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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 June 2016 (30.06.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/106239 A1

(51) **International Patent Classification**:
A01N 63/00 (2006.01) *C12N 9/14* (2006.01)
C12N 1/21 (2006.01) *C07H 21/04* (2006.01)
C12N 15/00 (2006.01)

(21) **International Application Number**:
PCT/US2015/067161

(22) **International Filing Date**:
21 December 2015 (21.12.2015)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:
62/096,507 23 December 2014 (23.12.2014) US

(71) **Applicant**: **THE REGENTS OF THE UNIVERSITY OF CALIFORNIA** [US/US]; 1111 Franklin Street, 12th Floor, Oakland, California 94607 (US).

(72) **Inventors**: **NUNEZ, James K.**; 731A Stanley Hall, MC #220, Berkeley, California 94720 (US). **DOUDNA, Jennifer A.**; 708A Stanley Hall, MC #220, Berkeley, California 94720 (US).

(74) **Agent**: **BORDEN, Paula A.**; 1900 University Avenue, Suite 200, East Palo Alto, California 94303 (US).

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

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

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

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

WO 2016/106239 A1

(54) **Title**: METHODS AND COMPOSITIONS FOR NUCLEIC ACID INTEGRATION

(57) **Abstract**: The disclosure provides methods and compositions for the integration (insertion) of a donor DNA molecule into a target DNA molecule. In general, the methods include contacting a target DNA molecule with a linear donor DNA molecule and a Cas1 protein, where the target DNA molecule includes an AT-rich region (e.g., in some cases positioned 5 and within 50 nucleotides of a region that forms a DNA cruciform structure), where the contacting is not in a bacterial or archaeal cell (e.g., the contacting is in vitro outside of a cell, inside of a eukaryotic cell, etc.), and provides for integration of the donor DNA molecule into the target DNA molecule.

METHODS AND COMPOSITIONS FOR NUCLEIC ACID INTEGRATION

CROSS-REFERENCE

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 62/096,507, filed December 23, 2014, which application is incorporated herein by reference in its entirety.

INCORPORATION BY REFERENCE OF SEQUENCE LISTING PROVIDED AS A TEXT FILE

[0002] A Sequence Listing is provided herewith as a text file, "BERK-270WO SeqList_ST25.txt" created on December 21, 2015 and having a size of 268 KB. The contents of the text file are incorporated by reference herein in their entirety.

INTRODUCTION

[0003] Prokaryotic adaptive immunity relies on clustered regularly interspaced short palindromic repeats (CRISPRs) together with CRISPR associated (Cas) proteins to detect and destroy foreign nucleic acids. CRISPR loci contain an A-T-rich leader sequence followed by repetitive sequence elements that flank spacer segments, each about 30 base pairs (bp) in length, that are transcribed to produce precursor CRISPR RNAs (pre-crRNAs). Spacers are frequently virus- or plasmid-derived, although "self"-derived spacers from the host chromosome are present in some CRISPR loci. After pre-crRNA processing and assembly with Cas proteins, the resulting surveillance complexes target and cleave foreign nucleic acids bearing sequences complementary to the crRNA spacer sequence. There is a need in the art for compositions and methods that utilize a Cas protein to facilitate the integration of foreign DNA into target DNA.

Publications

[0004] Nunez et al., Nat Struct Mol Biol. 2014 Jun;21(6):528-34, Epub 2014 May 4; Yosef et al., Nucleic Acids Res. 2012 Jul;40(12):5569-76; Krupovic et al., BMC Biol. 2014 May 19;12:36; Swarts et al., PLoS One. 2012;7(4):e35888, Epub 2012 Apr 27; Datsenko et al., Nat Commun. 2012 Jul 10;3:945; Barrangou et al., Science. 2007 Mar 23;315(5819):1709-12.

SUMMARY

[0005] The disclosure provides methods and compositions for the integration (insertion) of a donor DNA molecule into a target DNA molecule. In general, the methods include contacting a

target DNA molecule with a linear donor DNA molecule and a Cas1 protein, where the target DNA molecule includes an AT-rich region (e.g., in some cases positioned 5' and within 50 nucleotides of a region that forms a DNA cruciform structure), where the contacting is not in a bacterial or archaeal cell (e.g., the contacting is *in vitro* outside of a cell, inside of a eukaryotic cell, etc.), and provides for integration of the donor DNA molecule into the target DNA molecule. In some cases, the contacting is performed in the presence of a Cas2 protein. In some cases, the contacting includes introducing into a target cell: (i) the Cas1 protein, or a nucleic acid having nucleotides that encode the Cas1 protein; and (ii) a linear donor DNA molecule. In some cases, the linear donor DNA molecule (e.g., introduced into a cell) has a length in a range of from 10 to 500 nucleotides (nt) (e.g., 35 to 500 nt). In some cases, the linear donor DNA molecule (e.g., introduced into a cell) has a length that is greater than 35 nucleotides (nt). In some cases, the linear donor DNA molecule includes a 3' overhang with a length of from 1 to 6 nucleotides. In some cases, the method includes introducing into the target cell a Cas2 protein, or a nucleic acid having nucleotides that encode a Cas2 protein. In some cases, the method is performed in the presence of an integration host factor (IHF) protein. Thus, in some cases, the method includes introducing into the target cell an IHF protein, or a nucleic acid comprising nucleotides that encode an IHF protein (e.g., an expression vector in which the nucleotides encoding the IHF protein are operably linked to a promoter, e.g., a promoter operable in a eukaryotic cell).

[0006] In some cases, the method includes introducing the Cas1 protein and/or the Cas2 protein, into a target cell. In some cases, a Cas1 protein and a linear donor DNA molecule are introduced into a target cell as a targeting composition including the Cas1 protein and the linear donor DNA molecule. In some cases, the targeting composition further includes a Cas2 protein. In some cases, the Cas1 protein and/or the Cas2 protein is a protein that is isolated (e.g., purified) from a cell. In some cases, the Cas1 protein and/or the Cas2 protein, has an affinity tag. In some cases, the method includes, prior to the contacting step, a step of isolating the Cas1 protein and/or the Cas2 protein from a cell. In some such cases, the Cas1 protein and/or the Cas2 protein has an affinity tag during the isolating step. In some cases, the method includes a step of removing one or more affinity tags (e.g., via cleavage) prior to the contacting step.

[0007] In some cases, the method includes a step of introducing into a target cell a nucleic acid comprising a nucleotide sequence that encodes the Cas1 protein and/or a nucleic acid comprising a nucleotide sequence that encodes a Cas2 protein. In some cases, a nucleotide sequence that encodes the Cas1 protein and a nucleotide sequence that encodes the Cas2 protein are present on the same nucleic acid molecule (e.g., on a recombinant expression vector). In some cases, the

nucleotide sequence that encodes the Cas1 protein and/or the nucleotide sequence that encodes the Cas2 protein is operably linked to a promoter that is operable in the target cell.

[0008] In some embodiments, the target DNA molecule does not contain a leader sequence from a naturally existing CRISPR locus. In some cases, the target DNA molecule does not contain a repeat sequence from a naturally existing CRISPR locus. In some cases, the target DNA does not contain a leader sequence or a CRISPR repeat sequence from a naturally existing CRISPR locus. In some cases, the target DNA molecule does not contain a naturally existing CRISPR locus.

[0009] Kits are also provided for practicing the subject methods.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] **Figures 1A-E** provide evidence related the Cas1-Cas2 complex integrating protospacers *in vitro*.

[0011] **Figures 2A-H** provide evidence related to half-site integration, full-site integration, and pCRISPR topoisomer products. Note, the gel depicted in panel E was pre-stained with ethidium bromide (EtBr) while the gel depicted in panel F was post-stained with EtBr.

[0012] **Figures 3A-E** provide evidence related to integration requiring 3'-OH protospacer ends and supercoiled target DNA.

[0013] **Figures 4A-E** provide evidence related to protospacers specifically integrating into the CRISPR locus. Figure 4B: top (SEQ ID NO: 230); bottom (SEQ ID NO: 231). Figure 4E: top left (SEQ ID NO: 232); top right (SEQ ID NO: 233); bottom left (SEQ ID NO: 234); bottom right (SEQ ID NO: 235).

[0014] **Figure 5** provides a schematic of a model of protospacer integration during CRISPR-Cas adaptive immunity.

[0015] **Figures 6A-E** provide evidence related to the integration reaction being dependent on the presence of protospacers, low salt and divalent metal ions.

[0016] **Figures 7A-C** provide evidence related to Cas1 and Cas2 together providing for robust protospacer integration.

[0017] **Figures 8A-C** provide evidence related to the catalytic activity of Cas1 being required for integration.

[0018] **Figures 9A-D** provide evidence related to band X corresponding to topoisomers of pCRISPR.

[0019] **Figures 10A-D** provide evidence related to Cas1 catalyzing the disintegration of half-site integrated protospacers.Figure 10A: Strand A (SEQ ID NO: 236); Strand B (SEQ ID NO:

237); Strand C top (SEQ ID NO: 238); Strand C middle (SEQ ID NO: 239); Strand C bottom (SEQ ID NO: 240); Strand D (SEQ ID NO: 241).

[0020] **Figures 11A-E** provide evidence related to Cas1–Cas2 being able to integrate various lengths of double-stranded DNA with blunt and 3′-overhang ends into a supercoiled target plasmid.

[0021] **Figures 12A-C** provide evidence related to Tyr residues in the vicinity of the Cas1 active site. Figure 12B: top to bottom (SEQ ID NO: 242-246).

[0022] **Figures 13A-H** provide evidence related to high-throughput sequencing of integration products revealing sequence-specific integration. Figure 13D: top (SEQ ID NO: 247); bottom (SEQ ID NO: 248).

[0023] **Figures 14A-F** provide evidence related to Cas1 and Cas2 correctly orienting the protospacer during integration. Figure 14A: top (SEQ ID NO: 249); bottom (SEQ ID NO: 250). Figure 14C: top (SEQ ID NO: 251); bottom (SEQ ID NO: 252). Figure 14E: top (SEQ ID NO: 253); bottom (SEQ ID NO: 254).

[0024] **Figure 15** provides a schematic model of the CRISPR-Cas adaptive immunity pathway in *E. coli.*

[0025] **Figure 16** provides Cas1 and Cas2 protein sequences from various species.

[0026] **Figure 17** provides leader sequences and repeat sequences from various species.

DEFINITIONS

[0027] 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.

[0028] The term “oligonucleotide” refers to a polynucleotide of between 3 and 150 nucleotides of single- or double-stranded nucleic acid (e.g., DNA, RNA, or a modified nucleic acid). 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, transcribed (*in vitro* and/or *in vivo*), or chemically synthesized. 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.

[0029] 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 (stem 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. In some cases, a stem-loop structure forms where there is an inverted repeat sequence. The intervening sequence of nucleotides between the initial sequence and the reverse complement of an inverted repeat can be any length including zero, and when the intervening length is zero, the composite sequence is a palindromic sequence. The length of the intervening sequence of nucleotides (i.e., the number of intervening nucleotides) determines the size of the loop portion of the stem-loop. When referring to a double stranded nucleic acid molecule (e.g., a double stranded DNA molecule), a "cruciform structure" (e.g., a DNA cruciform structure) can be formed when both strands form a stem-loop structure at the same location in the molecule. For example, an inverted repeat sequence on one strand of a double stranded DNA will lead to a stem-loop structure in both strands (and therefore a cruciform structure) because the second strand is the reverse complement of the first strand.

[0030] By "hybridizable" or "complementary" or "substantially complementary" it is meant that a nucleic acid (e.g. RNA, DNA) 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. 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, for hybridization between two RNA molecules (e.g., dsRNA), and for hybridization of a DNA molecule with an RNA molecule: guanine (G) can also base pair 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. Thus, in the context of this disclosure, a guanine (G) is considered complementary to both a uracil (U) and to an adenine (A).

[0031] 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.

[0032] 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 complementarity, variables well known in the art. The greater the degree of complementarity 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 can become important (see Sambrook et al., supra, 11.7-11.8). Typically, the length for a hybridizable nucleic acid is 8 nucleotides or more (e.g., 10 nucleotides or more, 12 nucleotides or more, 15 nucleotides or more, 20 nucleotides or more, 22 nucleotides or more, 25 nucleotides or more, or 30 nucleotides or more). 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.

[0033] It is understood that the sequence of a 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 60% or more, 65% or more, 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 98% or more, 99% or more, 99.5% or more, or 100% sequence complementarity to a target region within the target nucleic acid sequence to which it will hybridize. 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 using any convenient method. Examples of methods include BLAST programs (basic local alignment search tools) and PowerBLAST programs (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).

[0034] 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.

[0035] "Binding" as used herein refers to a non-covalent interaction between macromolecules. 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 (K_d) 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^{-13} 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 K_d.

[0036] 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 domain), an RNA molecule (an RNA-binding domain) and/or a protein molecule (a protein-binding domain). In the case of a protein having a protein-binding domain, it can in some cases bind to itself (to form homodimers, homotrimers, etc.) and/or it can bind to one or more regions of a different protein or proteins.

[0037] 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-glycine, and asparagine-glutamine.

[0038] 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 ways. 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" followed by ".gov/BLAST"; "ebi.ac"followed by ".uk/Tools/msa/tcoffee/"; "ebi.ac." followed by "uk/Tools/msa/muscle/"; and "mafft.cbrc" followed by ".jp/alignment/software". See, e.g., Altschul et al. (1990), J. Mol. Bioi. 215:403-10.

[0039] 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, microRNA (miRNA), a "non-coding" RNA (ncRNA), a guide nucleic acid, etc.).

[0040] 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.

[0041] 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., guide nucleic acid) or a coding sequence (e.g., Cas9 polypeptide, or Cas9 polypeptide) and/or regulate translation of an encoded polypeptide.

[0042] 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.

[0043] The term "naturally-occurring" or “unmodified” or “wild type” 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 wild type (and 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, the chimeric polypeptide includes amino acid sequences that are derived from different polypeptides. A chimeric polypeptide can comprise modified and/or naturally-occurring polypeptide sequences. Similarly, "chimeric" in the context of a polynucleotide encoding a chimeric polypeptide includes nucleotide sequences derived from different coding regions.

[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. The heterologous polypeptide sequence may exhibit an activity (e.g., enzymatic activity) that will also be exhibited by the chimeric protein. 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.

