIKELGSDILKEYPVENTQLQNEKLYLYYLQNGRDMYVDQEL
DINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKK
MKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKVGFIKRQLVETRQIT
KHVAQILDSRMNTKYDENDKLIREVRVITLKSKLVSDFRKDFQFYKVREIN
NYHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIG
KATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATV
RKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGF
DSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKDPIDFLEAKGY
KEVRKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLA
SHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVL
SAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLD
ATLIHQSITGLYETRIDLSQLGGD

FIGURE 4

A. Sequence identities relative to: *S. pyogenes* Cas9/Csn1

		Sequence identities - MUSCLE alignment		
		Full-length	*Domain 1*	*Domain 2*
species	*reference sequence*	% indentity	% indentity	% identity
Streptococcus pyogenes M1 GAS	NP_269215	100.0	100.0	100.0
Streptococcus pyogenes MGAS5005	YP_282132.1	99.9	99.4	100.0
Listeria innocua Clip11262	NP_472073	54.3	60.0	64.9
Pasteurella multocida subsp. multocida str. Pm70	NP_246064.1	19.7	29.0	25.9
Streptococcus thermophilus LMD-9 Csn1-A	YP_820832	59.2	75.6	72.4
Streptococcus thermophilus LMD-9 Csn1-B	YP_820161.1	20.6	27.3	26.8
Neisseria meningitidis Z2491	YP_002342100.1	20.2	33.6	28.1
Streptococcus mutans UA159	NP_721764	64.9	78.1	74.1
Streptococcus gordonii str. Challis substr. CH1	YP_001450662.1	19.8	28.2	27.0
Campylobacter jejuni subsp. jejuni NCTC 11168	YP_002344900.1	19.6	30.3	26.3
Treponema denticola ATCC 35405	NP_970941	32.5	47.3	38.8

B. Sequence identities relative to: *N. meningitidis* Cas9/Csn1

		Sequence identities - MUSCLE alignment		
		Full-length	*Domain 1*	*Domain 2*
species	reference sequence	% indentity	% indentity	% identity
Streptococcus pyogenes M1 GAS	NP_269215	20.2	33.6	28.1
Streptococcus pyogenes MGAS5005	YP_282132.1	20.3	34.5	28.1
Listeria innocua Clip11262	NP_472073	18.8	33.6	25.9
Pasteurella multocida subsp. multocida str. Pm70	NP_246064.1	64.3	72.1	69.0
Streptococcus thermophilus LMD-9 Csn1-A	YP_820832	19.6	35.3	25.9
Streptococcus thermophilus LMD-9 Csn1-B	YP_820161.1	25.8	35.7	35.1
Neisseria meningitidis Z2491	YP_002342100.1	100.0	100.0	100.0
Streptococcus mutans UA159	NP_721764	19.2	36.1	25.5
Streptococcus gordonii str. Challis substr. CH1	YP_001450662.1	25.3	37.5	35.8
Campylobacter jejuni subsp. jejuni NCTC 11168	YP_002344900.1	34.7	45.0	41.0
Treponema denticola ATCC 35405	NP_970941	18.8	31.5	25.5

FIGURE 5

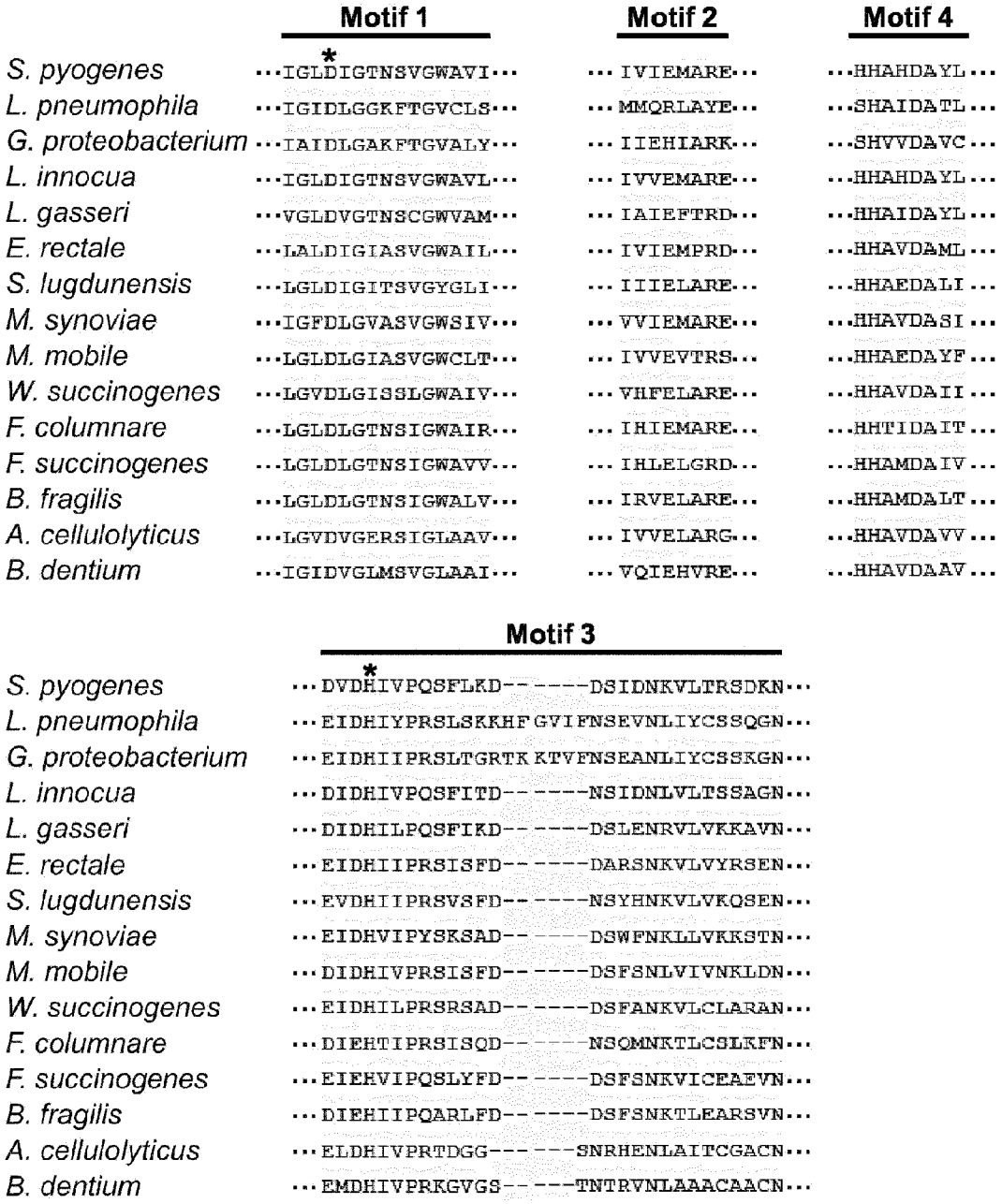
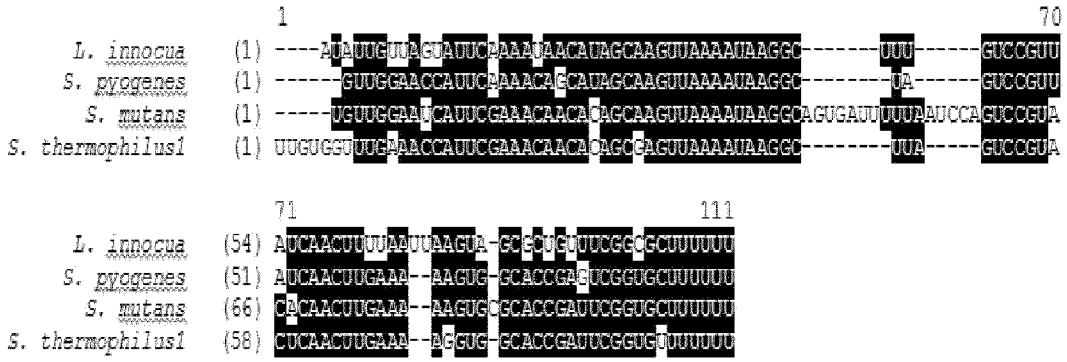

Motif 1
Motif 2
Motif 4
S. pyogenes ···IGLDIGTNSVGWAVI··· ···IVIEMARE··· ···HHAHDAYL···
L. pneumophila ···IGIDLGGKFTGVCLS··· ···MMQRLAYE··· ···SHAIDATL···
G. proteobacterium ···IAIDLGAKFTGVALY··· ···IIEHIARK··· ···SHVVDAVC···
L. innocua ···IGLDIGTNSVGWAVL··· ···IVVEMARE··· ···HHAHDAYL···
L. gasseri ···VGLDVGTNSCGWVAM··· ···IAIEFTRD··· ···HHAIDAYL···
E. rectale ···LALDIGIASVGWAIL··· ···IVIEMPRD··· ···HHAVDAML···
S. lugdunensis ···LGLDIGITSVGYGLI··· ···IIIELARE··· ···HHAEDALI···
M. synoviae ···IGFDLGVASVGWSIV··· ···VVIEMARE··· ···HHAVDASI···
M. mobile ···LGLDLGIASVGWCLT··· ···IVVEVTRS··· ···HHAEDAYF···
W. succinogenes ···LGVDLGISSLGWAIV··· ···VHFELARE··· ···HHAVDAII···
F. columnare ···LGLDLGTNSIGWAIR··· ···IHIEMARE··· ···HHTIDAIT···
F. succinogenes ···LGLDLGTNSIGWAVV··· ···IHLELGRD··· ···HHAMDAIV···
B. fragilis ···LGLDLGTNSIGWALV··· ···IRVELARE··· ···HHAMDALT···
A. cellulolyticus ···LGVDVGERSIGLAAV··· ···IVVELARG··· ···HHAVDAVV···
B. dentium ···IGIDVGLMSVGLAAI··· ···VQIEHVRE··· ···HHAVDAAV···
Motif 3
S. pyogenes ···DVDHIVPQSFLKD----DSIDNKVLTRSDKN···
L. pneumophila ···EIDHIYPRSLSKKHFGVIFNSEVNLIYCSSQGN···
G. proteobacterium ···EIDHIIPRSLTGRTKKTVFNSEANLIYCSSKGN···
L. innocua ···DIDHIVPQSFITD-----NSIDNLVLTSSAGN···
L. gasseri ···DIDHILPQSFIKD----DSLENRVLVKKAVN···
E. rectale ···EIDHIIPRSISFD----DARSNKVLVYRSEN···
S. lugdunensis ···EVDHIIPRSVSFD----NSYHNKVLVKQSEN···
M. synoviae ···EIDHVIPYSKSAD----DSWFNKLLVKKSTN···
M. mobile ···DIDHIVPRSISFD----DSFSNLVIVNKLDN···
W. succinogenes ···EIDHILPRSRSAD----DSFANKVLCLARAN···
F. columnare ···DIEHTIPRSISQD----NSQMNKTLCSLKFN···
F. succinogenes ···EIEHVIPQSLYFD----DSFSNKVICEAEVN···
B. fragilis ···DIEHIIPQARLFD----DSFSNKTLEARSVN···
A. cellulolyticus ···ELDHIVPRTDGG---SNRHENLAITCGACN···
B. dentium ···EMDHIVPRKGVGS---TNTRVNLAAACAACN···

FIGURE 6

A


1
70
L. innocua (1)
S. pyogenes (1)
S. mutans (1)
S. thermophilus1 (1)
71
111
L. innocua (54)
S. pyogenes (51)
S. mutans (66)
S. thermophilus1 (58)

B

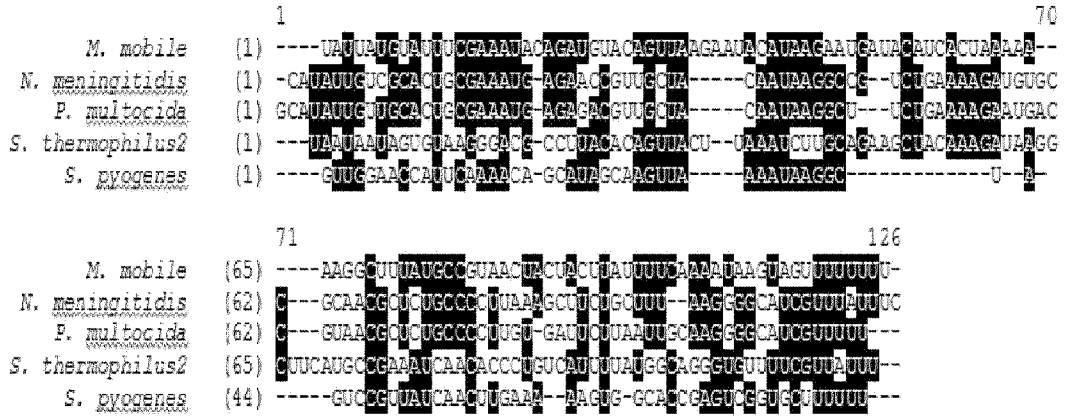
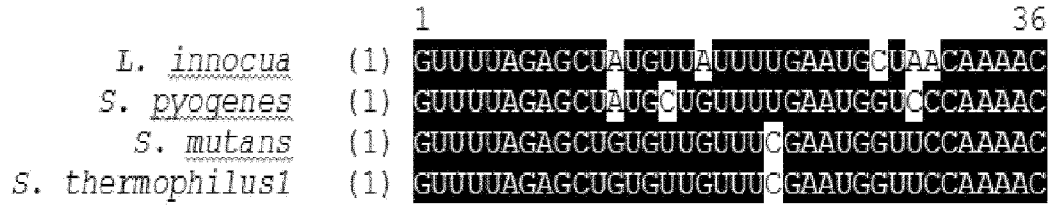

1
70
M. mobile (1)
N. meningitidis (1)
P. multocida (1)
S. thermophilus2 (1)
S. pyogenes (1)
71
126
M. mobile (65)
N. meningitidis (62)
P. multocida (62)
S. thermophilus2 (65)
S. pyogenes (44)

FIGURE 7

A


1
36
L. innocua (1) GUUUUAGAGCUAUGUUAUUUUGAAUGCUAACAAAAC
S. pyogenes (1) GUUUUAGAGCUAUGCUGUUUUGAAUGGUCCCAAAAC
S. mutans (1) GUUUUAGAGCUGUGUUGUUUCGAAUGGUUCCAAAAC
S. thermophilus1 (1) GUUUUAGAGCUGUGUUGUUUCGAAUGGUUCCAAAAC

B

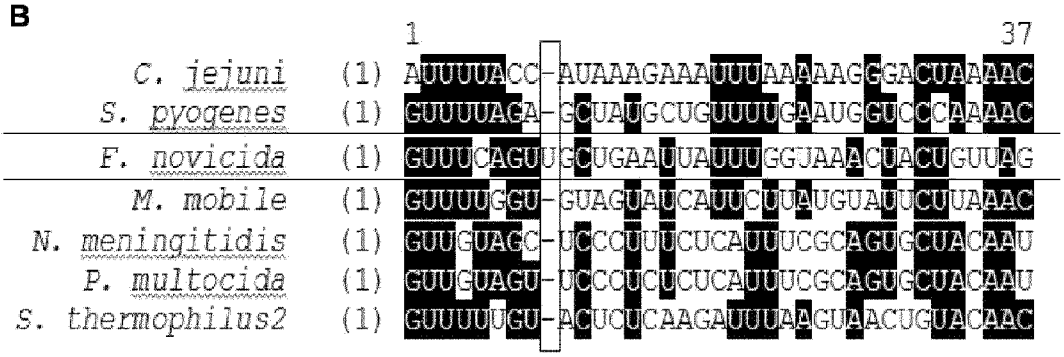
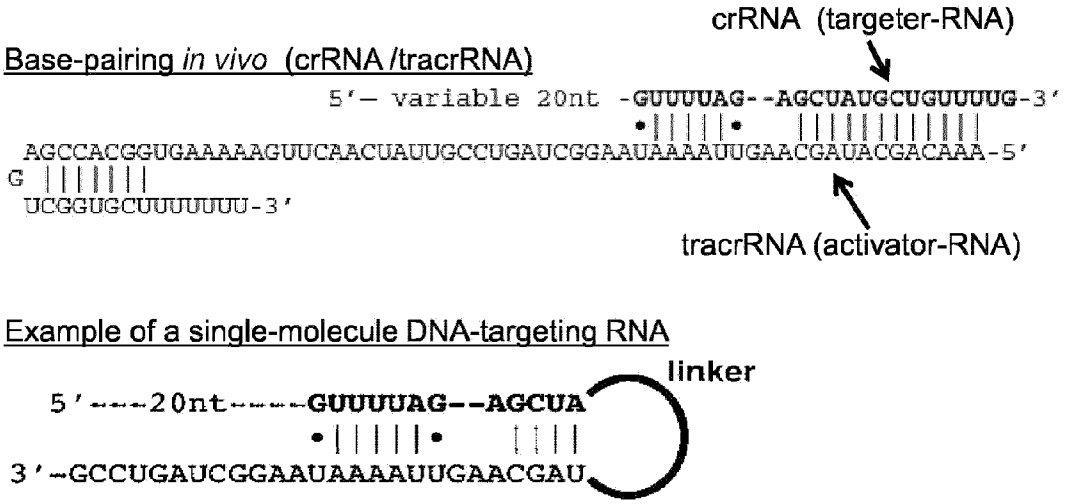

1
37
C. jejuni (1) AUUUUACC-AUAAAGAAAUUUAAAAAGGGACUAAAAC
S. pyogenes (1) GUUUUAGA-GCUAUGCUGUUUUGAAUGGUCCCAAAAC
F. novicida (1) GUUUCAGUUGCUGAAUUAUUUGGUAAACUACUGUUAG
M. mobile (1) GUUUUGGU-GUAGUAUCAUUCUUAUGUAUUCUUAAAC
N. meningitidis (1) GUUGUAGC-UCCCUUUCUCAUUUCGCAGUGCUACAAU
P. multocida (1) GUUGUAGU-UCCCUCUCUCAUUUCGCAGUGCUACAAU
S. thermophilus2 (1) GUUUUUGU-ACUCUCAAGAUUUAAGUAACUGUACAAC

FIGURE 8

STRAIN	NUMBER OF CRISPRs	CAS54 CRISPR identifier[a]	CRISPR REPEAT:tracrRNA BASEPAIRING[b]
Streptococcus pyogenes SF370	2	NC_002737_1	5′ GUUUUAG--AGCUAUGCUGUUUUGAAUGGUCCCAAAAC 3′ 3′ AAAUUGAACGAUACGACAAAACUUACCAAGGUUGUU 5′
Streptococcus mutans UA159	1	NC_004350_1	GUUUUAG--AGCUGUGUUGUUUCGAAUGGUUCCAAAAC AAAUUGAACGACACAACAAAGCUUACUAAGGUUGUG
Streptococcus thermophilus LMD-9	3	NC_008532_5	GUUUUAG--AGCUGUGUUGUUUCGAAUGGUUCCAAAAC AAAUUGAGCGACACAACAAAGCUUACCAAAGUUUGG
		NC_008532_2	GUUUUUGUACUCU-CAAGAUUUAAGUAACUGUACAAC AAACAUCGAAGACGUUCUAAAUUCAUUGACACAUUC
Listeria innocua Clip11262	1	NC_003212_2	GUUUUAG--AGCUAUGUUAUUUUGAAUGCUAACAAAAC AAAUUGAACGAUACAAUAAAACUUAUGAUUGUUAUA
Treponema denticola ATCC35405	1	NC_002967_1	GUUUGAG--AGUUGUGUAAUUUAAGAUGGAUCUCAAAC AACUUGAGCAACACAUUAAAUUCUACCUAGAAUUUA
Neisseria meningitidis Z2491	2	NC_003116_10	GUUGUAGCUCCCUUUCUCAUUUCGCAGUGCUACAAU UAACAUCGUUGCCAAGAGUAAAGCGUCACGCUGUUA
Streptococcus gordonii str. Challis substr. CH1	1	NC_009785_2	GUUUUGUACUCU-CAAGAUUUAAGUAACUGUACAAC AAACAUCGAAGACGUUCUAAAUUCAUUGACACAUUC
Bifidobacterium bifidum S17	1	NC_014616_1	GUUUCA-AUGCCUGUCAGAUCAAUGACUUUGACCAC AGUUAAAUACGGACAGUCCAGUUACUGGAACUAGUA
Lactobacillus salivarius UCC118	1	NC_007929_1	GUUUCAGAAGUAUGUUAAAUCAAUAAGGUUAAGACC AAGUUGAGUCUUACAAUUUAGUUACUCCAGUUUUGG
Francisella tularensis subsp. *novicida* U112	2	NC_008601_1[c]	CUAACAGUAGUUUACCAAAUAAUUCAGCAACUGAAAC UUGUGUUCAUGUAUGGUUUAUUAGAUUGUUGACUUUG CUAACAGUAGUUUACCAAAUAAUUCA-GCAACUGAAAC UUUAAUAUUUACAUGGUUUAUUAAUUACG-AGACAUUA
Legionella pneumophila str. Paris	1	NC_006368_1	CCAAUAAUCCCUCAUCUAAAAAUCCA-ACCACUGAAAC AUUUAAUCUUUAGUAGAUUUAAAGCUAUGG-GACUUUA

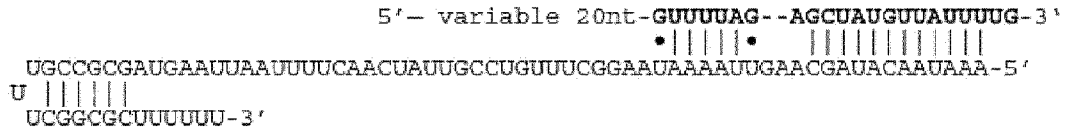
FIGURE 9

Streptococcus pyogenes



Listeria innocua

Base-pairing *in vivo* (crRNA/tracrRNA)



Example of a single-molecule DNA-targeting RNA

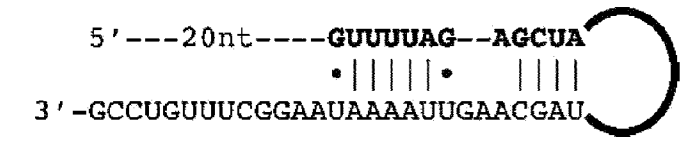


FIGURE 10

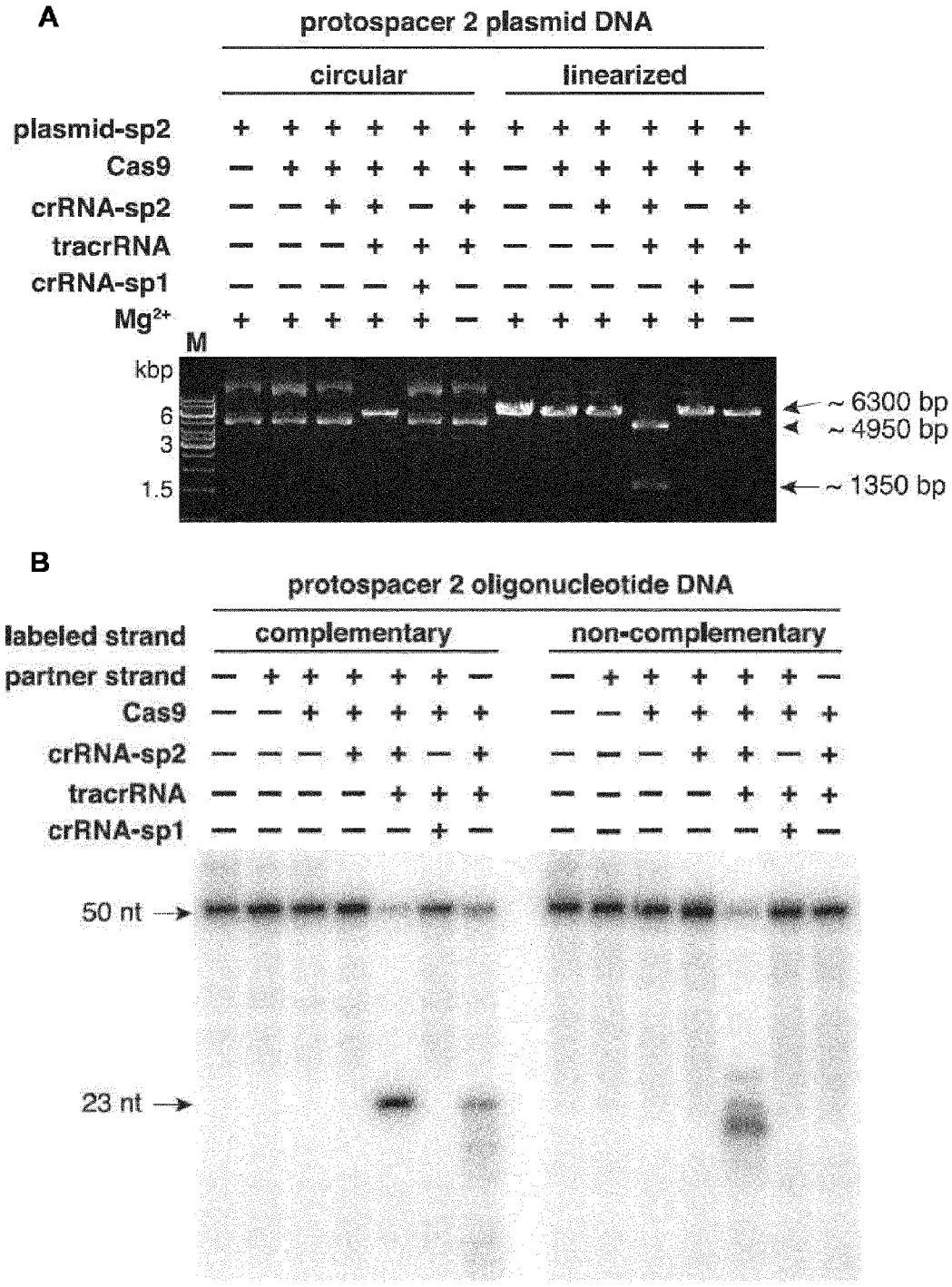

A
protospacer 2 plasmid DNA
circular
linearized
plasmid-sp2
Cas9
crRNA-sp2
tracrRNA
crRNA-sp1
Mg2+
M
kbp
6
3
1.5
~ 6300 bp
~ 4950 bp
~ 1350 bp
B
protospacer 2 oligonucleotide DNA
labeled strand
complementary
non-complementary
partner strand
Cas9
crRNA-sp2
tracrRNA
crRNA-sp1
50 nt
23 nt

FIGURE 10

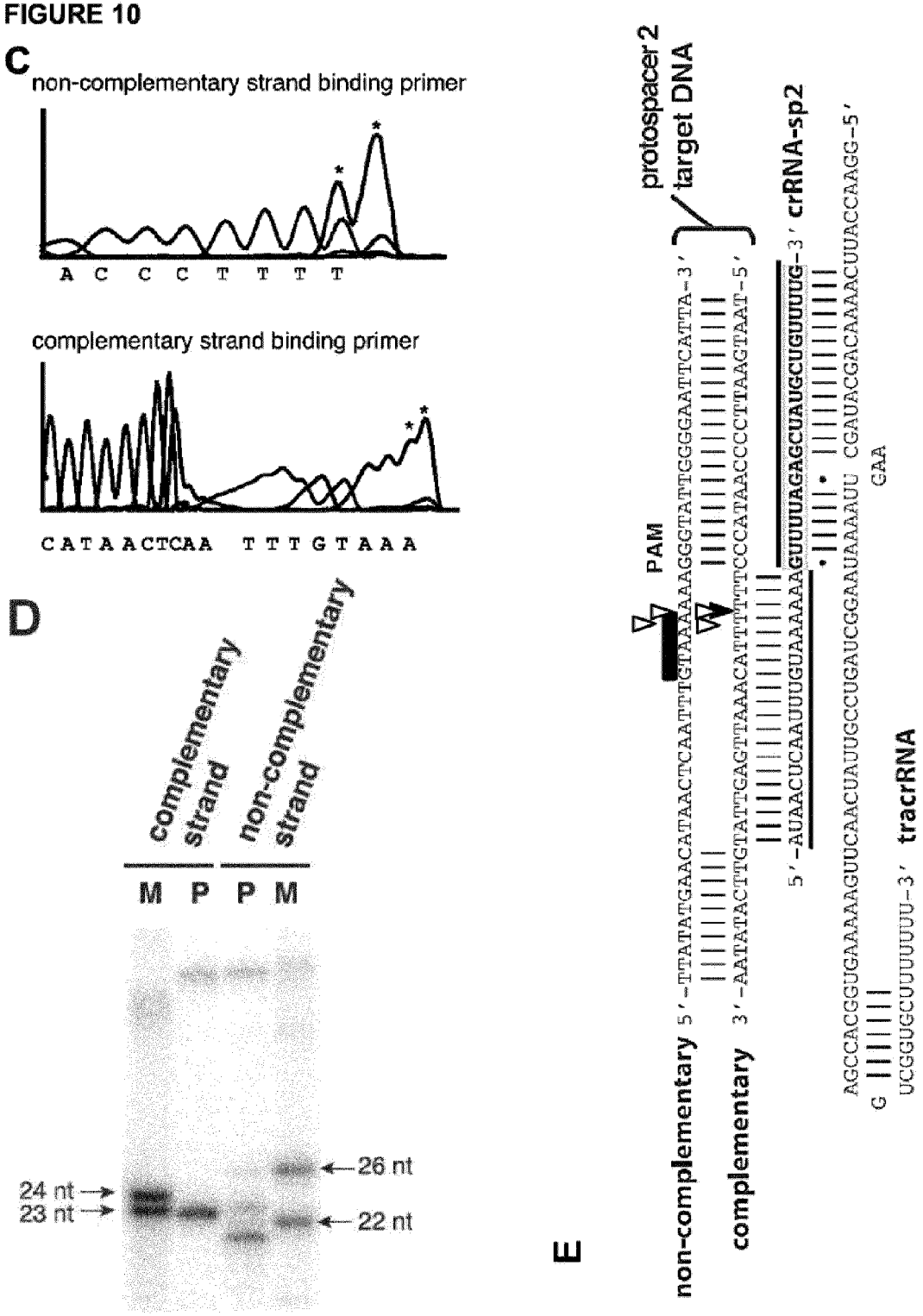

C
non-complementary strand binding primer
A C C C T T T T
complementary strand binding primer
CATAACTCAA TTTGTAAA
D
complementary strand
non-complementary strand
M P P M
24 nt
23 nt
26 nt
22 nt
E
protospacer2
target DNA
PAM
crRNA-sp2
non-complementary
complementary
tracrRNA

FIGURE 11

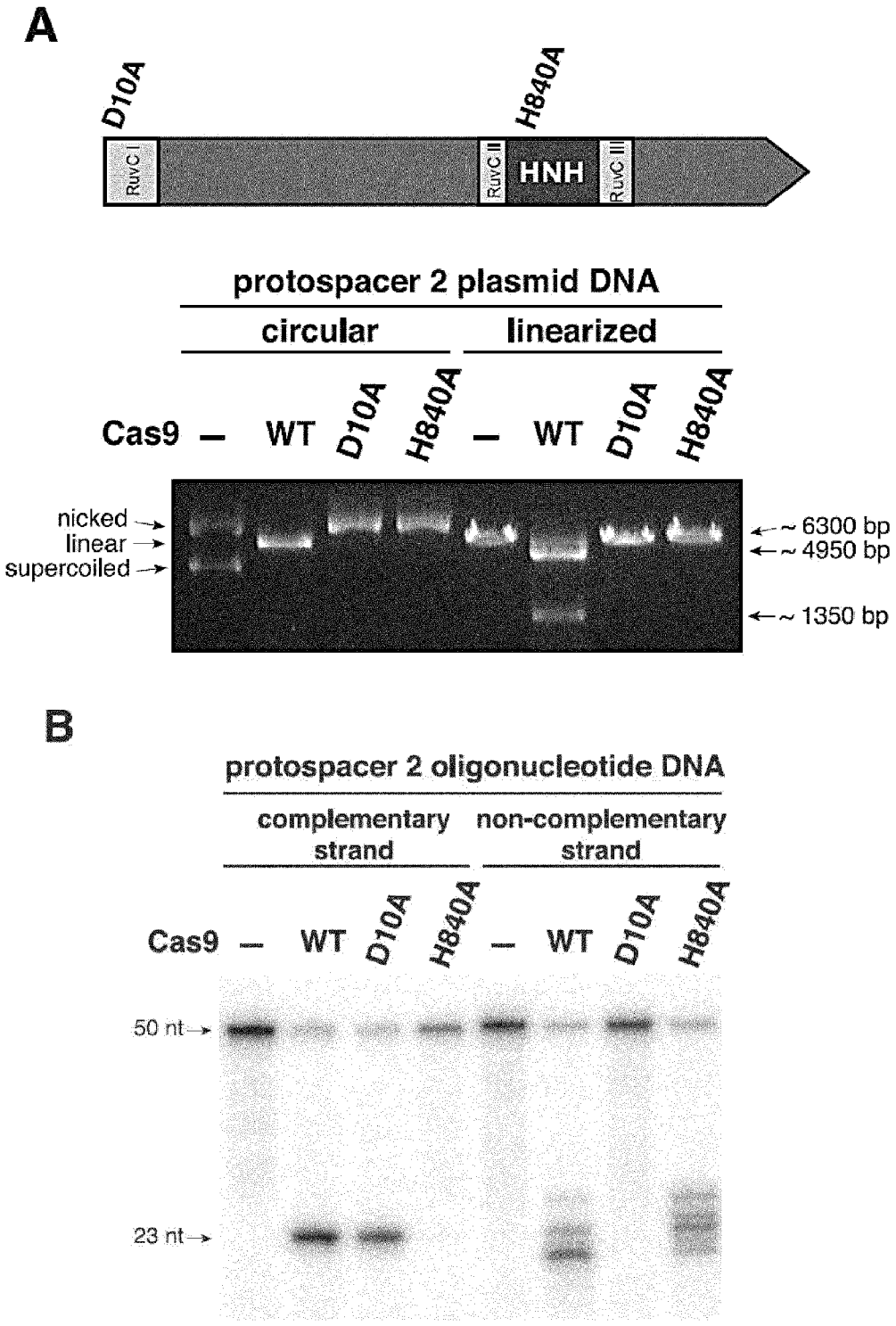

A
D10A
H840A
RuvC I
RuvC II
HNH
RuvC III
protospacer 2 plasmid DNA
circular
linearized
Cas9
–
WT
D10A
H840A
–
WT
D10A
H840A
nicked
linear
supercoiled
~ 6300 bp
~ 4950 bp
~ 1350 bp
B
protospacer 2 oligonucleotide DNA
complementary strand
non-complementary strand
Cas9
–
WT
D10A
H840A
–
WT
D10A
H840A
50 nt
23 nt