[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 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 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 "host cell" or “target cell” as used herein, denotes an in vivo or in vitro eukaryotic cell, a prokaryotic cell (e.g., bacterial or archaeal cell), or a cell from a multicellular organism (e.g., a cell line) cultured as a unicellular entity, which eukaryotic or prokaryotic cells can be, or have been, used as recipients for a nucleic acid, and include the progeny of the original cell which has been transformed by the nucleic acid. It is understood that the progeny of a single cell

may not necessarily be completely identical in morphology or in genomic or total DNA complement as the original parent, due to natural, accidental, or deliberate mutation. A "recombinant host cell" (also referred to as a "genetically modified host cell") is a host cell into which has been introduced a heterologous nucleic acid, e.g., an expression vector. For example, a subject bacterial host cell is a genetically modified bacterial host cell by virtue of introduction into a suitable bacterial host cell of an exogenous nucleic acid (e.g., a plasmid or recombinant expression vector) and a subject eukaryotic host cell is a genetically modified eukaryotic host cell (e.g., a mammalian germ cell), by virtue of introduction into a suitable eukaryotic host cell of an exogenous nucleic acid.

[0055] 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.

[0056] 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.).

[0057] 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 (Thomson et. al, Science. 1998 Nov 6;282(5391):1145-7) 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, as well as the term embryonic germ stem cells (EGSC), which are another example of a PSC. 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.

[0058] By "embryonic stem cell" (ESC) is meant a PSC that was isolated from an embryo, typically from the inner cell mass of the blastocyst. ESC lines are listed in the NIH Human Embryonic Stem Cell Registry, e.g. hESBGN-01, hESBGN-02, hESBGN-03, hESBGN-04 (BresaGen, Inc.); HES-1, HES-2, HES-3, HES-4, HES-5, HES-6 (ES Cell International); Miz-hES1 (MizMedi Hospital-Seoul National University); HSF-1, HSF-6 (University of California at San Francisco); and H1, H7, H9, H13, H14 (Wisconsin Alumni Research Foundation (WiCell Research Institute)). 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, the disclosures of which are incorporated herein by reference. Methods for proliferating hESCs in the undifferentiated form are described in WO 99/20741, WO 01/51616, and WO 03/020920.

[0059] 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, the disclosures of which are incorporated herein by reference.

[0060] 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, the disclosures of which are incorporated herein by reference. 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.

[0061] 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.

[0062] 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.

[0063] 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.

[0064] 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.

[0065] In some instances, a component (e.g., a donor DNA molecule, a protein component (e.g., a Cas1 and/or a Cas2 protein), and the like) includes a label moiety. The terms "label", "detectable label", or "label moiety" as used herein refer to any moiety that provides for signal detection and may vary widely depending on the particular nature of the assay. Label moieties of interest include both directly detectable labels (direct labels)(e.g., a fluorescent label) and indirectly detectable labels (indirect labels)(e.g., a binding pair member). A fluorescent label can be any fluorescent label (e.g., a fluorescent dye (e.g., fluorescein, Texas red, rhodamine, ALEXAFLUOR® labels, and the like), a fluorescent protein (e.g., GFP, EGFP, YFP, RFP, CFP, YFP, cherry, tomato, tangerine, and any fluorescent derivative thereof), etc.). Suitable detectable (directly or indirectly) label moieties for use in the methods include any moiety that is detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical, chemical, or other means. For example, suitable indirect labels include biotin (a binding pair member), which can be bound by streptavidin (which can itself be directly or indirectly labeled). Labels can also include: a radiolabel (a direct label)(e.g., ^{3}H, ^{125}I, ^{35}S, ^{14}C, or ^{32}P); an enzyme (an indirect label)(e.g., peroxidase, alkaline phosphatase, galactosidase, luciferase, glucose oxidase, and the like); a fluorescent protein (a direct label)(e.g., green fluorescent protein, red fluorescent protein, yellow fluorescent protein, and any convenient derivatives thereof); a metal label (a direct label); a colorimetric label; a binding pair member; and the like. By "partner of a binding pair" or "binding pair member" is meant one of a first and a second moiety, wherein the first and the second moiety have a specific binding affinity for each other. Suitable binding pairs include, but are not limited to: antigen/antibodies (for example, digoxigenin/anti-digoxigenin, dinitrophenyl (DNP)/anti-DNP, dansyl-X-anti-dansyl, fluorescein/anti-fluorescein, lucifer yellow/anti-lucifer yellow, and rhodamine anti-rhodamine), biotin/avidin (or biotin/streptavidin) and calmodulin binding protein (CBP)/calmodulin. Any binding pair member can be suitable for use as an indirectly detectable label moiety.

[0066] Any given component, or combination of components can be unlabeled, or can be detectably labeled with a label moiety. In some cases, when two or more components are labeled, they can be labeled with label moieties that are distinguishable from one another.

[0067] 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), the disclosures of which are incorporated herein by reference.

[0068] 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.

[0069] 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.

[0070] 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. All publications mentioned herein are incorporated herein by reference to disclose and describe the methods and/or materials in connection with which the publications are cited.

[0071] 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 "a cell" includes a plurality of such cells and reference to "the protein" includes reference to one or more proteins 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.

[0072] 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.

[0073] 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

[0074] The disclosure provides methods and compositions for the integration (insertion) of a donor DNA molecule into a target DNA molecule. In general, the methods include contacting a target DNA molecule with a linear donor DNA molecule and a Cas1 protein, where the target DNA molecule includes an AT-rich region (e.g., in some cases positioned 5' and within 50 nucleotides of a region that forms a DNA cruciform structure), where the contacting is not in a bacterial or archaeal cell (e.g., the contacting is *in vitro* outside of a cell, inside of a eukaryotic cell, etc.), and provides for integration of the donor DNA molecule into the target DNA molecule. In some cases, the contacting is performed in the presence of a Cas2 protein. "Cas1" protein refers to CRISPR associated (Cas) protein 1, and "Cas2" protein refers to CRISPR associated (Cas) protein 2.

Proteins

Cas1 protein

[0075] A Cas1 polypeptide (used interchangeably with the term "Cas1 protein") used in the methods described herein can be any Cas1 protein (i.e., a Cas1 protein from any species). Cas1 proteins have an N-terminal β-sheet domain and a C-terminal α-helical domain [Wiedenheft, B. *et al.* , *Structure* **17**, 904–912 (2009); Babu, M. *et al., Mol. Microbiol.* **79**, 484–502 (2011); Kim et al, *Biochem. Biophys. Res. Commun.* **441**, 720–725 (2013)].

[0076] In some embodiments, a Cas1 protein is from an archaeal microorganism. In some embodiments, a Cas1 protein is from a Euryarchaeota microorganism. In some embodiments, a Cas1 protein is from a Crenarchaeota microorganism. In some embodiments, a Cas1 protein is from a bacterium. In some embodiments, a Cas1 polypeptide has an amino acid sequence having 60% or more amino acid sequence identity (e.g.,. 60% or more, 65% or more, 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 98% or more, 99% or more, 99.5% or more, or 100% amino acid sequence identity) to a Cas1 protein amino acid sequence as set forth in any one of SEQ ID NOs: 28-86 (e.g., see Figure 16). In certain embodiments, Cas1 protein includes an amino acid sequence as set forth in any one of SEQ ID NOs: 28-86.

[0077] In certain embodiments, Cas1 protein may be a "functional derivative" of a naturally occurring Cas1 protein. A "functional derivative" of a native sequence polypeptide is a compound having a qualitative biological property in common with a native sequence polypeptide. "Functional derivatives" include, but are not limited to, fragments of a native sequence and derivatives of a native sequence polypeptide and its fragments, provided that they have a biological activity in common with a corresponding native sequence polypeptide. The term "derivative" encompasses both amino acid sequence variants of polypeptide, covalent modifications, and fusions thereof. A "fusion" polypeptide is a polypeptide comprising a polypeptide or portion (e.g., one or more domains) thereof fused or bonded to heterologous polypeptide (e.g., an affinity tag).

[0078] "Cas1 protein" encompasses a full-length Cas1 polypeptide, an enzymatically active fragment of a Cas1 polypeptide, and enzymatically active derivatives of a Cas1 polypeptide or fragment thereof. Suitable derivatives of a Cas1 polypeptide or a fragment thereof include but are not limited to mutants, fusions, covalent modifications of Cas1 protein or a fragment thereof. Cas1 protein which includes Cas1 protein or a fragment thereof, as well as derivatives of Cas1 protein or a fragment thereof, may be obtainable from a cell or synthesized chemically or by a combination of these two procedures. The cell may be a cell that naturally produces Cas1 protein, or a cell that naturally produces Cas1 protein and is genetically engineered to produce the endogenous Cas1 protein at a higher expression level or to produce a Cas1 protein from an exogenously introduced nucleic acid, which nucleic acid encodes a Cas1 that is same or different from the endogenous Cas1. In some case, the cell does not naturally produce Cas1 protein and is genetically engineered to produce a Cas1 protein.

Cas2 protein

[0079] A Cas2 polypeptide (used interchangeably with the term "Cas2 protein") used in the methods described herein can be any Cas2 protein (i.e., a Cas2 protein from any species). Cas2

proteins form symmetrical homodimers with a core ferredoxin fold [Beloglazova, N. *et al., J. Biol. Chem.* ***283****, 20361–20371 (2008)*; Nam, K.H. *et al.* , *J. Biol. Chem.* **287**, 35943–35952 (2012); and Samai, P., et al., *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1552–1556 (2010)].

[0080] In some embodiments, a Cas2 protein is from an archaeal microorganism. In some embodiments, a Cas2 protein is from a Euryarchaeota microorganism. In some embodiments, a Cas2 protein is from a Crenarchaeota microorganism. In some embodiments, a Cas2 protein is from a bacterium. In some embodiments, a Cas2 polypeptide has an amino acid sequence having 60% or more amino acid sequence identity (e.g.,. 60% or more, 65% or more, 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 98% or more, 99% or more, 99.5% or more, or 100% amino acid sequence identity) to a Cas2 protein amino acid sequence as set forth in any one of SEQ ID NOs: 87-161 (e.g., see Figure 16). In certain embodiments, Cas2 protein includes an amino acid sequence as set forth in any one of SEQ ID NOs: 87-161.

[0081] In certain embodiments, Cas2 protein may be a "functional derivative" of a naturally occurring Cas2 protein. A "functional derivative" of a native sequence polypeptide is a compound having a qualitative biological property in common with a native sequence polypeptide. "Functional derivatives" include, but are not limited to, fragments of a native sequence and derivatives of a native sequence polypeptide and its fragments, provided that they have a biological activity in common with a corresponding native sequence polypeptide. The term "derivative" encompasses both amino acid sequence variants of polypeptide, covalent modifications, and fusions thereof. A "fusion" polypeptide is a polypeptide comprising a polypeptide or portion (e.g., one or more domains) thereof fused or bonded to heterologous polypeptide (e.g., an affinity tag).

[0082] "Cas2 protein" encompasses a full-length Cas2 polypeptide, an enzymatically active fragment of a Cas2 polypeptide, and enzymatically active derivatives of a Cas2 polypeptide or fragment thereof. Suitable derivatives of a Cas2 polypeptide or a fragment thereof include but are not limited to mutants, fusions, covalent modifications of Cas2 protein or a fragment thereof. Cas2 protein which includes Cas2 protein or a fragment thereof, as well as derivatives of Cas2 protein or a fragment thereof, may be obtainable from a cell or synthesized chemically or by a combination of these two procedures. The cell may be a cell that naturally produces Cas2 protein, or a cell that naturally produces Cas2 protein and is genetically engineered to produce the endogenous Cas2 protein at a higher expression level or to produce a Cas2 protein from an exogenously introduced nucleic acid, which nucleic acid encodes a Cas2 that is same or different

from the endogenous Cas2. In some case, the cell does not naturally produce Cas2 protein and is genetically engineered to produce a Cas2 protein.

[0083] Mutants (variants) of Cas1 protein and/or Cas2 proteins may be generated by performing conservative substitutions. By conservative substitutions is intended combinations such as those from the following groups: gly, ala; val, ile, leu; asp, glu; asn, gln; ser, thr; lys, arg; and phe, tyr. Amino acids that are not present in the same group are "substantially different" amino acids. In certain cases, the conserved residues may not be substituted and the substitutions limited to the non-conserved residues.

[0084] A subject Cas1 and/or Cas2 protein can be a variant protein by virtue of being fused to a heterologous sequence. For example, a Cas1 and/or a Cas2 protein can have a label (e.g., as defined above, e.g., can have an affinity tag, can be fused to a fluorescent protein, can include a fluorescent dye label, and the like). In some cases, a subject Cas1 and/or Cas2 protein includes (i.e., is fused to) a heterologous sequence that provides for subcellular localization (e.g., a nuclear localization signal (NLS) for targeting to the nucleus; a mitochondrial localization signal for targeting to the mitochondria; a chloroplast localization signal for targeting to a chloroplast; an ER retention signal; and the like). In some cases, a subject Cas1 and/or Cas2 protein includes 2 or more, 3 or more, 4 or more, or 5 or more NLSs. In some cases, an NLS is located at or near (e.g., within 75 amino acids, 50 amino acids, or 30 amino acids) the N-terminus and/or at or near (e.g., within 75 amino acids, 50 amino acids, or 30 amino acids) the C-terminus.

[0085] A Cas1 protein and/or Cas2 protein can be provided as a protein. For example, in some cases, a target DNA molecule is contacted in vitro, outside of a cell with a Cas1 protein and/or a Cas2 protein. In some cases, a Cas1 protein and/or a Cas2 protein is introduced into a cell. In some cases, a Cas1 protein and/or a Cas2 protein is introduced into a cell in a composition that also includes a donor DNA molecule.

[0086] In some embodiments, the Cas1 protein and/or a Cas2 protein can be purified (isolated) from an organism. The organism (e.g., a bacterial cell, an archaeal cell) may be producing the Cas1 protein from an endogenous gene or from an exogenous gene. The exogenous gene may be present in the organism transiently or stably. For example, a polynucleotide encoding a Cas1 protein and/or a polynucleotide encoding a Cas2 protein can be introduced into a suitable expression vector. The expression vector can be introduced into a suitable cell, and a Cas1 and/or Cas2 protein can be isolated. Cas1 protein and/or Cas2 protein may be recovered and purified from recombinant cell cultures by any convenient method, e.g., including ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, high performance

liquid chromatography, affinity chromatography, protein G affinity chromatography, for example, hydroxyapatite chromatography and lectin chromatography, etc.