FIGURE 12

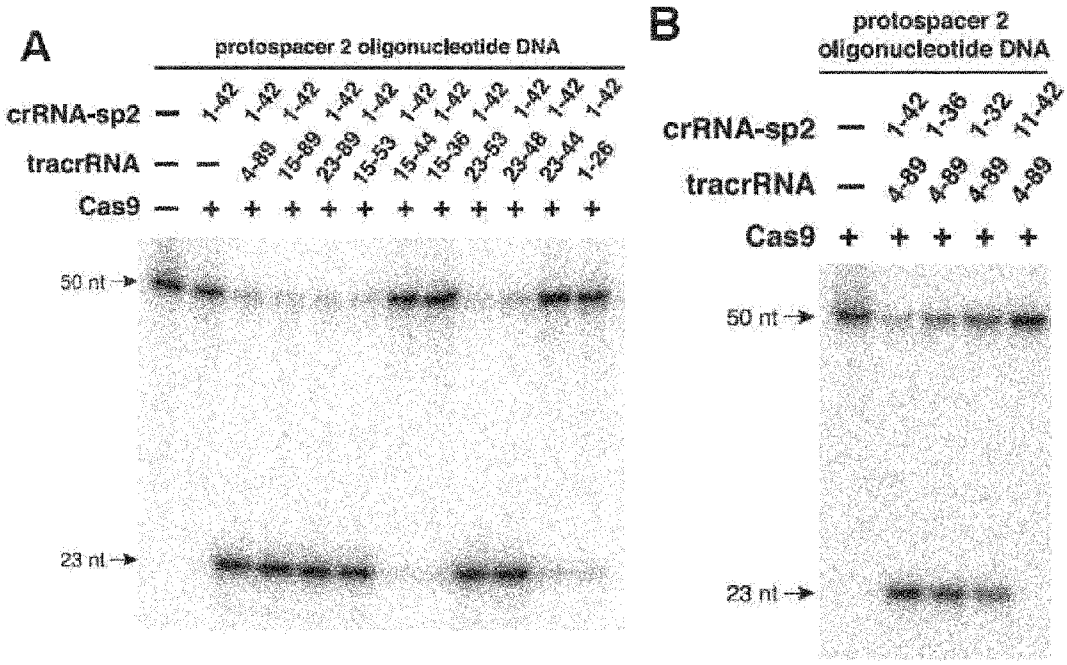

A
protospacer 2 oligonucleotide DNA
crRNA-sp2 — 1-42 1-42 1-42 1-42 1-42 1-42 1-42 1-42 1-42 1-42 1-42
tracrRNA — — 4-89 15-89 23-89 15-53 15-44 15-35 23-53 23-48 23-44 1-26
Cas9 — + + + + + + + + + + +
50 nt
23 nt
B
protospacer 2 oligonucleotide DNA
crRNA-sp2 — 1-42 1-36 1-32 11-42
tracrRNA — 4-89 4-89 4-89 4-89
Cas9 + + + + +
50 nt
23 nt

FIGURE 12

C

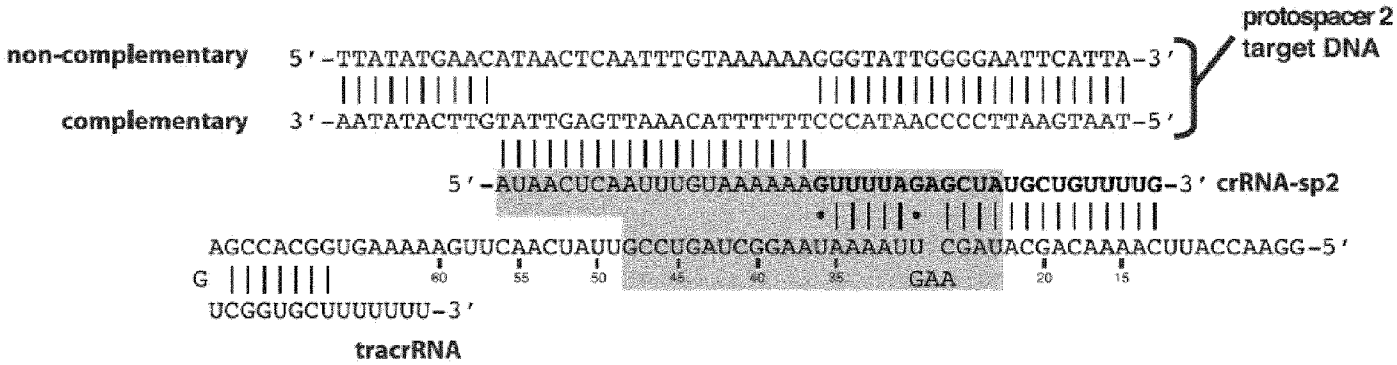
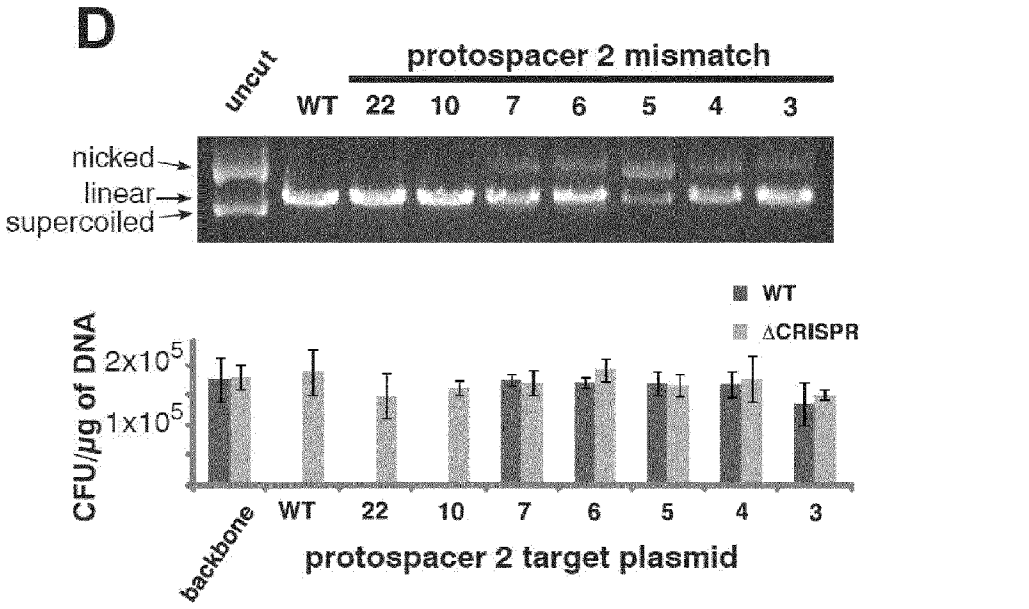
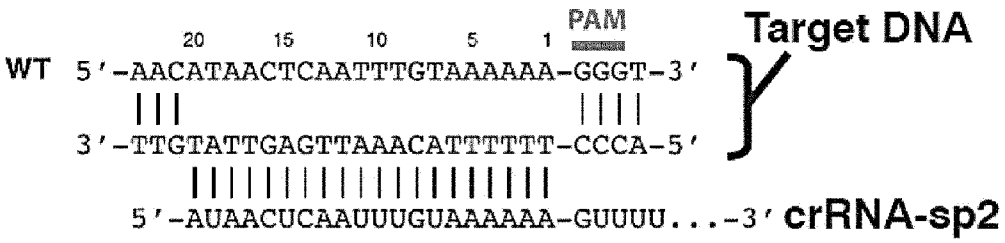


FIGURE 12





mismatched targets

22 5′-AGCATAACTCAATTTGTAAAAAA-3′

10 5′-AACATAACTCAATCTGTAAAAAA-3′

7 5′-AACATAACTCAATTTGAAAAAAA-3′

6 5′-AACATAACTCAATTTGTTAAAAA-3′

5 5′-AACATAACTCAATTTGTATAAAA-3′

4 5′-AACATAACTCAATTTGTAATAAA-3′

3 5′-AACATAACTCAATTTGTAAATAA-3′

FIGURE 12

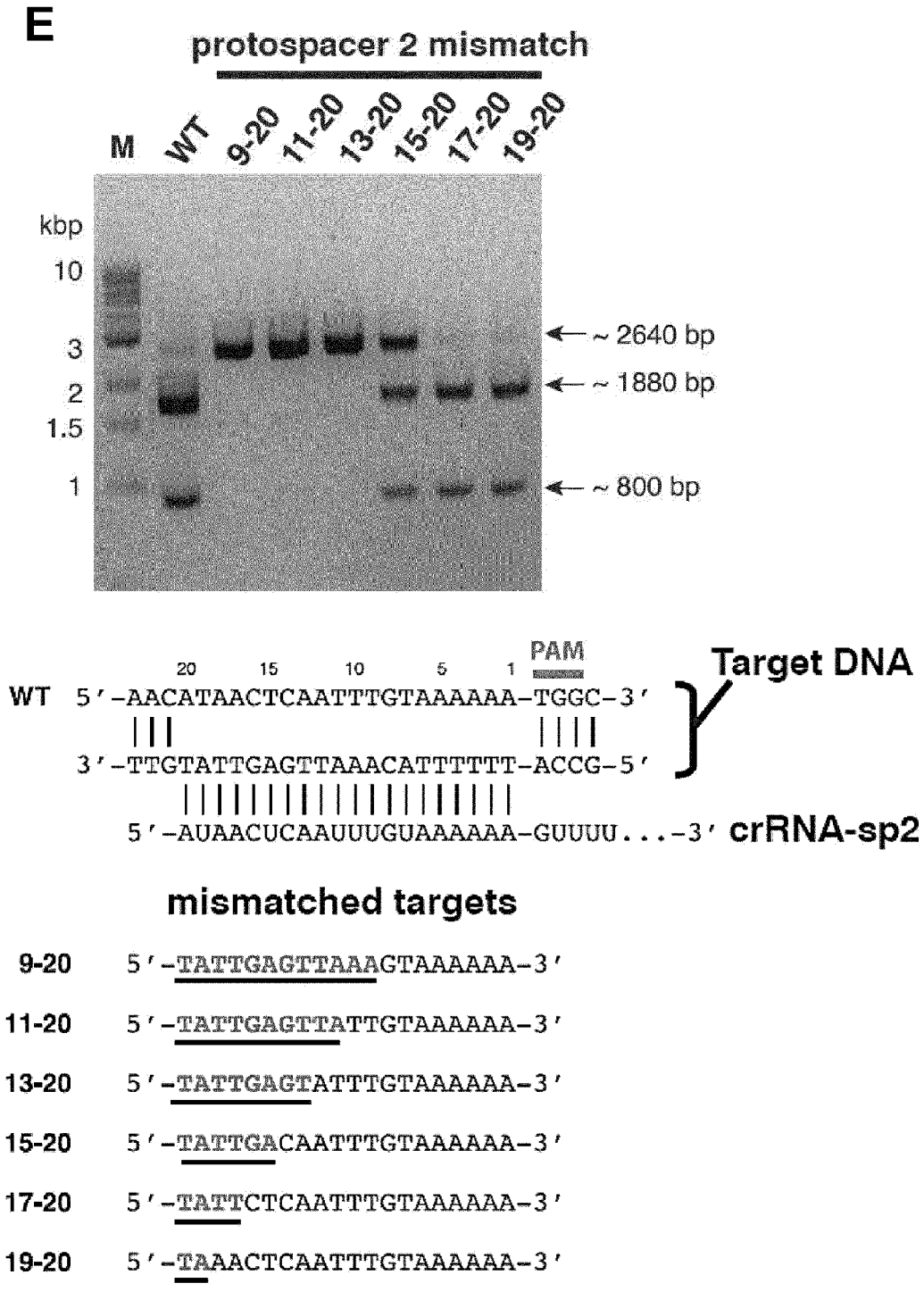

E
protospacer 2 mismatch
M
WT
9-20
11-20
13-20
15-20
17-20
19-20
kbp
10
3
2
1.5
1
~ 2640 bp
~ 1880 bp
~ 800 bp
PAM
20
15
10
5
1
Target DNA
WT 5′-AACATAACTCAATTTGTAAAAAA-TGGC-3′
3′-TTGTATTGAGTTAAACATTTTTT-ACCG-5′
5′-AUAACUCAAUUUGUAAAAAA-GUUUU...-3′
crRNA-sp2
mismatched targets
9-20 5′-TATTGAGTTAAAGTAAAAAA-3′
11-20 5′-TATTGAGTTATTGTAAAAAA-3′
13-20 5′-TATTGAGTATTTGTAAAAAA-3′
15-20 5′-TATTGACAATTTGTAAAAAA-3′
17-20 5′-TATTCTCAATTTGTAAAAAA-3′
19-20 5′-TAAACTCAATTTGTAAAAAA-3′

FIGURE 13

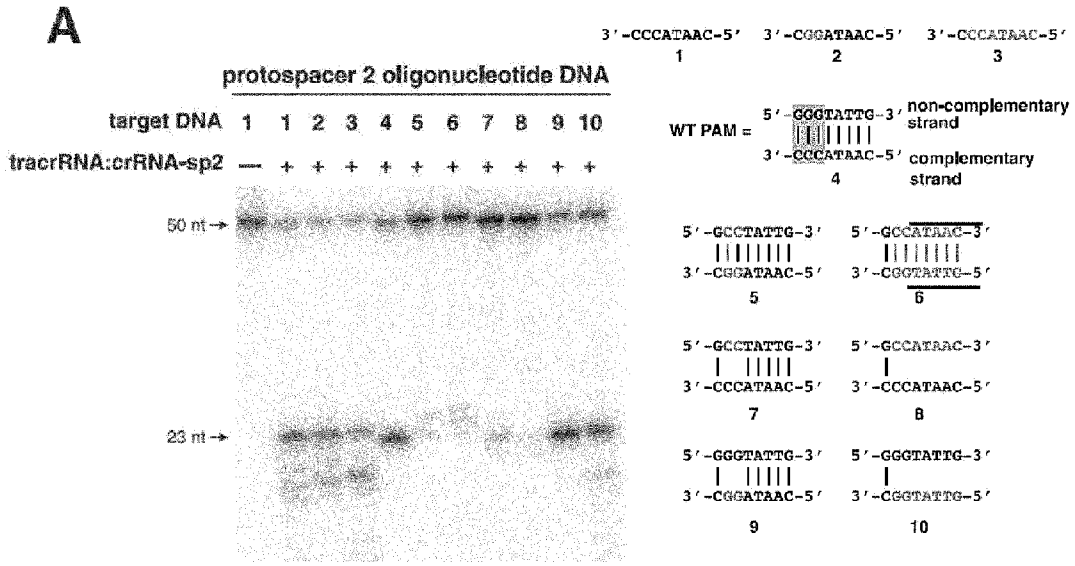

A
protospacer 2 oligonucleotide DNA
target DNA 1 1 2 3 4 5 6 7 8 9 10
tracrRNA:crRNA-sp2 — + + + + + + + + + +
50 nt
23 nt
3'-CCCATAAC-5' 3'-CGGATAAC-5' 3'-CCCATAAC-5'
1 2 3
WT PAM = 5'-GGGTATTG-3' non-complementary strand
3'-CCCATAAC-5' complementary strand
4
5'-GCCTATTG-3' 5'-GCCATAAC-3'
3'-CGGATAAC-5' 3'-CGGTATTG-5'
5 6
5'-GCCTATTG-3' 5'-GCCATAAC-3'
3'-CCCATAAC-5' 3'-CCCATAAC-5'
7 8
5'-GGGTATTG-3' 5'-GGGTATTG-3'
3'-CGGATAAC-5' 3'-CGGTATTG-5'
9 10

FIGURE 13

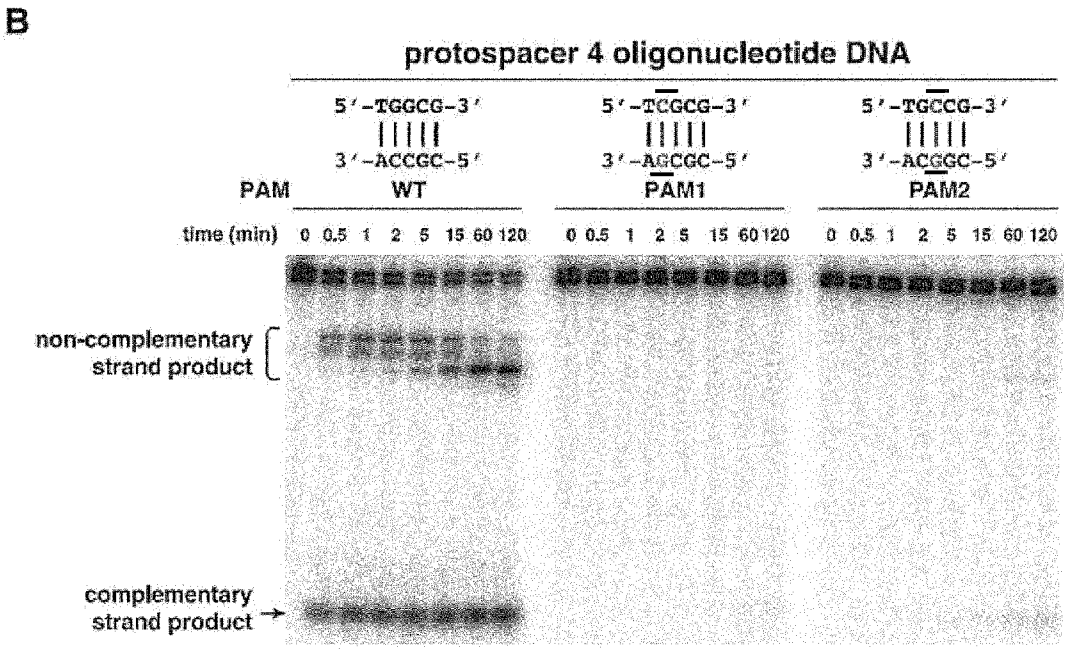

B
protospacer 4 oligonucleotide DNA
5′-TGGCG-3′
3′-ACCGC-5′
PAM
WT
5′-TCGCG-3′
3′-AGCGC-5′
PAM1
5′-TGCCG-3′
3′-ACGGC-5′
PAM2
time (min)
0 0.5 1 2 5 15 60 120
0 0.5 1 2 5 15 60 120
0 0.5 1 2 5 15 60 120
non-complementary strand product
complementary strand product

FIGURE 13

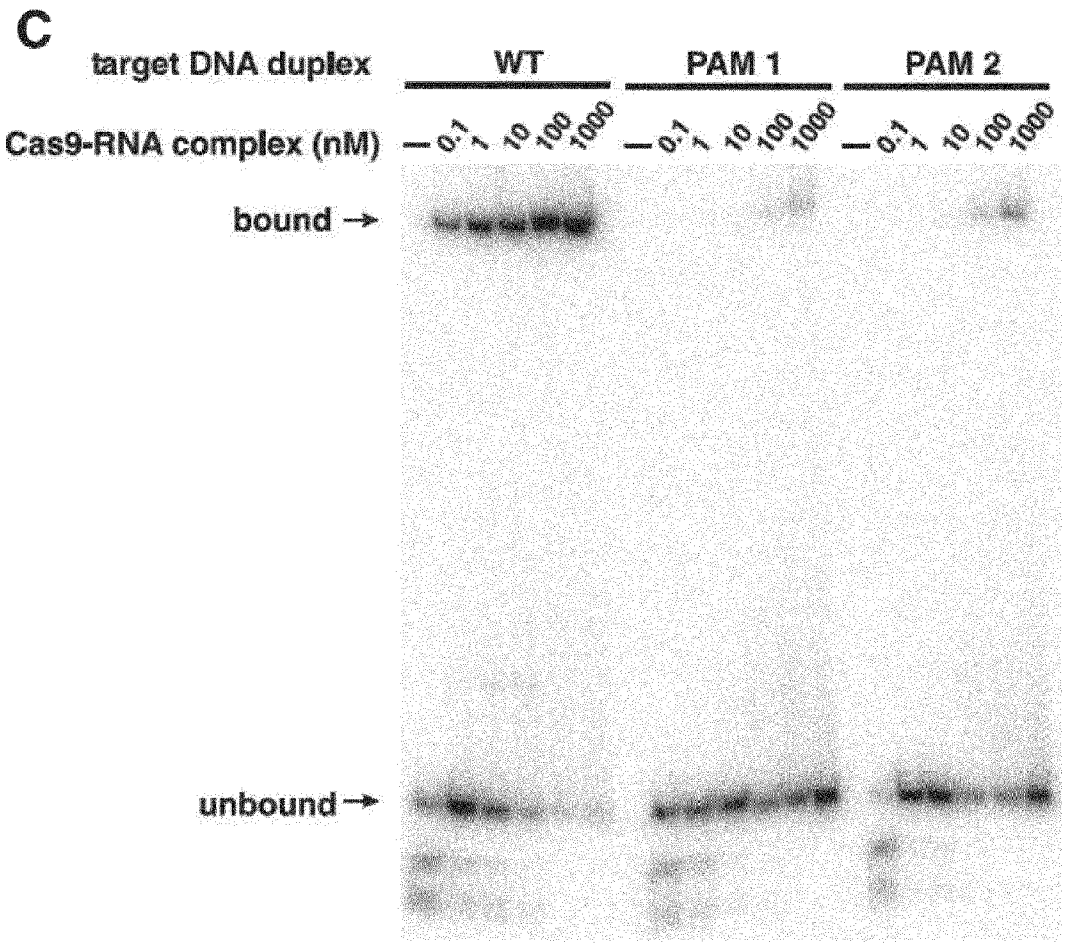

C
target DNA duplex
WT
PAM 1
PAM 2
Cas9-RNA complex (nM)
—
0.1
1
10
100
1000
bound
unbound

FIGURE 14

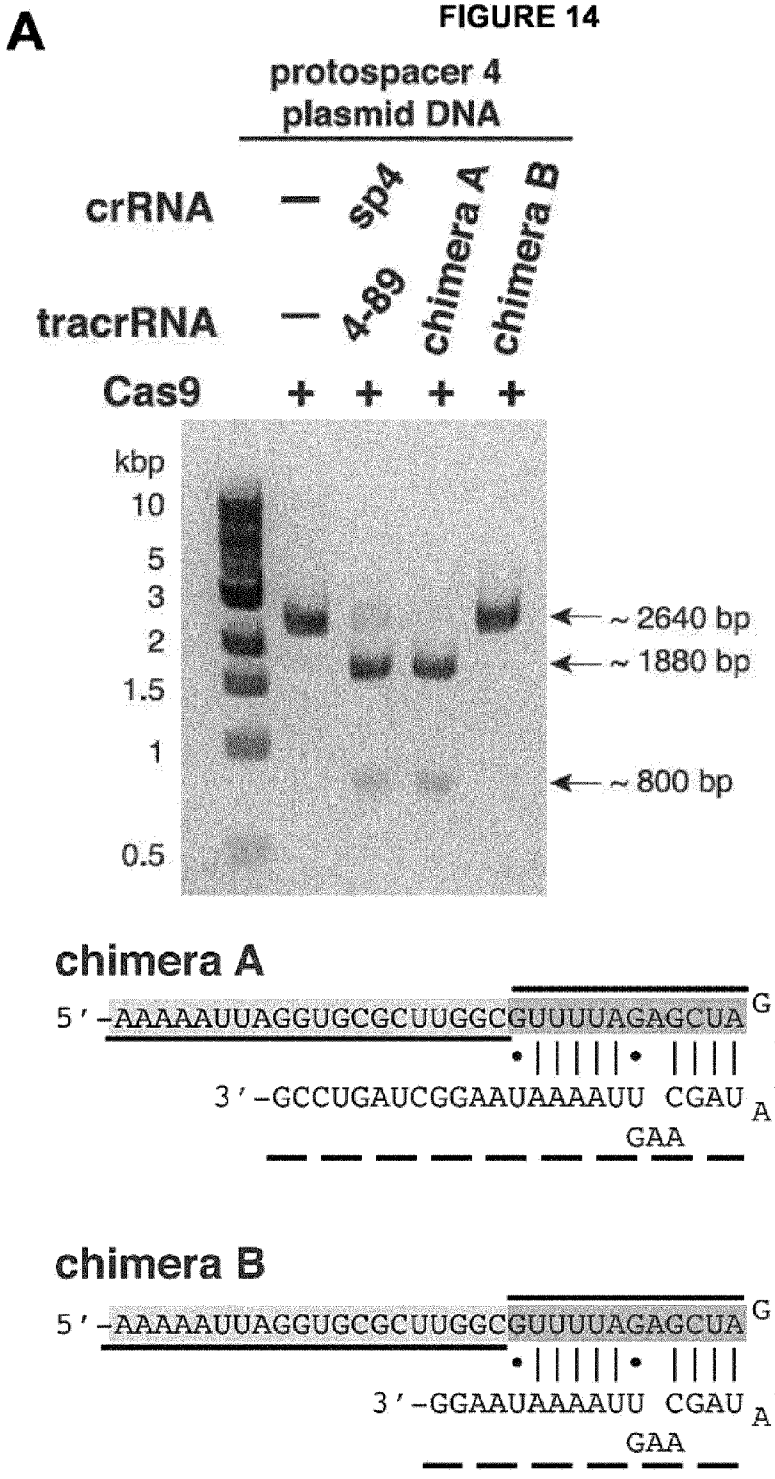

A
protospacer 4
plasmid DNA
crRNA
tracrRNA
Cas9
—
sp4
4-89
chimera A
chimera B
+
kbp
10
5
3
2
1.5
1
0.5
~ 2640 bp
~ 1880 bp
~ 800 bp
chimera A
5′-AAAAAUUAGGUGCGCUUGGCGUUUUAGAGCUA
3′-GCCUGAUCGGAAUAAAAUU CGAU
GAA
chimera B
5′-AAAAAUUAGGUGCGCUUGGCGUUUUAGAGCUA
3′-GGAAUAAAAUU CGAU
GAA

FIGURE 14

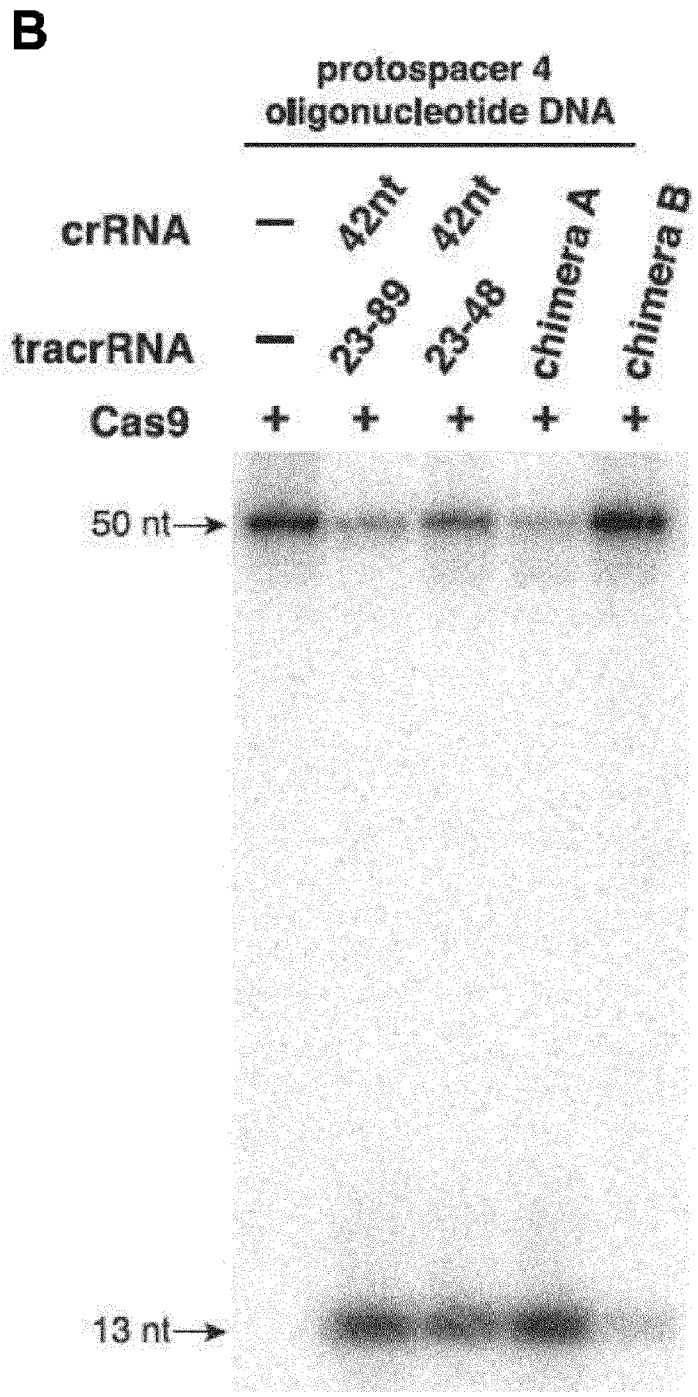

B
protospacer 4
oligonucleotide DNA
crRNA
tracrRNA
Cas9
–
–
+
42nt
23-89
+
42nt
23-48
+
chimera A
+
chimera B
+
50 nt
13 nt

FIGURE 14

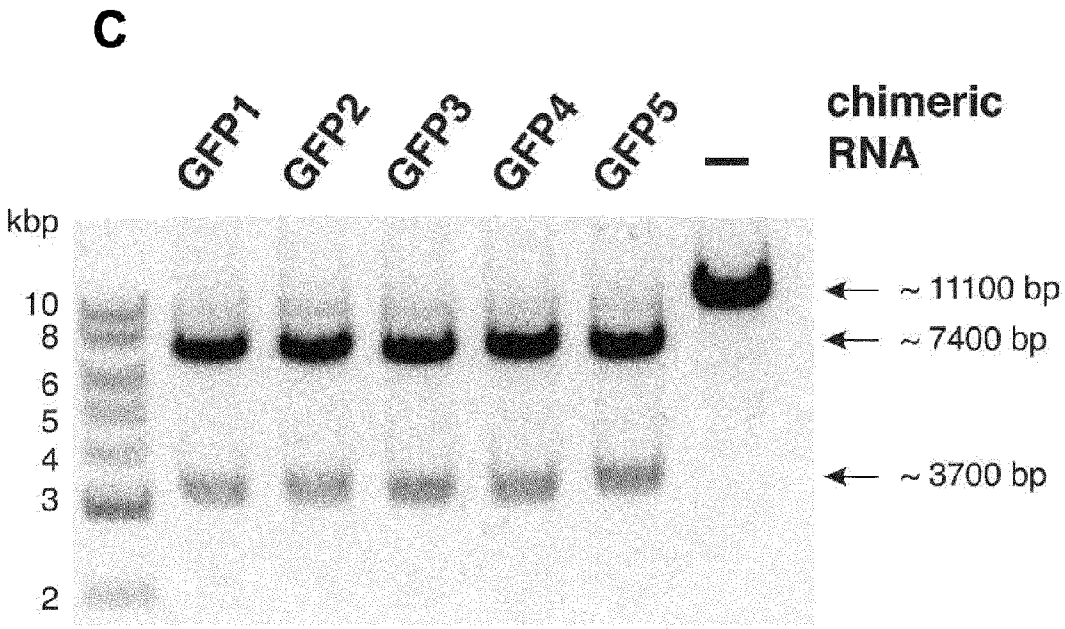

C
GFP1
GFP2
GFP3
GFP4
GFP5
—
chimeric RNA
kbp
10
8
6
5
4
3
2
~ 11100 bp
~ 7400 bp
~ 3700 bp

FIGURE 15

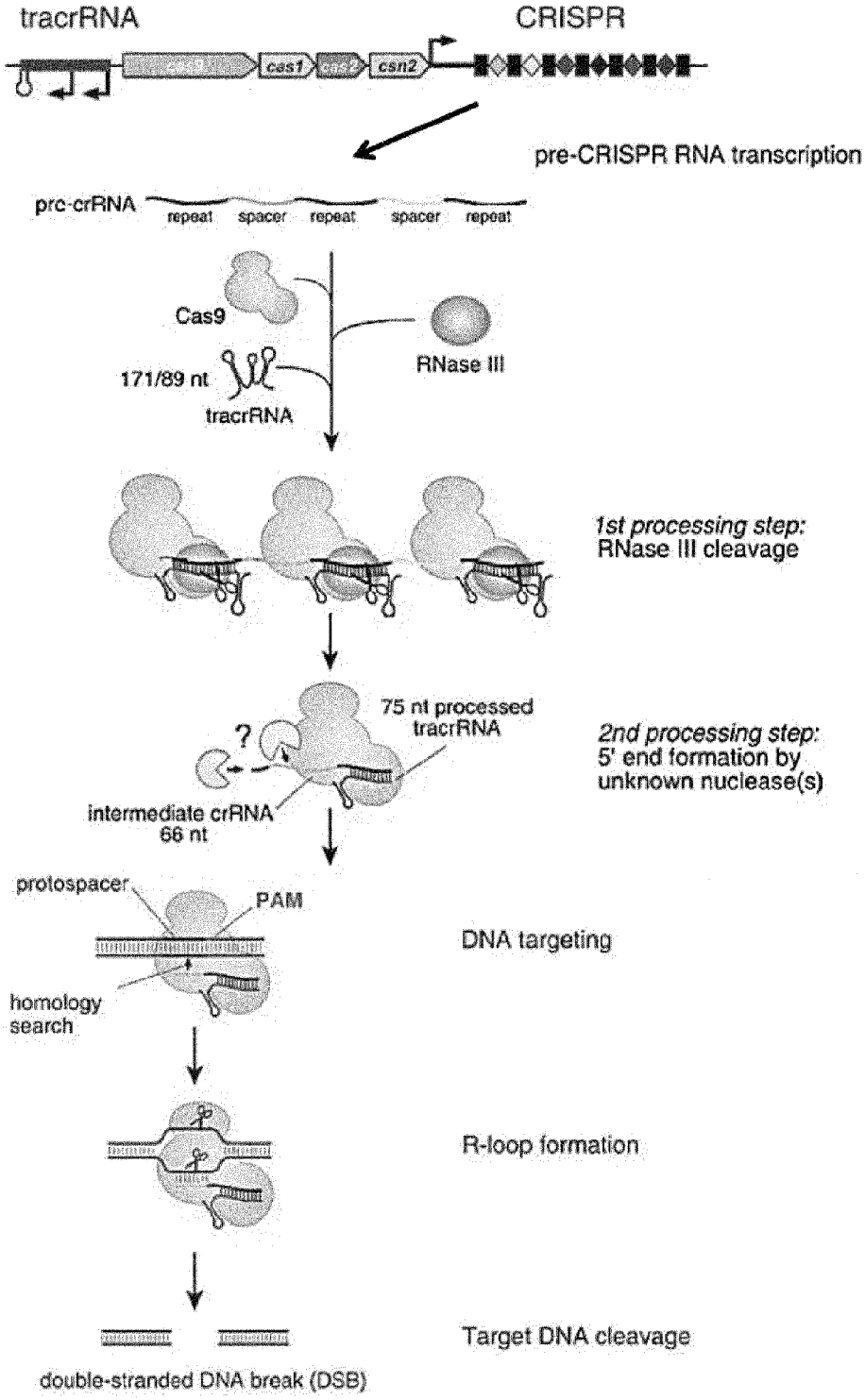

tracrRNA
CRISPR
cas9
cas1
cas2
csn2
pre-CRISPR RNA transcription
pre-crRNA
repeat
spacer
repeat
spacer
repeat
Cas9
171/89 nt
tracrRNA
RNase III
1st processing step:
RNase III cleavage
75 nt processed
tracrRNA
?
intermediate crRNA
66 nt
2nd processing step:
5' end formation by
unknown nuclease(s)
protospacer
PAM
homology
search
DNA targeting
R-loop formation
Target DNA cleavage
double-stranded DNA break (DSB)

FIGURE 16

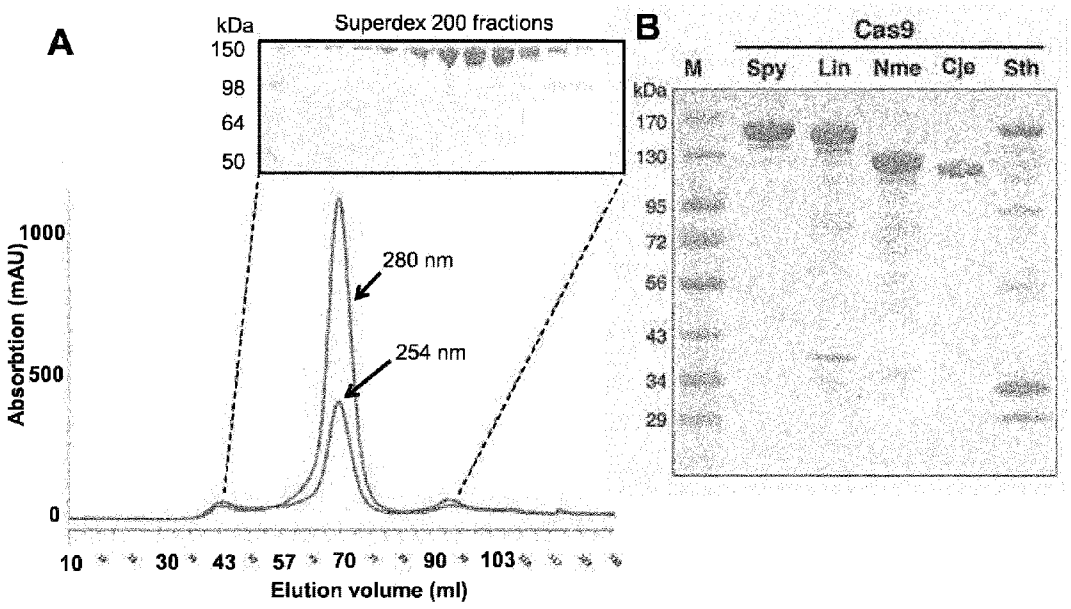

A
kDa
Superdex 200 fractions
150
98
64
50
1000
500
0
Absorbtion (mAU)
280 nm
254 nm
10
30
43
57
70
90
103
Elution volume (ml)
B
Cas9
M
Spy
Lin
Nme
Cje
Sth
kDa
170
130
95
72
56
43
34
29

FIGURE 17

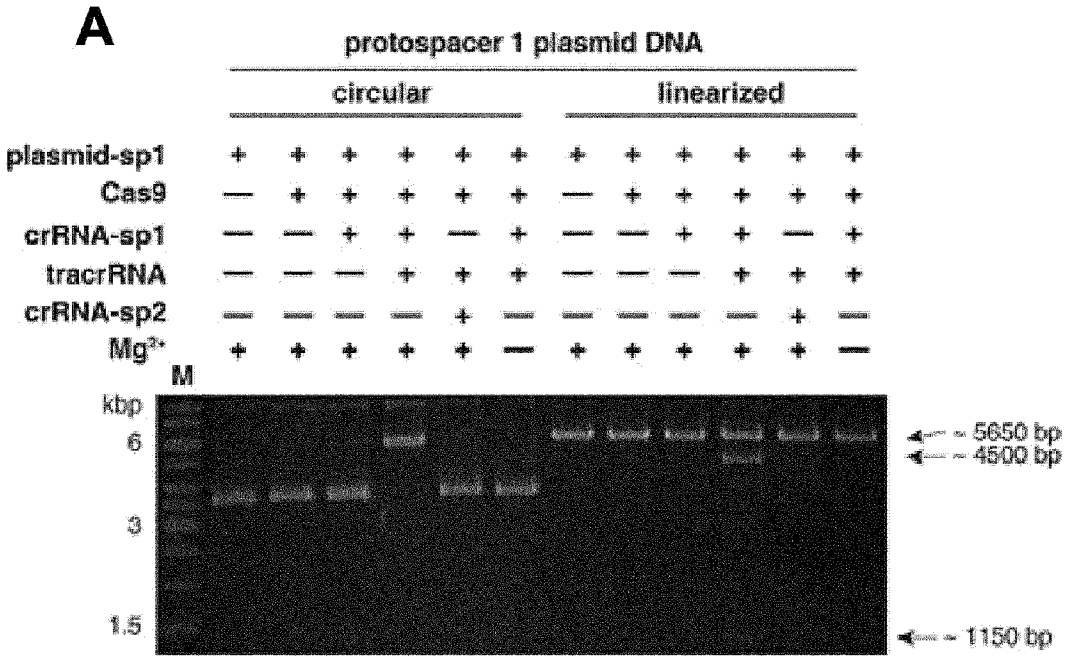
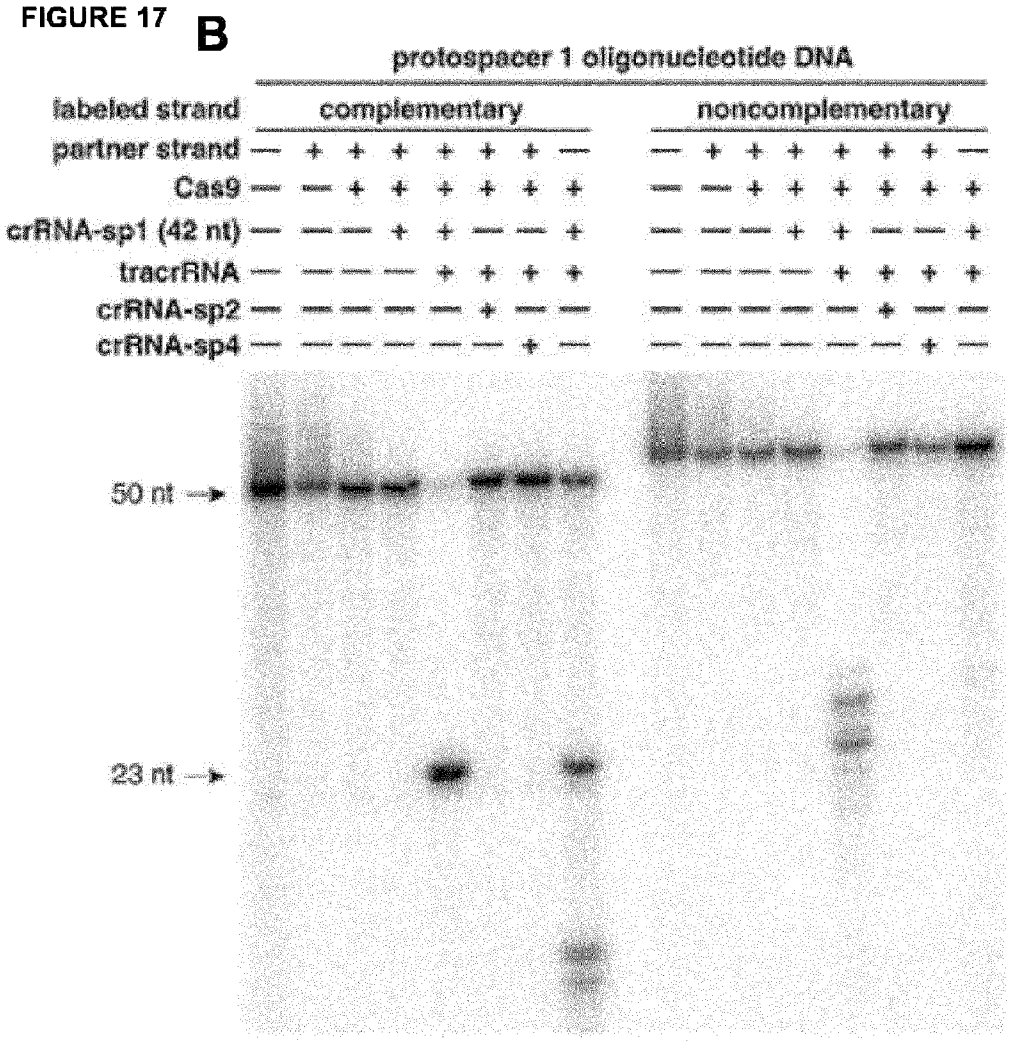

A
protospacer 1 plasmid DNA
circular
linearized
plasmid-sp1
Cas9
crRNA-sp1
tracrRNA
crRNA-sp2
Mg2+
M
kbp
6
3
1.5
~ 5650 bp
~ 4500 bp
~ 1150 bp

FIGURE 17


B
protospacer 1 oligonucleotide DNA
labeled strand
complementary
noncomplementary
partner strand
Cas9
crRNA-sp1 (42 nt)
tracrRNA
crRNA-sp2
crRNA-sp4
50 nt
23 nt

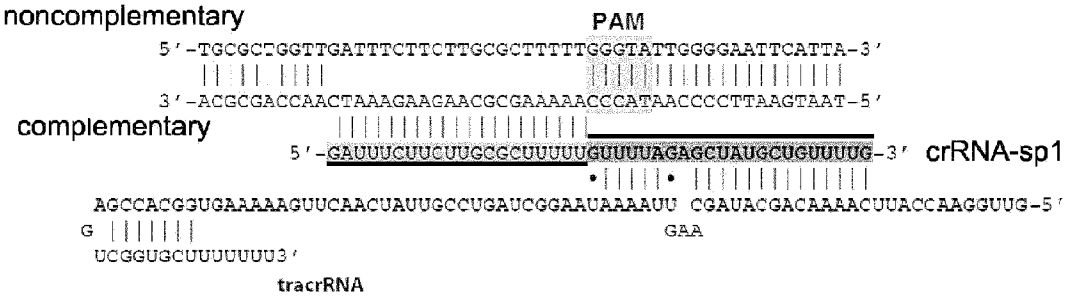

noncomplementary
PAM
5′-TGCGCTGGTTGATTTCTTCTTGCGCTTTTTGGGTATTGGGGAATTCATTA-3′
3′-ACGCGACCAACTAAAGAAGAACGCGAAAAACCCATAACCCCTTAAGTAAT-5′
complementary
5′-GAUUUCUUCUUGCGCUUUUUGUUUUAGAGCUAUGCUGUUUUG-3′
crRNA-sp1
AGCCACGGUGAAAAAGUUCAACUAUUGCCUGAUCGGAAUAAAAUU CGAUACGACAAAACUUACCAAGGUUG-5′
G
GAA
UCGGUGCUUUUUUU3′
tracrRNA

FIGURE 17

C

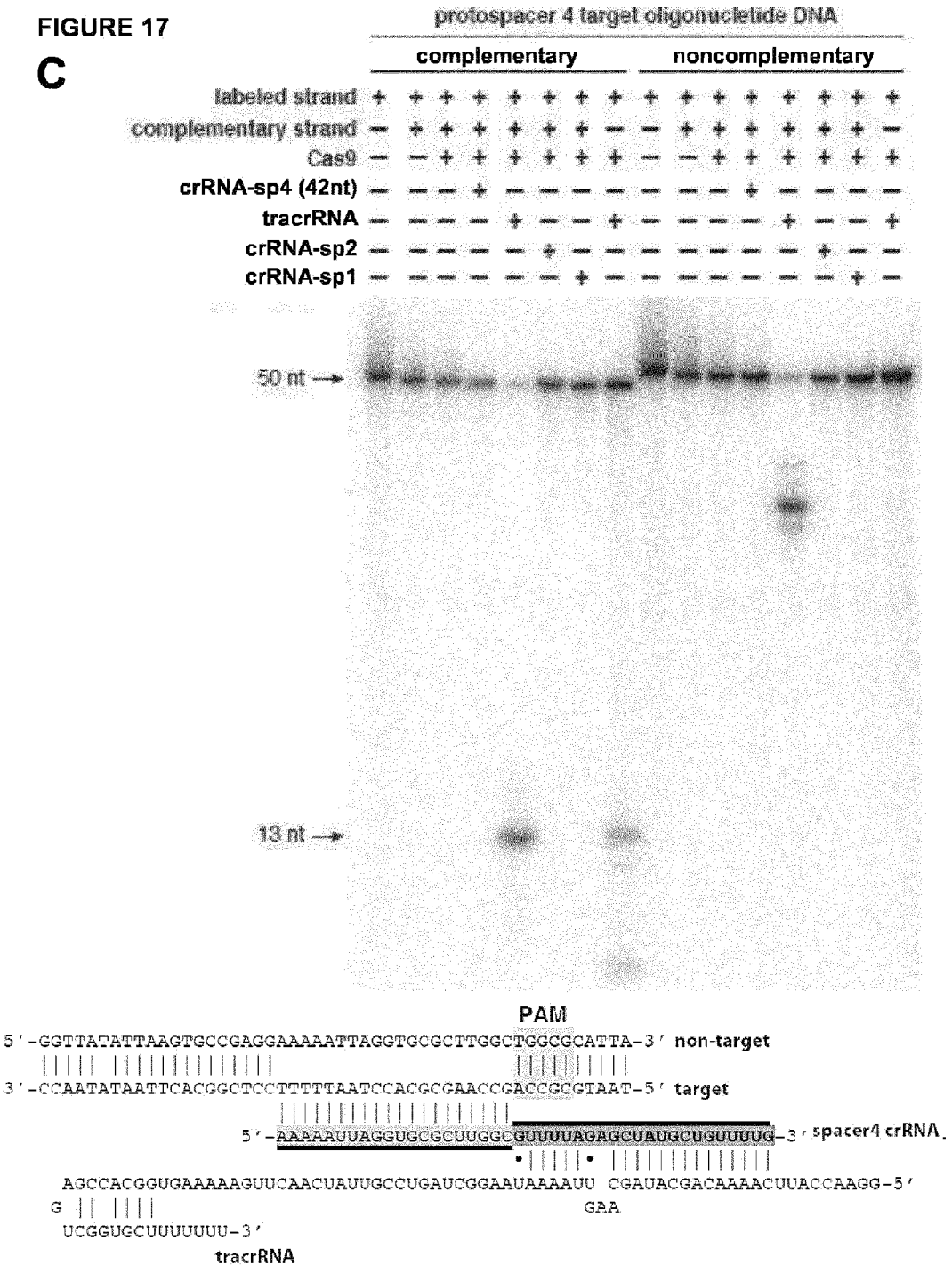

protospacer 4 target oligonucleotide DNA
complementary
noncomplementary
labeled strand
complementary strand
Cas9
crRNA-sp4 (42nt)
tracrRNA
crRNA-sp2
crRNA-sp1
50 nt
13 nt
PAM
5′-GGTTATATTAAGTGCCGAGGAAAAATTAGGTGCGCTTGGCTGGCGCATTA-3′ non-target
3′-CCAATATAATTCACGGCTCCTTTTTAATCCACGCGAACCGACCGCGTAAT-5′ target
5′-AAAAAUUAGGUGCGCUUGGCGUUUUAGAGCUAUGCUGUUUUG-3′ spacer4 crRNA
AGCCACGGUGAAAAAGUUCAACUAUUGCCUGAUCGGAAUAAAAUU CGAUACGACAAAACUUACCAAGG-5′
G
GAA
UCGGUGCUUUUUUU-3′
tracrRNA

Figure 18

A

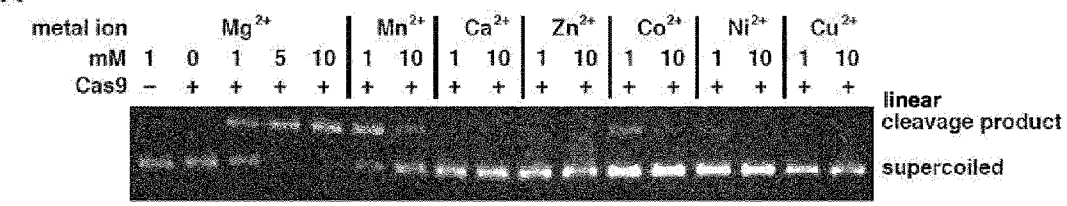

metal ion
Mg2+
Mn2+
Ca2+
Zn2+
Co2+
Ni2+
Cu2+
mM 1 0 1 5 10 1 10 1 10 1 10 1 10 1 10 1 10
Cas9 – + + + + + + + + + + + + + + + +
linear cleavage product
supercoiled

Figure 18

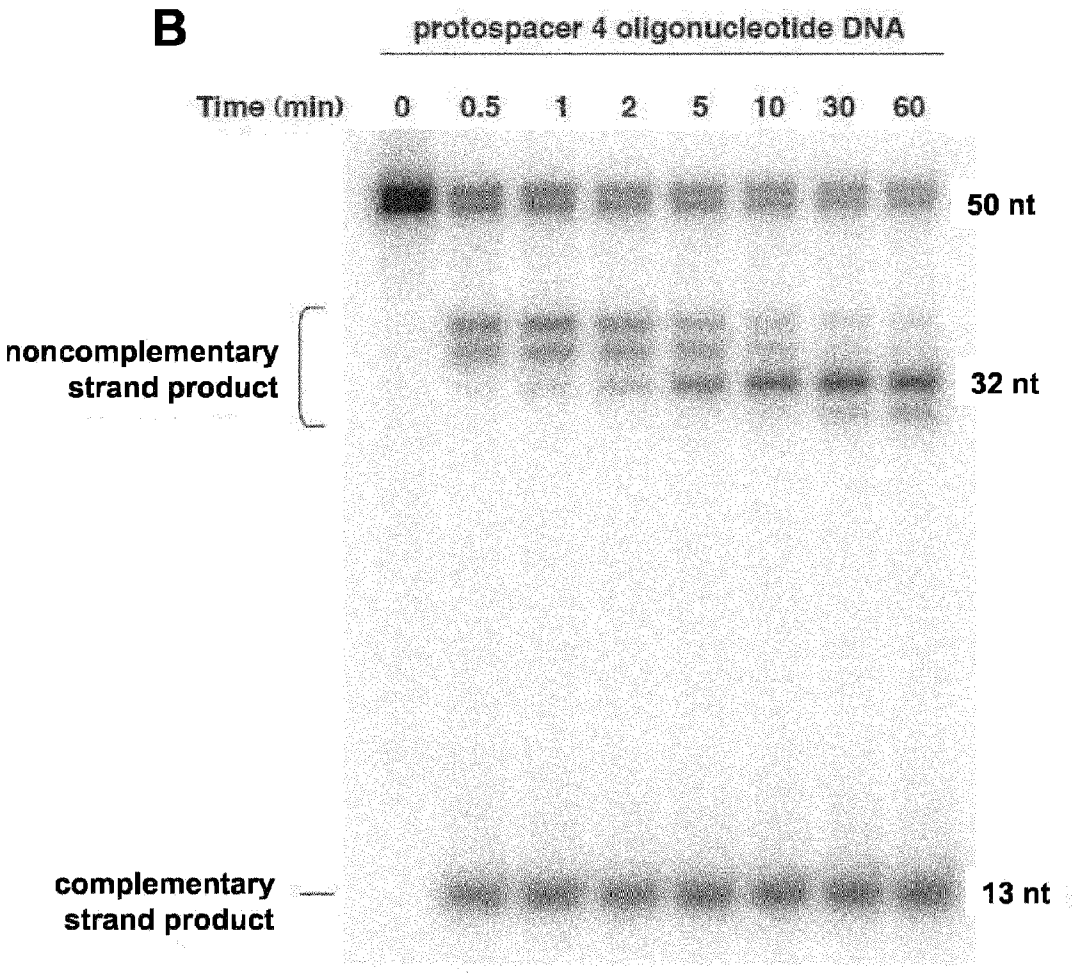

B
protospacer 4 oligonucleotide DNA
Time (min)
0
0.5
1
2
5
10
30
60
50 nt
noncomplementary strand product
32 nt
complementary strand product
13 nt

Figure 19

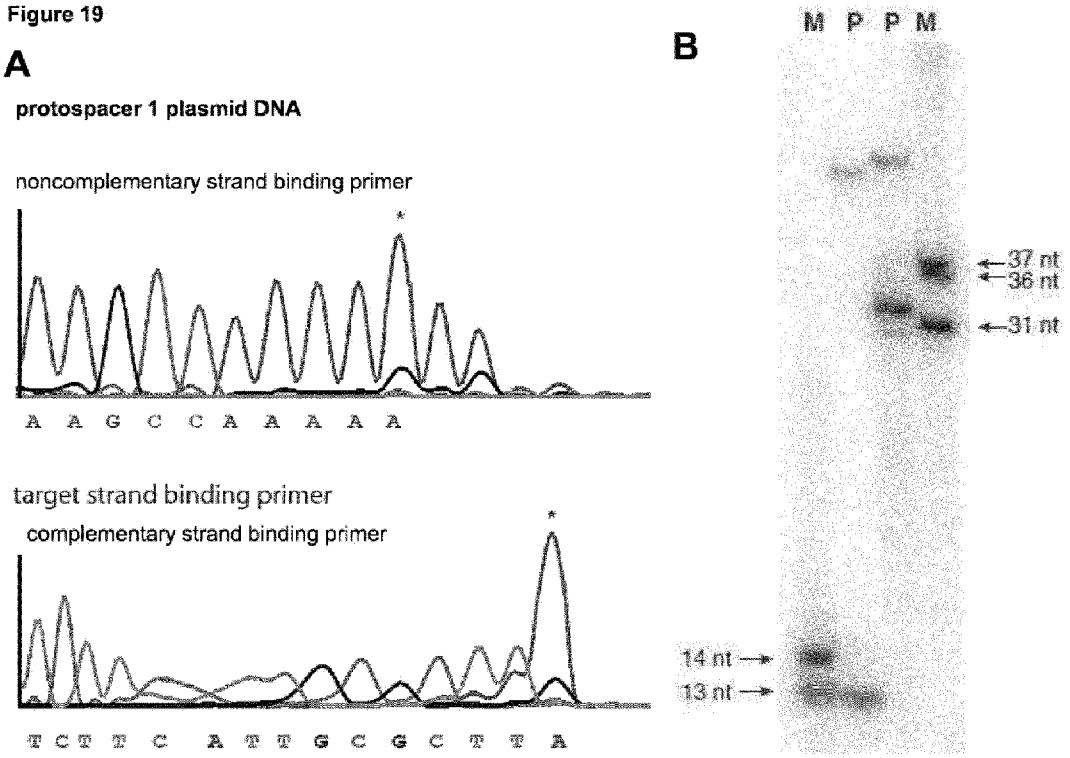
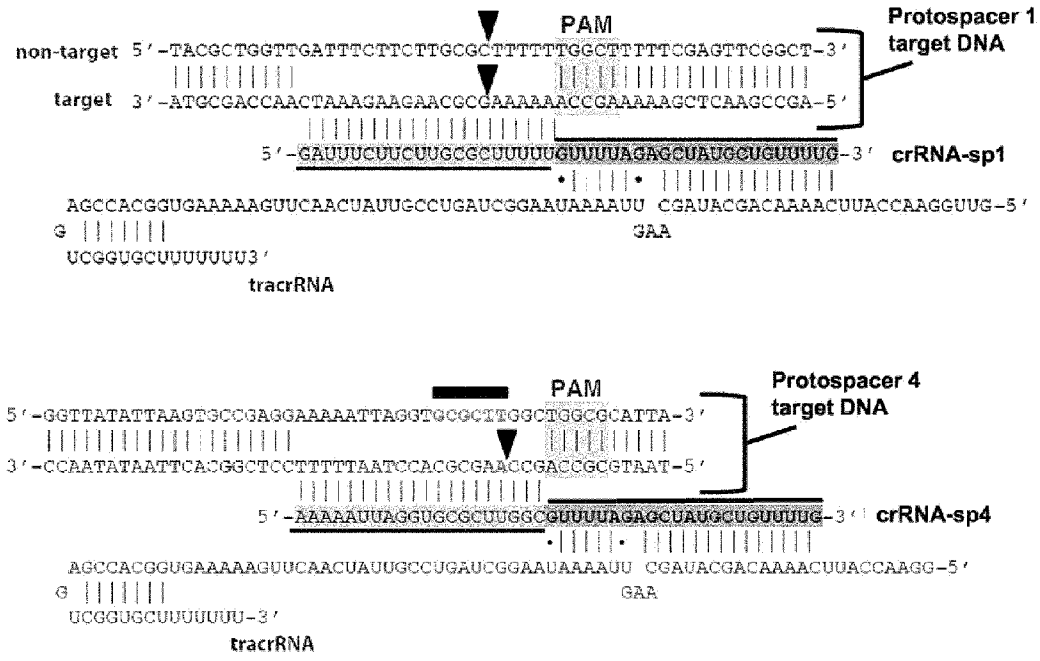

A
protospacer 1 plasmid DNA
noncomplementary strand binding primer
A A G C C A A A A A
target strand binding primer
complementary strand binding primer
T C T T C A T T G C G C T T A
B
M P P M
37 nt
36 nt
31 nt
14 nt
13 nt

Figure 19

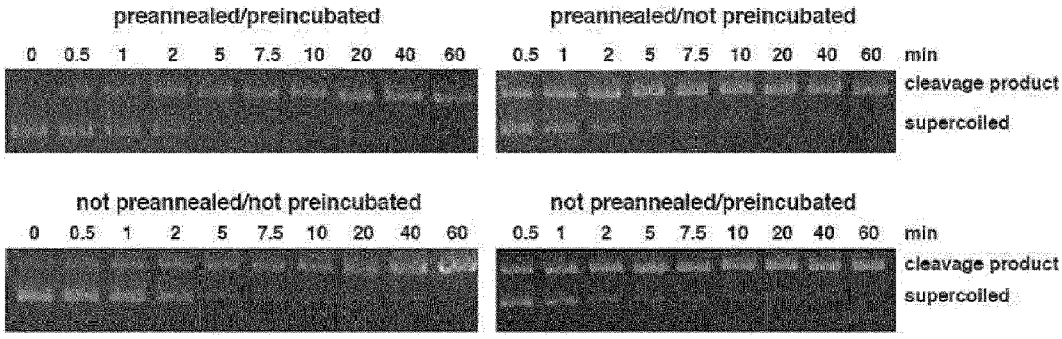
C

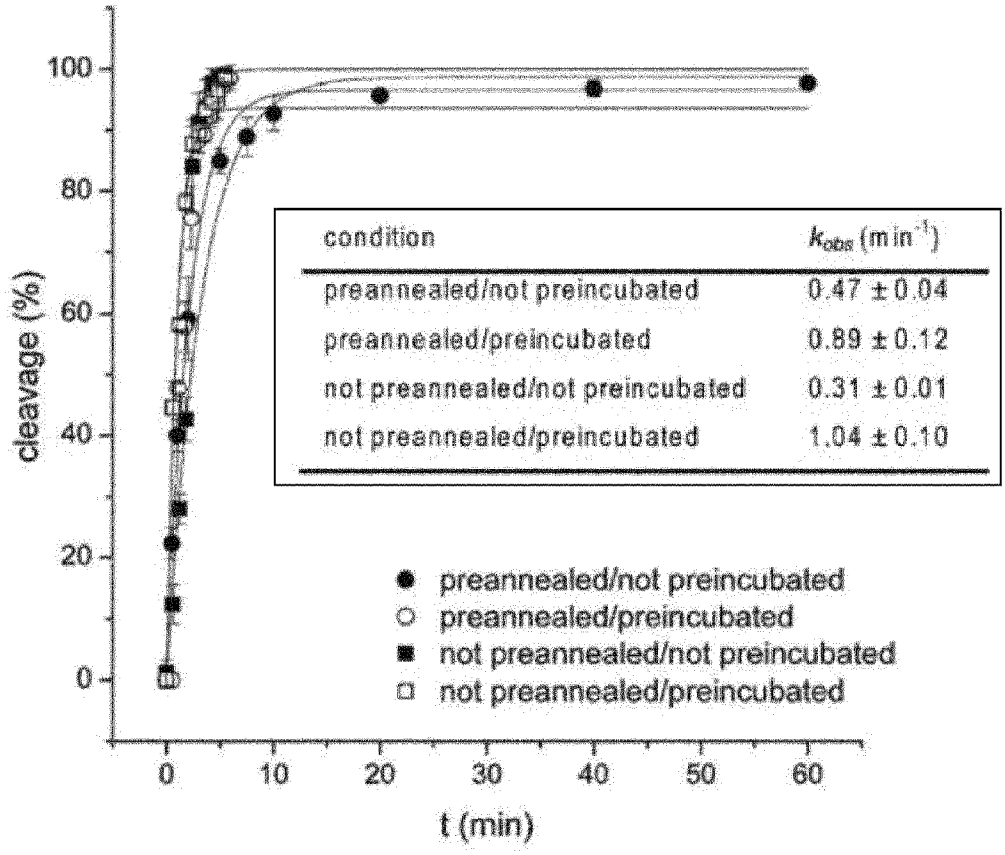

PAM
Protospacer 1
target DNA
non-target 5′-TACGCTGGTTGATTTCTTCTTGCGCTTTTTTGGCTTTTTCGAGTTCGGCT-3′
target 3′-ATGCGACCAACTAAAGAAGAACGCGAAAAAACCGAAAAAGCTCAAGCCGA-5′
5′-GAUUUCUUCUUGCGCUUUUUGUUUUAGAGCUAUGCUGUUUUG-3′ crRNA-sp1
AGCCACGGUGAAAAAGUUCAACUAUUGCCUGAUCGGAAUAAAAUU CGAUACGACAAAACUUACCAAGGUUG-5′
G
GAA
UCGGUGCUUUUUUU3′
tracrRNA

PAM
Protospacer 4
target DNA
5′-GGTTATATTAAGTGCCGAGGAAAAATTAGGTGCGCTTGGCTGGCGCATTA-3′
3′-CCAATATAATTCACGGCTCCTTTTTAATCCACGCGAACCGACCGCGTAAT-5′
5′-AAAAAUUAGGUGCGCUUGGCGUUUUAGAGCUAUGCUGUUUUG-3′ crRNA-sp4
AGCCACGGUGAAAAAGUUCAACUAUUGCCUGAUCGGAAUAAAAUU CGAUACGACAAAACUUACCAAGG-5′
G
GAA
UCGGUGCUUUUUUU-3′
tracrRNA

Figure 20

A





condition	k_{obs} (min^{-1})
preannealed/not preincubated	0.47 ± 0.04
preannealed/preincubated	0.89 ± 0.12
not preannealed/not preincubated	0.31 ± 0.01
not preannealed/preincubated	1.04 ± 0.10

Figure 20

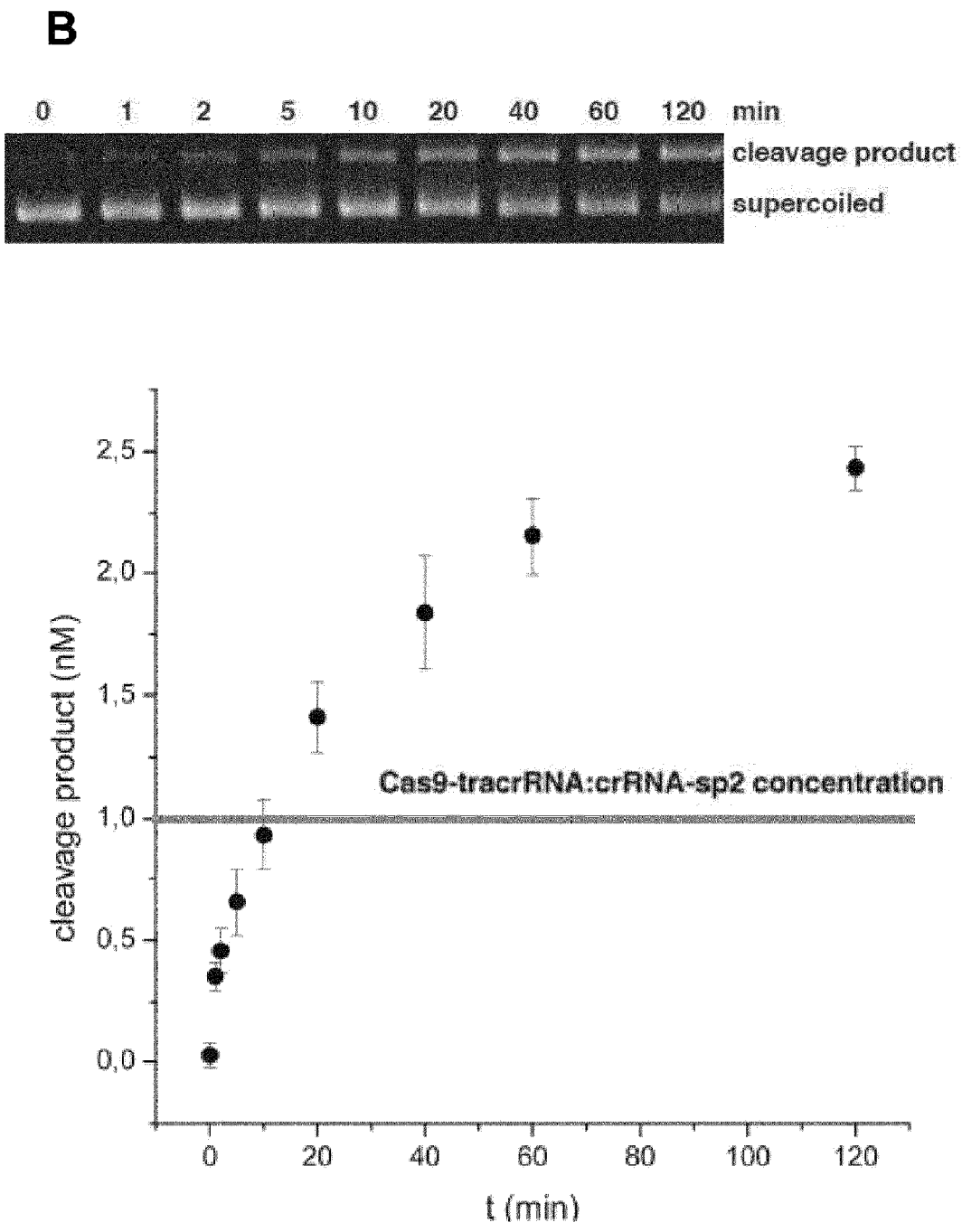

B
0
1
2
5
10
20
40
60
120
min
cleavage product
supercoiled
cleavage product (nM)
2,5
2,0
1,5
1,0
0,5
0,0
Cas9-tracrRNA:crRNA-sp2 concentration
0
20
40
60
80
100
120
t (min)