[0087] Cas1 protein and/or Cas2 protein may also be recovered from: products of purified cells, whether directly isolated or cultured; products of chemical synthetic procedures; and products produced by recombinant techniques from a prokaryotic or eukaryotic host, including, for example, bacterial, yeast higher plant, insect, and mammalian cells.

[0088] In some cases, a Cas1 protein and/or a Cas2 protein includes a label (e.g., an affinity tag) that can be used to facilitate purification. In some cases, the label can be removed. For example, in some cases, a cleavage site (e.g., a tobacco etch virus (TEV) protease cleavage site) existing between the affinity tag and the rest of the protein can be used for cleavage (removal of the label) prior to use of the protein.

[0089] In some cases, a subject method includes a step of isolating (purifying) a Cas1 protein and/or a Cas2 protein prior to contacting a target DNA molecule with the protein. In some cases, contacting a target DNA molecule with a Cas1 and/or a Cas2 protein includes introducing into a cell one or nucleic acids (e.g., RNA, DNA) that include nucleotide sequences encoding a Cas1 protein and/or a Cas2 protein.

Nucleic Acids

Donor DNA molecule

[0090] A subject donor DNA molecule is a linear DNA molecule (in some cases double stranded, in some cases single stranded). A subject donor DNA molecule is a linear molecule (e.g., not a circular molecule such as a plasmid DNA). A donor DNA molecule can have any desired sequence. In some cases, the 3' most nucleotide on at least one end of the donor DNA molecule is a C. In some cases, the 3' most nucleotide on one and only one end of the donor DNA molecule is a C. In some cases, the 3' most nucleotide on at least one end of the donor DNA molecule is a G. In some cases, the 3' most nucleotide on one and only one end of the donor DNA molecule is a G. In some cases, the 3' most nucleotide on at least one end of the donor DNA molecule is an A. In some cases, the 3' most nucleotide on one and only one end of the donor DNA molecule is an A. In some cases, the 3' most nucleotide on at least one end of the donor DNA molecule is a T. In some cases, the 3' most nucleotide on one and only one end of the donor DNA molecule is a T.

[0091] In some cases, the linear donor DNA molecule has a length in a range of from 10 to 1000 nucleotides (nt) (e.g., 15 to 500, 20 to 500, 30 to 500, 33 to 500, 35 to 500, 40 to 500, 45 to 500, 50 to 500, 15 to 250, 20 to 250, 30 to 250, 33 to 250, 35 to 250, 40 to 250, 45 to 250, 50 to 250, 15 to 150, 20 to 150, 30 to 150, 33 to 150, 35 to 150, 40 to 150, 45 to 150, 50 to 150, 15 to

100, 20 to 100, 30 to 100, 33 to 100, 35 to 100, 40 to 100, 45 to 100, 50 to 100, 15 to 50, 20 to 50, 30 to 50, 33 to 50, 35 to 50, 40 to 50, or 45 to 50 nt). In some cases, a subject method includes introducing into a cell a subject linear donor DNA molecule. In some cases, a donor DNA molecule includes a label (e.g., as defined above, e.g., a biotin label, a fluorescent dye, etc.).

[0092] In some cases, the linear donor DNA molecule includes a 3'-overhang. For example, in some cases, the linear donor DNA molecule includes a 3'-overhang having a length in a range of from 1 to 6 nucleotides (nt) (e.g., 1 to 5 nt, 1 to 4 nt, 1 to 3 nt, 1 to 2 nt, 2 to 6 nt, 2 to 5 nt, 2 to 4 nt, 2 to 3 nt, 3 to 6 nt, 3 to 5 nt, 3 to 4 nt, 4 to 6 nt, 4 to 5 nt, 5 to 6 nt, 1 nt, 2 nt, 3 nt, 4 nt, 5 nt, or 6 nt). In some cases, the linear donor DNA molecule does not have a 3'-overhang. Thus, in some cases, the linear donor DNA molecule includes a 3'-overhang having a length in a range of from 0 to 6 nucleotides (nt) (e.g., 0 to 5 nt, 0 to 4 nt, 0 to 3 nt, 0 to 2 nt, 0 to 1 nt, 1 to 6 nt, 1 to 5 nt, 1 to 4 nt, 1 to 3 nt, 1 to 2 nt, 2 to 6 nt, 2 to 5 nt, 2 to 4 nt, 2 to 3 nt, 3 to 6 nt, 3 to 5 nt, 3 to 4 nt, 4 to 6 nt, 4 to 5 nt, 5 to 6 nt, 1 nt, 2 nt, 3 nt, 4 nt, 5 nt, or 6 nt).

Target DNA molecule

[0093] A subject donor DNA molecule is any supercoiled target DNA (e.g., a plasmid DNA, chromosomal DNA, etc.). As shown in the working examples below, a Cas1 protein (sometimes in combination with a Cas2 protein) biases insertion of a linear donor DNA molecules to a region abutting a region of the target DNA molecule having an AT-rich region (e.g., a leader sequence from a CRISPR locus) upstream of a region that forms a DNA cruciform structure (e.g., a repeat sequence from a CRISPR locus).

[0094] By "AT-rich" is meant greater than 40% AT content (e.g., 41% or more, 45% or more, 50% or more, 51% or more, 52% or more, 53% or more, 54% or more, 55% or more, 56% or more, 57% or more, 58% or more, 59% or more, or 60% or more AT content) over a stretch of at least 40 base pairs (e.g., at least 50 base pairs, at least 60 base pairs, at least 65 base pairs, at least 70 base pairs, at least 75 base pairs, at least 80 base pairs, etc.). Naturally existing leader sequences from CRISPR loci are set forth in SEQ ID NOs: 163-188. In some cases, the entire leader sequence from a naturally occurring leader sequence is not necessary. For example, in some cases, 60 bp is sufficient size for an AT-rich region. Thus, in some cases, an AT-rich region as referred to herein includes a sequence that has 60% or more sequence identity (65% or more, 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 98% or more, 99% or more, or 100% sequence identity) with a leader sequence from a CRISPR locus (e.g., a leader sequence set forth in any of SEQ ID NOs: 163-188) over a stretch of at least 40 base pairs (e.g., at least 50 base pairs, at least 60 base pairs, at least 65 base pairs, at least 70 base

pairs, at least 75 base pairs, at least 80 base pairs, etc.), and/or that is similar to such a sequence with regard to AT content.

[0095] In some cases, a subject AT-rich region is positioned 5' (upstream of) and within 50 nucleotides (nt) (e.g., within 40 nt, 30 nt, 20 nt, 15 nt, 10 nt, 5 nt, 2 nt, or immediately adjoining) of a region that forms a DNA cruciform structure. In some cases, a subject AT-rich region is not positioned within 50 nucleotides (nt) (e.g., within 40 nt, 30 nt, 20 nt, 15 nt, 10 nt, 5 nt, 2 nt, or immediately adjoining) of a region that forms a DNA cruciform structure. In some cases, a subject AT-rich region is a naturally occurring AT-rich region positioned 5' (upstream of) and within 50 nucleotides (nt) (e.g., within 40 nt, 30 nt, 20 nt, 15 nt, 10 nt, 5 nt, 2 nt, or immediately adjoining) of a region that forms a DNA cruciform structure, and the region that forms a DNA cruciform structure is not naturally found near or in the same species as the AT-rich region. For example, in some cases a subject target DNA molecule includes a naturally occurring AT-rich region and a naturally occurring cruciform-form region, but those regions to do not naturally occur together (e.g., within 50 nucleotides of one another).

[0096] A target DNA molecule can be present in a living cell, or can be isolated from a living cell. For example, the DNA substrate can be present in a cell lysate.

[0097] In some cases, a subject target DNA molecule includes an AT-rich region positioned 5' (upstream of) and within 50 nucleotides (nt) (e.g., within 40 nt, 30 nt, 20 nt, 15 nt, 10 nt, 5 nt, 2 nt, or immediately adjoining) of a region that forms a DNA cruciform structure. The donor DNA molecule is inserted next to the DNA cruciform structure (e.g., at or near the base of the stem structure) (e.g., as described in the working examples, one strand of a double stranded donor DNA is inserted upstream of the cruciform structure on one strand (strand A) of the target DNA molecule, and the other strand of the donor DNA is inserted downstream of the cruciform structure on the other strand (strand B) of the target DNA molecule). In some cases, a subject target DNA molecule includes an AT-rich region that is not positioned 5' (upstream of) and within 50 nucleotides (nt) (e.g., within 40 nt, 30 nt, 20 nt, 15 nt, 10 nt, 5 nt, 2 nt, or immediately adjoining) of a region that forms a DNA cruciform structure (i.e., in some cases a subject target DNA molecule includes an AT-rich region that is not 5' of and within 50 nt of a region that forms a DNA cruciform structure.

[0098] As noted above, when referring to a double stranded nucleic acid molecule (e.g., a double stranded DNA molecule), a "cruciform structure" (e.g., a DNA cruciform structure) can be formed when both strands form a stem-loop structure at the same location in the molecule. For example, an inverted repeat sequence on one strand of a double stranded DNA will lead to a stem-loop structure in both strands (and therefore a cruciform structure can form) because the

second strand is the reverse complement of the first strand. In some cases, the length of the upper and lower stem of the DNA cruciform structure is in a range of from 3 to 30 base pairs (bp)(e.g., 5 to 25 bp, 5 to 20 bp, 5 to 15 bp, 5 to 10 bp, 5 to 7 bp, 3 to 25 bp, 3 to 20 bp, 3 to 15 bp, 3 to 10 bp, 3 to 7 bp, or 3 to 5 bp). In some cases, the complementarity in the stem region of the stem-loop is 70% or greater (e.g., 80% or greater, 90% or greater, 95% or greater, 98% or greater, 99% or greater, 99.5% or greater, or 100%). In some cases, region that forms a DNA cruciform structure includes a sequence that has 60% or more sequence identity (65% or more, 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 98% or more, 99% or more, or 100% sequence identity) with a repeat sequence from a CRISPR locus (e.g., a repeat sequence set forth in any of SEQ ID NOs: 189-209), and/or that is structurally similar to such a sequence.

Using nucleic acids

[0099] In some cases, a subject method includes a step of introducing into a target cell (e.g., a eukaryotic cell) one or more nucleic acids (e.g., a subject donor DNA molecule, a nucleic acid that includes nucleotide sequences encoding a Cas1 protein and/or a Cas2 protein, etc.). Methods of introducing a nucleic acid into a cell are known in the art and any convenient method can be used (e.g., electroporation, lipofection, nucleofection, injection, viral vectors, etc.). In some cases, a subject DNA molecule is introduced into a cell in a composition that also includes a Cas1 protein and/or a Cas2 protein.

[00100] When one or more nucleic acids are used that include nucleotides encoding a Cas1 and/or a Cas2 protein, the sequence encoding the Cas1 and/or the Cas2 protein can be codon-optimized. A sequence encoding any suitable Cas1 and/or Cas2 protein can be codon optimized. As a non-limiting example, if the intended host cell were a mouse cell, then a mouse codon-optimized nucleotide sequence encoding a Cas1 and/or Cas2 (or variant thereof) would be suitable. While codon optimization is not required, it is acceptable and may be preferable in certain cases.

[00101] In some embodiments, one or more of the above nucleic acids a recombinant expression vector. 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, a recombinant retroviral construct, etc.

[00102] 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.

[00103] 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.

[00104] 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).

[00105] In some embodiments, a nucleotide sequence encoding a Cas1 protein and/or a Cas2 protein 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) (e.g, in cases where a Cas1 protein and/or a Cas2 protein will be isolated/purified prior to the contacting step). In some embodiments, a nucleotide sequence encoding a as1 protein and/or a Cas2 protein is operably linked to multiple control elements that allow expression of the nucleotide sequence encoding a as1 protein and/or a Cas2 protein in both prokaryotic and eukaryotic cells.

[00106] 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. In some cases, a promoter is chosen to achieve a desirable level expression (e.g., which in some cases can be as high as possible, whereas in some cases may be above or below a desirable threshold, e.g., to achieve the desired goal while reducing off-target effects). 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 Cas1 and/or Cas2 protein, thus resulting in one nor more chimeric polypeptides.

[00107] In some embodiments, a nucleotide sequence encoding a Cas1 and/or a Cas2 protein is operably linked to an inducible promoter. In some embodiments, a nucleotide sequence encoding a Cas1 and/or a Cas2 protein is operably linked to a constitutive promoter.

[00108] 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.

Nucleic acid modifications

[00109] In some embodiments, a subject nucleic acid (e.g., a donor DNA molecule) 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

[00110] Examples of suitable nucleic acids containing modifications include nucleic acids containing modified backbones or non-natural internucleoside linkages. Nucleic acids (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.

[00111] 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 oligonucleotides 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.

[00112] In some embodiments, a subject nucleic acid 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.

[00113] Also suitable are nucleic acids having morpholino backbone structures as described in, e.g., U.S. Pat. No. 5,034,506. For example, in some embodiments, a subject nucleic acid 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.

[00114] 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

[00115] A subject nucleic acid 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.

[00116] 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.

[00117] 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.

[00118] 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.

[00119] 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).

[00120] The synthesis and preparation of the LNA monomers adenine, cytosine, guanine, 5-methyl-cytosine, 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

[00121] A subject nucleic acid 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-$N(CH_3)_2$.

[00122] 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

[00123] A subject nucleic acid 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-amino-

adenine, 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).

[00124] 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

[00125] Another possible modification of a subject nucleic acid 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.

[00126] 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-O-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.\

[00127] 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 the amino terminus of a polypeptide (e.g., a Cas1 and/or Cas2 protein). In some embodiments, a PTD is covalently linked to the carboxyl terminus of a polypeptide (e.g., a Cas1 and/or Cas2 protein). In some embodiments, a PTD is covalently linked to the carboxyl terminus and the amino terminus of a polypeptide (e.g., a Cas1 and/or Cas2 protein). In some cases a PTD includes a nuclear localization signal (NLS) (e.g, in some cases 2 or more, 3 or more, 4 or more, or 5 or more NLSs). In some embodiments, a PTD is covalently linked to a nucleic acid (e.g., a nucleic acid encoding a Cas1 and/or Cas2 protein, a donor DNA molecule, 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: 210); 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:211); Transportan GWTLNSAGYLLGKINLKALAALAKKIL (SEQ ID NO:212); KALAWEAKLAKALAKALAKHLAKALAKALKCEA (SEQ ID NO:213); and RQIKIWFQNRRMKWKK (SEQ ID NO:214). Exemplary PTDs include but are not limited to, YGRKKRRQRRR (SEQ ID NO:210), RKKRRQRRR (SEQ ID NO:215); 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:210); RKKRRQRR (SEQ ID NO:216); YARAAARQARA (SEQ ID NO:217); THRLPRRRRRR (SEQ ID NO:218); and GGRRARRRRRR (SEQ ID NO:219). 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.