FIGURE 21

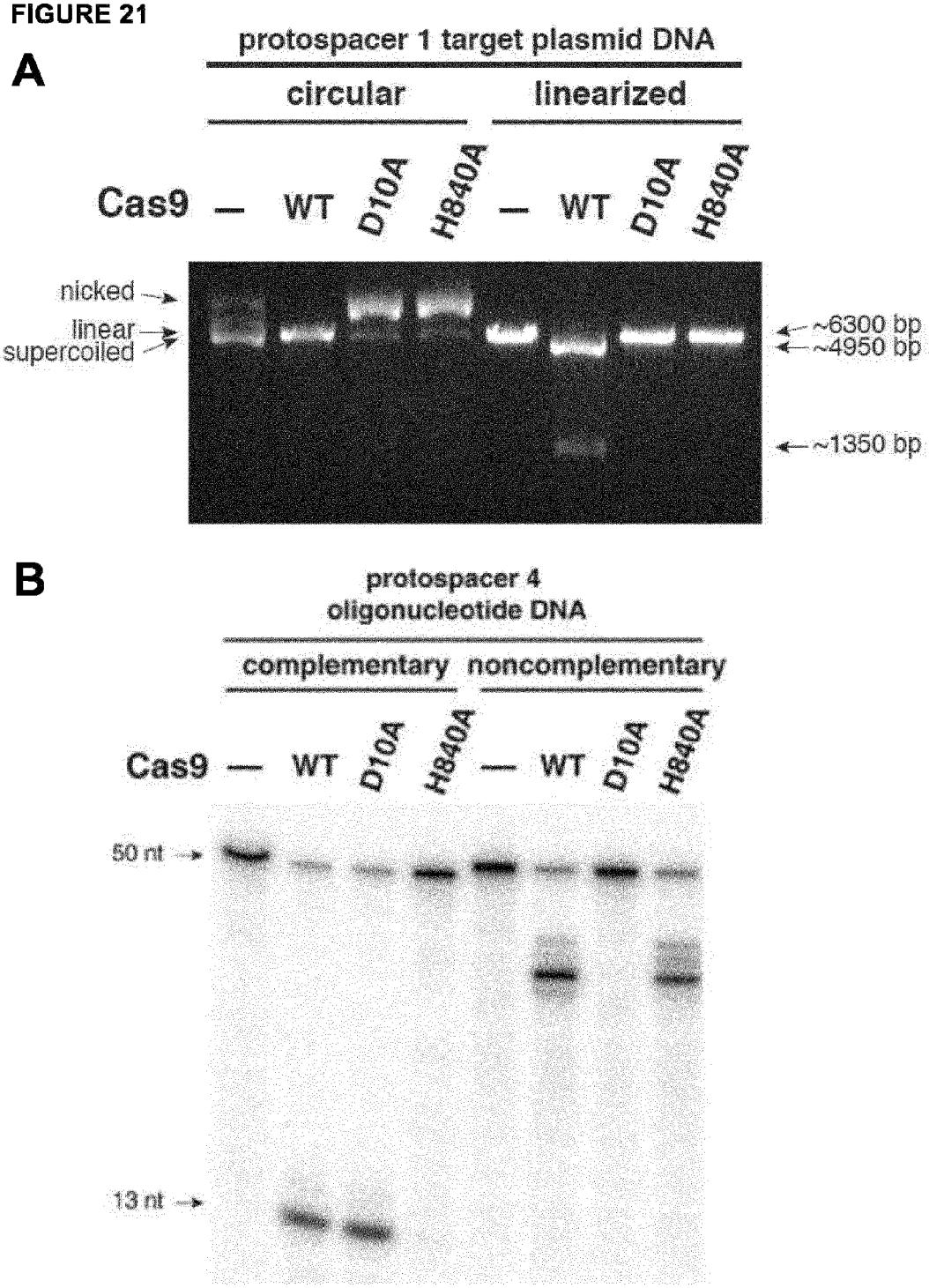

A
protospacer 1 target plasmid DNA
circular
linearized
Cas9
—
WT
D10A
H840A
—
WT
D10A
H840A
nicked
linear
supercoiled
~6300 bp
~4950 bp
~1350 bp
B
protospacer 4
oligonucleotide DNA
complementary
noncomplementary
Cas9
—
WT
D10A
H840A
—
WT
D10A
H840A
50 nt
13 nt

FIGURE 22

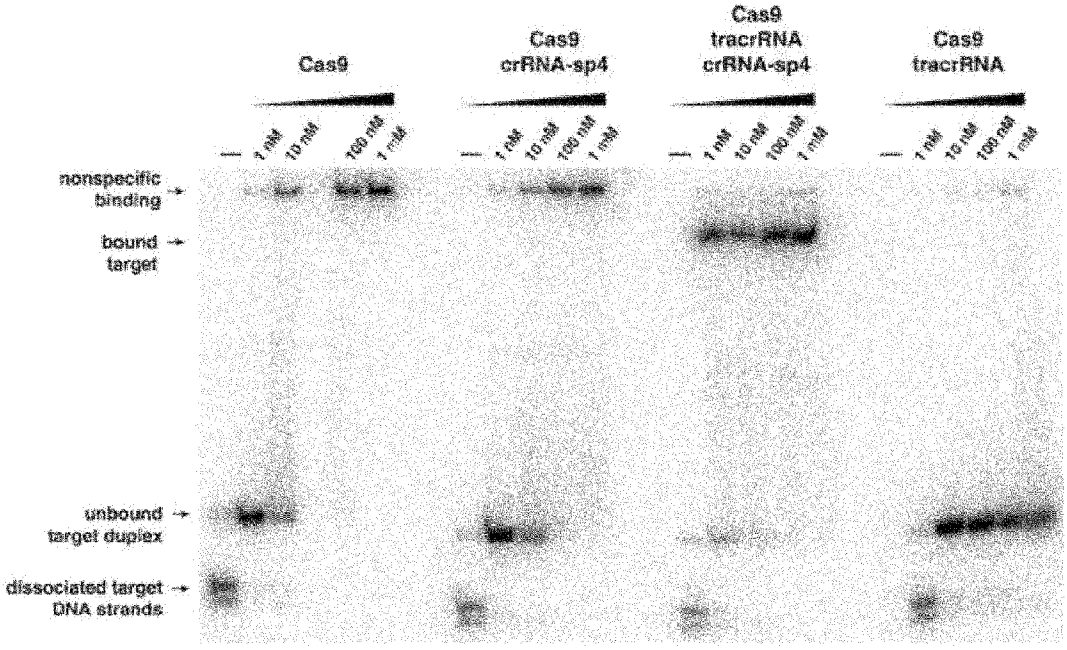

Cas9
Cas9 crRNA-sp4
Cas9 tracrRNA crRNA-sp4
Cas9 tracrRNA
—
1 nM
10 nM
100 nM
1 mM
nonspecific binding
bound target
unbound target duplex
dissociated target DNA strands

FIGURE 23

A

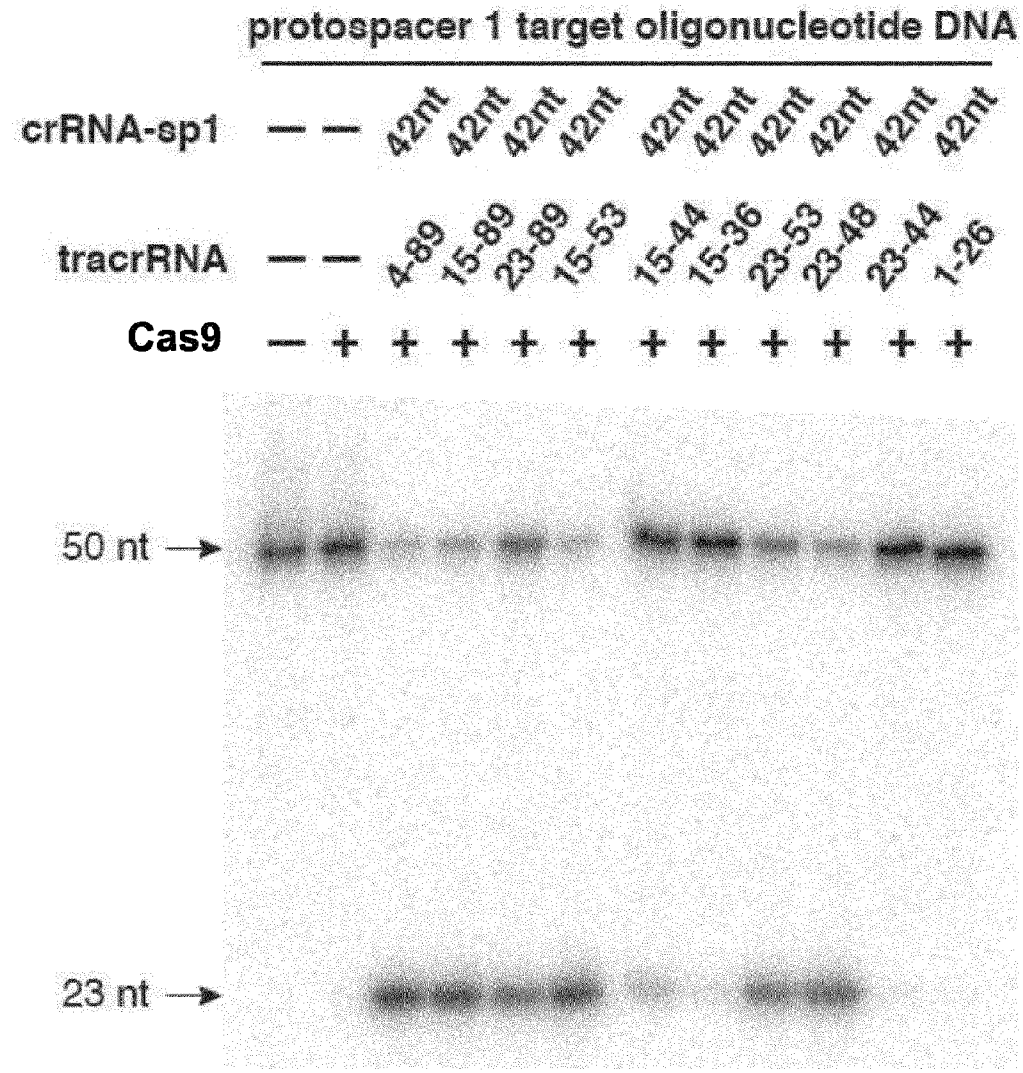

protospacer 1 target oligonucleotide DNA
crRNA-sp1
— — 42nt 42nt 42nt 42nt 42nt 42nt 42nt 42nt 42nt 42nt
tracrRNA
— — 4-89 15-89 23-89 15-53 15-44 15-36 23-53 23-48 23-44 1-26
Cas9
— + + + + + + + + + + +
50 nt
23 nt

FIGURE 23

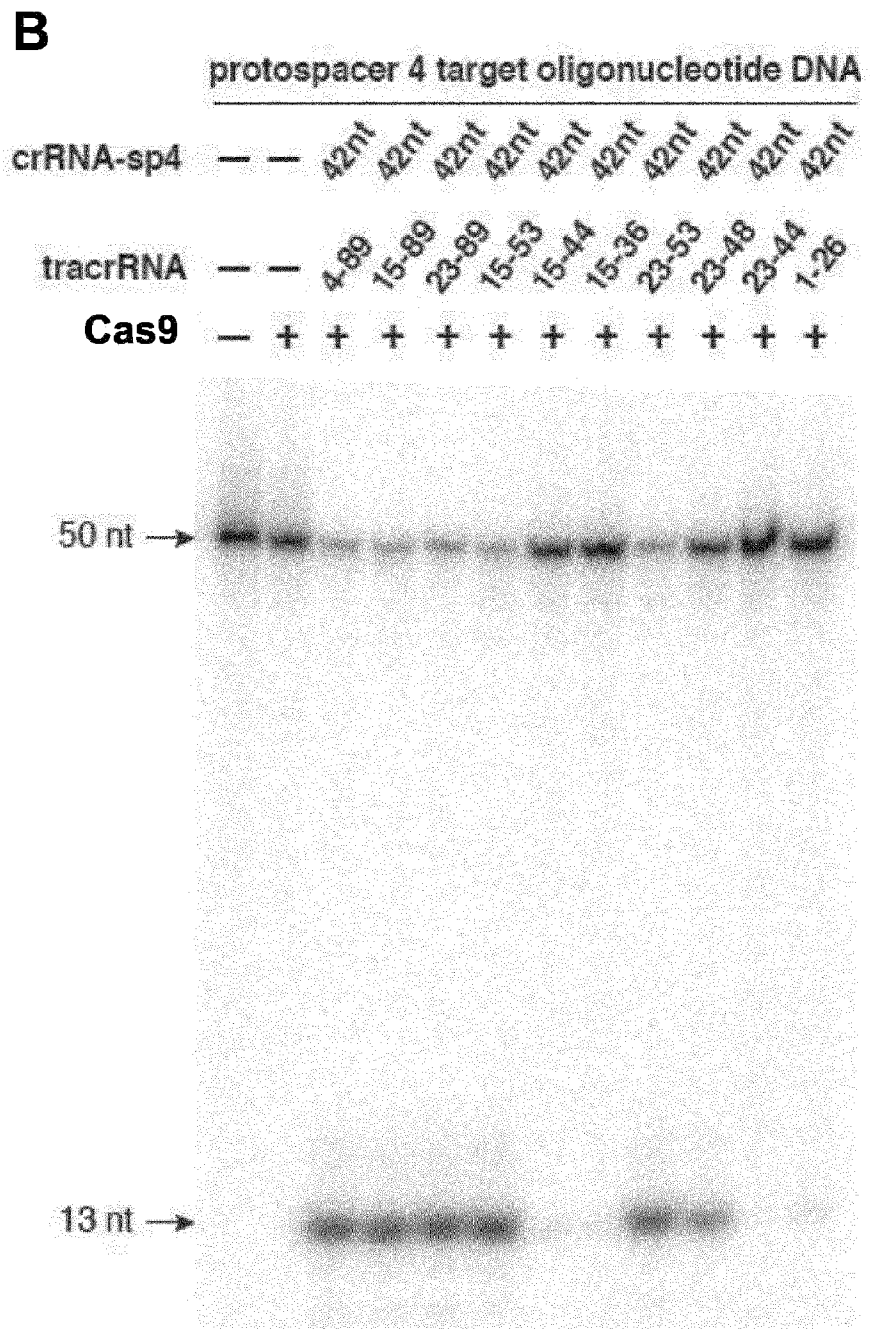
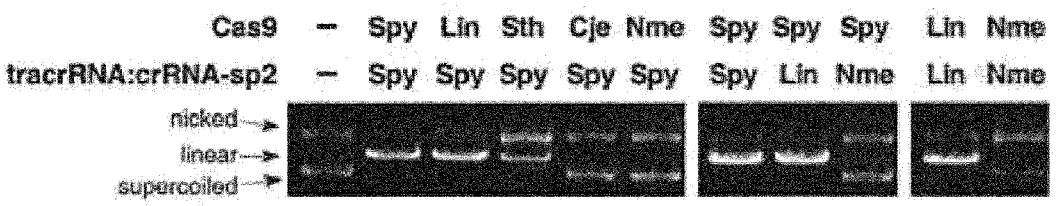

B
protospacer 4 target oligonucleotide DNA
crRNA-sp4
— — 42nt 42nt 42nt 42nt 42nt 42nt 42nt 42nt 42nt 42nt
tracrRNA
— — 4-89 15-89 23-89 15-53 15-44 15-36 23-53 23-48 23-44 1-26
Cas9
— + + + + + + + + + + +
50 nt
13 nt

Figure 24

A


Cas9 – Spy Lin Sth Cje Nme Spy Spy Spy Lin Nme
tracrRNA:crRNA-sp2 – Spy Spy Spy Spy Spy Spy Lin Nme Lin Nme
nicked
linear
supercoiled

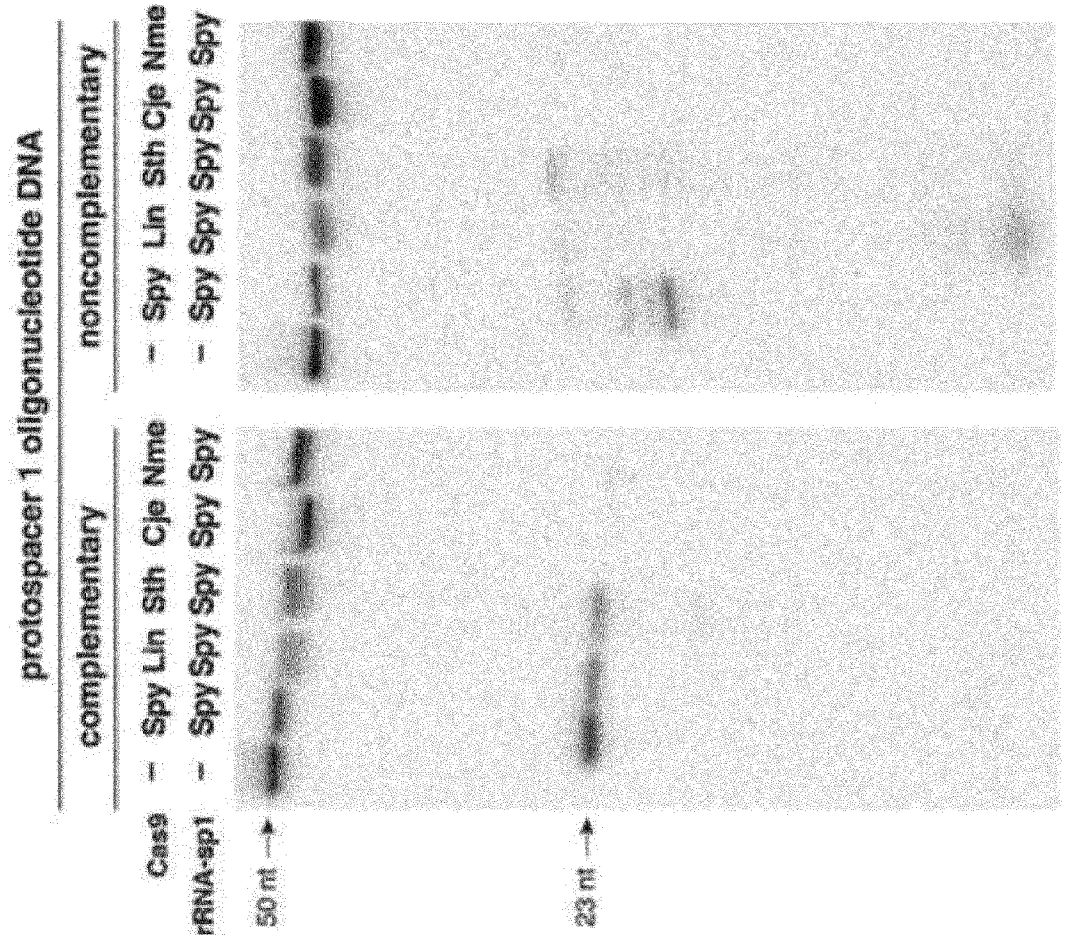

B
protospacer 1 oligonucleotide DNA
complementary
noncomplementary
Cas9 – Spy Lin Sth Cje Nme – Spy Lin Sth Cje Nme
tracrRNA:crRNA-sp1 – Spy Spy Spy Spy Spy – Spy Spy Spy Spy Spy
50 nt
23 nt

Figure 24

Figure 24

C

	S. pyogenes	L. innocua	S. thermophilus	C. jejuni	N. meningitidis
S. pyogenes	x	54	58	16	16
L. innocua	54	x	52	15	14
S. thermophilus	58	52	x	16	15
C. jejuni	16	15	16	x	32
N. meningitidis	16	14	15	32	x

Figure 24

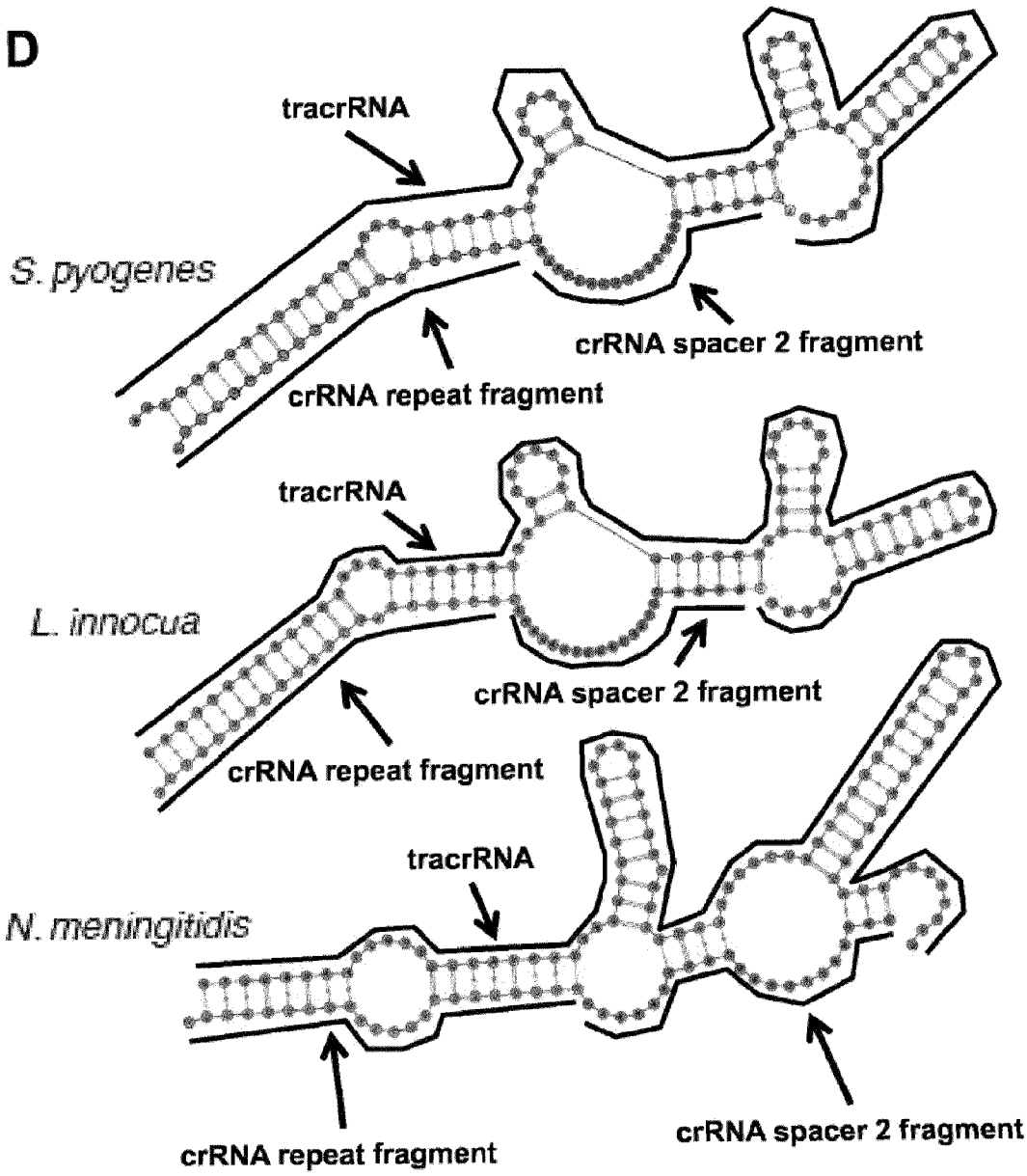

D
tracrRNA
S. pyogenes
crRNA spacer 2 fragment
crRNA repeat fragment
tracrRNA
L. innocua
crRNA spacer 2 fragment
crRNA repeat fragment
tracrRNA
N. meningitidis
crRNA repeat fragment
crRNA spacer 2 fragment

Figure 25

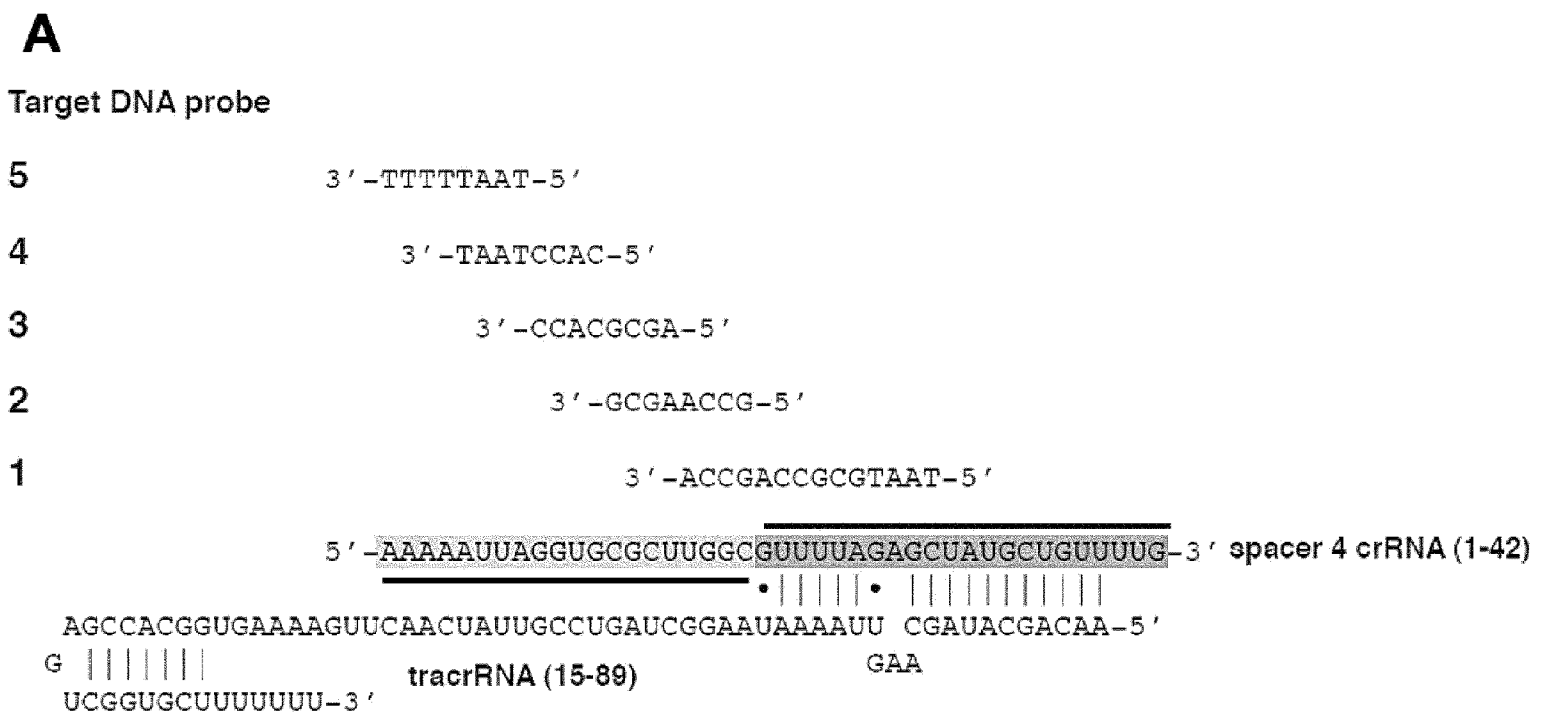


Figure 25

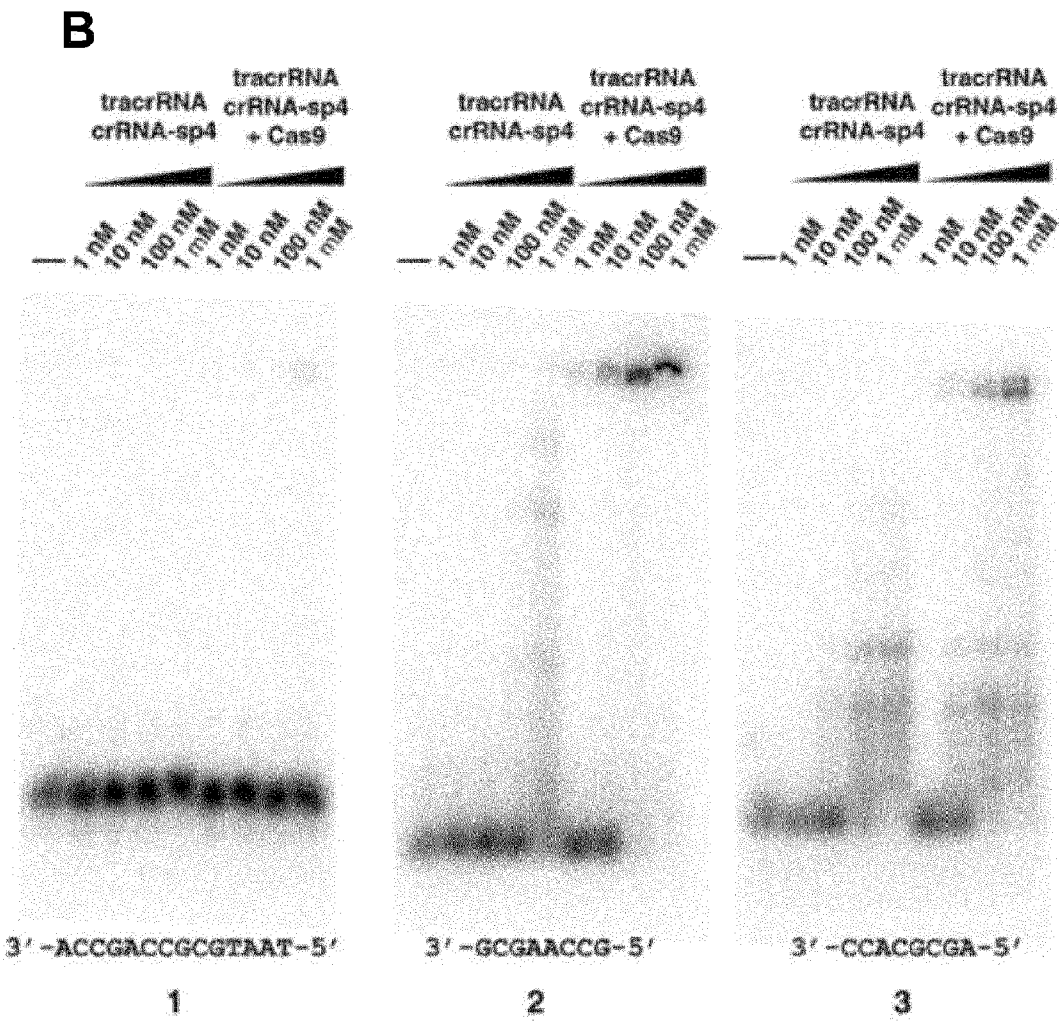
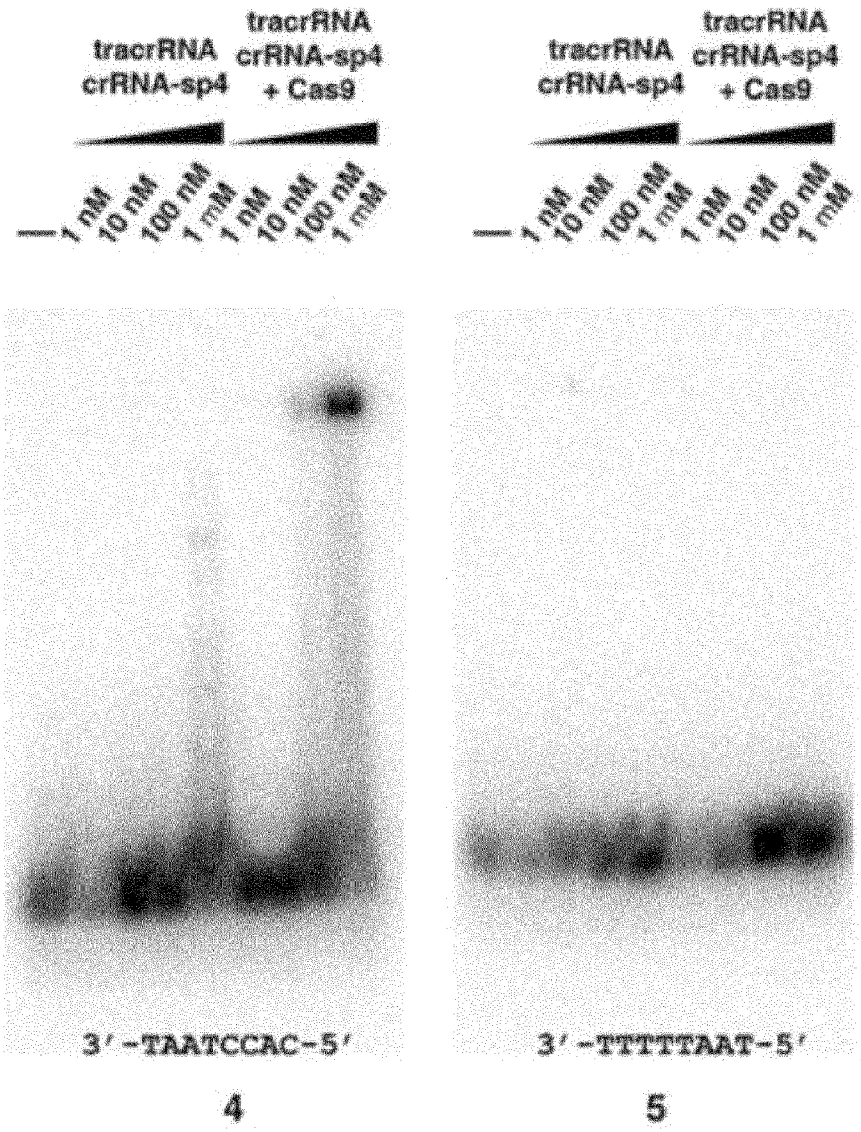

B
tracrRNA crRNA-sp4
tracrRNA crRNA-sp4 + Cas9
— 1 nM 10 nM 100 nM 1 mM 1 nM 10 nM 100 nM 1 mM
3′-ACCGACCGCGTAAT-5′
3′-GCGAACCG-5′
3′-CCACGCGA-5′
1
2
3

Figure 25

C



4 5

FIGURE 26

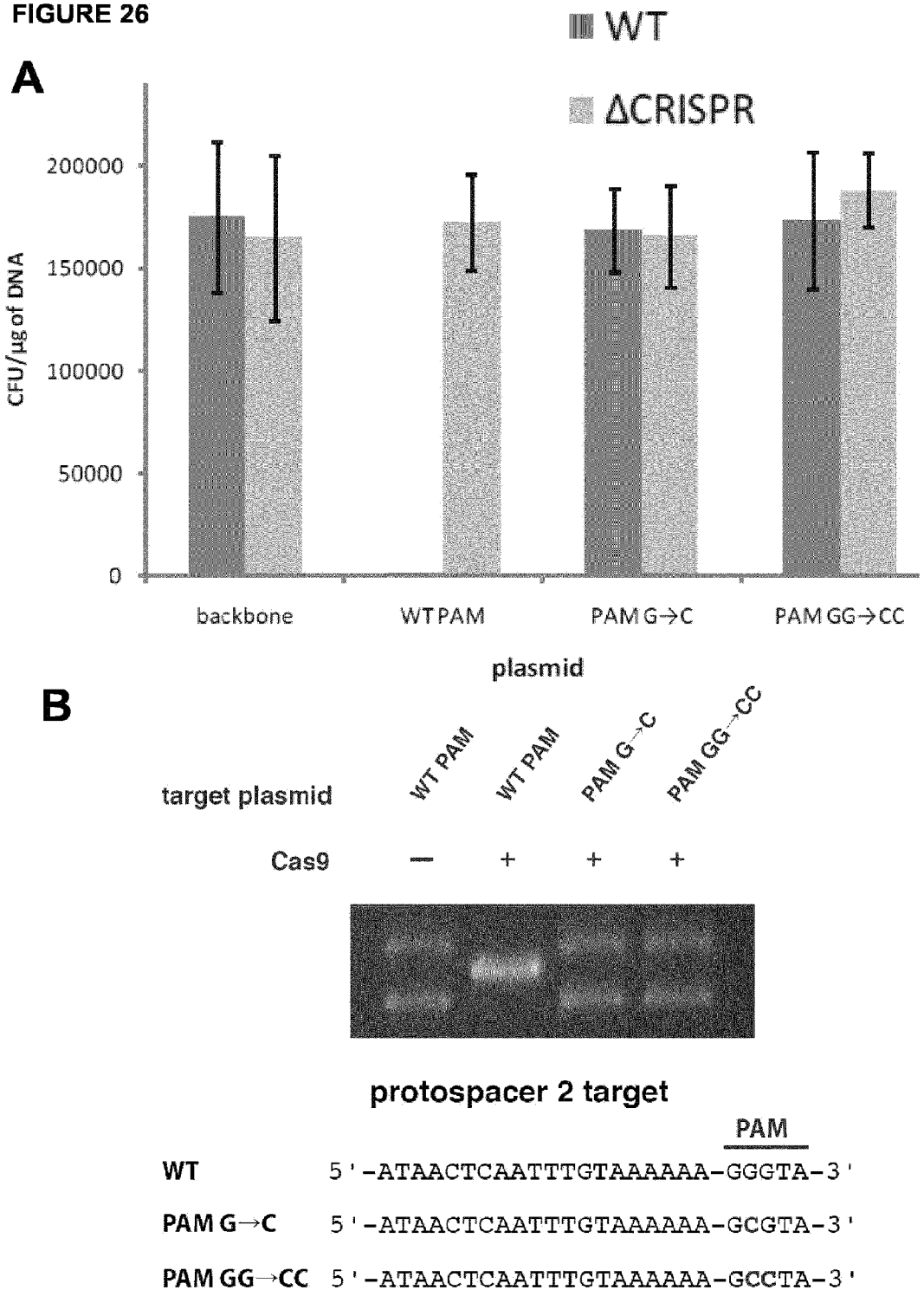
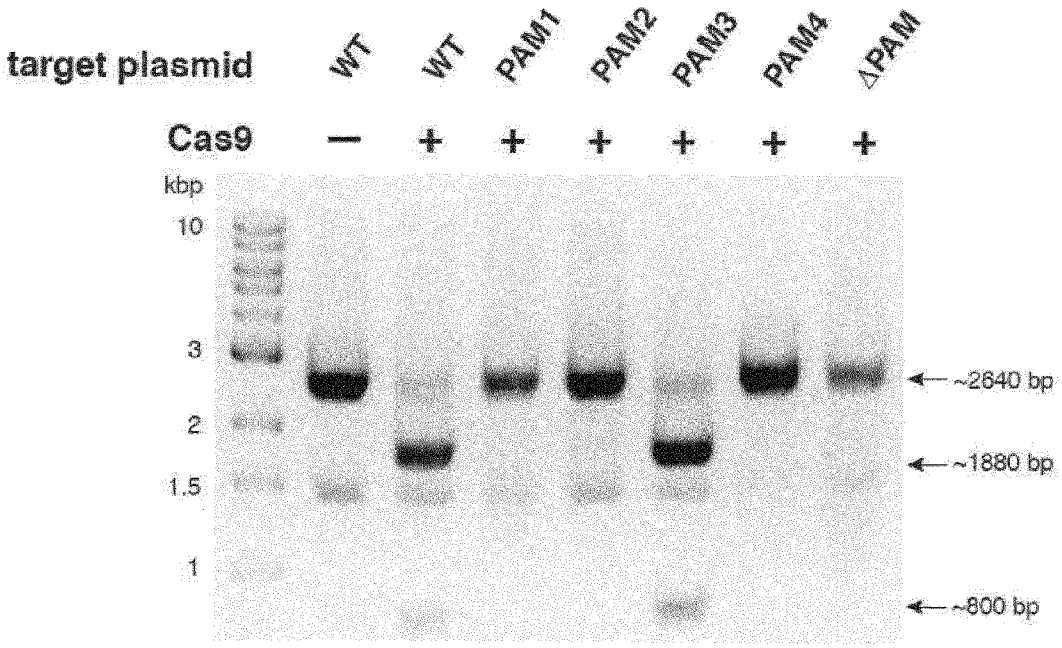

A
WT
ΔCRISPR
CFU/µg of DNA
200000
150000
100000
50000
0
backbone
WT PAM
PAM G→C
PAM GG→CC
plasmid
B
target plasmid
WT PAM
WT PAM
PAM G→C
PAM GG→CC
Cas9
—
+
+
+
protospacer 2 target
PAM
WT 5'-ATAACTCAATTTGTAAAAAA-GGGTA-3'
PAM G→C 5'-ATAACTCAATTTGTAAAAAA-GCGTA-3'
PAM GG→CC 5'-ATAACTCAATTTGTAAAAAA-GCCTA-3'

FIGURE 26

C



protospacer 4 target

PAM

WT	5'-AAAAATTAGGTGCGCTTGGC-TGGCGC-3'
PAM1	5'-AAAAATTAGGTGCGCTTGGC-TCGCGC-3'
PAM2	5'-AAAAATTAGGTGCGCTTGGC-TGCCGC-3'
PAM3	5'-AAAAATTAGGTGCGCTTGGC-TGGCCC-3'
PAM4	5'-AAAAATTAGGTGCGCTTGGC-TCCCCC-3'
DPAM	5'-AAAAATTAGGTGCGCTTGGC-T-3'

Figure 27

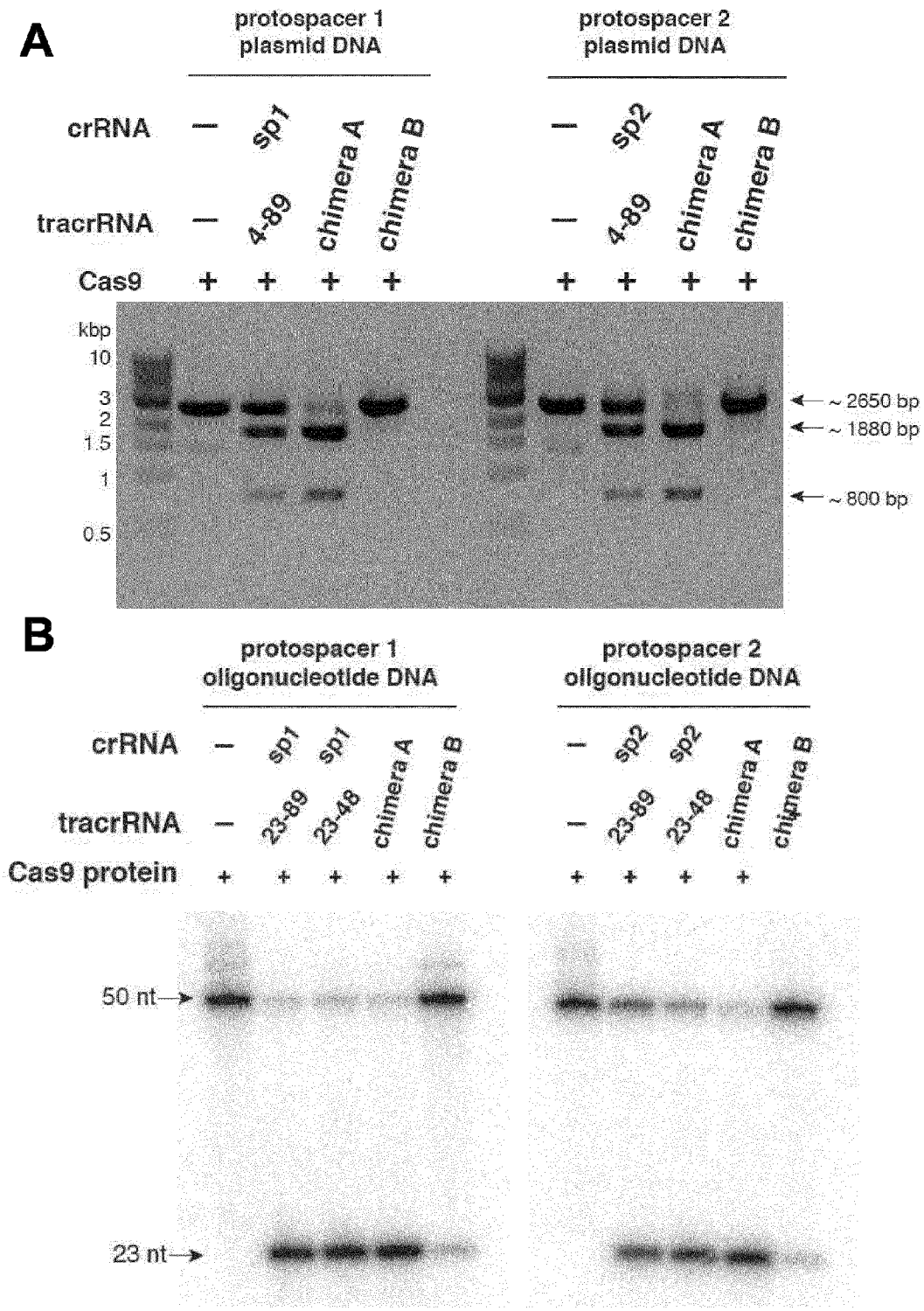

A
protospacer 1 plasmid DNA
protospacer 2 plasmid DNA
crRNA
tracrRNA
Cas9
–
sp1
sp2
4-89
chimera A
chimera B
+
kbp
10
3
2
1.5
1
0.5
~ 2650 bp
~ 1880 bp
~ 800 bp
B
protospacer 1 oligonucleotide DNA
protospacer 2 oligonucleotide DNA
crRNA
tracrRNA
Cas9 protein
23-89
23-48
50 nt
23 nt

Figure 27

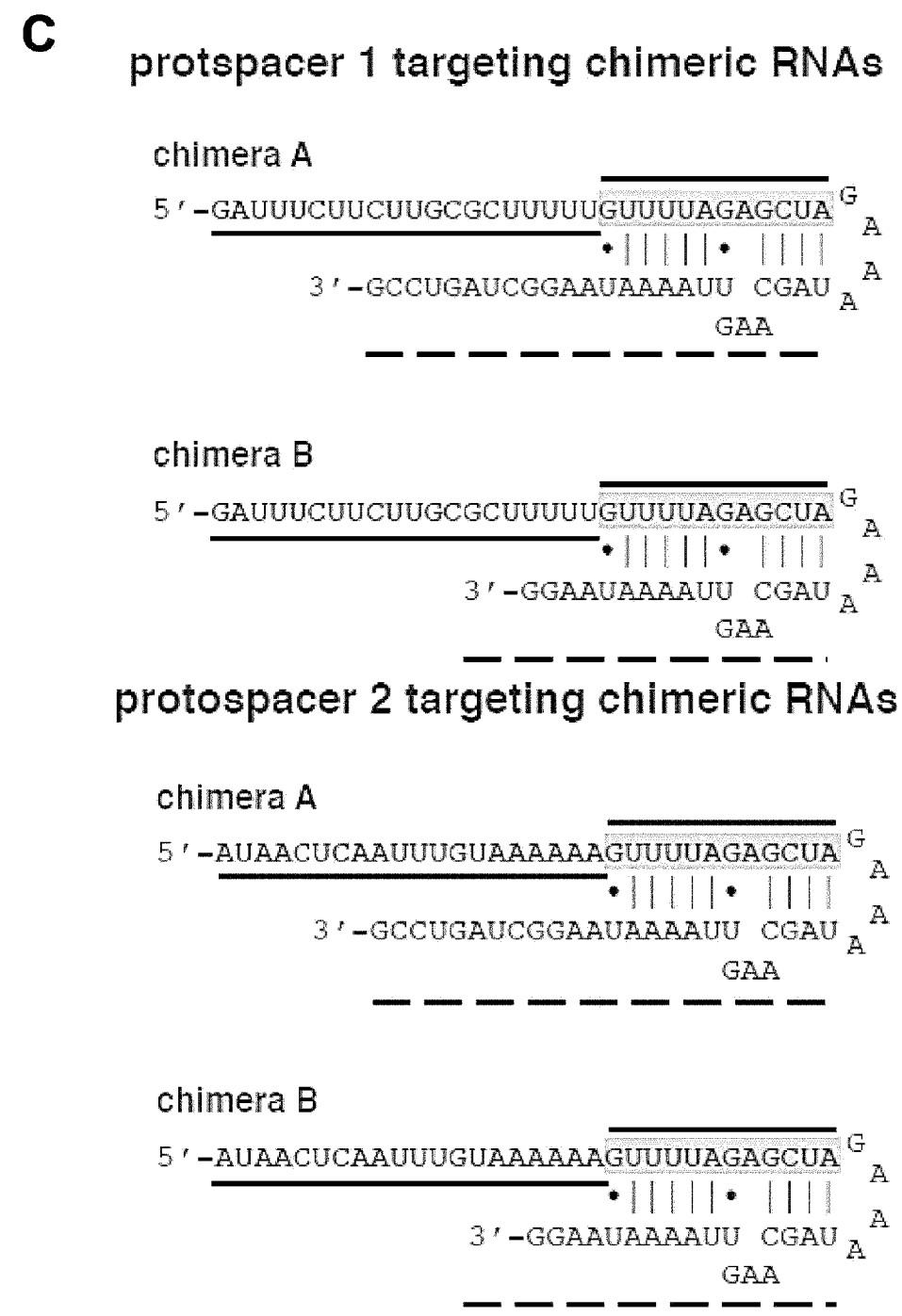

C
protspacer 1 targeting chimeric RNAs
chimera A
5′-GAUUUCUUCUUGCGCUUUUUGUUUUAGAGCUA G A A
3′-GCCUGAUCGGAAUAAAAUU CGAU A
GAA
chimera B
5′-GAUUUCUUCUUGCGCUUUUUGUUUUAGAGCUA G A A
3′-GGAAUAAAAUU CGAU A
GAA
protospacer 2 targeting chimeric RNAs
chimera A
5′-AUAACUCAAUUUGUAAAAAAGUUUUAGAGCUA G A A
3′-GCCUGAUCGGAAUAAAAUU CGAU A
GAA
chimera B
5′-AUAACUCAAUUUGUAAAAAAGUUUUAGAGCUA G A A
3′-GGAAUAAAAUU CGAU A
GAA

FIGURE 28

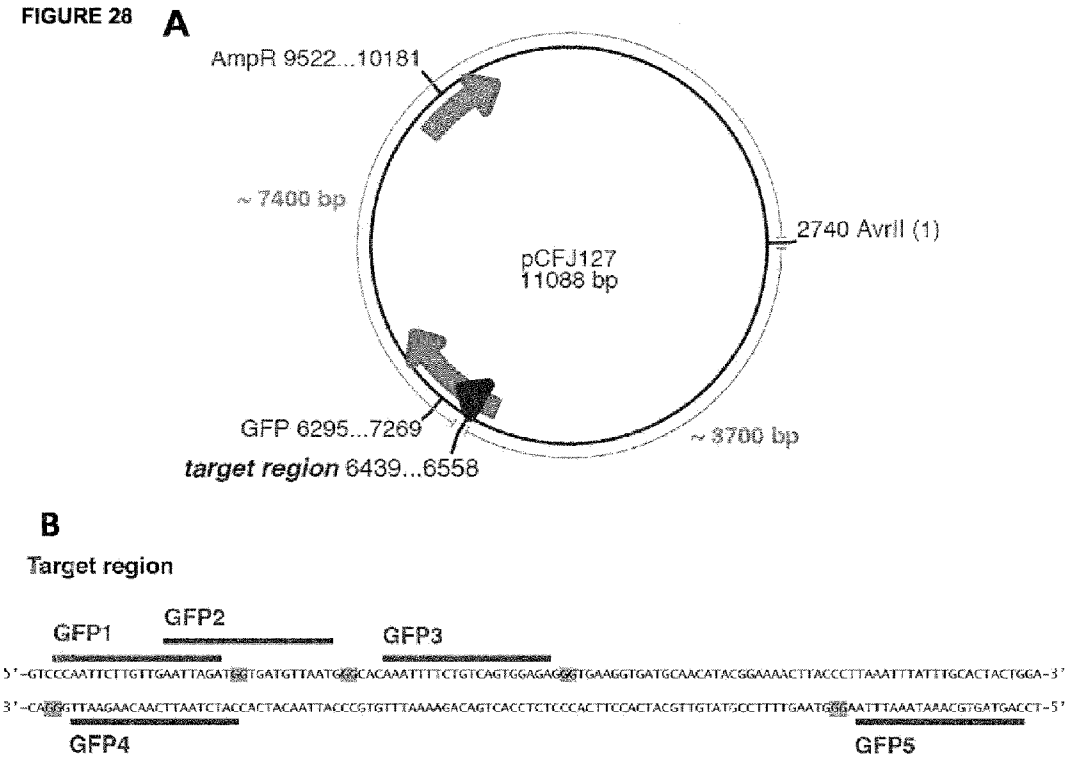

A
AmpR 9522...10181
~ 7400 bp
2740 AvrII (1)
pCFJ127
11088 bp
GFP 6295...7269
target region 6439...6558
~ 3700 bp
B
Target region
GFP1
GFP2
GFP3
5'-GTCCCAATTCTTGTTGAATTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACCCTTAAATTTATTTGCACTACTGGA-3'
3'-CAGGGTTAAGAACAACTTAATCTACCACTACAATTACCCGTGTTTAAAAGACAGTCACCTCTCCCACTTCCACTACGTTGTATGCCTTTTGAATGGGAATTTAAATAAACGTGATGACCT-5'
GFP4
GFP5

Figure 28

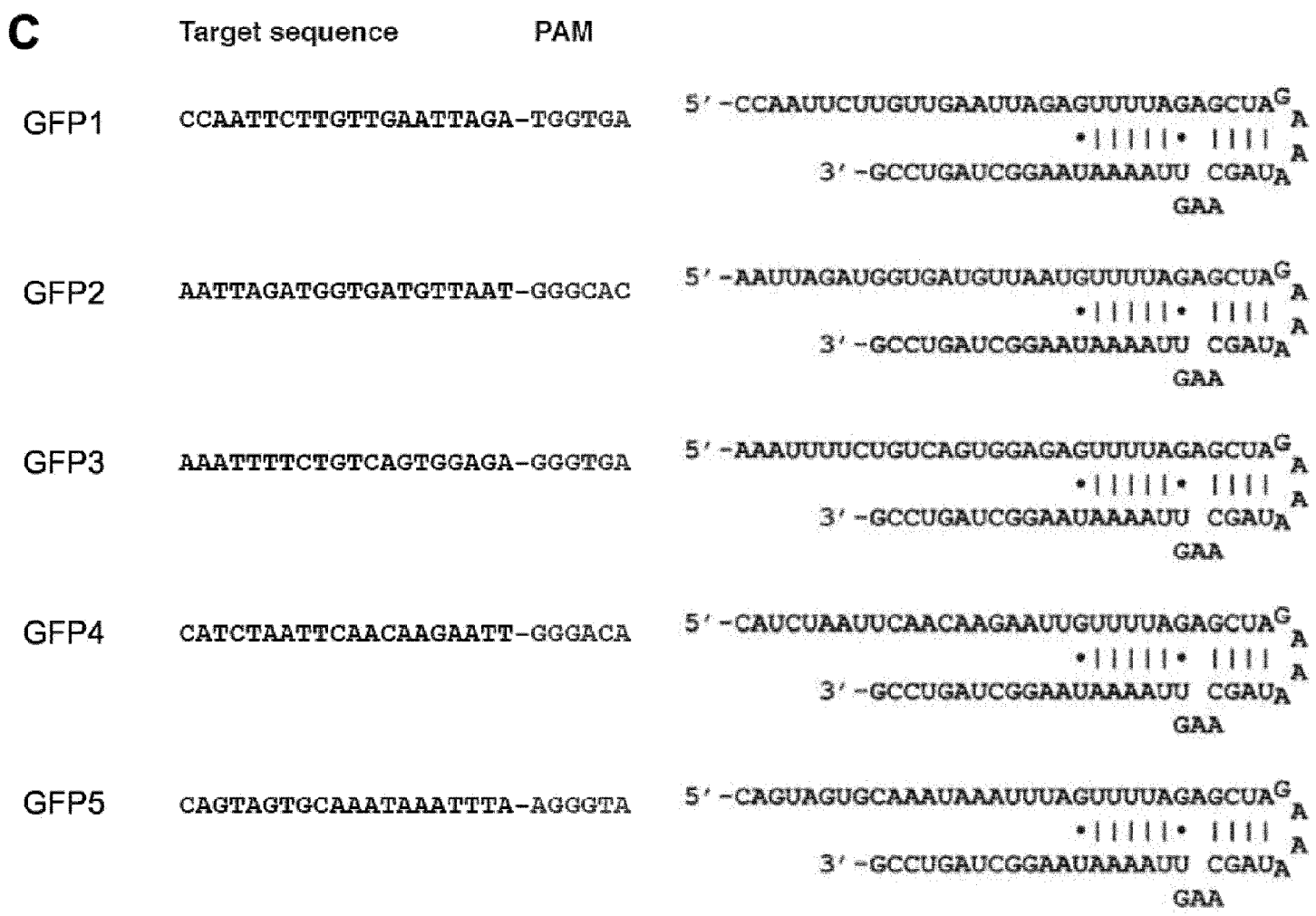


Figure 28

D

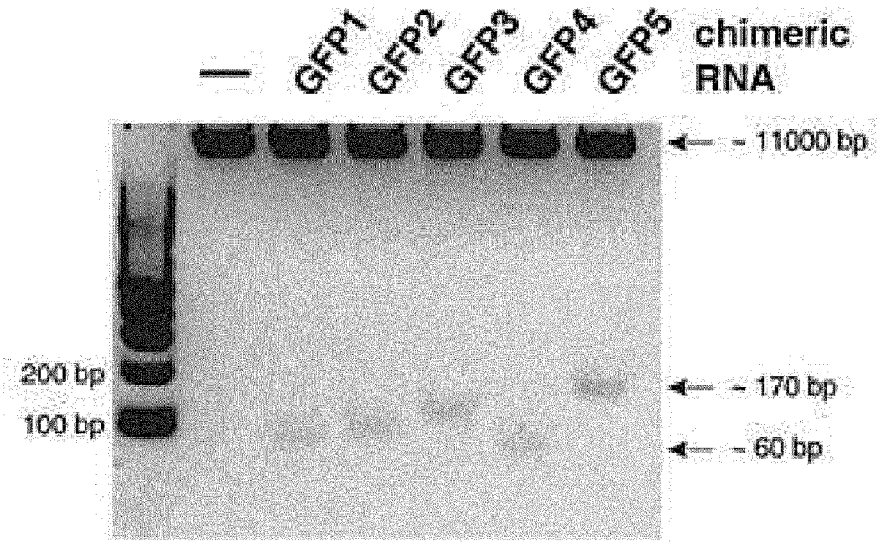

—
GFP1
GFP2
GFP3
GFP4
GFP5
chimeric RNA
~ 11000 bp
200 bp
100 bp
~ 170 bp
~ 60 bp

Figure 29

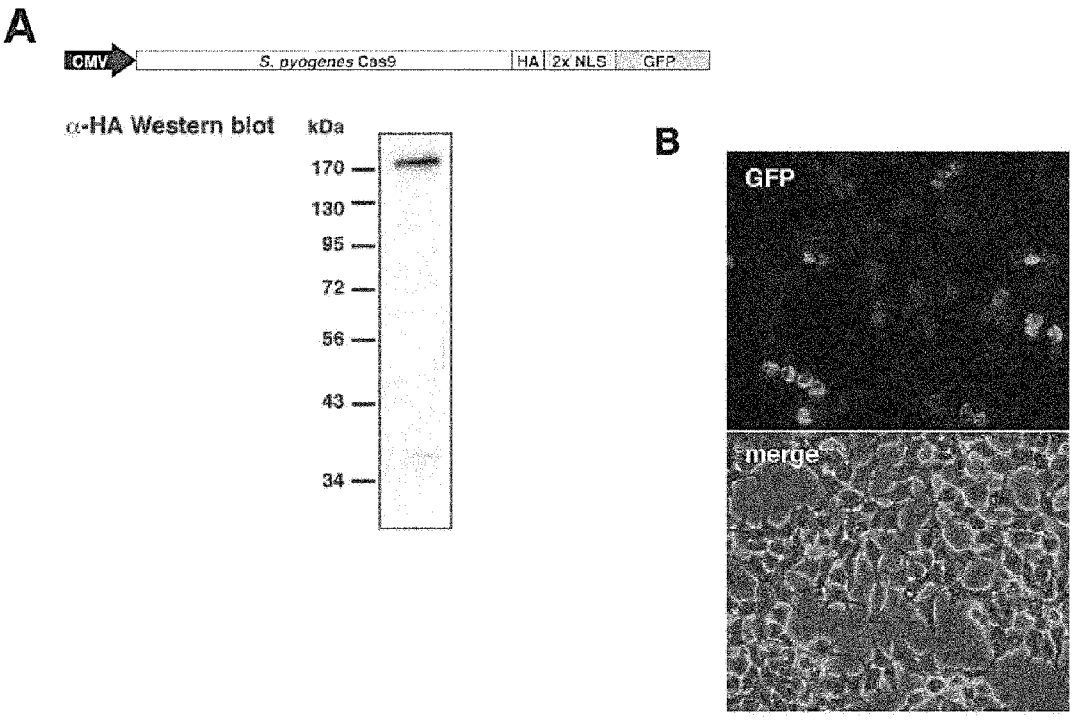

A
CMV
S. pyogenes Cas9
HA
2x NLS
GFP
α-HA Western blot
kDa
170
130
95
72
56
43
34
B
GFP
merge

Figure 29

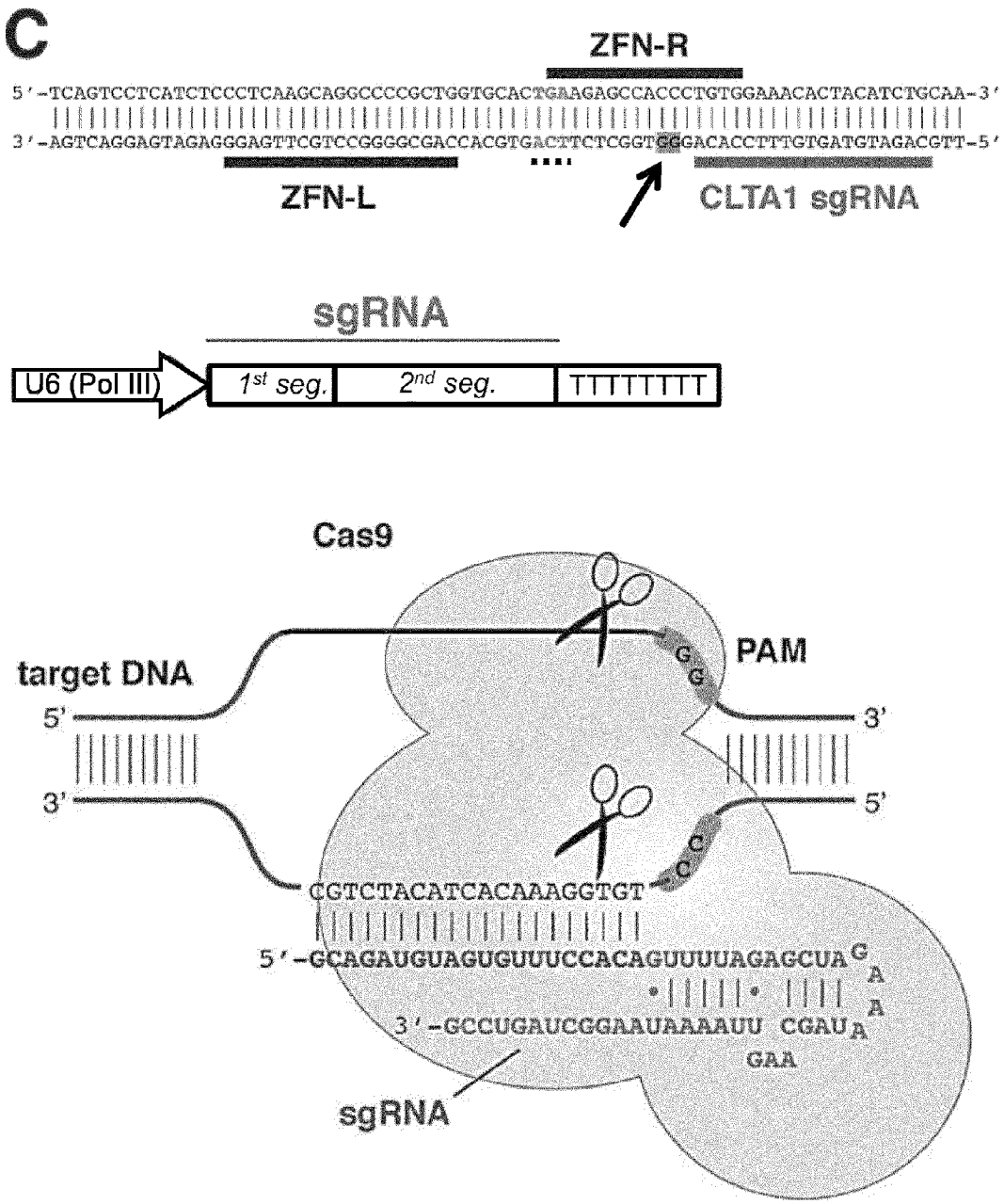

C
ZFN-R
5′-TCAGTCCTCATCTCCCTCAAGCAGGCCCCGCTGGTGCACTGAAGAGCCACCCTGTGGAAACACTACATCTGCAA-3′
3′-AGTCAGGAGTAGAGGGAGTTCGTCCGGGGCGACCACGTGACTTCTCGGTGGGACACCTTTGTGATGTAGACGTT-5′
ZFN-L
CLTA1 sgRNA
sgRNA
U6 (Pol III)
1st seg.
2nd seg.
TTTTTTTT
Cas9
target DNA
5'
3'
PAM
CGTCTACATCACAAAGGTGT
5′-GCAGAUGUAGUGUUUCCACAGUUUUAGAGCUA
3′-GCCUGAUCGGAAUAAAAUU CGAU
GAA
sgRNA

Figure 29

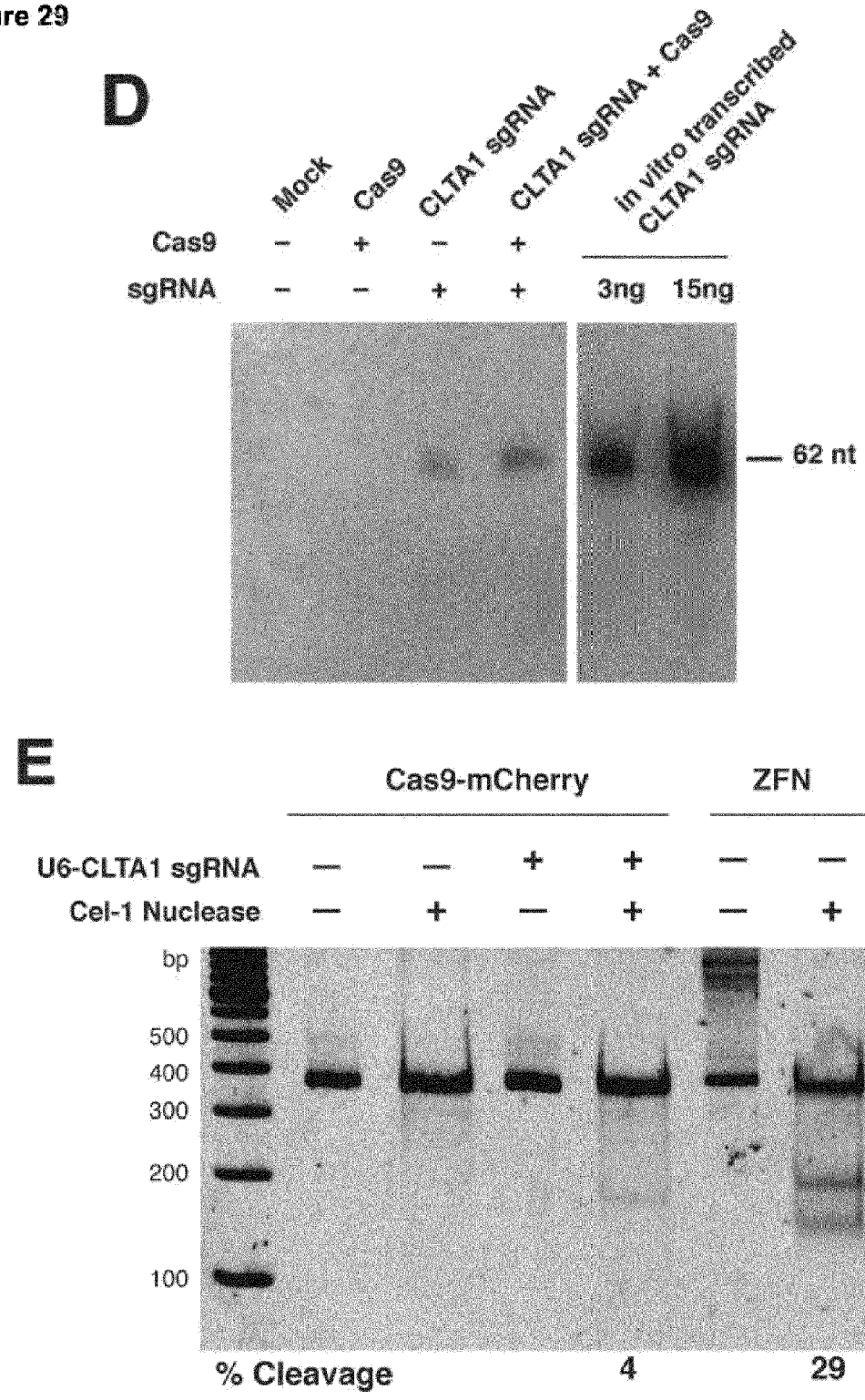

D
Mock
Cas9
CLTA1 sgRNA
CLTA1 sgRNA + Cas9
in vitro transcribed CLTA1 sgRNA
Cas9 – + – +
sgRNA – – + +
3ng 15ng
62 nt
E
Cas9-mCherry
ZFN
U6-CLTA1 sgRNA — — + + — —
Cel-1 Nuclease — + — + — +
bp
500
400
300
200
100
% Cleavage
4
29