[00128] In some cases, a subject method is performed in vitro outside of a cell. When performed in vitro outside of a cell, The duration of the contacting step may be 0.1 hour-48 hours, for example, from 0.1 hour to 0.2 hour, from 0.2 hour to 0.3 hour, from 0.3 hour to 0.5 hour, from 0.5 hour to 1 hour, from 0.3 hour to 46 hours, 0.5 hour-45 hours, 1 hour-40 hours, 2 hours-35 hours, 4 hours-30 hours, 6 hours-24 hours, 8 hours -20 hours, 10 hours-18 hours, or 12 hours-16 hours, such as, 0.3 hour, 0.5 hour, 1 hour, 3 hours, 10 hours, 13 hours, 16 hours, or 18 hours.

[00129] The amount of Cas1 protein and/or Cas2 protein that is employed is can be from 10 units/ml-50,000 units/ml, for example, from 20 units/ml-30,000 units/ml, 30 units/ml-10,000 units/ml, 50 units/ml-5000 units/ml, 100 units/ml-3000 units/ml, 200 units/ml-2000 units/ml, 300 units/ml-1000 units/ml, such as, 100 units/ml, 300 units/ml, 1000 units/ml, 2000 units/ml, 5000 units/ml, 10,000 units/ml, 20,000 units/ml, or 50,000 units/ml.

[00130] The temperature at which the method is carried out is can be from 4°C-50°C, for example, about 10 °C-45 °C, about 16 °C-40 °C, about 20 °C-37 °C, about 25 °C-35 °C, about 30 °C-33 °C, e.g., 10°C, 18°C, 25°C, 30°C, 37°C, or 45°C.

[00131] The contacting step may be carried out in conditions suitable for Cas1 mediated integration. In certain embodiments, the conditions suitable for Cas1 mediated integration are conditions in which a divalent metal ion such as magnesium (Mg^{2+}) is present. In some cases, the Mg^{2+} concentration can range from 1 mM-25 mM, for example, 1.5 mM-20 mM, 2 mM-15 mM, 2 mM-10 mM, 3 mM-8 mM, or 5 mM-6 mM, such as, 2 mM, 2.5 mM, 3 mM, or 5 mM.

[00132] In certain embodiments, the conditions suitable for Cas1 mediated integration are conditions in which a divalent metal ion such as Manganese (Mn^{2+}) is present. In some cases, the Mn^{2+} concentration can range from 1 mM-25 mM, for example, 1.5 mM-20 mM, 2 mM-15 mM, 2 mM-10 mM, 3 mM-8 mM, or 5 mM-6 mM, such as, 2 mM, 2.5 mM, 3 mM, or 5 mM.

[00133] Under the conditions suitable for Cas1 endonuclease activity, the pH typically ranges from about pH 4.5-pH 10, for example, pH 5-pH 8.5, pH 7-pH 8.5, or pH 7-pH 8, such as, pH 7, pH 7.5, pH 8, or pH 8.5.

IHF

[00134] In some embodiments, the subject methods, compositions, and/or kits include an integration host factor (IHF) protein, or a nucleic acid encoding an IHF protein. Thus, in some cases, a subject method (e.g., contacting a contacting a target DNA molecule with a donor DNA molecule and a Cas1 protein, contacting a contacting a target DNA molecule with a donor DNA molecule and a Cas1 protein and a Cas2 protein, and the like) is performed in the presence of an integration host factor (IHF) protein. In some cases, inclusion of IHF increases sequence specificity of integration into the target DNA molecule. In some cases, a subject method (e.g., contacting a contacting a target DNA molecule with a donor DNA molecule and a Cas1 protein, contacting a contacting a target DNA molecule with a donor DNA molecule and a Cas1 protein and a Cas2 protein, and the like) includes a step of introducing into a target cell an IHF protein, or a nucleic acid comprising a nucleotide sequence that encodes an IHF protein.

[00135] IHF can be made up of two separate subunits, an alpha subunit and a beta subunit. As a non-limiting, illustrative example, the IHF alpha subunit of *Escherichia coli* (str. K-12 substr. MG1655) is:
MALTKAEMSEYLFDKLGLSKRDAKELVELFFEEIRRALENGEQVKLSGFGNFDLRDKN QRPGRNPKTGEDIPITARRVVTFRPGQKLKSRVENASPKDE (SEQ ID NO: 255).

[00136] As a non-limiting, illustrative example, the IHF beta subunit of *Escherichia coli* (str. K-12 substr. MG1655) is:
MTKSELIERLATQQSHIPAKTVEDAVKEMLEHMASTLAQGERIEIRGFGSFSLHYRAPRT GRNPKTGDKVELEGKYVPHFKPGKELRDRANIYG (SEQ ID NO: 256). Any alpha/beta subunit combination can be used (e.g., including corresponding subunits, where both subunits

are from or are derived from the same species). In some cases, the IHF protein is from (or derived from) the same species that the Cas1 and/or Cas2 protein(s) is from or derived from.

[00137] In some cases, a suitable IHF comprises an alpha subunit comprising an amino acid sequence having at least 50%, at least 60%, at least 70%, at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, at least 99%, or 100%, amino acid sequence identity to the amino acid sequence set forth in SEQ ID NO:255. In some cases, a suitable IHF comprises an beta subunit comprising an amino acid sequence having at least 50%, at least 60%, at least 70%, at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, at least 99%, or 100%, amino acid sequence identity to the amino acid sequence set forth in SEQ ID NO:256.

Target cells

[00138] The subject methods can be performed outside of a bacterial or archaeal cell (e.g., not in a bacterial or archaeal cell). For example, subject methods can be performed in vitro outside of a cell, or can be performed in a eukaryotic cell. A target cell of interest can include a cell from any non-bacterial or archaeal organism (e.g. a cell of a single-cell eukaryotic organism, a plant cell, an algal cell, e.g., *Botryococcus braunii, Chlamydomonas reinhardtii, Nannochloropsis gaditana, Chlorella pyrenoidosa*, *Sargassum patens, C. agardh*, and the like, a fungal cell (e.g., a yeast cell), an animal cell, a cell from an invertebrate animal (e.g. fruit fly, cnidarian, echinoderm, nematode, etc.), a cell from a vertebrate animal (e.g., fish, amphibian, reptile, bird, mammal), a cell from a mammal, a cell from a rodent, a cell from a human, etc.).

[00139] Any type of cell may be of interest (e.g. a stem cell, e.g. an embryonic stem (ES) cell, an induced pluripotent stem (iPS) cell, a germ cell; a somatic cell, e.g. a fibroblast, a hematopoietic cell, a neuron, a muscle cell, a bone cell, a hepatocyte, a pancreatic cell; an in vitro or in vivo embryonic cell of an embryo at any stage, e.g., a 1-cell, 2-cell, 4-cell, 8-cell, etc. stage zebrafish embryo; etc.). Cells may be from established cell lines or they may be primary cells, where "primary cells", "primary cell lines", and "primary cultures" are used interchangeably herein to refer to cells and cells cultures that have been derived from a subject and allowed to grow in vitro for a limited number of passages, i.e. splittings, of the culture. For example, primary cultures are cultures that may have been passaged 0 times, 1 time, 2 times, 4 times, 5 times, 10 times, or 15 times, but not enough times go through the crisis stage. Typically, the primary cell lines of the present invention are maintained for fewer than 10 passages in vitro. Target cells are in many embodiments unicellular organisms, or are grown in culture.

[00140] If the cells are primary cells, they may be harvest from an individual by any convenient method. For example, leukocytes may be conveniently harvested by apheresis, leukocytapheresis, density gradient separation, etc., while cells from tissues such as skin, muscle,

bone marrow, spleen, liver, pancreas, lung, intestine, stomach, etc. are most conveniently harvested by biopsy. An appropriate solution may be used for dispersion or suspension of the harvested cells. Such solution will generally be a balanced salt solution, e.g. normal saline, phosphate-buffered saline (PBS), Hank's balanced salt solution, etc., conveniently supplemented with fetal calf serum or other naturally occurring factors, in conjunction with an acceptable buffer at low concentration, generally from 5-25 mM. Convenient buffers include HEPES, phosphate buffers, lactate buffers, etc. The cells may be used immediately, or they may be stored, frozen, for long periods of time, being thawed and capable of being reused. In such cases, the cells will usually be frozen in 10% DMSO, 50% serum, 40% buffered medium, or some other such solution as is commonly used in the art to preserve cells at such freezing temperatures, and thawed in a manner as commonly known in the art for thawing frozen cultured cells.

Kits

[00141] The present disclosure provides kits for carrying out a subject method. A subject kit can include one or more of (in any combination): a Cas1 protein, a nucleic acid having nucleotides encoding a Cas1 protein, a Cas2 protein, a nucleic acid having nucleotides encoding a Cas2 protein, and a subject linear DNA molecule. A kit can further include one or more additional reagents, where such additional reagents can be selected from: a dilution buffer; a reconstitution solution; a wash buffer; a control reagent; a control expression vector or RNA polynucleotide; a reagent for *in vitro* production of a Cas1 and/or Cas2 protein from DNA, and the like. The components of a subject kit can be in the same or different containers (in any desired combination).

[00142] In addition to above-mentioned components, a subject kit can further include instructions for using the components of the kit to practice the subject methods. The instructions for practicing the subject methods are generally recorded on a suitable recording medium. For example, the instructions may be printed on a substrate, such as paper or plastic, etc. As such, the instructions may be present in the kits as a package insert, in the labeling of the container of the kit or components thereof (i.e., associated with the packaging or subpackaging) etc. In other embodiments, the instructions are present as an electronic storage data file present on a suitable computer readable storage medium, e.g. CD-ROM, diskette, flash drive, etc. In yet other embodiments, the actual instructions are not present in the kit, but means for obtaining the instructions from a remote source, e.g. via the internet, are provided. An example of this embodiment is a kit that includes a web address where the instructions can be viewed and/or

from which the instructions can be downloaded. As with the instructions, this means for obtaining the instructions is recorded on a suitable substrate.

Utility

[00143] The subject compositions, kits, and methods find use for the integration of a donor DNA molecule into any desirable supercoiled target DNA molecule. The following uses are merely illustrative examples, and are by no means meant to limit the use of the subject methods. The compositions, kits, and methods can find use in vitro outside of a cell (e.g., to modify a plasmid DNA, to modify an isolated chromosomal DNA, etc.), and can find use inside of a eukaryotic cell (e.g., in vitro and/or in in vivo and/or ex vivo). The subject compositions, kits, and methods can be used to insert and/or modify a control element (e.g., a transcriptional control element such as an enhancer, a promoter, a transcription terminator, etc.). The subject compositions, kits, and methods can be used to modify a target gene (e.g., in some cases disrupting the expression of the target gene, in some cases, modifying the transcribed RNA, etc.). The subject compositions, kits, and methods can be used to modify a coding and/or a non-coding sequence (e.g., modify a gene coding sequence, modify a sequence that codes for a non-coding RNA such as a microRNA).

EXAMPLES

[00144] 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.

Example 1

Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity

[00145] Bacteria and archaea insert spacer sequences acquired from foreign DNAs into CRISPR loci to generate immunological memory. The *Escherichia coli* Cas1–Cas2 complex mediates

spacer acquisition, but the molecular mechanism of this process is not known. The data presented below show that the purified Cas1–Cas2 complex integrates linear DNA substrates into acceptor DNA to yield products similar to those generated by retroviral integrases and DNA transposases. Cas1 is the catalytic subunit, whereas Cas2 increases DNA integration activity. Integration occurs preferentially at the ends of CRISPR repeats and at sequences adjacent to cruciform structures abutting A-T rich regions. The results demonstrate the Cas1–Cas2 complex to be the minimal machinery required to catalyze spacer DNA acquisition and explain the significance of CRISPR repeats in providing sequence and structural specificity for Cas1–Cas2-mediated adaptive immunity.

METHODS

Cas1, Cas2 and DNA preparation

[00146] Cas1 and Cas2 from *E. coli* K12 (MG1655) were separately purified as previously described[15]. The proteins were stored in 100 mM KCl, 20 mM HEPES-NaOH, 5% glycerol and 1 mM TCEP at -80 °C prior to use. Single-stranded DNAs were synthesized (Integrated DNA Technologies). Double-stranded DNA protospacers were annealed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ or $MnCl_2$, 1 mM DTT, 10% DMSO by heating at 95 °C for 3 min and slow cooling to room temperature. The sequence of the 33 bp protospacer used in this study was shown to be the most acquired *in vivo* in *E. coli* K12 after M13 bacteriophage infection[13]: Strand 1 (5′-GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT-3′) (SEQ ID NO: 220) and Strand 2 (5′-ACGAGAAACACCAGAACGAGTAGTAAATTGGGC-3′)(SEQ ID NO: 221).

***In vitro* integration assays.**

[00147] The integration reactions were performed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ or $MnCl_2$, 1 mM DTT and 10% DMSO. All of the reactions were conducted with $MgCl_2$ unless otherwise noted. For reactions with the Cas1–Cas2 complex, separately purified Cas1 and Cas2 were pre-incubated for 20–30 min at 4 °C to allow complex formation. The protospacer DNAs were incubated with the protein(s) for 10–15 min at 4 °C, followed by the addition of the target pCRISPR or pUC19 plasmid DNA. The reactions were conducted at 37 °C for 1 h and quenched with DNA loading buffer containing a final concentration of 50 mM EDTA. The products were analyzed on 1.5% agarose gels pre-stained with ethidium bromide. All of the reactions, except those shown in Figure 1 and Figure 6a, c–e, were conducted with 75 nM protein, 200 nM protospacers and 7.5 nM pCRISPR to clearly visualize Band X from pCRISPR. Reactions in Figure 1 and Figure 6a, c, e were performed with 50 nM protospacers.