FIGURE 30

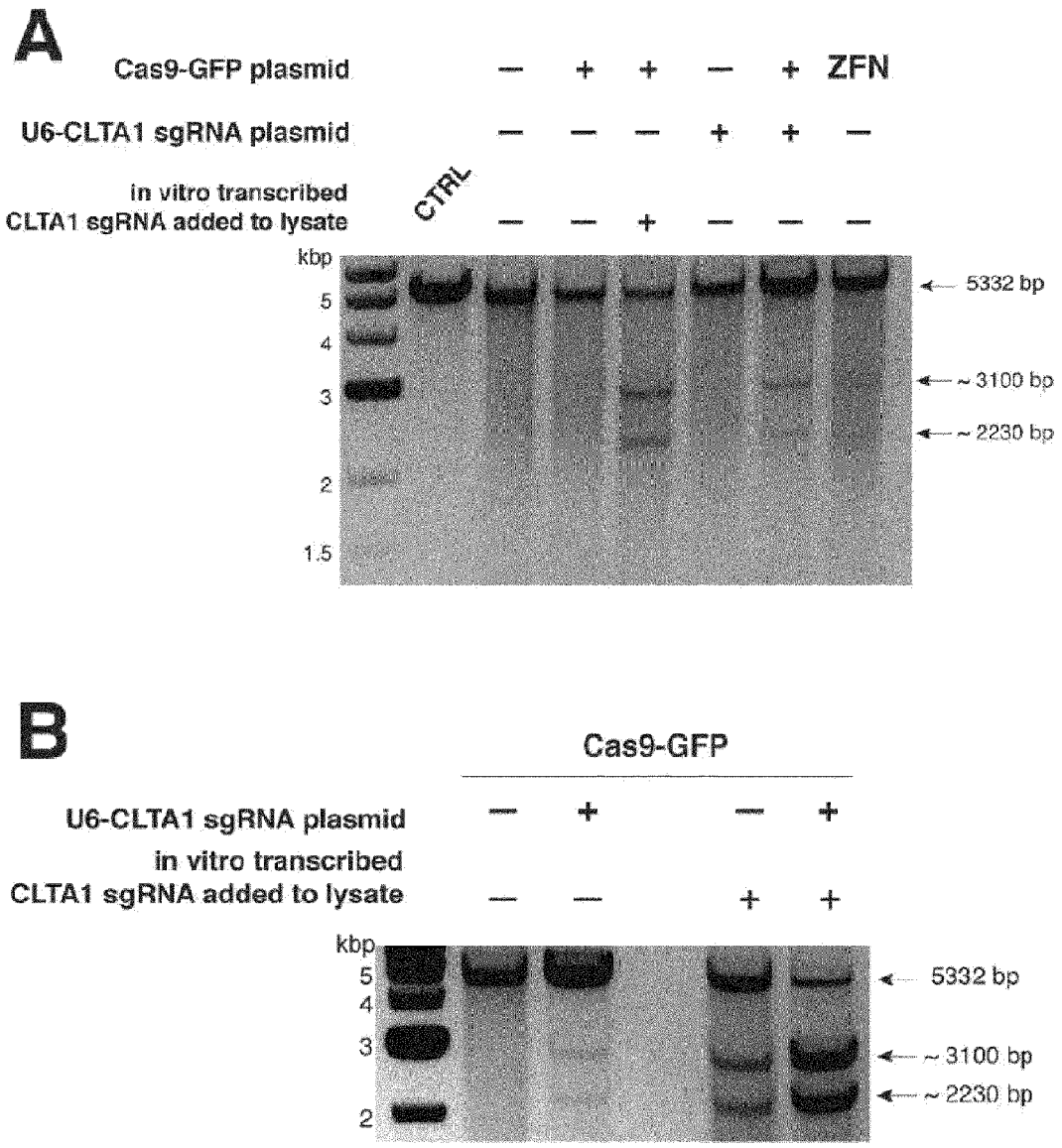

A
Cas9-GFP plasmid
U6-CLTA1 sgRNA plasmid
in vitro transcribed CLTA1 sgRNA added to lysate
CTRL
ZFN
kbp
5
4
3
2
1.5
5332 bp
~ 3100 bp
~ 2230 bp
B
Cas9-GFP
U6-CLTA1 sgRNA plasmid
in vitro transcribed CLTA1 sgRNA added to lysate
kbp
5
4
3
2
5332 bp
~ 3100 bp
~ 2230 bp

FIGURE 31

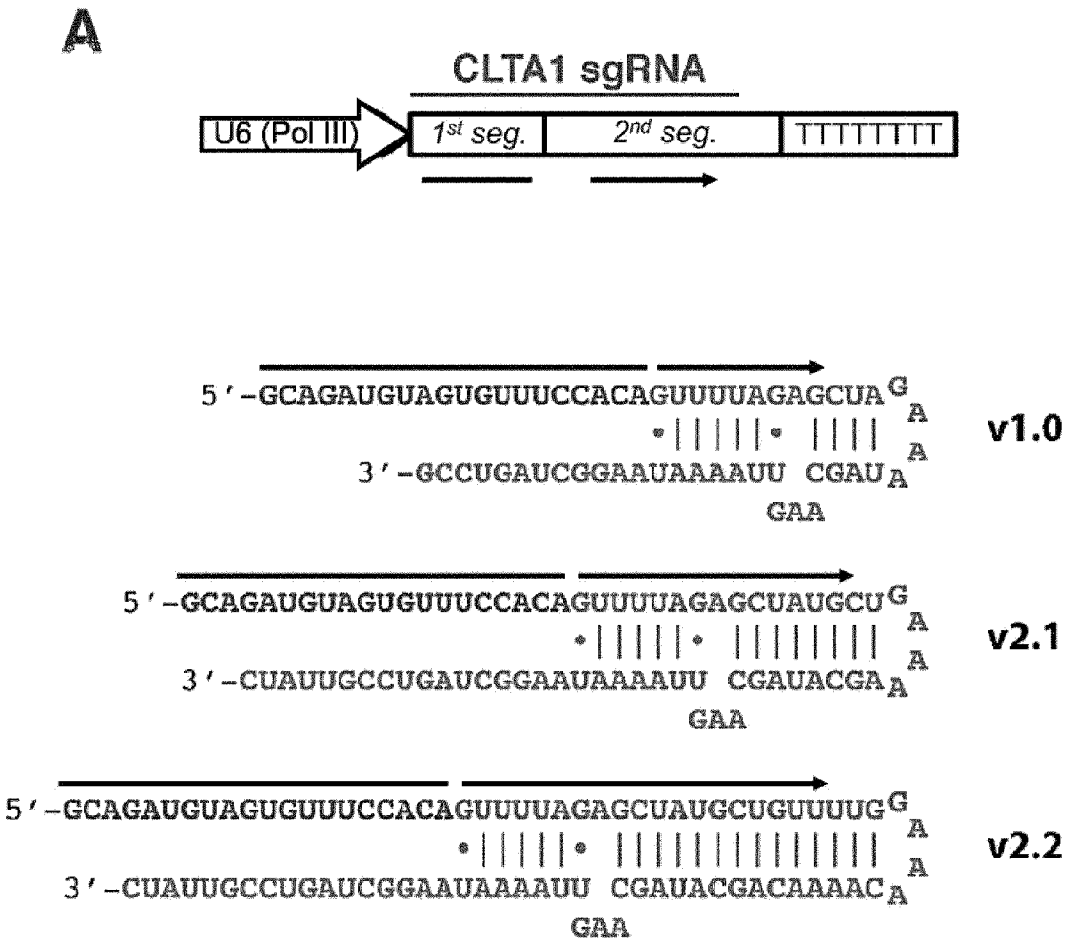

A
CLTA1 sgRNA
U6 (Pol III)
1st seg.
2nd seg.
TTTTTTTT
5′-GCAGAUGUAGUGUUUCCACAGUUUUAGAGCUA G A
3′-GCCUGAUCGGAAUAAAAUU CGAU A A
GAA
v1.0
5′-GCAGAUGUAGUGUUUCCACAGUUUUAGAGCUAUGCU G A
3′-CUAUUGCCUGAUCGGAAUAAAAUU CGAUACGA A A
GAA
v2.1
5′-GCAGAUGUAGUGUUUCCACAGUUUUAGAGCUAUGCUGUUUUG G A
3′-CUAUUGCCUGAUCGGAAUAAAAUU CGAUACGACAAAAC A A
GAA
v2.2

FIGURE 31

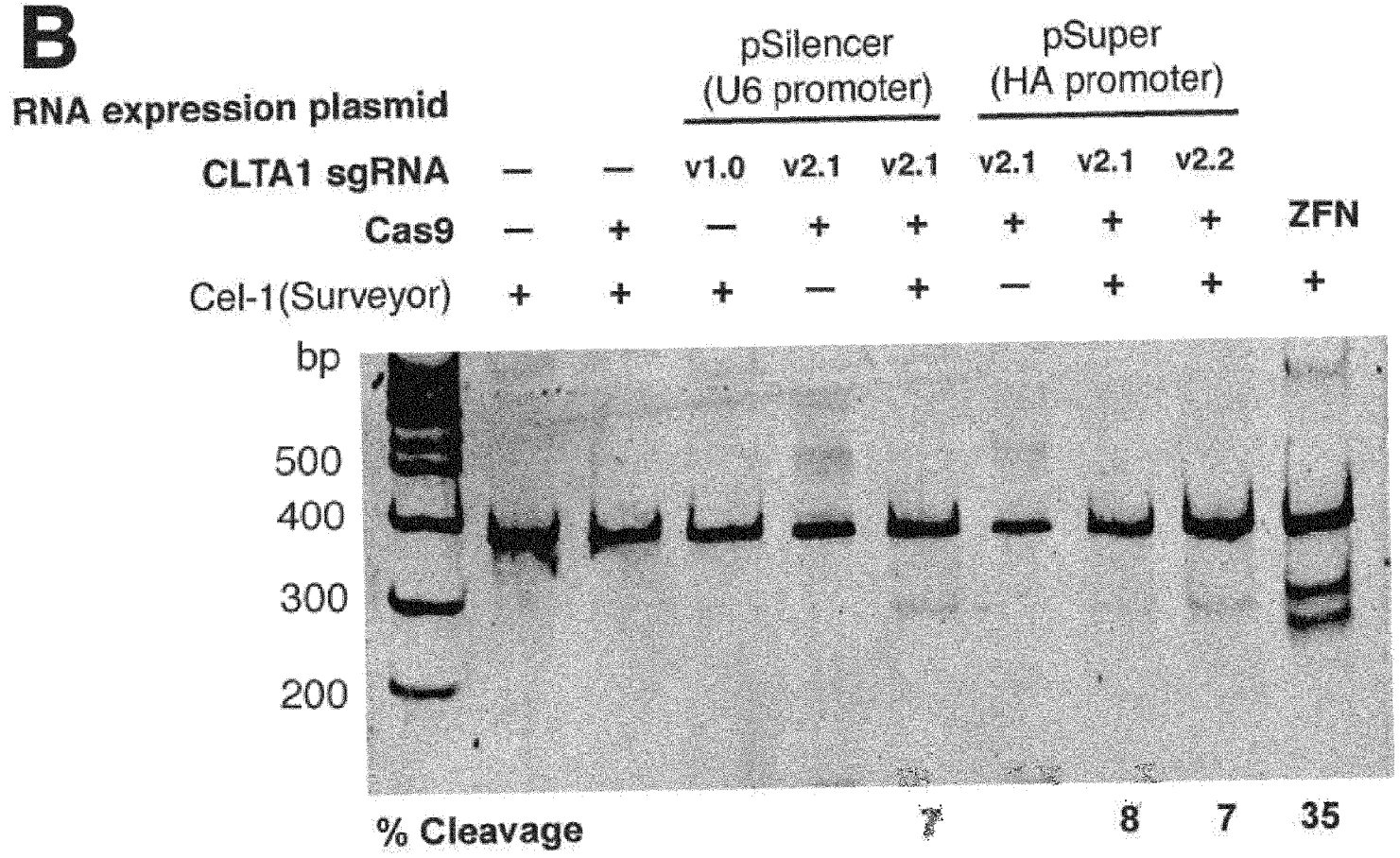


FIGURE 32 A

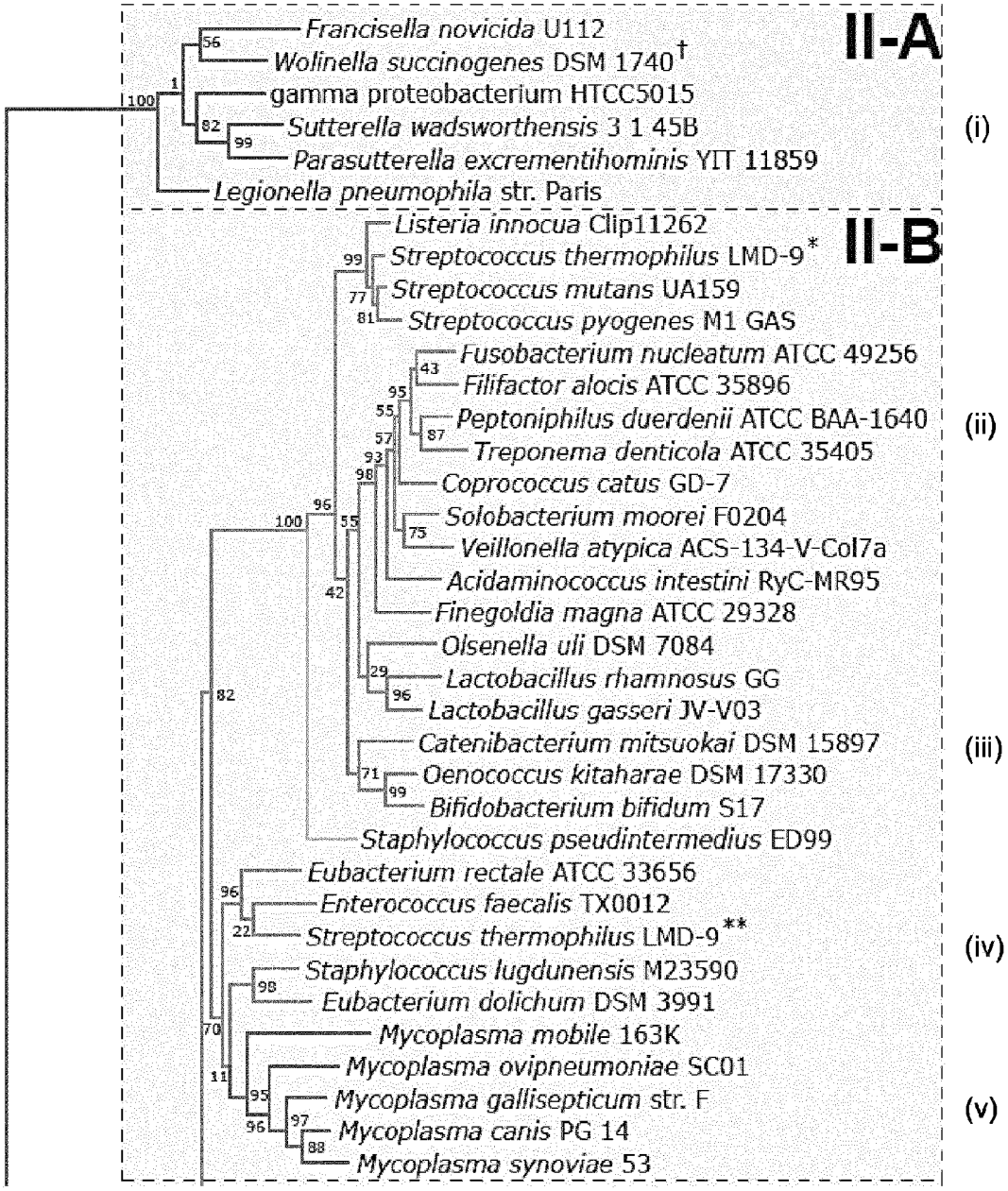

II-A
II-B
Francisella novicida U112
Wolinella succinogenes DSM 1740†
gamma proteobacterium HTCC5015
Sutterella wadsworthensis 3 1 45B
Parasutterella excrementihominis YIT 11859
Legionella pneumophila str. Paris
Listeria innocua Clip11262
Streptococcus thermophilus LMD-9*
Streptococcus mutans UA159
Streptococcus pyogenes M1 GAS
Fusobacterium nucleatum ATCC 49256
Filifactor alocis ATCC 35896
Peptoniphilus duerdenii ATCC BAA-1640
Treponema denticola ATCC 35405
Coprococcus catus GD-7
Solobacterium moorei F0204
Veillonella atypica ACS-134-V-Col7a
Acidaminococcus intestini RyC-MR95
Finegoldia magna ATCC 29328
Olsenella uli DSM 7084
Lactobacillus rhamnosus GG
Lactobacillus gasseri JV-V03
Catenibacterium mitsuokai DSM 15897
Oenococcus kitaharae DSM 17330
Bifidobacterium bifidum S17
Staphylococcus pseudintermedius ED99
Eubacterium rectale ATCC 33656
Enterococcus faecalis TX0012
Streptococcus thermophilus LMD-9**
Staphylococcus lugdunensis M23590
Eubacterium dolichum DSM 3991
Mycoplasma mobile 163K
Mycoplasma ovipneumoniae SC01
Mycoplasma gallisepticum str. F
Mycoplasma canis PG 14
Mycoplasma synoviae 53
(i)
(ii)
(iii)
(iv)
(v)

FIGURE 32 A (cont.)

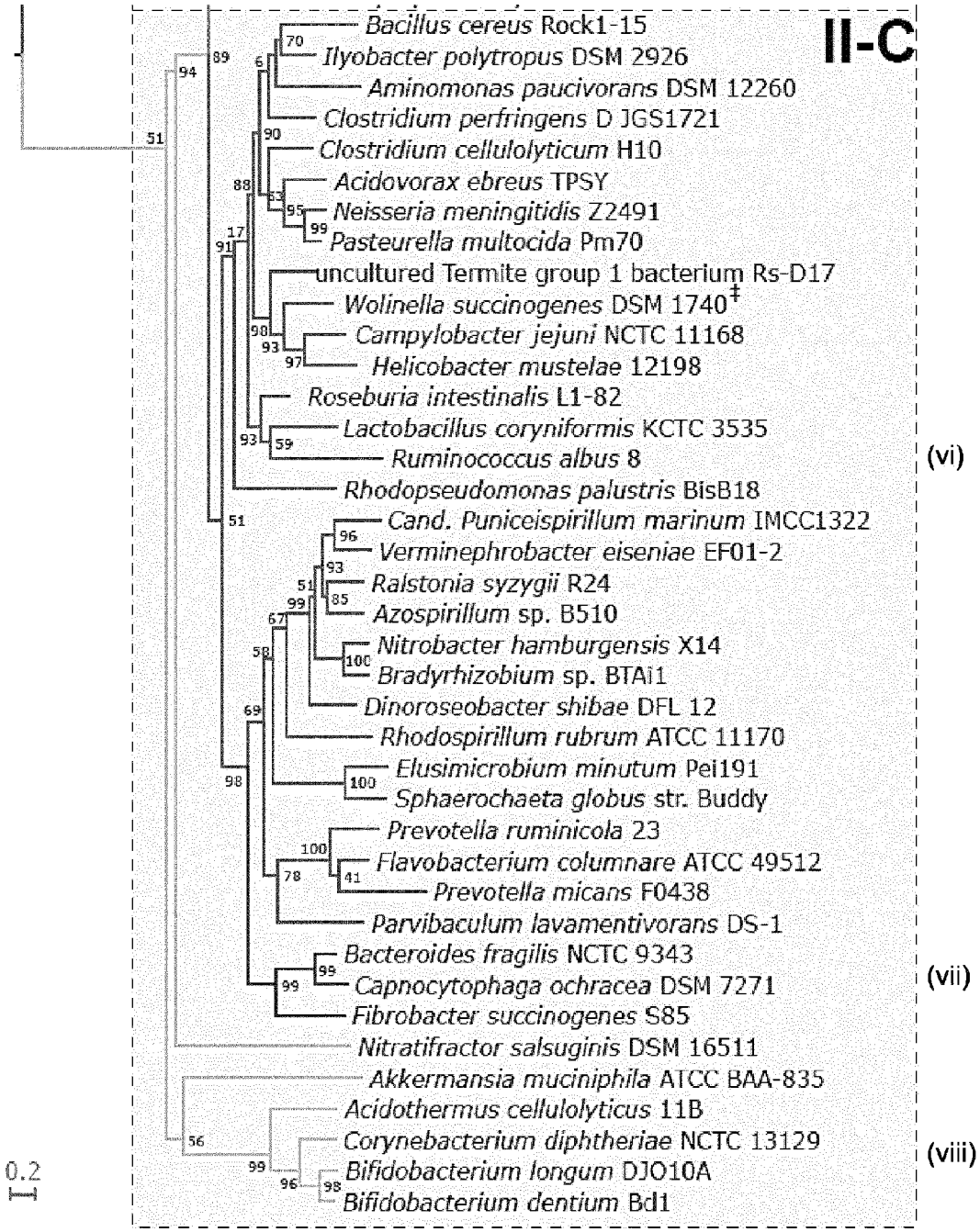


Figure 32 B

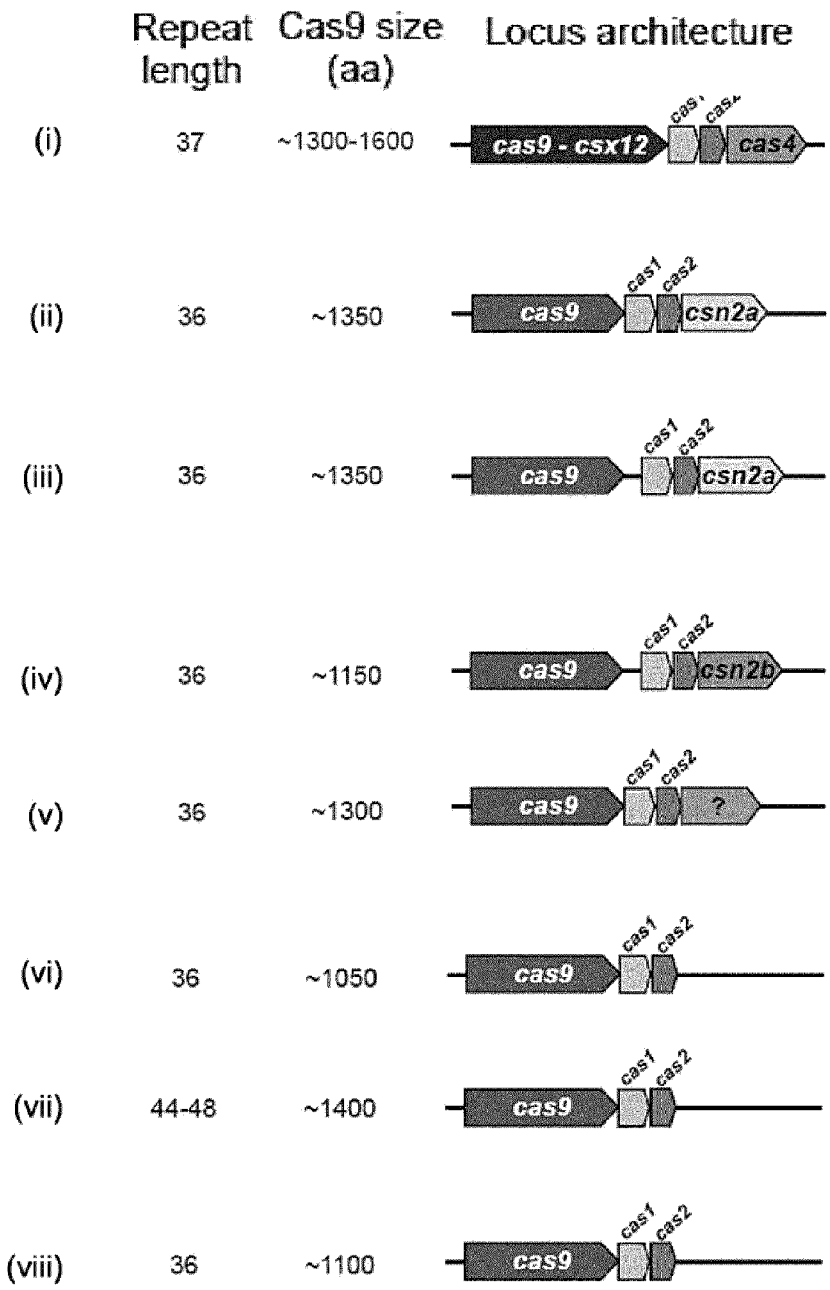
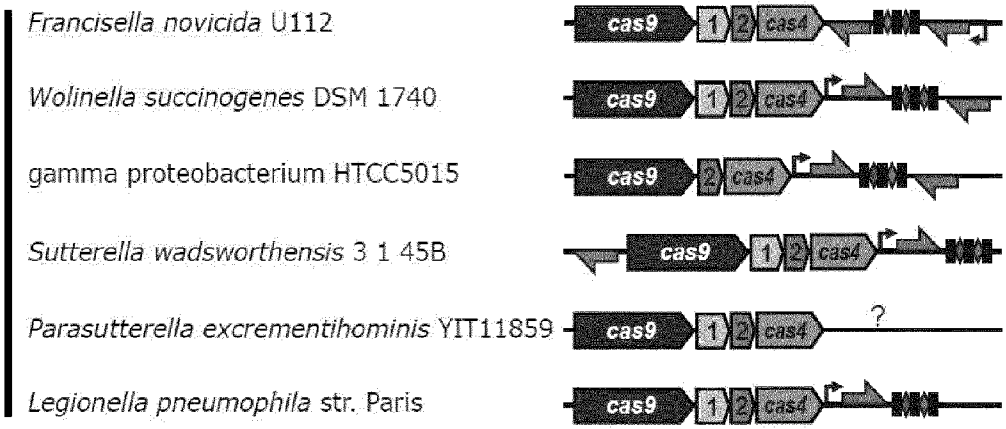

Repeat length
Cas9 size (aa)
Locus architecture
(i)
37
~1300-1600
cas9 - csx12
cas4
(ii)
36
~1350
cas9
cas1
cas2
csn2a
(iii)
36
~1350
cas9
cas1
cas2
csn2a
(iv)
36
~1150
cas9
cas1
cas2
csn2b
(v)
36
~1300
cas9
cas1
cas2
?
(vi)
36
~1050
cas9
cas1
cas2
(vii)
44-48
~1400
cas9
cas1
cas2
(viii)
36
~1100
cas9
cas1
cas2

Figure 33

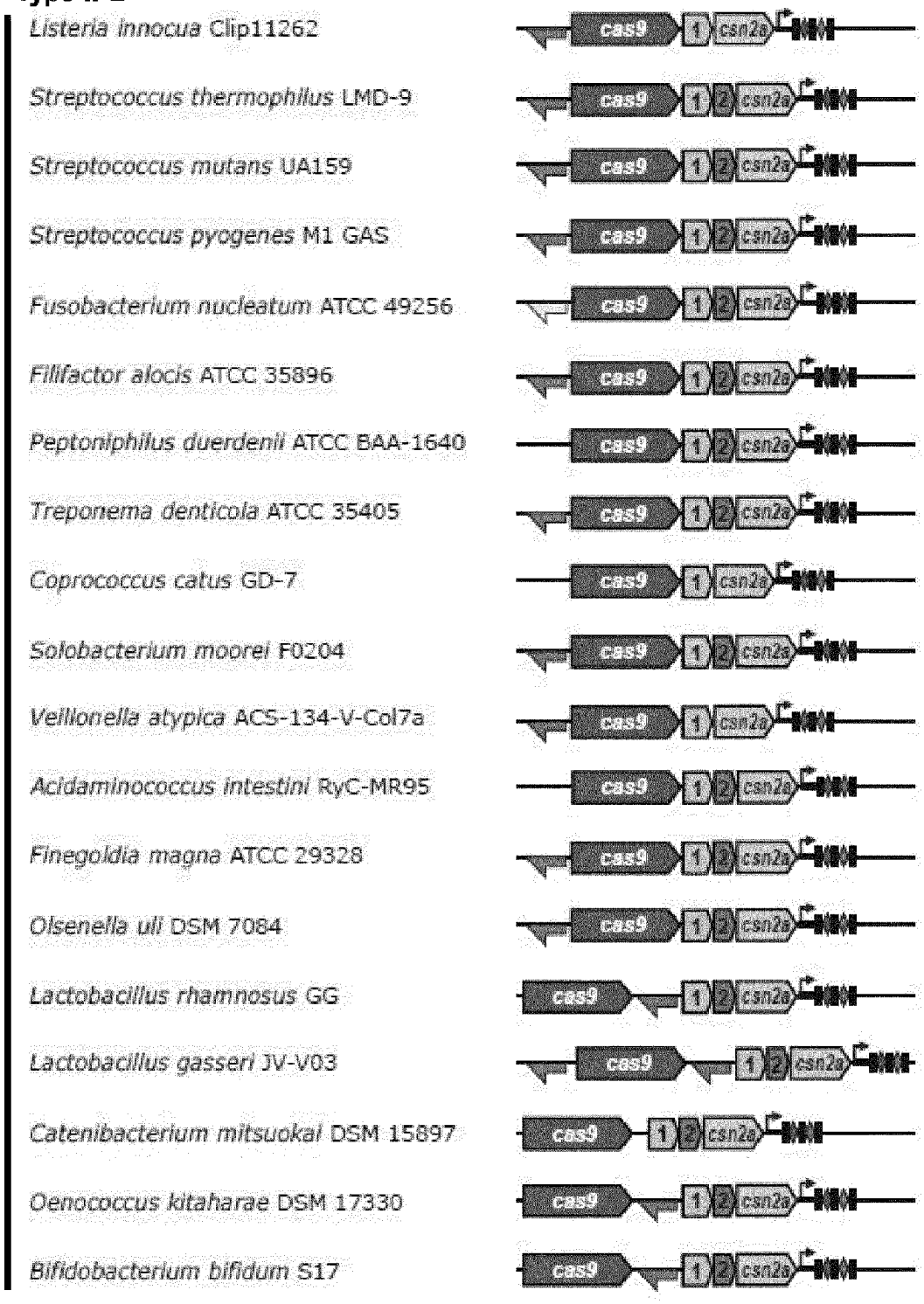
A

Type II-A



B **Figure 33**

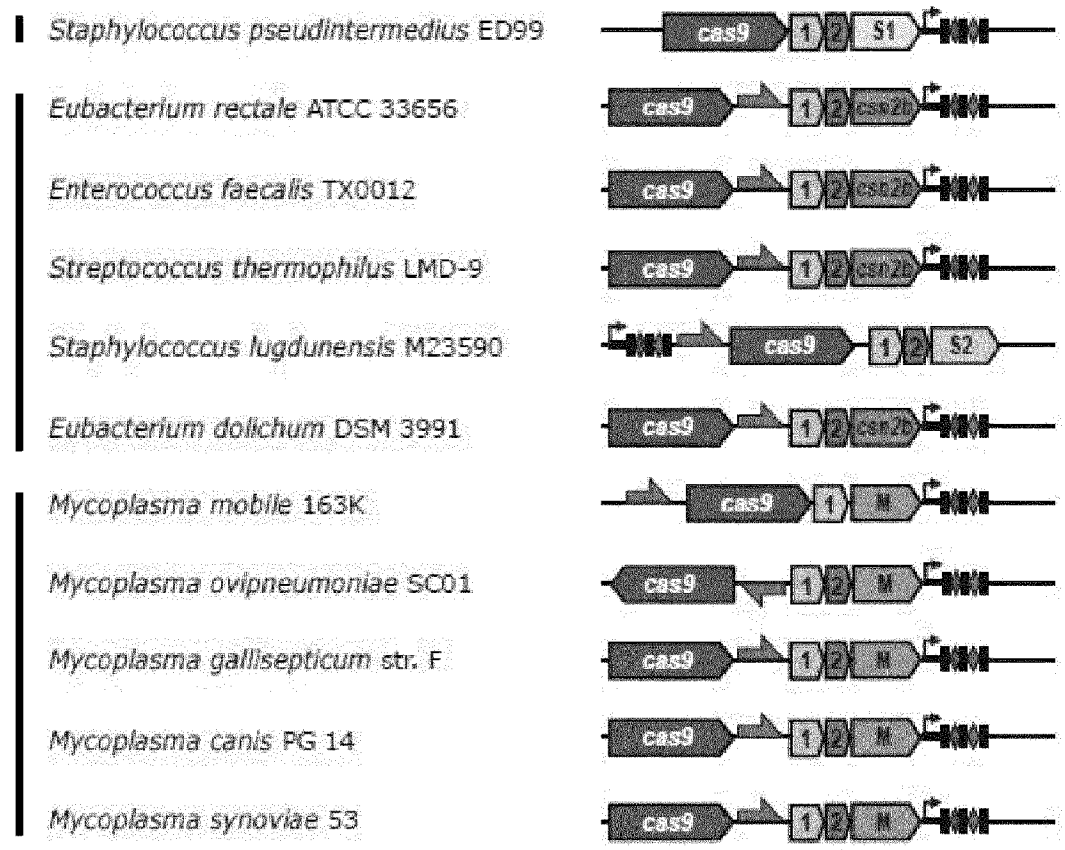
Type II-B



C

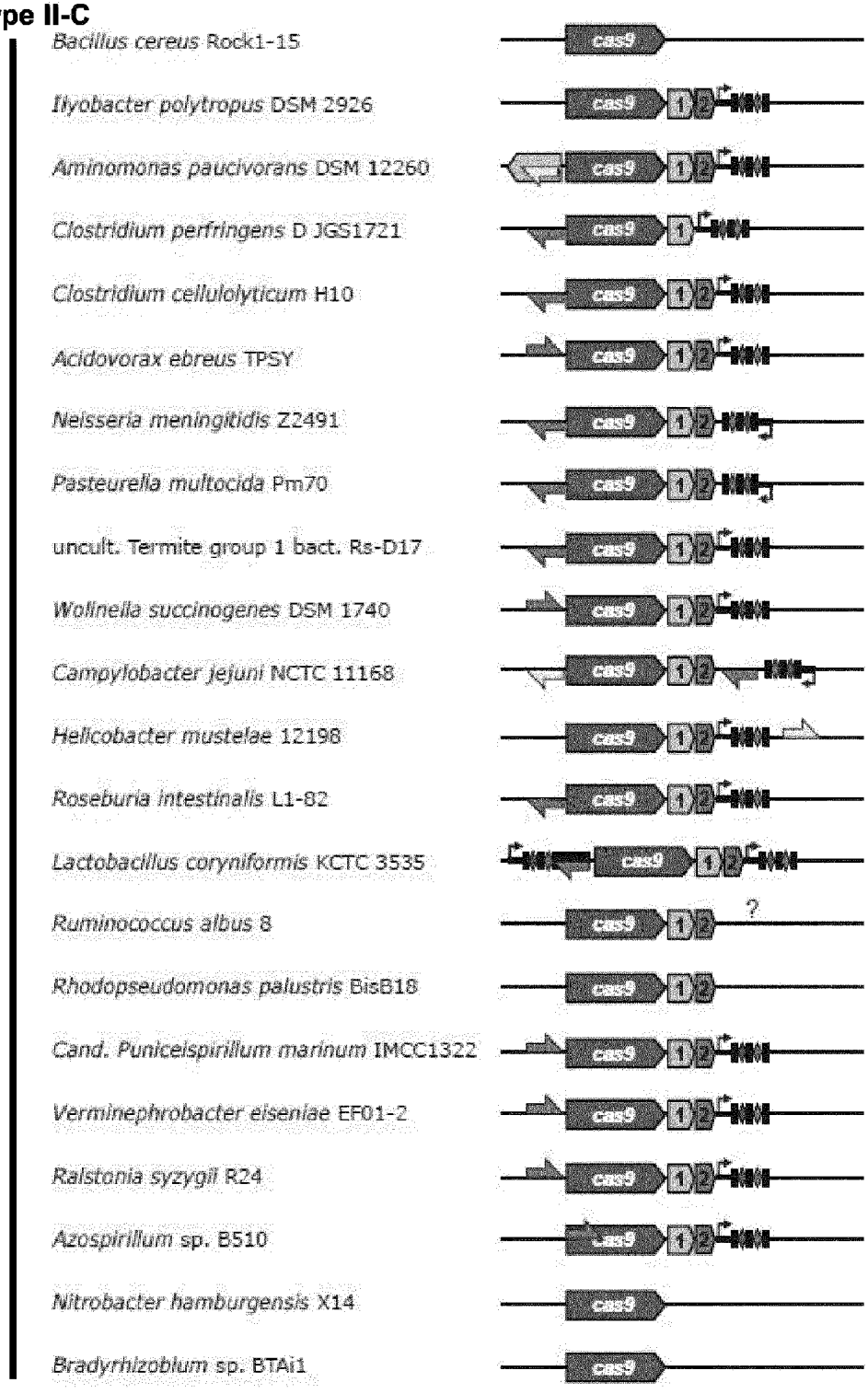
Figure 33

Type II-B (Continued)



D **Figure 33**

Type II-C



E

Figure 33

Type II-C (continued)

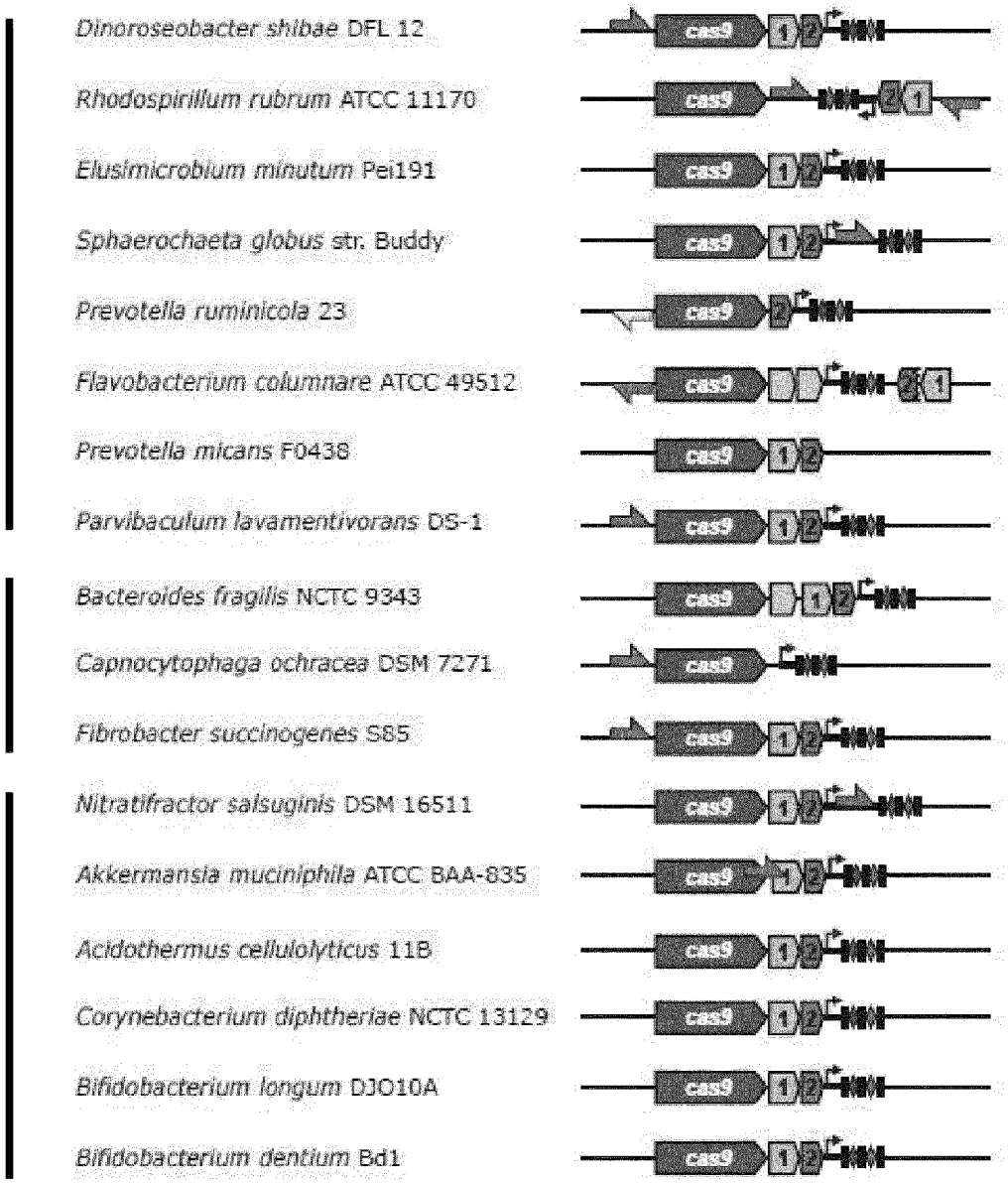


Figure 34

A

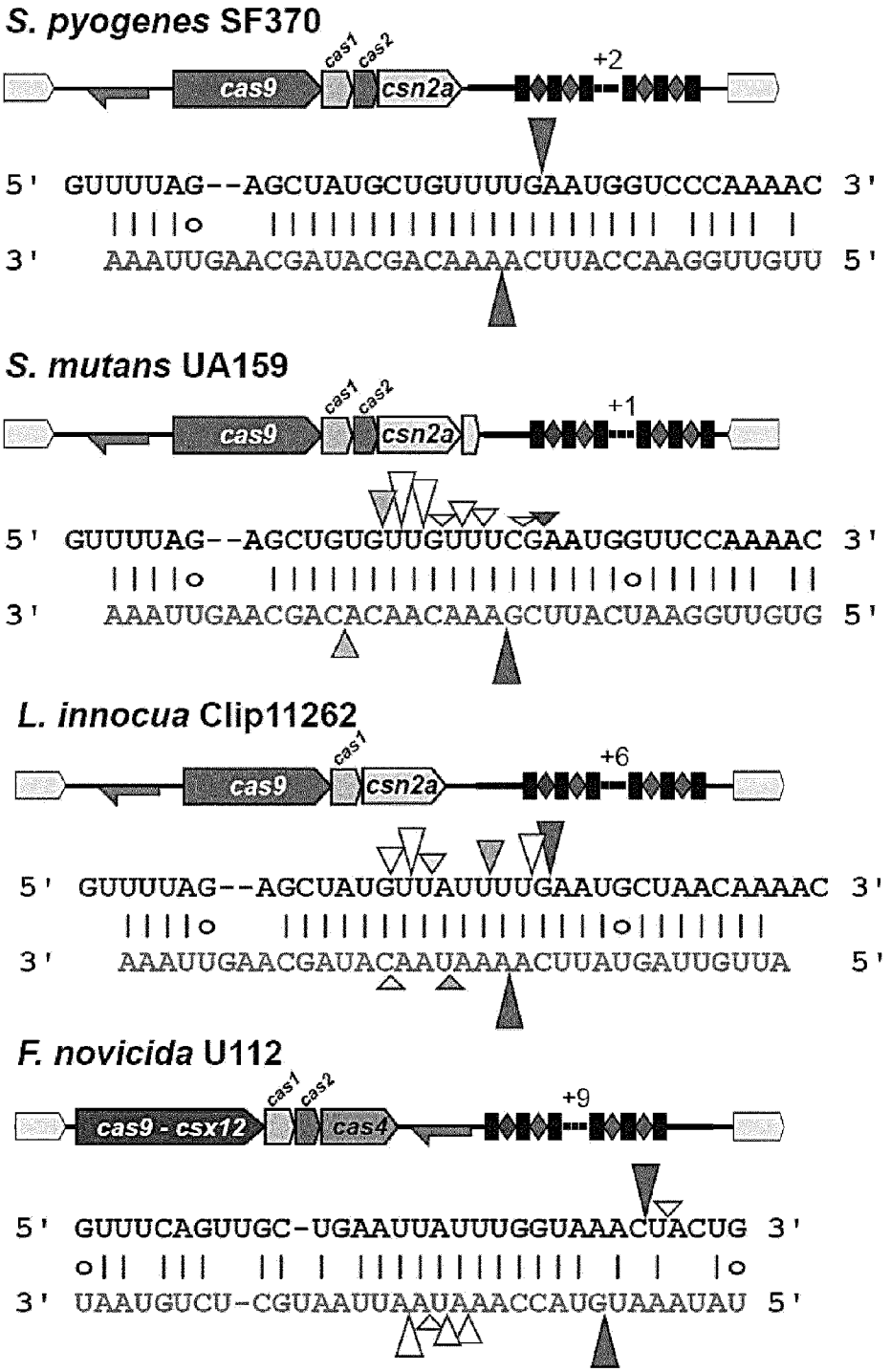

S. pyogenes SF370
cas9
cas1
cas2
csn2a
+2
5' GUUUUAG--AGCUAUGCUGUUUUGAAUGGUCCCAAAAC 3'
3' AAAUUGAACGAUACGACAAAACUUACCAAGGUUGUU 5'
S. mutans UA159
cas9
cas1
cas2
csn2a
+1
5' GUUUUAG--AGCUGUGUUGUUUCGAAUGGUUCCAAAAC 3'
3' AAAUUGAACGACACAACAAAGCUUACUAAGGUUGUG 5'
L. innocua Clip11262
cas9
cas1
csn2a
+6
5' GUUUUAG--AGCUAUGUUAUUUUGAAUGCUAACAAAAC 3'
3' AAAUUGAACGAUACAAUAAAACUUAUGAUUGUUA 5'
F. novicida U112
cas9 - csx12
cas1
cas2
cas4
+9
5' GUUUCAGUUGC-UGAAUUAUUUGGUAAACUACUG 3'
3' UAAUGUCU-CGUAAUUAAUAAACCAUGUAAAUAU 5'

Figure 34

B

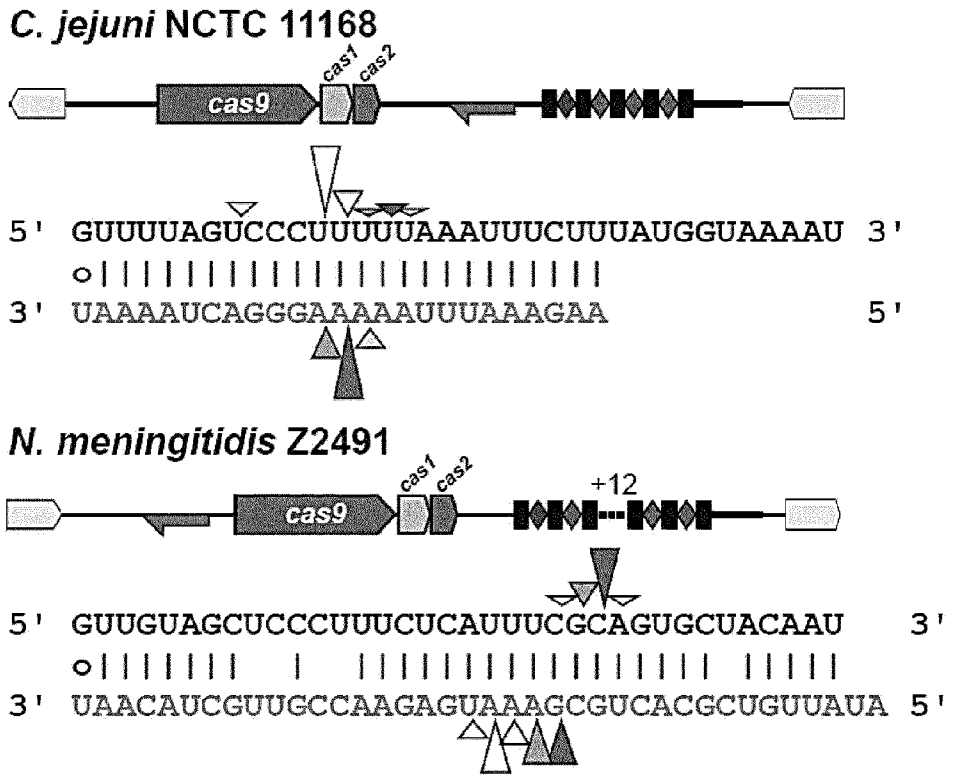

C. jejuni NCTC 11168
cas9
cas1
cas2
5' GUUUUAGUCCCUUUUUAAAUUUCUUUAUGGUAAAAU 3'
3' UAAAAUCAGGGAAAAAUUUAAAGAA 5'
N. meningitidis Z2491
cas9
cas1
cas2
+12
5' GUUGUAGCUCCCUUUCUCAUUUCGCAGUGCUACAAU 3'
3' UAACAUCGUUGCCAAGAGUAAAGCGUCACGCUGUUAUA 5'

FIGURE 35

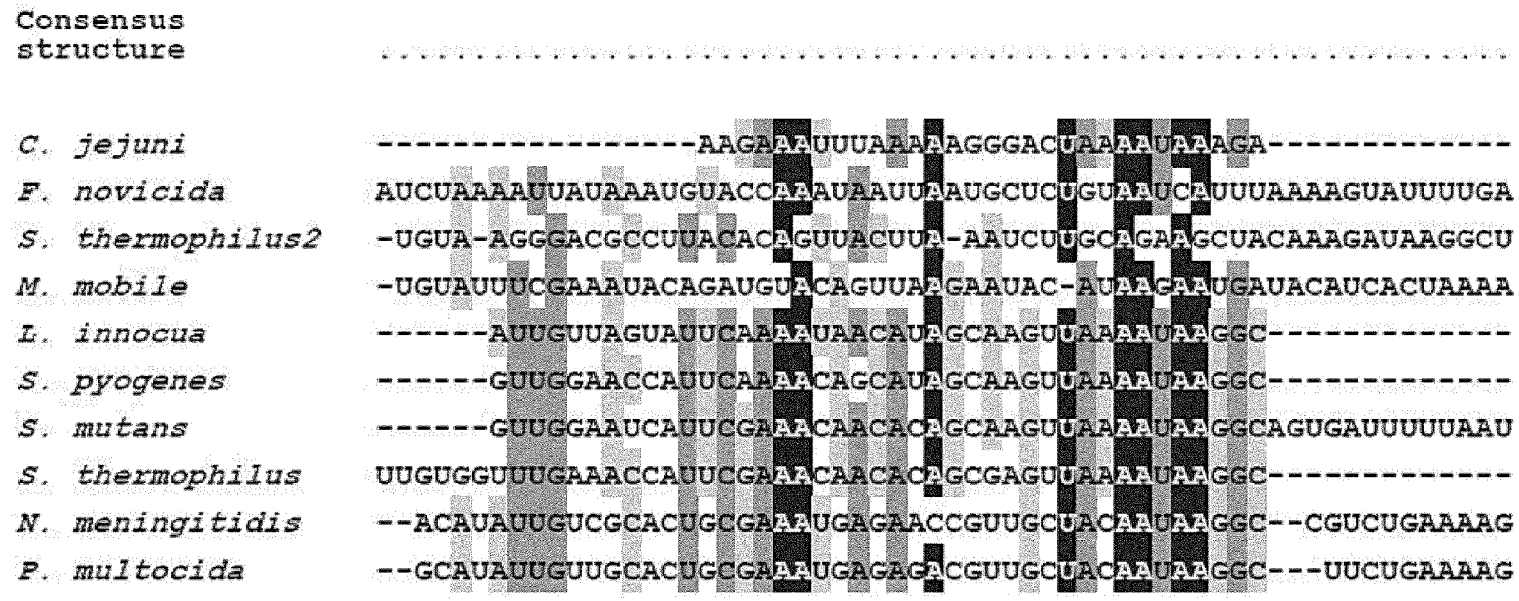

Consensus
structure
C. jejuni
F. novicida
S. thermophilus2
M. mobile
L. innocua
S. pyogenes
S. mutans
S. thermophilus
N. meningitidis
P. multocida

FIGURE 35 **Continued**

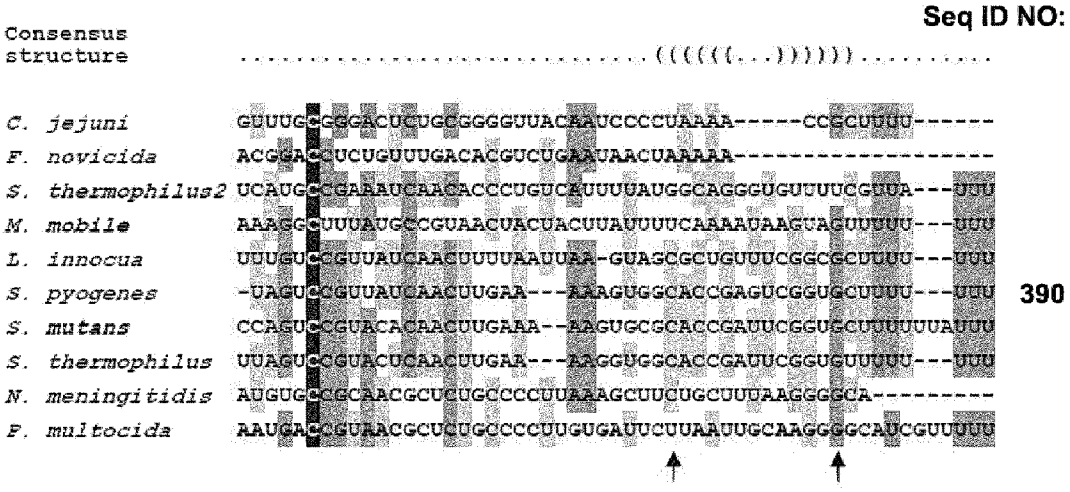


Figure 36

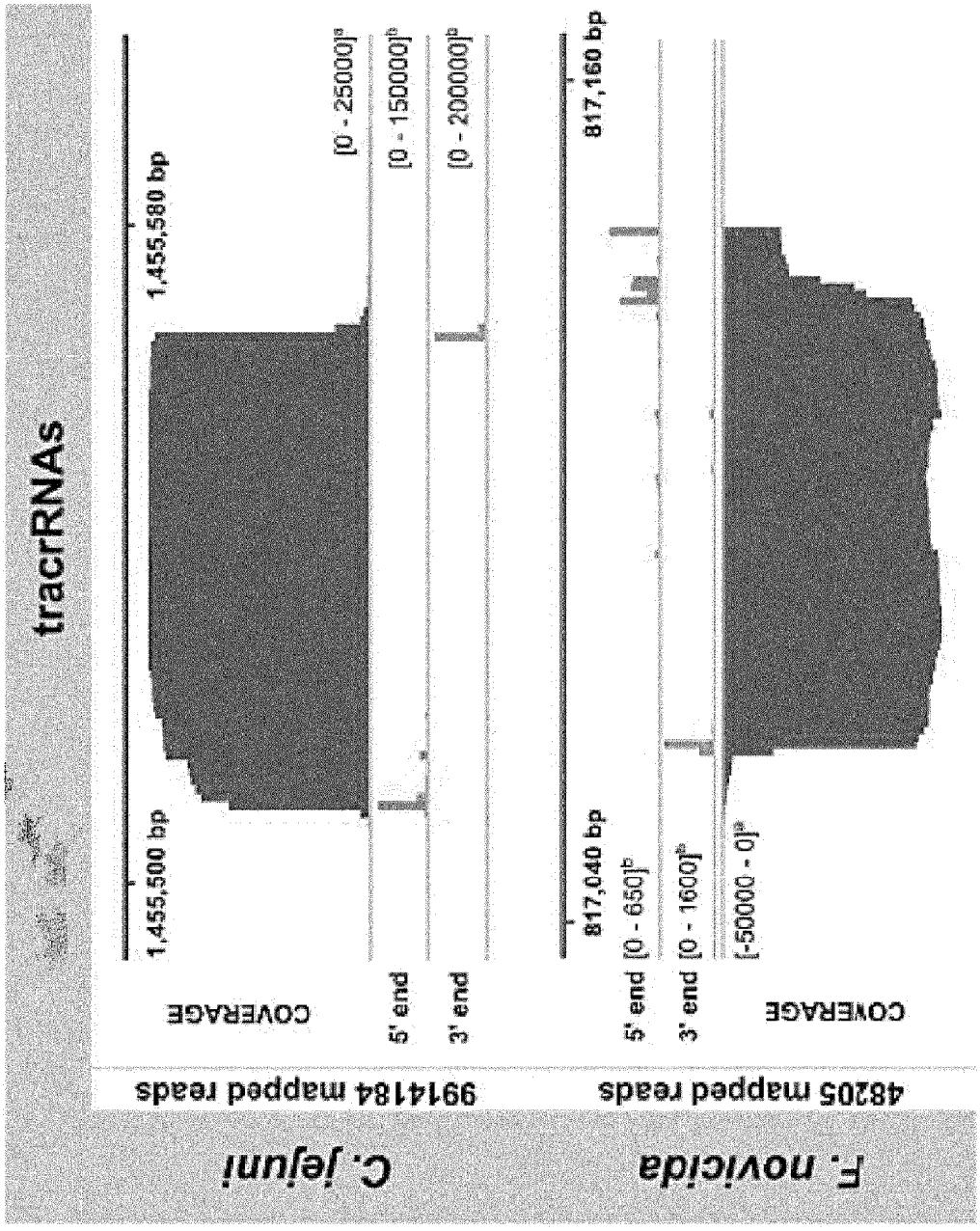

A
tracrRNAs
C. jejuni
F. novicida
9914184 mapped reads
48205 mapped reads
COVERAGE
1,455,500 bp
1,455,580 bp
[0 - 25000]a
[0 - 150000]b
[0 - 200000]b
5' end
3' end
817,040 bp
817,160 bp
5' end [0 - 650]b
3' end [0 - 1600]b
[-50000 - 0]a
COVERAGE

Figure 36

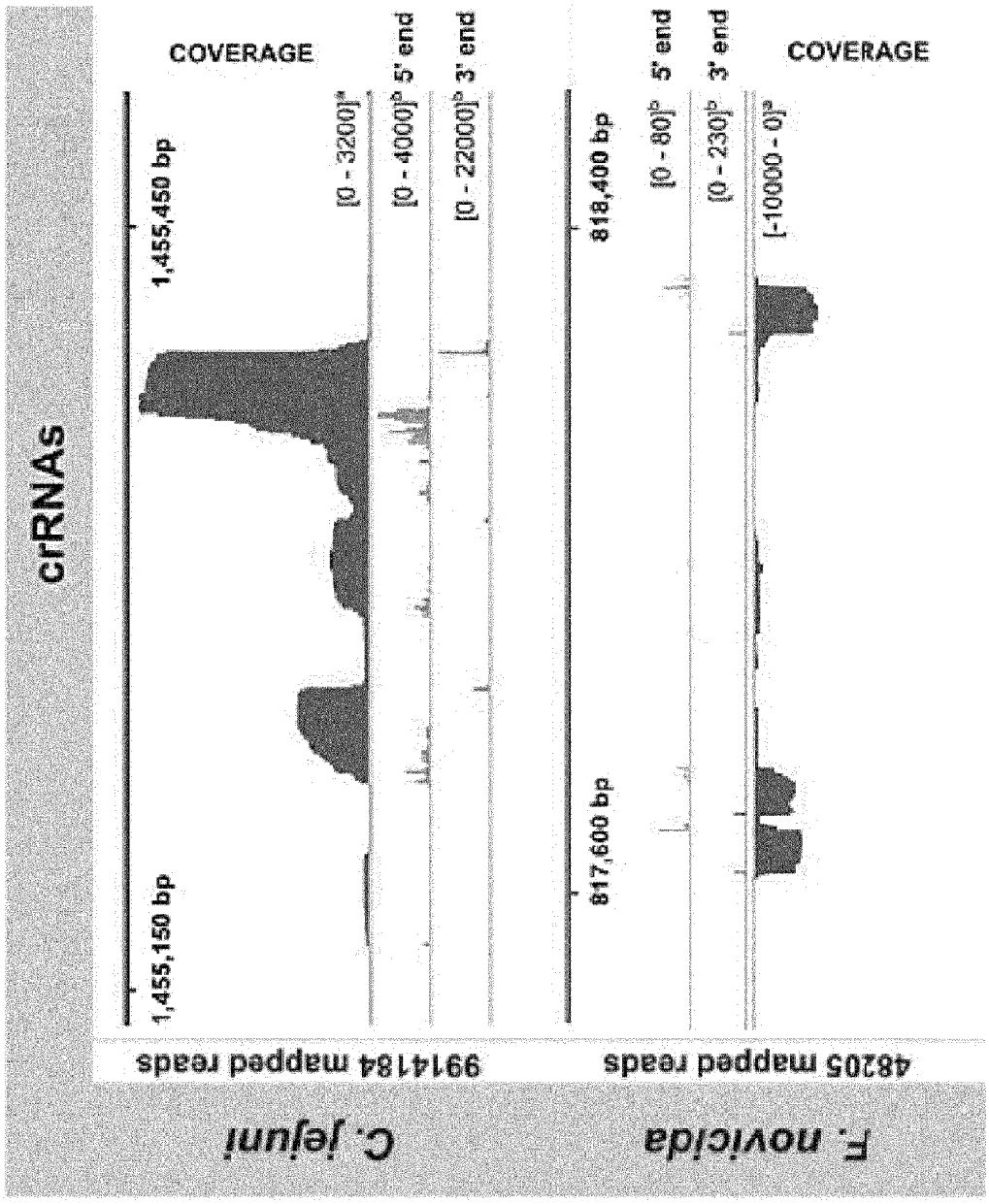
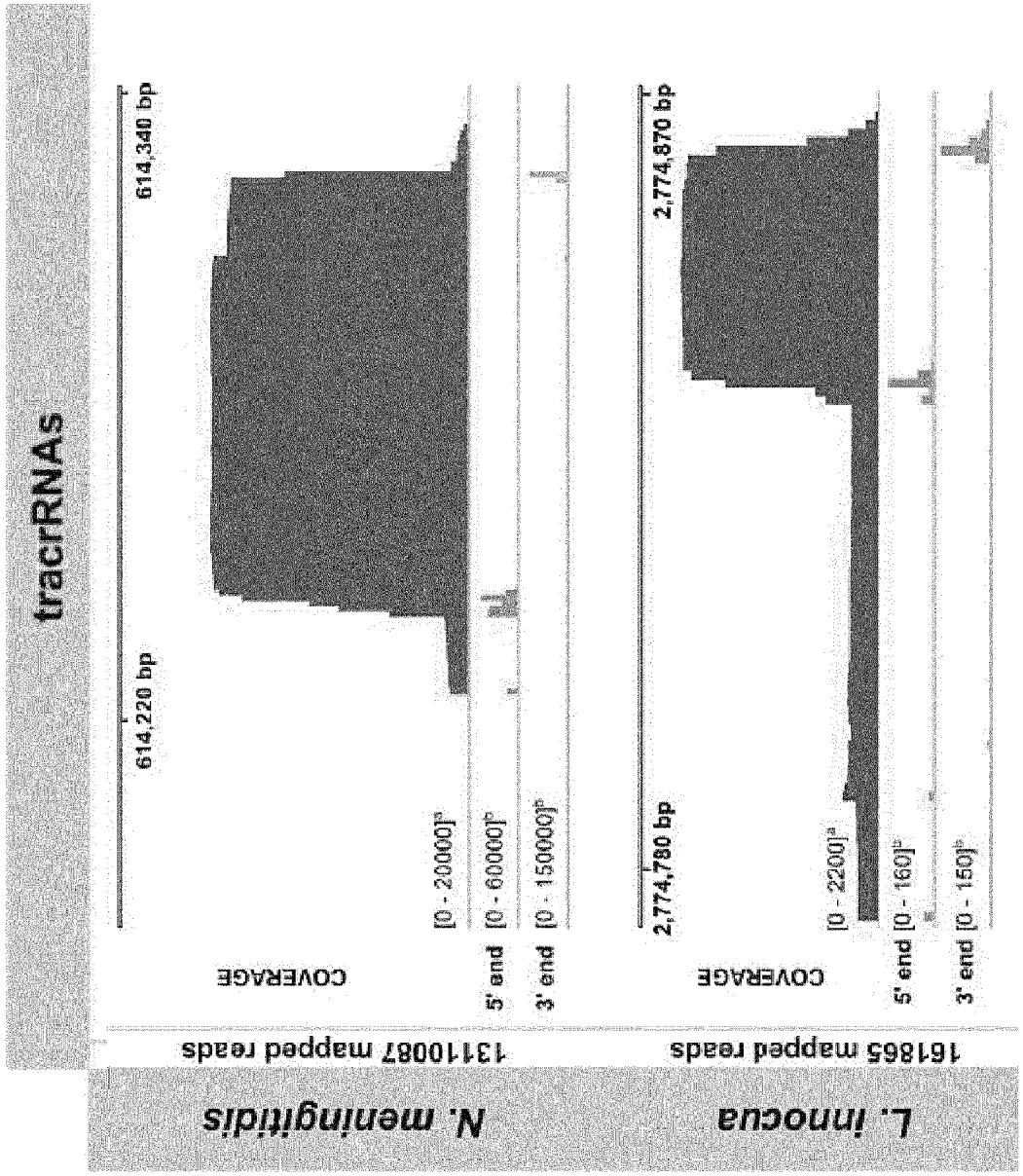

B
crRNAs
C. jejuni
F. novicida
9914184 mapped reads
48205 mapped reads
COVERAGE
1,455,150 bp
1,455,450 bp
817,600 bp
818,400 bp
[0 - 3200]a
[0 - 4000]b 5' end
[0 - 22000]b 3' end
[0 - 80]b 5' end
[0 - 230]b 3' end
[-10000 - 0]a
COVERAGE

Figure 36

C


tracrRNAs
N. meningitidis
1311087 mapped reads
614,220 bp
614,340 bp
COVERAGE
[0 - 20000]
5' end
[0 - 60000]
3' end
[0 - 150000]
L. innocua
161865 mapped reads
2,774,780 bp
2,774,870 bp
COVERAGE
[0 - 2200]
5' end
[0 - 160]
3' end
[0 - 150]

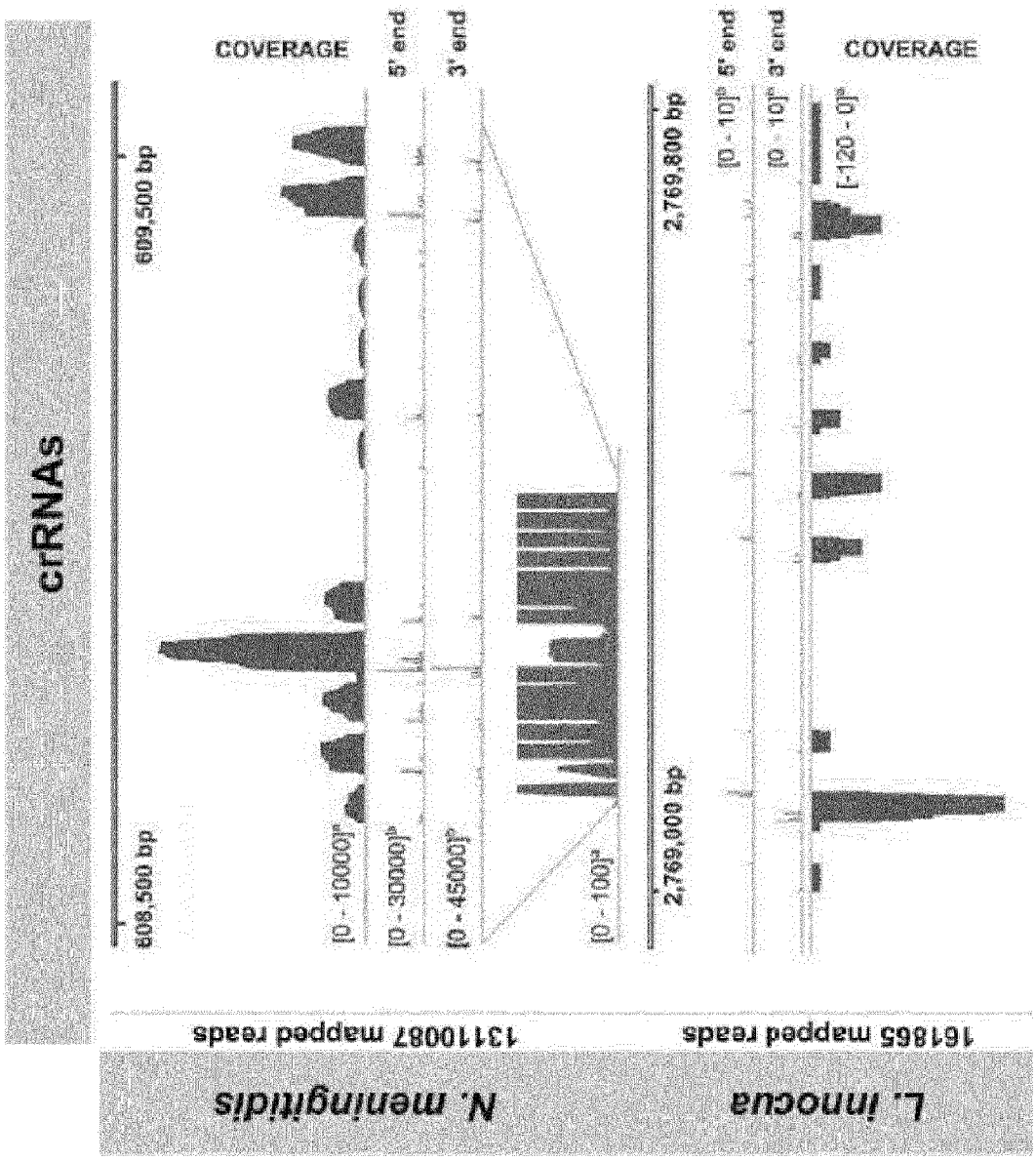

crRNAs
N. meningitidis
L. innocua
1311008? mapped reads
1618?5 mapped reads
COVERAGE
5' end
3' end
608,500 bp
609,500 bp
[0 - 10000]
[0 - 30000]
[0 - 45000]
[0 - 100]
2,769,000 bp
2,769,800 bp
[0 - 10]
[0 - 10]
[-120 - 0]
COVERAGE
5' end
3' end

Figure 36

D

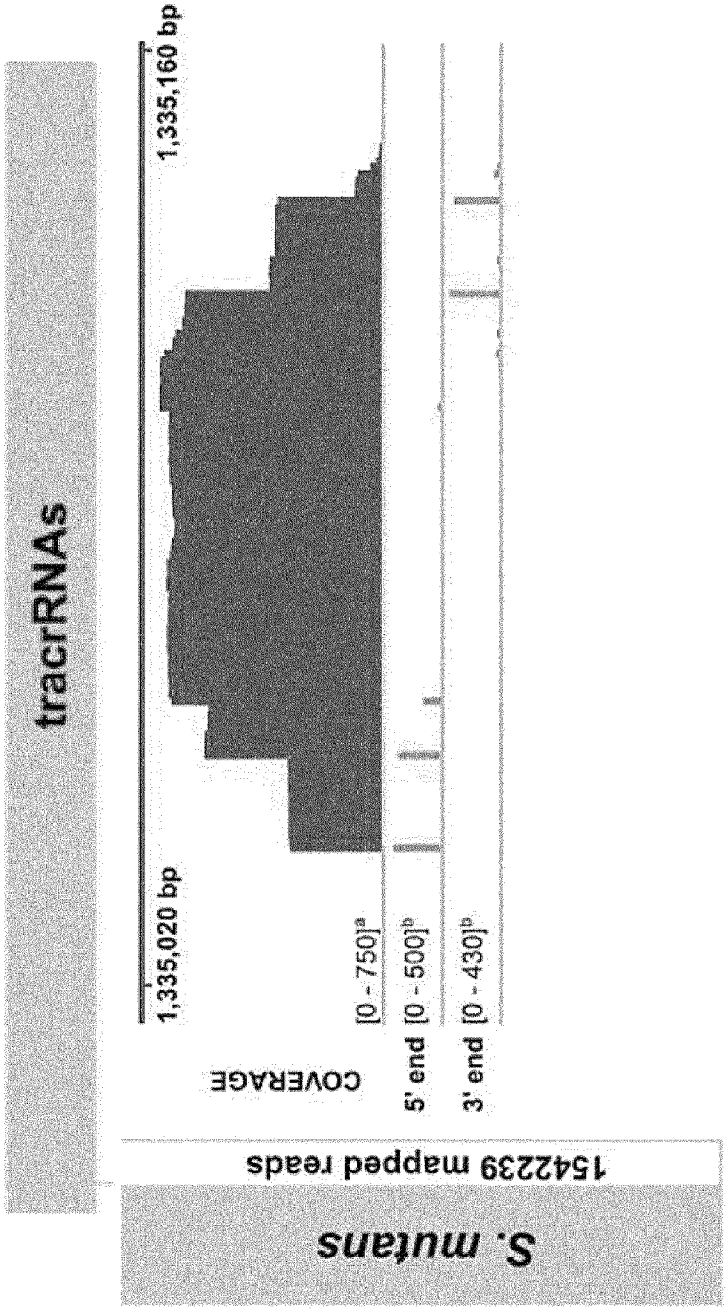

tracrRNAs
S. mutans
1542239 mapped reads
COVERAGE
1,335,020 bp
1,335,160 bp
[0 - 750]a
5' end [0 - 500]b
3' end [0 - 430]b