Radiolabeled protospacer integration assays

[00148] Pre-annealed double-stranded protospacer DNA substrates were 5′-radiolabeled using [γ-^{32}P]-ATP (PerkinElmer) and T4 polynucleotide kinase (New England Biolabs). Protospacers with 3′-PO_4 ends were 5′-radiolabeled using T4 polynucleotide kinase with 3′ phosphatase minus activity (New England Biolabs). The reactions were carried out in the same buffer as above. Unless otherwise noted, 200 nM of Cas1–Cas2 was first incubated with 20 nM protospacers at 4 °C for 10–15 min, followed by the addition of 200 ng (~5nM) of pCRISPR. The reactions were conducted at 37 °C for 1 h and quenched with 25 mM EDTA and 0.4% SDS. The DNAs were deproteinized with 30 μg of Proteinase K for 1 h at 37 °C and ethanol precipitated. The reactions were analyzed on 1.5% agarose gels. After electrophoresis, the gels were dried onto positively charged nylon transfer membrane (GE Healthcare) and imaged using Phosphor Screens (GE Healthcare). The restriction enzyme digest experiments were performed by first conducting the integration reaction, followed by addition of the respective enzymes (New England Biolabs), which were allowed to digest for an additional 1 h at 37 °C.

Disintegration assays

[00149] The four single stranded DNA substrates were annealed to form the Y DNA in a stepwise manner: 95 °C for 3 min, 65 °C for 20 min, 50 °C for 20 min, and gradual cooling to room temperature. The annealing reactions were analyzed on a 15% native polyacrylamide gel to confirm the formation of the Y DNA (Figure 10b). The disintegration assay was performed in the integration reaction buffer with 50 nM protein and 5 nM Y DNA at 37 °C for 1 h. For native polyacrylamide gel analysis, the reaction was quenched with DNA loading buffer with 50 mM EDTA and analyzed on 15% polyacrylamide gels. For denaturing polyacrylamide gel analysis, the reactions were quenched with formamide buffer and heating at 95 °C prior to loading on 15% 8M urea-polyacrylamide gels. The sequences of the four strands are as follows:

A (5′-GGCCCCAGTGCTGCAATGAT-3′) (SEQ ID NO: 222);

B (5′-GTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCC-3′) (SEQ ID NO: 223);

C (5′-GCCCAATTTACTACTCGTTCTGGTGTTTCTCGTACCGCGAGACCCACGCTCAC-3′) (SEQ ID NO: 224); and

D (5′- ACGAGAAACACCAGAACGAGTAGTAAATTGGGC-3′) (SEQ ID NO: 225).

High-throughput sequencing

[00150] The integration reaction was performed with 75 nM Cas1–Cas2, 200 nM protospacer and 7.5 nM pCRISPR in 20 mM HEPES, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$, 10% DMSO and 1 mM DTT. The DNAs were isolated by phenol-chloroform extraction and ethanol precipitation.

The excess protospacers were removed using 100K MWCO Amicon Ultra-0.5 ml centrifugal filters. The resulting integration products were digested into smaller DNA fragments using dsDNA Fragmentase (New England Biolabs) for 75 min at 37 °C and quenched at 65 °C for 15 min. Fragments were end repaired using T4 DNA Polymerase (NEB), Klenow (NEB) and T4 PNK (NEB) and A-tailed with Klenow exo (3′ to 5′ exo minus) (NEB). Adapters were ligated onto fragments using T4 DNA ligase (NEB) and cDNA libraries were amplified by PCR using Phusion (NEB). Libraries were sequenced on an Illumina HiSeq2500 on rapid run mode. The oligonucleotides used are: Universal adapter: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA CGCTCTTCCGATC*T-3′ (*phosphorothioate bond) (SEQ ID NO: 226); Indexed adapter: 5′-/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-index-ATCTCGTATGCCGTC TTCTGCTTG-3′) (SEQ ID NO: 227); PCR primers: 5′- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA-3′ (SEQ ID NO: 228, 5′-CAAGCAGAAGACGGCATACGAGAT-3′(SEQ ID NO: 229).

Computational analysis

[00151] For preprocessing, 3′ adapters were removed from raw Illumina reads using Cutadapt ("http://" followed by "code.google." followed by "com/p/cutadapt/"), discarding reads shorter than 15 nt. Reads containing integrated protospacer were selected using Cutadapt, requiring the presence of at least 10 nt of protospacer sequence with no errors. After creating Bowtie[46] indexes from fasta files of the pUC19 empty and pCRISPR plasmid sequences, these reads were mapped to the respective plasmids using Bowtie, allowing up to 2 mismatches and requiring unique mapping. Sequence motif analysis depicted in Figure 14 were generated using WebLogo, utilizing integration sites that are represented at least ten times in the sequencing data[47].

RESULTS

The Cas1–Cas2 complex integrates protospacer DNA *in vitro*

[00152] Experiments were designed to test whether the Cas1–Cas2 complex was sufficient to catalyze the mobilization of protospacer substrate DNA *in vitro*. DNA recombination assays were conducted using purified Cas1–Cas2 complex, 33 bp protospacer DNA and an acceptor "target" plasmid consisting of the pUC19 backbone with an inserted CRISPR locus (pCRISPR) (Fig 1a). Co-incubation of these reagents converted the supercoiled plasmid into three main products: relaxed and linear plasmid species, and a fast-migrating species we term Band X (Figure 1b, c and Figure 6a). Product formation was dependent on Cas1, Cas2 and the protospacer DNA, and was enhanced at low salt concentration and by the presence of Mg^{2+} or Mn^{2+} (Figure 6b-d). Target DNA was fully converted to products in the presence of ~26-fold

molar excess of protospacer DNA (Figure 6d). Little difference was observed in product DNA migration when reactions were post-treated with EDTA, EDTA and phenol-chloroform extraction or Proteinase K in the presence of EDTA and detergent (Figure 6e), indicating that product DNAs are unlikely to be bound to Cas1 and/or Cas2. Consistent with product DNA resulting from covalent integration of protospacer DNA into the plasmid, the relaxed and linear forms of pCRISPR became radiolabeled in reactions containing ^{32}P-labeled protospacer DNA (Figure 1d and Figure 7). Although Cas1 alone catalyzed a low level of protospacer integration in the presence of Mn^{2+}, the reaction was enhanced significantly by the presence of Cas2 (Figure 7b).

[00153] Cas1 active site mutants H208A and D221A were defective for protospacer integration *in vitro*, whereas the Cas2 E9Q active-site mutant supported integration (Figure 1c, e and Extended Figure 3). The Cas2 C-terminal β6–β7 deletion mutant, which is defective for complex formation with Cas1 and spacer acquisition *in vivo*, failed to support Cas1-mediated integrase activity (Figure 1c, e). These data show that the *in vitro* assay recapitulates the *in vivo* functions of Cas1 and Cas2 during spacer acquisition and that Cas2 likely functions primarily as a scaffold, perhaps by bridging the protospacer nucleoprotein complex and target DNA.

[00154] **Figure 1.** The Cas1–Cas2 complex integrates protospacers *in vitro*. **a,** Schematic of the *in vitro* integration assay. **b,** The presence of Cas1, Cas2 and a protospacer results in the conversion of the supercoiled pCRISPR into relaxed, linear and Band X products. **c,** Neither the Cas1 H208A active site mutant nor and the complex formation-defective Cas2 β6–β7 deletion mutant support the reaction. The Cas2 E9Q active site mutant (lane 5 from the marker) is as active as the wild-type. **d,** Salt- and metal-dependence of radiolabeled protospacer integration into pCRISPR. **e,** Same as **c** except using radiolabeled protospacers.

[00155] **Figure 6.** The integration reaction is dependent on the presence of protospacers, low salt and divalent metal ions. **a,** *In vitro* integration assay alongside EcoRI- and Nb.BbvCI nickase-treated pCRISPR. **b,** Salt-dependence assay using Cas1 or Cas2 only and Cas1+Cas2. The titration corresponds to 0, 25, 50, 100 and 200 nM KCl, on top of the salt carried in from the reaction reagents. **c,** Integration assays in the presence of 10 mM EDTA, Mg^{2+}, Mn^{2+} or no additive. **d,** Integration assays with increasing protospacer concentrations. **e,** A comparison of post-reaction treatments as indicated.

[00156] **Figure 7.** Cas1 requires Cas2 for robust protospacer integration. **a,** Schematic of the integration assays using ^{32}P-labeled protospacers. **b,** Integration assays in the presence of increasing protein and 10 mM $MnCl_2$. The titration corresponds to 0, 50, 100 and 200 nM protein. **c**, Same as **b** except in the presence of 10 mM $MgCl_2$.

[00157] **Figure 8.** The catalytic activity of Cas1 is required for integration. **a,** Close-up view of the Cas1 active site with the conserved residues shown in stick configurations (PDB 4P6I) **b,** Integration assays of purified Cas1 active site mutants complexed with wild type Cas2. **c,** The same as **b** except using radiolabeled protospacers.

Cas1–Cas2-catalyzed recombination leads to half-site and full-site integration as well as pCRISPR topoisomers

[00158] It was then tested whether the reaction products of Cas1–Cas2-mediated DNA integration resemble those formed by the strand transfer activity of retroviral integrases and cut-and-paste tranposases[23-26]. These enzymes generated two main products *in vitro* corresponding to half-site and full-site integration events (Figure 2a). The integration of one strand of donor DNA into one strand of the target yields half-site products, which co-migrate with relaxed plasmid DNA during agarose gel electrophoresis. Full-site products can co-migrate with relaxed or linear plasmid DNA depending on whether one or two copies of donor DNA are utilized during integration. Similar gel mobility of the slowly migrating DNA product generated by Cas1–Cas2 and Nb.BbvCI nickase-digested pCRISPR was observed, consistent with the slow-migrating relaxed DNA species corresponding to half-site products and/or products resulting from full-site integration of one protospacer molecule (Figure 6a). Digestion with EcoRI, which cuts pCRISPR once, converted the reaction products to linear DNAs (Figure 2b, compare lane 4 to lane 2, and Figure 2c). Tus, both the relaxed and Band X DNA products comprise unit-sized pCRISPR circles.

[00159] Band X did not become radiolabeled in reactions conducted with ^{32}P-labeled protospacer DNA. A time course analysis revealed relaxed DNA product formation within the first minute, followed by accumulation of Band X between 10 and 30 min (Figure 2d). To determine the properties of Band X, the purified product was analyzed in two different types of agarose gels – one pre-stained with ethidium bromide, similar to the gels presented thus far, and the other stained with ethidium bromide after electrophoresis (post-stained) (Figure 9a). Although Band X migrated as a single species in the pre-stained gel, a ladder of species that migrated faster than the relaxed products was observed in the post-stained gel (Figure 2e, f). These intermediates are reminiscent of plasmid topoisomers[27,28]. The same pre- and post-stained agarose gel analysis was performed on the entire integration reaction, generating similar results to those observed with purified Band X (Figure 9b, c). PCR analysis of various segments of pCRISPR using gel-purified Band X as the template yielded amplification products indistinguishable from those generated using unreacted supercoiled pCRISPR or relaxed integration products, supporting the conclusion that Band X corresponds to pCRISPR topoisomers (Figure 9d).

[00160] Band X might therefore arise from the excision of the protospacer from half-site integration products to yield fully enclosed pCRISPR in different states of supercoiling (Figure 2g). The process of disintegration has been previously observed in *in* vitro reactions with retroviral integrases and transposases[29,30]. To test this hypothesis, four single-stranded DNAs were annealed to produce a synthetic Y-structured DNA intermediate that mimics the half-site integration product (Figure 10a,b). The 5′-end of the extended protospacer arm was radiolabeled, such that the liberated 33 bp protospacer DNA could be detected following disintegration activity. Using this substrate Cas1 catalyzed disintegration activity either by itself or in the presence of Cas2 (Figure 2h). Disintegration activity was confirmed by radiolabeling the 20-nt target DNA strand and monitoring the formation of the joined 40 bp target DNA product (Figure 10c, d). Thus, Cas1–Cas2 integration and disintegration activities are similar to those of retroviral integrases and transposases, although the *in vivo* function of disintegration, if any, remains unknown.

[00161] **Figure 2.** Half-site, full-site integration and pCRISPR topoisomer products. **a,** Schematic of half-site and full-site integration products. **b,** Linearization of the integration products (lane 4). Lane 3 is the un-treated reaction products. **c,** Linearization of integration products from radiolabeled protospacer reactions. **d,** The time course reveals the initial formation of relaxed products, followed by Band X. The inset reveals the products detected using ^{32}P-labeled protospacers. **e,f,** Analysis of gel-purified relaxed and Band X on agarose gels pre-stained with ethidium bromide (**e**) or post-stained after electrophoresis (**f**). **g,** Schematic of the disintegration reaction. **h**, Native polyacrylamide gel analysis of the disintegration reaction.

[00162] **Figure 9.** Band X corresponds to topoisomers of pCRISPR. **a,** Agarose gel of purified relaxed and Band X integration products. **b,** Analysis of the total reaction products, after phenol chloroform extraction and ethanol precipitation, on a pre-stained agarose gel. **c,** Same as **b** except ethidium bromide staining was performed after electrophoresis. **d,** PCR amplification products of various segments of pCRISPR using the relaxed, Band X or pCRISPR template shown in **a**. The laddering effect of minor products using CRISPR locus primers likely reflects the propensity of CRISPR repeats to form DNA hairpins.

[00163] **Figure 10.** Cas1 catalyzes the disintegration of half-site integrated protospacers. **a,** Schematic of the four strands constituting the Y DNA substrate used in the disintegration assays. **b,** Native polyacrylamide gel analysis of the annealing products with either Strand A or Strand C radiolabeled. **c,** Native polyacrylamide gel analysis of disintegration assay products using Y DNA substrates with Strand A labeled. **d,** Denaturing gel analysis of the disintegration assay products with Strand A labeled

Integration requires 3′-OH protospacer ends

[00164] The DNA protospacer and target DNA requirements for integration were next investigated. Single-stranded protospacer DNA failed to support the reaction (Figure 3a, b). The Cas1–Cas2 complex accommodated various protospacer lengths *in vitro* despite the strict 33 bp requirement for spacer acquisition *in vivo* (Figure 11a). This suggests that the length requirement *in vivo* is pre-determined before integration by an unknown mechanism. The Cas1–Cas2 complex integrated DNA substrates with blunt-ends or with 3′-overhangs up to 5 nt in length (Figure 11b). In contrast to retroviral integrases[31], substrates with 5′-overhangs were nonviable (Figure 11b).

[00165] Retroviral integration and transposition reactions proceed via nucleophilic attack of DNA 3′-OH groups at target DNA phosphodiester bonds[31,32]. Phosphorylation of both 3′-ends of the protospacer ablated integration, whereas phosphorylation of only one 3′ end strongly limited integration (Figure 3a, b). By analogy to known integrase enzyme mechanisms, DNA integration could proceed by Cas1-catalyzed direct nucleophilic attack of the substrate 3′-OH on the target DNA, or by formation of a Cas1–DNA intermediate, as occurs in the serine and tyrosine families of recombinases[33]. Based on available crystal structures of Cas1[17-19], there are four tyrosine residues in the vicinity of the Cas1 active site that could be involved in forming such a covalent intermediate (Figure 12a,b). Four mutants of Cas1 were constructed in which each tyrosine was individually changed to alanine. The purified Cas1 mutant proteins supported protospacer integration *in vitro* at levels comparable to wild type Cas1–Cas2 (Figure 12c). Thus, the integration reaction likely proceeds via direct nucleophilic attack of protospacer 3′-OH ends onto the target DNA phosphodiester bonds, a mechanism previously hypothesized to occur *in vivo*[34].

[00166] **Figure 3.** Integration requires 3′-OH protospacer ends and supercoiled target DNA. **a,b,** Integration assays using single-stranded DNAs and either –OH or $-PO_4$ at the 3′ or 5′ ends of (**a**) unlabeled or (**b**) radiolabeled protospacers. S1 corresponds to one strand of the protospacer and S2 corresponds to the complementary strand. **c,** Comparison of protospacer integration into different DNA targets. **d,e,** Restriction enzyme digestion of pCRISPR, either in a pUC19 (**d**) or pACYC backbone (**e**), after the integration assay detects integration into the CRISPR fragment (green arrows).

[00167] **Figure 11.** Cas1–Cas2 can integrate various lengths of double-stranded DNA with blunt- or 3′-overhang ends into a supercoiled target plasmid. **a,** Integration assays using the indicated lengths of protospacer DNA. **b,** Integration assays using varying 5′ or 3′ overhang lengths. **c,d,** A comparison of integration assays using pCRISPR or Nb.BbvCI-nicked pCRISPR target. **e,** Integration assay using different target plasmids with or without a CRISPR locus.

[00168] **Figure 12.** Cas1 tyrosine mutants support integration activity *in vitro*. **a,** A close-up of the Cas1 active site with the tyrosine residues labeled in blue. **b,** Structure-based sequence alignment of Cas1 proteins, highlighting the tyrosine residues mutated to alanine in this study. **c,** Radiolabeled protospacer integration assay of Cas1 tyrosine mutants complexed with WT Cas2.

Protospacer integration requires supercoiled target DNA and favors the CRISPR locus

[00169] Cas1 and Cas2 overexpression leads to site-selective spacer acquisition proximal to the leader end of the CRISPR locus, a result consistent with observations in native populations of CRISPR-containing bacteria[13-15,35]. To determine what drives such site-specific integration, various forms of the pCRISPR plasmid DNA were tested to determine target DNA requirements. Integration requires target DNA supercoiling, as neither relaxed nor linear pCRISPR, nor the isolated 1 kb CRISPR locus, supported integration (Figure 3c and Figure 11c,d).

[00170] As a control, supercoiled pUC19 DNA was used, the parental plasmid of pCRISPR that lacks a CRISPR locus, and integration products upon incubation with Cas1 and Cas2 in the presence of protospacer DNA were observed (Figure 3c and Figure 11e). This finding raised two possibilities: 1) *in vitro* spacer integration is non-specific with respect to target DNA sequence or 2) structures and/or sequence(s) favoring integration are present in the pUC19 plasmid. To determine if integration preferentially occurred at the CRISPR locus of pCRISPR, products of radiolabeled reactions were double-digested to separate the CRISPR locus (960 bp) from the pUC19 plasmid backbone (~2.27 kb). Suggestive of CRISPR-specific integration, the ^{32}P-radiolabel migrated solely with the CRISPR locus fragment (Figure 3d). The same result was observed when the experiment was conducted using a target plasmid containing the CRISPR locus and a different backbone sequence (pACYC) (Figure 3e).

The CRISPR repeats provide specificity for integration

[00171] To determine the exact sites of protospacer integration in these reactions, high-throughput sequencing was performed of reaction products that resulted from using either pCRISPR or the parental pUC19 vector as the target of integration. Reaction products were fragmented, end-repaired, adapter ligated, PCR amplified and analyzed by Illumina sequencing (Figure 13a). Of the 7,866 protospacer-pCRISPR junctions retrieved, ~71% mapped to the CRISPR locus (Figure 4a and Figure 13b). Analysis of the DNA integration sites within the CRISPR locus revealed spacer sequence insertion into the borders of each repeat, with the most preferred site at the first repeat adjacent to the leader (Figure 4b). The minus strand of each repeat (the bottom strand in Figure 4a,b that runs 5′-to-3′ towards the leader sequence) is also highly preferred, highlighting the role of CRISPR repeats in providing sequence specificity for the Cas1–Cas2 complex (Figure 4b). Sequence alignment of the integration sites revealed strong

preference for sequences resembling the CRISPR repeat on both strands of pCRISPR, further supporting the selection of CRISPR repeat borders by the Cas1–Cas2 complex (Figure 13d-f).

[00172] The most frequent integration site in the pUC19 control plasmid mapped to the *amp* resistance gene adjacent to the A-T rich promoter sequence (~8.8% of 5,524 total retrieved junctions, Figure 4c and Figure 13c). An inverted repeat sequence with a propensity to form a DNA cruciform[36] occurs 9 nt adjacent to this integration site (plus strand sequence: 5′-TTCAATATTATTGAA-3′; SEQ ID NO:///), suggesting that potential DNA cruciform formation adjacent to A-T rich sequences is important for protospacer integration. Sequence analysis of pUC19 target sites revealed the propensity for a G nucleotide to occur at the -2 and +1 positions of the protospacer insertion site, similar to the preferred pCRISPR sites, but otherwise little sequence specificity surrounding the integration sites relative to the pCRISPR repeats (Figure 13g, h). This observation implies that pCRISPR repeat sequence selectivity stems from the unique structural features of these sites, such as their ability to form cruciforms (Figure 4a, b, e).

[00173] In *E. coli*, newly acquired spacers harbor a 5′ G as the first nucleotide flanking the leader-proximal end of the repeats, which originates from the last nucleotide of the AAG PAM in the foreign DNA[13-15,37-39]. Such positional specificity is critical for crRNA-guided interference, as a mutation in this position of the corresponding crRNA disrupts PAM binding and subsequent target destruction[40-42]. The sequencing data here was used to determine if the Cas1–Cas2 complex preferentially utilized the terminal 3′ C or T of protospacer DNA during integration (see Figure 4b for protospacer sequence). About 73% of all integration events into pCRISPR utilized the C 3′-OH end, and there was a strong preference for this nucleotide to attack the minus strand of the repeat sequence (Figure 4b, d, e). A similar nucleotide bias was observed in the pUC19 target plasmid sequence data (Figure 4d). This preference positions the G at the 5′ end of the protospacer substrate as the first nucleotide of the newly integrated spacer in the CRISPR locus (Figure 5). When protospacer DNAs lacking a 3′ C or bearing 3′ C on both ends were used, the preference for integration into the minus strand of the CRISPR locus was significantly decreased (Figure 14). This observation results in the loss of preferential orientation of the protospacer after full integration. Thus, the Cas1–Cas2 complex plays a critical role in correctly orienting the C 3′-OH end of protospacer DNA substrates for incorporation within the CRISPR locus.

[00174] **Figure 4. Protospacers are specifically integrated into the CRISPR locus. a,** Mapped integration sites along pCRISPR. **b,** Magnified view of the integration sites along the ~1 kb CRISPR locus. The cyan peaks represent positions where the 3′ T of the protospacer DNA was integrated whereas the black peaks represent the C 3′-OH integration events. The protospacer sequence is depicted above the plot. **c,** Mapped integration sites along pUC19 empty vector. **d,**

Comparison of C 3′-OH or T 3′-OH selection in the total reads from pCRISPR and pUC19 empty targets (Chi-square test, *p<0.0001). **e,** Schematic of DNA cruciform formation of the repeat sequences. The orange arrows depict the cleavage site on the plus and minus strands, based on the integration sites in (**b**).

[00175] **Figure 5.** Model of protospacer integration during CRISPR–Cas adaptive immunity. The first nucleophilic attack occurs on the minus strand of the first repeat, distal to the leader, by the C 3′-OH end of the protospacer. After half-site intermediate formation, the second integration event occurs on the opposite strand at the leader-repeat border. The resulting single-stranded DNA gaps are repaired by yet uncharacterized mechanisms and the protospacer is fully integrated with the G as the first nucleotide at its 5′ end. The asterisk denotes the duplication of the first repeat, as previously observed *in vivo* (refs 13-15).

[00176] **Figure 13** High-throughput sequencing of integration products reveals sequence-specific integration. **a,** Schematic of the workflow for high-throughput sequencing analysis of the integration sites. **b,** Raw map of the total reads along pCRISPR before collapsing into single peaks of protospacer-pCRISPR junctions depicted in Figure 4. **c,** Same as **b**, except for the pUC19 target. **d,** Sequence of the leader-end of the CRISPR locus in *E. coli*. **e,f,** WebLogo analysis from the -5 to +5 positions surrounding the protospacer integration sites on the (**e**) plus and (**f**) minus of pCRISPR. The arrow points to the nucleotide that is covalently joined to the protospacer. **g, h,** Same as **e,f**, except for the pUC19 target.

[00177] **Figure 14.** Cas1–Cas2 correctly orients the protospacer DNA during integration. Mapped integration sites along the CRISPR locus of pCRISPR when using protospacer DNA with nucleotide ends (**a**) "wild type" 3′ C and 3′ T, (**c**) 3′ A and 3′ T, and (**e**) 3′ C and 3′ C. The red arrow in **c** and **e** points to the nucleotide change in the protospacer DNA compared to the "wild type" sequence in **a**. The protospacer DNA 3′ nucleotide and the CRISPR locus strand biases in **a, c, e** are plotted in **b, d** and **f,** respectively, as percentages of integration events within the CRISPR locus. The black and clear bars represent the (-) and (+) strands of the CRISPR locus, respectively. NS corresponds to not significant and *p<0.0001 by Chi-square test.

Mechanism of protospacer integration during CRISPR–Cas immunity

[00178] The results presented here explain the mechanistic basis for foreign DNA acquisition during CRISPR–Cas adaptive immunity (Figure 5). The Cas1–Cas2 complex catalyzes integration of protospacers preferentially at the leader-end of the CRISPR locus. Intriguingly, the experiments herein show that Cas1–Cas2 also functions to select the terminal C 3′-OH as the attacking nucleophile to preferentially target the minus strand of CRISPR repeats, resulting in the 5′ G on the opposite strand of the protospacer becoming the first nucleotide of the newly

integrated spacer. This orientation bias, previously observed *in vivo*[39], is a key step during immunity for productive downstream foreign DNA targeting by the Cascade complex and Cas3 effector nuclease (Figure 15). Interestingly, the presence of the complete AAG PAM in the protospacer is not required for *in vitro* integration, suggesting that a highly specific selection or processing step occurs *in vivo* to exclude the AA nucleotides from the mature protospacer prior to integration.

[00179] **Figure 15.** Model of the CRISPR–Cas adaptive immunity pathway in *E. coli.* Mature double-stranded protospacers bearing a 3′ C-OH are site-specifically integrated into the leader-end of the CRISPR locus. Correct protospacer integration (left) results in the 5′G/3′C as the first nucleotide of the spacer, proximal to the leader. After transcription of the CRISPR locus and subsequent crRNA processing, foreign DNA destruction is initiated by strand-specific recognition of the 3′-TTC-5′ PAM sequence in the target strand by the crRNA-guided Cascade complex. Incorrect protospacer integration (right) cannot initiate foreign DNA destruction due to the inability for the crRNA to recognize the strand with the 3′-TTC-5′ PAM. Thus, foreign DNA interference during CRISPR–Cas adaptive immunity relies on the Cas1–Cas2 complex for correctly orienting the protospacer during integration.

[00180] A two-step integration mechanism is suggested during spacer acquisition, in which the C 3′-OH first attacks the minus strand of the CRISPR repeat to produce a half-site intermediate (Figure 5). The 3′-OH on the opposite strand of the integrating DNA then attacks the target DNA 28 bp away on the opposite side of the repeat on the plus strand, leading to full integration of the protospacer (Figure 5). As evidenced by agarose gel electrophoresis and high throughput sequencing, most of the *in vitro* products are half-site integration intermediates. The conclusion is that the *in vitro* system predominantly traps the first step of a two-step integration mechanism. We posit that the second nucleophilic attack is greatly accelerated *in vivo* in the presence of cellular factors and high Cas1–Cas2 concentrations. This model is consistent with spacer integration intermediates that are observed *in vivo*, in which protospacers are integrated such that staggered cleavage at each end of the repeat generates single-stranded gaps that ensure repeat duplication[34].

[00181] The results highlight the fundamental role of repeat sequences at multiple stages of CRISPR–Cas adaptive immunity. In addition to their role in creating structures within nascent CRISPR transcripts that ensure correct RNA processing during crRNA maturation[45], the repeats also operate at the DNA level to recruit the Cas1–Cas2 complex for sequence- and structure-specific protospacer integration. This recruitment involves transient DNA cruciform formation within the CRISPR inverted repeats that occurs as a function of target DNA supercoiling[46].

[00182] References

1 Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709-1712, doi:10.1126/science.1138140 (2007).

2 van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nature reviews. Microbiology 12, 479-492, doi:10.1038/nrmicro3279 (2014).

3 Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of molecular evolution 60, 174-182, doi:10.1007/s00239-004-0046-3 (2005).

4 Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151, 2551-2561, doi:10.1099/mic.0.28048-0 (2005).

5 Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151, 653-663, doi:10.1099/mic.0.27437-0 (2005).

6 Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends in genetics : TIG 26, 335-340, doi:10.1016/j.tig.2010.05.008 (2010).

7 Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes & development 22, 3489-3496, doi:10.1101/gad.1742908 (2008).

8 Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science 329, 1355-1358, doi:10.1126/science.1192272 (2010).