Figure 36

E

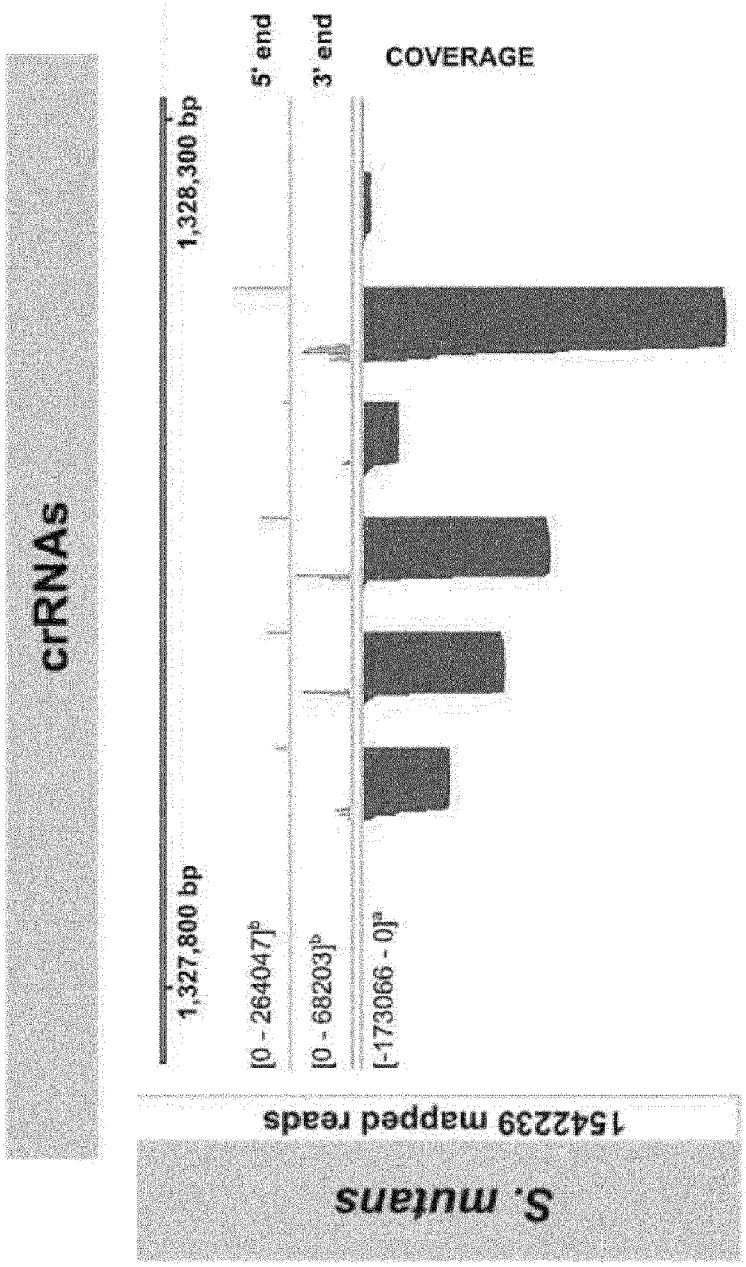

crRNAs
S. mutans
1542239 mapped reads
1,327,800 bp
1,328,300 bp
5' end
3' end
COVERAGE
[0 - 264047]b
[0 - 68203]b
[-173066 - 0]a

Figure 36
F

Figure 37 A

sRNA	Strand	Size mature form	Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
C. jejuni NCTC 11168 (NC_002163.1); total mapped reads: 9914184													
crRNA 4 spacers	→	crRNA1	38	1455167	1455202	781	0,0079	AGTTTTAAAAGAGCTTGGCGCTTGTTTAGT CCCTTTT	1455162	A	1	1455198	T	5
									1455165	G	2	1455199	C	5
									1455166	C	3	1455200	C	16
									1455167	A	415	1455201	C	72
									1455168	G	19	1455202	T	355
									1455169	T	7	1455203	T	82
									1455170	T	75	1455204	T	30
									1455171	T	24	1455205	T	112
									1455172	T	9	1455206	T	5
									1455173	T	11			
									1455174	A	4			
		crRNA2	38	1455231	1455268	2658	0,0298	CAAAGTTTCATTAGTTGAATTTAACTGTTTTA GTCCCTTTT	1455227	T	4	1455264	T	3
			29	1455240	1455268				1455228	T	1	1455265	C	14
									1455230	C	37	1455266	C	60
									1455231	C	208	1455267	C	240
									1455232	A	72	1455268	T	1304
									1455233	A	34	1455269	T	348
									1455234	A	59	1455270	T	29
									1455235	G	280	1455271	T	44
									1455236	T	188	1455272	T	15
									1455237	T	130	1455273	A	2
									1455238	T	36			
									1455239	C	259			
									1455240	A	317			
									1455241	T	148			
									1455242	T	288			
									1455243	A	157			
									1455244	G	40			

Figure 37 B

sRNA	Strand	Size mature form		Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
C. jejuni NCTC 11168 (NC_002163.1), total mapped reads: 9914184														
crRNA 4 spacers		crRNA3	35	1455300	1455334	4729	0,0477	AAGAATGAGGATTATGATATTTTACAGTTTTA GTCCCTTTT	1456202	T	16	1456329	G	21
									1455293	C	62	1455330	T	479
									1455294	T	5	1455331	C	23
									1455295	A	12	1455332	C	48
									1455296	C	23	1455333	C	179
									1455297	A	367	1455334	T	1226
									1455298	A	143	1455335	T	220
									1455209	G	563	1455336	T	54
									1456300	A	1014	1455337	T	49
									1455301	A	415	1455338	[illegible]	55
									1455302	T	242	1455339	A	1
									1455303	G	580			
									1455304	A	55			
									1455305	G	2			
		crRNA4	26	1455376	1455401	17404	0,1755	GTGTGCTAAAAAAAATGGACTTAAATGTTTTA GTCCCTT[illegible]	1455360	T	6	1455395	C	63
									1455361	G	2	1455399	C	645
									1455362	A	9	1455400	T	9844
									1455363	G	261	1455401	T	1518
									1455364	T	51	1455402	T	481
									1455365	G	598	1455403	T	703
									1455366	T	77	1455404	[illegible]	578
									1455367	G	552	1455405	A	4
									1455368	C	579			
									1455369	T	901			
									1455370	A	816			
									1455371	A	227			
									1455372	A	262			
									1455373	A	1072			
									1455374	A	1269			
									1455375	A	1991			
									1455376	A	2204			
									1455377	A	1978			
									1455378	T	371			

Figure 37 C

sRNA	Strand	Size mature form		Region of interest		Reads	Coverage [%]	Sequence	5' end read number			3' end read number		
C. jejuni NCTC 11168 (NC_002163.1), total mapped reads: 9814184														
tracrRNA	→	tracrRNA1	65	1455502	1455566	833820	8,4105	AAGAAATTTAA[illegible]GGGACTAAAATAAAGAGT TTGCGGGATCTTTTGGGGTTTCAATTCTTTTAA AACCGCTTTT	1455498	T	1	1455565	G	12371
		tracrRNA2	58	1455509	1455566				1455497	A	31	1455566	C	713283
									1455498	A	27	1455567	T	74594
									1455499	G	24	1455568	T	10412
									1455500	A	19	1455569	T	9580
									1455501	A	232	1455570	T	426
									1455502	A	435	1455571	A	91
									1455503	T	369	1455572	A	97
									1455504	T	253	1455573	A	542
									1455505	T	85			
									1455506	A	65			
									1455507	A	193			
									1455508	A	33472			
									1455509	A	515001			
									1455510	A	1311879			
									1455511	G	16444			
									1455512	G	9390			
									1455513	G	1053			

Figure 37 D

sRNA	Strand	Size mature form	Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
F. novicida U112 (NC_008601.1), total mapped reads: 48205													
crRNA 13 spacers	→	crRNA5	52	817656	817607	117	0,2427	ATAACTTTAATTAATATTTTACAAATTTTCATT TGCTGAATTATTTGGTAAACCT	817612	T	10	817657	A	3
									817611	T	10	817656	C	55
									817607	A	53	817655	T	13
									817606	T	7	817653	C	12
												817652	T	1
		crRNA6	56	817627	817682	116	0,2406	GGCAAGTTTTGTATGGTCATATAGGAGTATTT CAGTTGCTGAATTATTTGGTAAACCT	817685	A	4	817629	A	4
									817682	G	23	817628	A	3
									817681	G	13	817627	C	43
									817679	A	2	817626	T	20
									817677	G	4			
		crRNA7	66	817699	817764	11	0,0228	AGTTATAGGAGTACCTATCTTTTGAGTGTGG CAAATAGTTTCAGTTGCTGAATTATTTGGTA AACCT	817764	A	1	817699	A	2
									817763	G	1			
									817754	T	4			
									817745	T	2			
		crRNA8	53	817845	817897	24	0,0498	ATGCTTTTAAACTACTGATATATACGTTTCAG TTGCTGAATTATTTGGTAAACCT	817897	A	3	817845	C	7
tracrRNA	←	tracrRNA1	74	817065	817138	2808	5,8251	GTACCAAATAATTAATGCTCTGTAATCATTTA AAAGTATTTCGAACCGAGCTCTGTTTGACACG TCTGAATAACTAAAAA	817140	A	2	817066	A	28
		tracrRNA2	64	817065	817128				817139	T	2	817065	C	1328
									817138	G	615	817064	T	440
									817136	A	9	817063	A	7
									817135	C	2	817062	A	16
									817134	C	37	817061	A	19
									817133	A	26	817060	A	14
									817132	A	7	817059	A	6
									817131	A	328	817058	G	5
									817130	T	355	817057	C	32
									817129	A	185	817056	A	10
									817128	A	484			
									817127	T	24			
									817126	I	55			
									817125	A	32			
									817124	A	19			

Figure 37 E

sRNA	Strand	Size mature form	Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
N. meningitidis AZ2491 (NC_003116.1), total mapped reads: 13110087													
crRNA 18 spacers	→	crRNA1	48	608456	608503	1346	0,0103	TATCCATTCCCAACCGGAAATTAAGTTGTAGC TGGTTTTCTGATTTGCAGT	608453	T	30	608500	T	25
									608454	G	105	608501	C	77
									608455	T	84	608502	G	175
									608456	A	259	608503	C	293
									608457	T	4	608504	A	78
									608458	C	21	608505	G	505
									608459	C	175	608507	T	1
									608460	A	118			
									608461	T	39			
		crRNA2	50	608520	608569	685	0,0052	GCTTTTTTATACAGCTCGCTTTCTTTGTTGTAG CTTCCTTTCTTATTTCTTGAGT	608517	T	5	608564	T	8
									608518	C	41	608565	T	16
									608519	T	7	608566	T	8
									608520	G	61	608567	C	31
									608521	C	21	608568	G	101
									608522	C	44	608569	C	173
									608523	T	21	608570	A	43
									608524	T	31	608571	G	6
									608525	T	17	608572	T	2

Figure 37 F

sRNA	Strand	Size mature form	Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
N. meningitidis AZ2991 (NC_003116.1), total mapped reads: 13119087													
		crRNA3	50	608586	608635	12402	0,0946	TAAACTTTCTCTTGCCACCCAATCTTCTAG CTCCATTTCCCATTTCCAGT	608583	T	6	608631	T	76
									608584	G	3	608632	T	114
									608585	G	513	608633	C	414
									608586	T	2248	608634	G	1510
									608587	A	2091	608635	C	2210
									608588	A	188	608636	A	297
									608589	A	233	608637	G	52
									608590	G	744	608638	T	8
									608591	G	1152			
		crRNA4	49	608653	608701	26361	0,2011	TAACTTTGACCCGTGTCCAATCCAGTTAGTTAG AGCCCCCTTCCTATTTCCCAGT	608646	T	7	608697	T	295
									608647	C	127	608698	T	1678
									608648	T	203	608699	C	1167
									608649	T	377	608700	G	4910
									608650	T	2751	608701	C	4966
									608651	A	477	608702	A	764
									608652	A	1334	608703	G	62
									608653	C	11084	608704	T	4
									608654	T	3865			
									608655	T	382			

Figure 37 G

sRNA	Strand	Size mature form	Region of interest	Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
N. meningitidis AZ2491 (NC_003116.1), total mapped reads: 13110087												
crRNA5		48	608719	608767	28747	0,2193	AACCCACTAAATTTTGCAAATGCGGTTGTAGC TACCTTTCTCATTCCAGCT	608717	G	20	608764	T	428
								608718	C	1171	608765	C	1316
								608719	A	9128	608766	G	3040
								608720	A	118	608767	C	8431
								608721	C	59	608768	A	1048
								608722	C	420	608769	G	102
								608723	C	1009	608770	T	11
								608724	A	557			
crRNA6		50	608784	608833	121014	0,9231	TTTTTTTGTACTGTTGTTTGAACGAGTTGTAG CTCTCTTTCTCATTTCCCAGT	608781	G	42	608830	T	2295
								608782	C	630	608831	C	7060
								608783	C	10039	608832	G	13664
								608784	T	28025	608833	C	44473
								608785	T	11430	608834	A	6853
								608786	T	8248	608835	G	892
								608787	T	3848	608836	T	69
crRNA7		52	608848	608899	24611	0,1877	TTCGTTTCAGATAGCAAACGCAGTAGTGTTGT AGCTCTCTTTCTCATTTCCCAGT	608846	A	28	608896	T	369
								608847	A	631	608897	C	1114
								608848	T	11039	608898	G	3441
								608849	T	753	608899	C	9818
								608850	C	239	608900	A	1559
								608851	G	2205	608901	G	137
								608852	T	1910	608902	T	8
								608853	T	826			
		84	608916	608999	901	0,0069	ATATTAGGGTGGGCAACTGGTACAGTTTGTAG CTCTCTTTCTCATTTCCCAGTTTCCACAATCCC GGATATGACGGTGGGCAACT	608913	C	6	608996	A	6
								608914	G	3	608997	A	2
								608915	G	2	608998	C	2
								608916	A	266	608999	T	253
								608917	T	17	609000	G	171
								608918	A	98	609001	G	24
								608919	T	26	609004	T	16

Figure 37 H

sRNA	Strand	Size mature form	Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
N. meningitidis AZ2491 (NC_003116.1), total mapped reads: 13110087													
	crRNA10	48	609049	609097	5027	0,0383	CTTTGTATTGACTCAMGTGCTTGTTGTACC TCUCTTCTCATTTC*CAGT	609046	T	89	609094	T	81
								609047	C	37	609095	C	218
								609048	G	847	609096	G	845
								609049	C	1614	609097	C	1237
								609050	T	583	609098	A	219
								609051	T	208	609099	G	38
								609052	T	104	609101	G	7
								609053	T	18			
	crRNA11	52	609112	609163	23711	0,1732	ATTCGTCGATGATGGAAACTCGAGCATGTTGT AGCTCCCTTTCTCATTTC*CAGT	609109	A	58	609160	T	540
								609110	G	331	609161	C	1263
								609111	T	99	609162	G	2859
								609112	A	10627	609163	C	4382
								609113	T	1234	609164	A	552
								609114	T	66	609165	G	52
								609115	C	191	609166	T	15
								609116	G	4906			
								609117	T	136			
	crRNA12	52	609170	609229	5067	0,0386	TTGCCAGTGCTAAAACCGCACCCGCTTGTTGT AGCTCCCTTTCTCATTTC*CAGT	609175	G	4	609226	T	57
								609176	G	3	609227	C	189
								609177	G	42	609228	G	576
								609178	T	997	609229	C	1248
								609179	A	72	609230	A	321
								609180	G	584	609231	G	33
								609181	C	154	609232	T	2
								609182	C	314			
								609183	A	38			

Figure 37

sRNA	Strand	Size mature form	Region of interest	Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
N. meningitidis A22491 (NC_003116.1), total mapped reads: 13110037												
crRNA13	51	609245	609295	4666	0,0356	ATAGAAATACGTACGCCGAGTAATTAGTTGTA GCTCCCTTTTTCCATTTCGCAGT	609243	A	6	609262	T	65
							609244	A	62	609293	C	201
							609245	A	1311	609294	A	475
							609246	T	141	609295	C	1029
							609247	A	116	609296	A	229
							609248	G	44	609297	G	25
										609298	T	1
crRNA14	51	609311	609361	7147	0,0545	TTTTTGTAATTGTTCTGCCTTTTTAGTTGTA GCTCCCTTTTTCCATTTCGCAGT	609308	T	12	609358	T	136
							609309	T	22	609359	C	442
							609310	C	207	609360	G	858
							609311	T	1190	609361	C	2355
							609312	T	774	609362	A	540
							609313	T	577	609363	G	69
							609314	T	195	609364	T	5
							609315	T	37			
crRNA15	50	609378	609427	49818	0,3800	ACGGCGGAAACCATTGCCACAAAACGTTGTAG GTCCCTTTCCCATTTCGCAGT	609375	C	319	609424	T	532
							609376	C	7253	609425	C	1414
							609377	C	7249	609426	G	2563
							609378	A	19015	609427	C	7449
							609379	C	547	609428	A	1854
							609380	G	307	609429	G	210
							609381	G	190	609430	T	10

Figure 37 J

sRNA	Strand	Size mature form	Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
N. meningitidis A22491 (NC_003116.1); total mapped reads: 13110087													
		crRNA16	50	609444	609493	42398	0,3234	AATAATAACCAATACACGATGTTAAGTTGTA GATCCCTTTCTCATTTCCCAGT	609441	A	8	609490	T	579
									609442	A	548	609491	C	2197
									609443	A	3668	609492	G	6466
									609444	A	4650	609493	C	13217
									609445	T	1428	609494	A	2438
									609446	A	2861	609495	G	209
									609447	A	1322	609496	T	12
									609448	T	297			
tracrRNA	---	tracrRNA1	103	614102	614324	208318	1,5890	TTGTCTTATTTATATAAAATAAATATTATTGAC TTATCATCTCACAACGGCTAGAATCCCAAACA TATTGTCGCACTGCCAATTGAGAACCGTTGCT ACAATAAGCCGTCTGAAAAGATGTGCCTCAA CCCTCTGCCCTTTAAAGCTTCTTGTTTAAGCG GGATCGTTTAT	614158	C	1	614319	G	311
		tracrRNA2	109	614225	614324				614161	A	1	614320	G	109
		tracrRNA3	82	614243	614324				614162	T	2	614321	G	253
									614164	G	1	614322	G	150
									614223	A	596	614323	C	41244
									614224	C	39	614324	A	128531
									614225	A	14761	614325	T	167
									614226	T	169	614326	C	4197
									614227	A	276	614327	G	30
									614239	C	200	614328	T	122
									614240	G	42037	614329	T	865
									614241	A	39965	614330	T	1132
									614242	A	22890	614331	A	1081
									614243	A	51189	614332	T	2788
									614244	T	17711	614333	T	3758
									614245	G	4192	614334	T	77
									614246	A	488	614335	C	18
												614336	G	3
												614337	G	3
												614338	T	1

Figure 37 **K**

sRNA	Strand	Size mature form	Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
L. innocua Clip11262 (NC_003212.1), total mapped reads: 161865 (Note: low quality of the RNA library)													
crRNA 10 spacers	- -	crRNA1	35	2769606	2769640	2	0,0012	GGTAACTTTGCCTAGCATAGCTTTTAGAGCTAT GT	2769640	G	1	2769606	T 1
		crRNA2	22	2769540	2769561	2	0,0012	TATTATGTTTTAGAGCTTATGT	2769561	C	1	2769540	T 1
									2769560	A	1		
		crRNA3	24	2769468	2769491	3	0,0019	GAGTTTTAGAGCTATGT ATTTTG	2769491	G	3	2769468	G 1
		crRNA4	27	2769402	2769428	7	0,0043	TTATAGTTTTAGAGCTATGT ATTTTG	2769428	T	5	2769407	A 1
									2769427	T	2	2769406	T 1
												2769405	T 1
												2769403	T 1
												2769402	G 2
		crRNA5	26	2769337	2769362	5	0,0031	TAAATGTTTTAGAGCTATGT ATTTT	2769362	T	3	2769339	T 1
									2769360	A	2	2769337	T 2
		crRNA6	23	2769142	2769164	2	0,0012	TACAACTTTTAGAGCTATGT AT	2769164	T	1	2769143	A 1
									2769163	A	1	2769142	T 1
		crRNA9	30	2769072	2769101	19	0,0117	TTCATGTTGTTTTAGAGCTATGT AT TTG	2769101	T	8	2769079	T 4
									2769100	T	1	2769078	T 1
									2769099	C	4	2769075	T 2
									2769098	A	3	2769073	T 3
									2769097	T	3	2769072	G 8
									2769096	G	2		
		crRNA10	28	2769000	2769027	19	0,0117	GTTTTAGAGCTATGCTATTTCGAATACT	2769027	G	1	2769000	T 1

Figure 37 L

sRNA	Strand	Size mature form		Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
L. innocua Clip12262 (NC_003212.1); total mapped reads: 161865 (Note: low quality of the RNA library)														
tracrRNA	→	tracrRNA1	90	2774774	2774863	367	0,2267	ATTGTTAGTATTCAAAATAACATAGCAAGTTA AAATAAGGCTTTGTCCGTTATCAACTTTTAAT TAAGTAGCGCTGTTTCGGCGCTTTTTTT	2774774	A	34	2774861	T	2
		tracrRNA2	76	2774788	2774863				2774783	A	1	2774862	T	47
		tracrRNA3	68	2774796	2774863				2774786	C	1	2774863	T	150
									2774787	A	1	2774864	T	67
									2774788	A	22	2774865	T	30
									2774794	C	1	2774866	G	15
									2774795	A	1	2774867	T	6
									2774796	T	5			
									2774797	A	2			
									2774799	C	5			
									2774801	A	1			

Figure 37 **M**

sRNA	Strand	Size mature form	Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
S. mutans UA159 (NC_004350.2): total mapped reads: 1642239													
crRNA 5 spacers	←	crRNA1	38	1328162	1328199	267104	17,3192	GCCATTAATTAATAGTGAGTTTTAGAGCTG TGTTGTTCGA	1328201	A	8	1328166	G	18547
									1328200	C	178	1328165	T	41345
									1328199	G	264047	1328164	T	53386
									1328198	C	191	1328163	G	9084
									1328197	C	167	1328162	T	55197
									1328196	A	117	1328161	T	28333
												1328160	T	4240
												1328159	C	16236
												1328158	G	26742
												1328157	A	5573
												1328156	A	17
		crRNA2	36	1328098	1328133	26578	1,7233	GCTAGCCAGTTAGTGCTCTGTTTTAGAGCTG TGTTGTTCGA	1328135	C	4	1328101	T	37
									1328134	A	13	1328100	G	899
									1328133	G	25595	1328099	T	5395
									1328132	C	212	1328098	T	11268
									1328131	T	25	1328097	G	301
									1328130	A	62	1328096	T	1453
									1328129	G	25	1328095	T	1755
												1328094	T	447
												1328093	C	1302
												1328092	G	1996
												1328091	A	670
												1328090	A	5

Figure 37

N

sRNA	Strand	Size mature form	Region of interest	Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
S. mutans UA159 (NC_004350.2); total mapped reads: 1542239												
crRNA3	34	1328034	1328067	138134	8,9567	TGTTGTGTCATCATAGTTAGGTTTTAGAGCTG TGTTGTTTCAA	1328069	G	4	1328036	G	2380
							1328068	C	66	1328035	T	733
							1328067	T	134381	1328034	G	60203
							1328066	G	609	1328033	T	37212
							1328065	T	321	1328032	T	24520
										1328031	G	1003
										1328030	T	889
										1328029	T	619
										1328028	T	124
										1328027	C	371
										1328026	G	546
										1328025	A	145
crRNA4	35	1327967	1328001	104705	6,7892	CRATTAGACAATAGRCAAACGTTTTAGAGCTG TGTTGTTTCAA	1328003	T	8	1327969	T	348
							1328002	T	36	1327968	G	8506
							1328001	C	101017	1327967	T	59363
							1328000	A	681	1327966	T	25845
							1327999	A	216	1327965	G	756
										1327964	T	3053
										1327963	T	1433
										1327962	T	1802
										1327961	C	864
										1327960	G	839
										1327959	A	322

Figure 37

O

sRNA	Strand	Size mature form	Region of interest		Reads	Coverage (%)	Sequence	5' end read number			3' end read number		
S. mutans UA159 (NC_004350.2); total mapped reads: 1542239													
	crRNA5	37	1327899	1327936	83999	4,1497	TTCGGACATGACTTGCCACAGTTTTAGAACTG TCTTGTTATAGAA	1327940	A	4	1327902	G	902
								1327937	T	8	1327901	T	3296
								1327936	A	17	1327900	T	12381
								1327935	T	62587	1327899	G	19843
								1327934	T	1029	1327898	T	2984
								1327933	C	108	1327897	T	14180
								1327932	G	19	1327896	T	2081
								1327931	G	36	1327895	C	2079
											1327894	G	5258
											1327893	A	839
tracrRNA →		tracrRNA1	102	1335040	1335141	1289	0,0842	GTTGGAATCATTCGAAACAACAAGCAAGTTA AAATAAGGCAGTGATTTTTAATCCAGTCGTA CACAACTTGAAAAAGTGGCACCGAGTCGGTG CTTTTTTATTT	1335038	G	1	1335140	T	13
		tracrRNA2	88	1335054	1335141				1335040	G	486	1335141	T	64
		tracrRNA3	80	1335062	1335141				1335041	T	6	1335142	T	29
									1335042	T	1	1335143	A	13
									1335051	T	4	1335144	T	6
									1335053	G	3	1335145	T	13
									1335054	A	415	1335146	T	1
									1335055	A	2	1335149	T	1
									1335057	C	1			
									1335058	A	1			
									1335062	C	168			
									1335063	A	15			

A

Figure 38

Cluster	SEQ ID NO:
1	2, 3, 4, 5, 6, 7, 8, 15, 23, 24, 25, 36, 37, 38, 39, 41, 71, 74, 105, 116, 136, 138, 166, 177, 180, 183, 193, 204, and 232
2	83, 75, 156, 96, 121, 235, 208, 238, 127, 182, 134, 119, 246, 153, 202
3	101, 168, 48, 226, 216, 210, 120, 102, 176, 57, 108, 79, 1, 245
4	219, 135, 53, 62, 240, 165, 217, 82, 212, 19, 40, 18, 194
5	84, 21, 150, 221, 111, 76, 47, 59, 77, 112, 198, 147
6	90, 91, 214, 92, 152, 98, 243, 197, 32, 227, 162
7	103, 187, 223, 151, 158, 126
8	88, 167, 13, 164, 184, 123
9	58, 73, 195, 148, 31, 33
10	206, 188, 211, 161, 205, 44
11	50, 54, 78, 106, 174
12	209, 220, 146, 157
13	70, 154, 100, 117
14	128, 144, 118, 129
15	131, 66, 149, 145
16	89, 169, 163
17	141, 49, 72
18	196, 114, 86
19	55, 27, 215
20	228, 234
21	160, 213
22	207, 237
23	230, 94
24	200, 247
25	133, 143
26	64, 68
27	20, 45
28	60, 56
29	99, 52

B

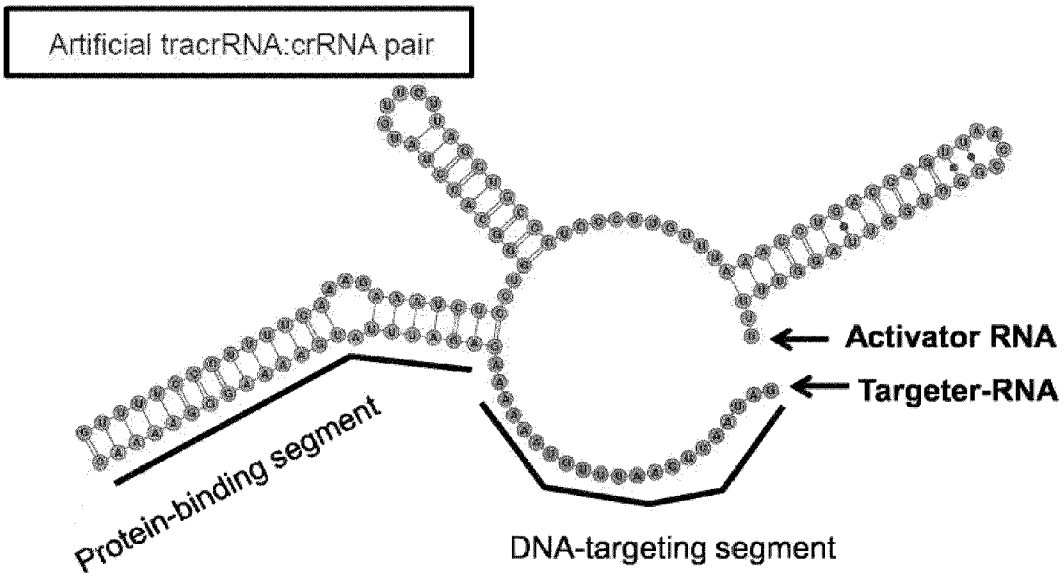
Figure 38

Cluster	SEQ ID NO:
30	244, 185
31	43
32	189
33	170
34	11
35	107
36	14
37	236
38	12
39	17
40	239
41	61
42	85
43	191
44	22

Cluster	SEQ ID NO:
45	233
46	122
47	16
48	242
49	203
50	26
51	137
52	199
53	34
54	201
55	178
56	42
57	190
58	81

Cluster	SEQ ID NO:
59	218
60	65
61	171
62	97
63	63
64	46
65	225
66	10
67	173
68	51
69	142
70	69
71	28
72	139
73	80
74	172
75	115
76	229
77	175
78	181

FIGURE 39 A


Artificial tracrRNA:crRNA pair
Activator RNA
Targeter-RNA
Protein-binding segment
DNA-targeting segment

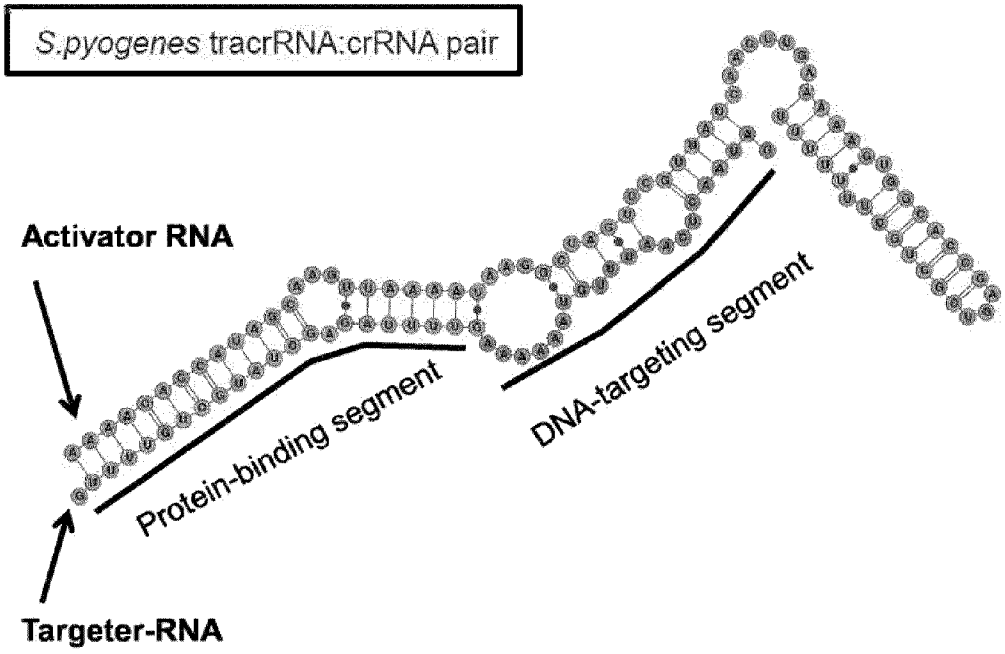
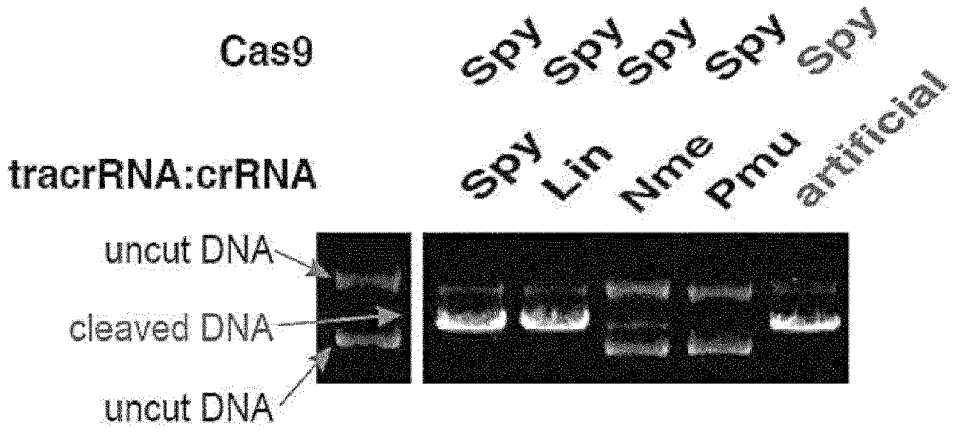

S.pyogenes tracrRNA:crRNA pair
Activator RNA
Protein-binding segment
DNA-targeting segment
Targeter-RNA

FIGURE 39 B


Cas9
Spy
Spy
Spy
Spy
Spy
tracrRNA:crRNA
Spy
Lin
Nme
Pmu
artificial
uncut DNA
cleaved DNA
uncut DNA

SEKVENSLISTE

Sekvenslisten er udeladt af skriftet og kan hentes fra det Europæiske Patent Register.

The Sequence Listing was omitted from the document and can be downloaded from the European Patent Register.