9 Deltcheva, E. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 471, 602-607, doi:10.1038/nature09886 (2011).

10 Brouns, S. J. et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321, 960-964, doi:10.1126/science.1159689 (2008).

11 Garneau, J. E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468, 67-71, doi:10.1038/nature09523 (2010).

12 Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816-821, doi:10.1126/science.1225829 (2012).

13 Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic acids research 40, 5569-5576, doi:10.1093/nar/gks216 (2012).

14 Datsenko, K. A. et al. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nature communications 3, 945, doi:10.1038/ncomms1937 (2012).

15 Swarts, D. C., Mosterd, C., van Passel, M. W. & Brouns, S. J. CRISPR interference directs strand specific spacer acquisition. PloS one 7, e35888, doi:10.1371/journal.pone.0035888 (2012).

16 Nunez, J. K. et al. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nature structural & molecular biology 21, 528-534, doi:10.1038/nsmb.2820 (2014).

17 Wiedenheft, B. et al. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure 17, 904-912, doi:10.1016/j.str.2009.03.019 (2009).

18 Babu, M. et al. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Molecular microbiology 79, 484-502, doi:10.1111/j.1365-2958.2010.07465.x (2011).

19 Kim, T. Y., Shin, M., Huynh Thi Yen, L. & Kim, J. S. Crystal structure of Cas1 from Archaeoglobus fulgidus and characterization of its nucleolytic activity. Biochemical and biophysical research communications, doi:10.1016/j.bbrc.2013.10.122 (2013).

20 Beloglazova, N. et al. A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. The Journal of biological chemistry 283, 20361-20371, doi:10.1074/jbc.M803225200 (2008).

21 Samai, P., Smith, P. & Shuman, S. Structure of a CRISPR-associated protein Cas2 from Desulfovibrio vulgaris. Acta crystallographica. Section F, Structural biology and crystallization communications 66, 1552-1556, doi:10.1107/S1744309110039801 (2010).

22 Nam, K. H. et al. Double-stranded endonuclease activity in Bacillus halodurans clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. The Journal of biological chemistry 287, 35943-35952, doi:10.1074/jbc.M112.382598 (2012).

23 Li, M. & Craigie, R. Processing of viral DNA ends channels the HIV-1 integration reaction to concerted integration. The Journal of biological chemistry 280, 29334-29339, doi:10.1074/jbc.M505367200 (2005).

24 Cherepanov, P. LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro. Nucleic acids research 35, 113-124, doi:10.1093/nar/gkl885 (2007).

25 Hare, S. et al. A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75. PLoS pathogens 5, e1000259, doi:10.1371/journal.ppat.1000259 (2009).

26 Yang, J. Y., Jayaram, M. & Harshey, R. M. Positional information within the Mu transposase tetramer: catalytic contributions of individual monomers. Cell 85, 447-455 (1996).

27 DiNardo, S., Voelkel, K. A., Sternglanz, R., Reynolds, A. E. & Wright, A. Escherichia coli DNA topoisomerase I mutants have compensatory mutations in DNA gyrase genes. Cell 31, 43-51 (1982).

28 Pruss, G. J., Manes, S. H. & Drlica, K. Escherichia coli DNA topoisomerase I mutants: increased supercoiling is corrected by mutations near gyrase genes. Cell 31, 35-42 (1982).

29 Chow, S. A., Vincent, K. A., Ellison, V. & Brown, P. O. Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. Science 255, 723-726 (1992).

30 Au, T. K., Pathania, S. & Harshey, R. M. True reversal of Mu integration. The EMBO journal 23, 3408-3420, doi:10.1038/sj.emboj.7600344 (2004).

31 Engelman, A., Mizuuchi, K. & Craigie, R. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. Cell 67, 1211-1221 (1991).

32 Mizuuchi, K. & Adzuma, K. Inversion of the phosphate chirality at the target site of Mu DNA strand transfer: evidence for a one-step transesterification mechanism. Cell 66, 129-140 (1991).

33 Curcio, M. J. & Derbyshire, K. M. The outs and ins of transposition: from mu to kangaroo. Nature reviews. Molecular cell biology 4, 865-877, doi:10.1038/nrm1241 (2003).

34 Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, U. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. Nucleic acids research 42, 7884-7893, doi:10.1093/nar/gku510 (2014).

35 Tyson, G. W. & Banfield, J. F. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environmental microbiology 10, 200-207, doi:10.1111/j.1462-2920.2007.01444.x (2008).

36 Sheflin, L. G. & Kowalski, D. Altered DNA conformations detected by mung bean nuclease occur in promoter and terminator regions of supercoiled pBR322 DNA. Nucleic acids research 13, 6137-6154 (1985).

37 Goren, M. G., Yosef, I., Auster, O. & Qimron, U. Experimental definition of a clustered regularly interspaced short palindromic duplicon in Escherichia coli. Journal of molecular biology 423, 14-16, doi:10.1016/j.jmb.2012.06.037 (2012).

38 Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. & Severinov, K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. RNA biology 10, 716-725, doi:10.4161/rna.24325 (2013).

39 Shmakov, S. et al. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic acids research 42, 5907-5916, doi:10.1093/nar/gku226 (2014).

40 Deveau, H. et al. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. Journal of bacteriology 190, 1390-1400, doi:10.1128/JB.01412-07 (2008).

41 Semenova, E. et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proceedings of the National Academy of Sciences of the United States of America 108, 10098-10103, doi:10.1073/pnas.1104144108 (2011).

42 Westra, E. R. et al. Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. PLoS genetics 9, e1003742, doi:10.1371/journal.pgen.1003742 (2013).

43 Craigie, R. & Bushman, F. D. HIV DNA integration. Cold Spring Harbor perspectives in medicine 2, a006890, doi:10.1101/cshperspect.a006890 (2012).

44 Nowotny, M. Retroviral integrase superfamily: the structural perspective. EMBO reports 10, 144-151, doi:10.1038/embor.2008.256 (2009).

45 Hochstrasser, M. L. & Doudna, J. A. Cutting it close: CRISPR-associated endoribonuclease structure and function. Trends in Biochemical Sciences (2014).

46 Palecek, E. Local supercoil-stabilized DNA structures. Critical reviews in biochemistry and molecular biology 26, 151-226, doi:10.3109/10409239109081126 (1991).

[00183] While the present invention has been described with reference to the specific embodiments thereof, it should be understood by those skilled in the art that various changes may be made and equivalents may be substituted without departing from the true spirit and scope of the invention. In addition, many modifications may be made to adapt a particular situation, material, composition of matter, process, process step or steps, to the objective, spirit and scope of the present invention. All such modifications are intended to be within the scope of the claims appended hereto.

CLAIMS

What is claimed is:

1. A method of nucleic acid integration, the method comprising:

contacting a target DNA molecule with a donor DNA molecule and a Cas1 protein, wherein:

(a) the target DNA molecule comprises an AT-rich region; and

(b) the donor DNA molecule is linear,

wherein said contacting is not in a bacterial or archaeal cell, and provides for integration of the donor DNA molecule into the target DNA molecule.

2. The method according to claim 1, wherein said contacting is performed in the presence of a Cas2 protein.

3. The method according to claim 1 or claim 2, wherein said contacting is *in vitro* outside of a cell.

4. The method according to claim 1 or claim 2, wherein said contacting comprises introducing into a target cell: (i) the Cas1 protein, or a nucleic acid comprising nucleotides that encode the Cas1 protein; and (ii) the linear donor DNA molecule, wherein the target cell comprises the target DNA molecule.

5. The method according to claim 4, wherein the method comprises introducing into the target cell a Cas2 protein, or a nucleic acid comprising nucleotides that encode a Cas2 protein.

6. The method according to claim 4 or claim 5, wherein the method comprises introducing one or more of: the Cas1 protein and a Cas2 protein, into the target cell.

7. The method according to any of claims 4 to 6, wherein the Cas1 protein and the linear donor DNA molecule are introduced into the target cell as a targeting composition comprising the Cas1 protein and the linear donor DNA molecule.

8. The method according to claim 7, wherein the targeting composition comprises a Cas2 protein.

9. The method according to any of claims 1 to 8, wherein one or more of: the Cas1 protein and the Cas2 protein, is a protein that is isolated from a cell.

10. The method according to any of claims 1 to 9, wherein one or more of: the Cas1 protein and the Cas2 protein, comprises an affinity tag.

11. The method according to any of claims 1 to 10, wherein the method comprises, prior to said contacting, isolating one or more of: the Cas1 protein and the Cas2 protein, from a cell.

12. The method according to claim 11, wherein one or more of: the Cas1 protein and the Cas2 protein, comprises an affinity tag during said isolating.

13. The method according to claim 12, wherein the method comprises a step of removing one or more affinity tags prior to said contacting.

14. The method according to any of claims 4 to 13, wherein the method comprises introducing into the target cell a nucleic acid comprising a nucleotide sequence that encodes the Cas1 protein and/or a nucleic acid comprising a nucleotide sequence that encodes a Cas2 protein.

15. The method according to claim 14, wherein the nucleotide sequence that encodes the Cas1 protein and the nucleotide sequence that encodes the Cas2 protein are present on the same nucleic acid molecule.

16. The method according to claim 14 or claim 15, wherein one or more of: the nucleotide sequence that encodes the Cas1 protein and the nucleotide sequence that encodes the Cas2 protein, is operably linked to a promoter that is operable in the target cell.

17. The method according to any of claims 4 to 16, wherein one or more of: the nucleotide sequence that encodes the Cas1 protein and the nucleotide sequence that encodes the Cas2 protein, is codon optimized for expression in the target cell.

18. The method according to any of claims 1-17, wherein said contacting is performed in the presence of an integration host factor (IHF) protein.

19. The method according to any of claims 1-18, wherein the method comprises introducing into the target cell an IHF protein, or a nucleic acid comprising nucleotides that encode an IHF protein.

20. The method according to any of claims 4 to 17, wherein the target cell is selected from: a eukaryotic cell, a plant cell, a cell of a single-celled eukaryotic organism, a fungal cell, an animal cell, a vertebrate cell, an invertebrate cell, a frog cell, a fish cell, a rodent cell, a mammalian cell, and a human cell.

21. The method according to any of claims 1 to 18, wherein the target DNA molecule is a chromosome.

22. The method according to any of claims 1 to 18, wherein the target DNA molecule is a supercoiled plasmid DNA.

23. The method according to any of claims 1 to 20, wherein the target DNA molecule does not contain a leader sequence or a CRISPR repeat sequence from a naturally existing CRISPR locus.

24. The method according to any of claims 1 to 21, wherein the target DNA molecule does not contain a naturally existing CRISPR locus.

25. The method according to any of claims 1 to 24, wherein the AT-rich region is positioned 5' and within 50 nucleotides of a region that forms a DNA cruciform structure.

26. The method according to claim 25, wherein the length of an upper and lower stem of the DNA cruciform structure is in a range of from 5 to 30 base pairs.

27. The method according to any of claims 1 to 26, wherein the linear donor DNA molecule comprises a 3' overhang with a length of from 1 to 6 nucleotides.

28. The method according to any of claims 1 to 27, wherein the linear donor DNA molecule has a length in a range of from 10 to 500 nucleotides (nt).

29. The method according to any of claims 1 to 27, wherein the linear donor DNA molecule has a length of from 35 to 500 nucleotides (nt).

30. The method according to any of claims 1 to 29, wherein the 3' most nucleotide on at least one end of the donor DNA molecule is a C.

31. The method according to any of claims 1 to 30, wherein the Cas1 protein comprises an amino acid sequence having 70% or more sequence identity with a Cas1 protein amino acid sequence set forth in any one of SEQ ID NOs: 28-86.

32. The method according to any of claims 1 to 31, wherein said contacting is performed in the presence of a Cas2 protein that comprises an amino acid sequence having 70% or more sequence identity with a Cas2 protein amino acid sequence set forth in any one of SEQ ID NOs: 87-161.

33. The method according to any of claims 1 to 32, wherein the Cas1 protein comprises an amino acid sequence modification relative to the corresponding naturally occurring Cas1 protein, and/or the Cas2 protein comprises an amino acid sequence modification relative to the corresponding naturally occurring Cas2 protein.

34. The method according to any of claims 1 to 33, wherein one or more of: the Cas1 protein and the Cas2 protein, is a fusion protein comprising one or more heterologous fusion partners.

35. The method according to claim 34, wherein one or more of: the Cas1 protein and the Cas2 protein, is a fusion protein comprising one or more heterologous fusion partners selected from: a fluorescent protein, a subcellular localization signal; and an affinity tag.

36. The method according to any of claims 1 to 35, wherein one or more of: the Cas1 protein and the Cas2 protein, comprises one or more nuclear localization signals (NLSs).

37. A composition comprising:

(a) a Cas1 protein or a nucleic acid comprising a nucleotide sequence that encodes a Cas1 protein; and

(b) at least one of:

(i) a donor DNA molecule, wherein the donor DNA molecule is a linear DNA molecule; and

(ii) a Cas2 protein or a nucleic acid comprising a nucleotide sequence that encodes a Cas2 protein.

38. The composition according to claim 37, wherein the linear donor DNA molecule has a length of from 10 to 500 nucleotides (nt).

39. The composition according to claim 38, wherein the linear donor DNA molecule has a length of from 35 to 500 nucleotides (nt).

40. The composition according to any of claims 37 to 39, wherein the 3' most nucleotide on at least one end of the linear donor DNA molecule is a C.

41. The composition according to any of claims 37 to 40, wherein the Cas1 protein comprises an amino acid sequence modification relative to the corresponding naturally occurring Cas1 protein, and/or the Cas2 protein comprises an amino acid sequence modification relative to the corresponding naturally occurring Cas2 protein.

42. The composition according to any of claims 37 to 41, wherein one or more of: the Cas1 protein and the Cas2 protein, is a fusion protein comprising one or more heterologous fusion partners.

43. The composition according to any of claims 37 to 42, wherein one or more of: the Cas1 protein and the Cas2 protein, is a fusion protein comprising one or more heterologous fusion partners selected from: a fluorescent protein, a subcellular localization signal; and an affinity tag.

44. The composition according to any of claims 37 to 43, wherein one or more of: the Cas1 protein and the Cas2 protein, comprises one or more nuclear localization signals (NLSs).

45. A kit comprising:

(a) a Cas1 protein or a nucleic acid comprising a nucleotide sequence that encodes a Cas1 protein; and

(b) at least one of:

(i) a donor DNA molecule, wherein the donor DNA molecule is a linear DNA molecule; and

(ii) a Cas2 protein or a nucleic acid comprising a nucleotide sequence that encodes a Cas2 protein.

wherein (a) and (b) are present in the same or separate containers.

46. The kit according to claim 45, wherein (a) and (b) are present in separate containers.

47. The kit according to claim 45 or claim 46, wherein the linear donor DNA molecule has a length of from 10 to 500 nucleotides (nt).

48. The kit according to any of claims 45 to 47, wherein the linear donor DNA molecule has a length of from 35 to 500 nucleotides (nt).

49. The kit according to any of claims 45 to 48, wherein the 3' most nucleotide on at least one end of the linear donor DNA molecule is a C.

50. The kit according to any of claims 45 to 49, wherein the Cas1 protein comprises an amino acid sequence modification relative to the corresponding naturally occurring Cas1 protein, and/or the Cas2 protein comprises an amino acid sequence modification relative to the corresponding naturally occurring Cas2 protein.

51. The kit according to any of claims 45 to 50, wherein one or more of: the Cas1 protein and the Cas2 protein, is a fusion protein comprising one or more heterologous fusion partners.

52. The kit according to any of claims 45 to 51, wherein one or more of: the Cas1 protein and the Cas2 protein, is a fusion protein comprising one or more heterologous fusion partners selected from: a fluorescent protein, a subcellular localization signal; and an affinity tag.

53. The kit according to any of claims 45 to 52, wherein one or more of: the Cas1 protein and the Cas2 protein, comprises one or more nuclear localization signals (NLSs).

54. An isolated Cas1 protein, or a nucleic acid encoding a Cas1 protein, wherein the Cas1 protein comprises an amino acid sequence modification relative to the corresponding naturally occurring Cas1 protein..

55. The isolated Cas1 protein, or a nucleic acid encoding a Cas1 protein according to claim 54, wherein one or more of: the Cas1 protein and the Cas2 protein, is a fusion protein comprising one or more heterologous fusion partners.

56. The isolated Cas1 protein, or a nucleic acid encoding a Cas1 protein according to claim 54 or claim 55, wherein the Cas1 protein is a fusion protein comprising one or more heterologous fusion partners selected from: a fluorescent protein, a subcellular localization signal; and an affinity tag.

57. The isolated Cas1 protein, or a nucleic acid encoding a Cas1 protein according to any of claims 54 to 56, wherein the Cas1 protein comprises one or more nuclear localization signals (NLSs).

58. An isolated Cas2 protein, or a nucleic acid encoding a Cas2 protein, wherein the Cas2 protein comprises an amino acid sequence modification relative to the corresponding naturally occurring Cas2 protein..

59. The isolated Cas2 protein, or a nucleic acid encoding a Cas2 protein according to claim 58, wherein one or more of: the Cas2 protein and the Cas2 protein, is a fusion protein comprising one or more heterologous fusion partners.

60. The isolated Cas2 protein, or a nucleic acid encoding a Cas2 protein according to claim 58 or claim 59, wherein the Cas2 protein is a fusion protein comprising one or more heterologous fusion partners selected from: a fluorescent protein, a subcellular localization signal; and an affinity tag.

61. The isolated Cas2 protein, or a nucleic acid encoding a Cas2 protein according to any of claims 58 to 60, wherein the Cas2 protein comprises one or more nuclear localization signals (NLSs).

Figure 1

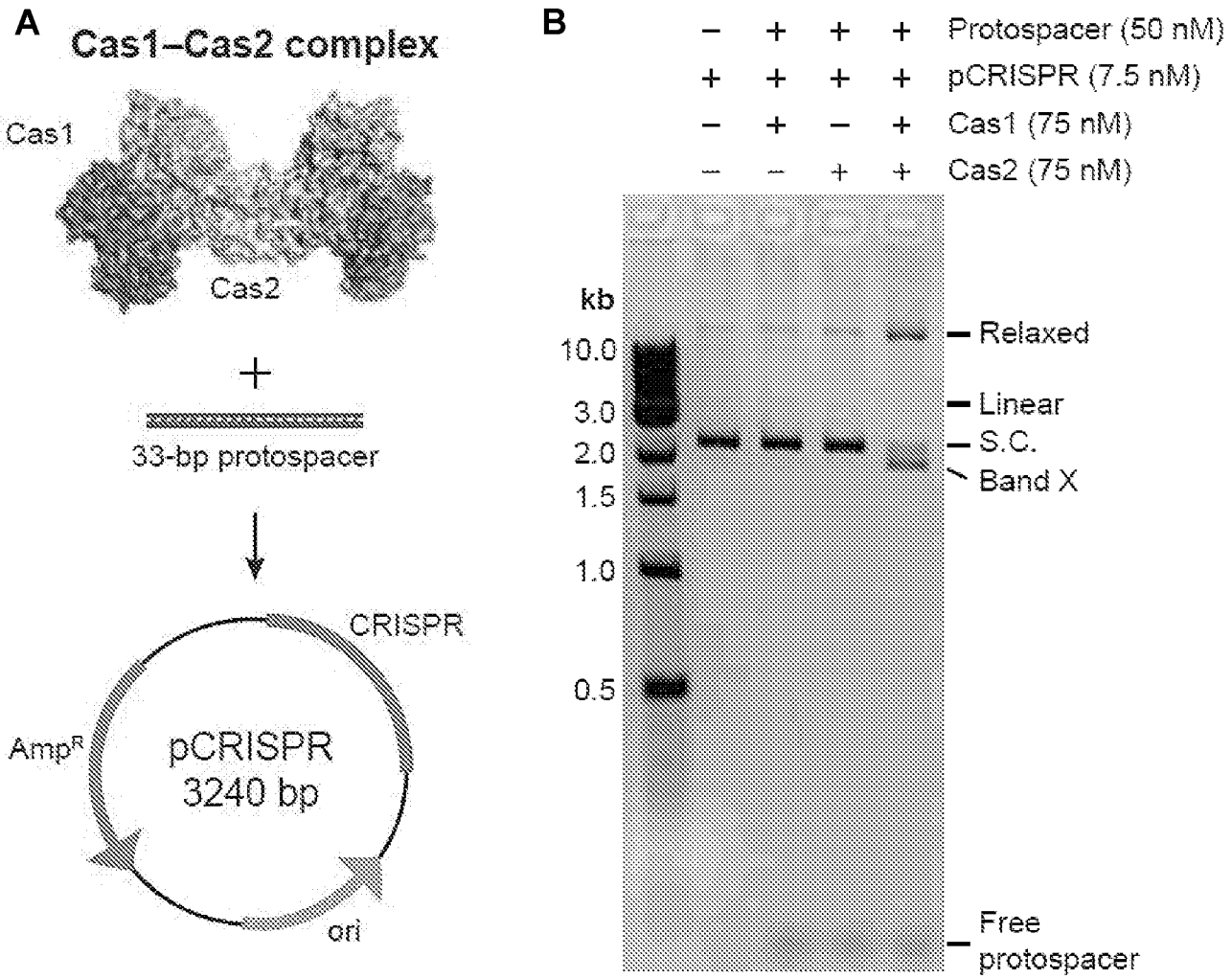

A
Cas1–Cas2 complex
Cas1
Cas2
+
33-bp protospacer
CRISPR
AmpR
pCRISPR
3240 bp
ori
B
– + + + Protospacer (50 nM)
+ + + + pCRISPR (7.5 nM)
– + – + Cas1 (75 nM)
– – + + Cas2 (75 nM)
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
S.C.
Band X
Free protospacer

Figure 1

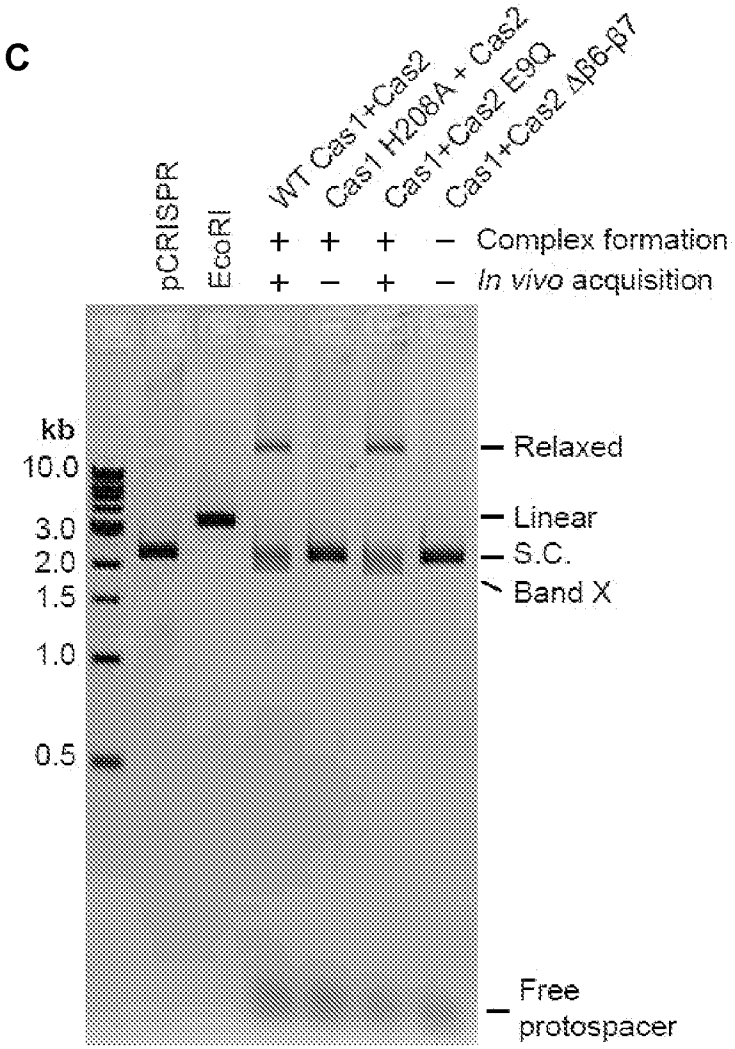

C
pCRISPR
EcoRI
WT Cas1+Cas2
Cas1 H208A + Cas2
Cas1+Cas2 E9Q
Cas1+Cas2 Δβ6-β7
+ + + – Complex formation
+ – + – In vivo acquisition
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
S.C.
Band X
Free protospacer

Figure 1

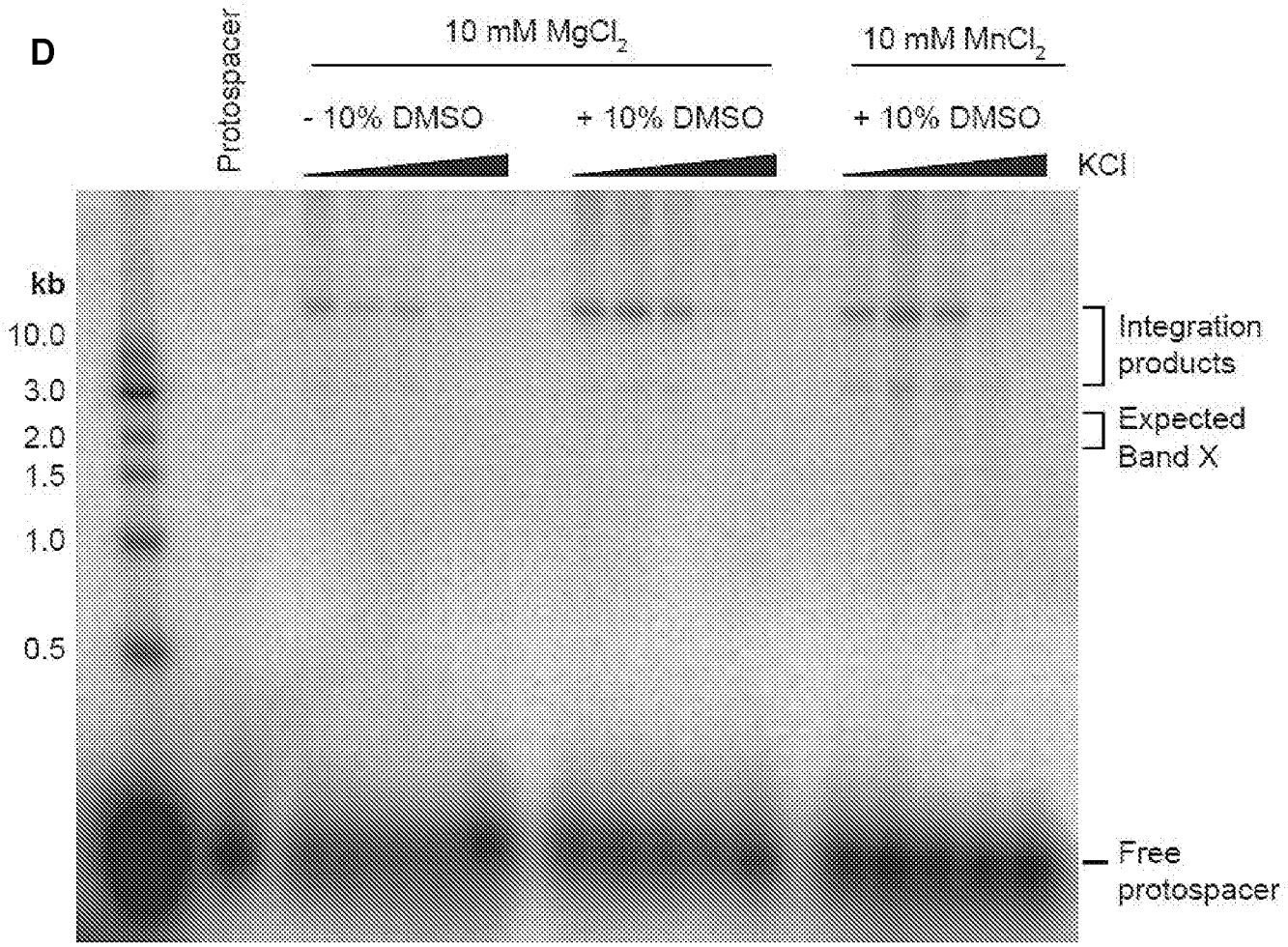

D
Protospacer
10 mM MgCl2
10 mM MnCl2
- 10% DMSO
+ 10% DMSO
+ 10% DMSO
KCl
kb
10.0
3.0
2.0
1.5
1.0
0.5
Integration products
Expected Band X
Free protospacer

Figure 1

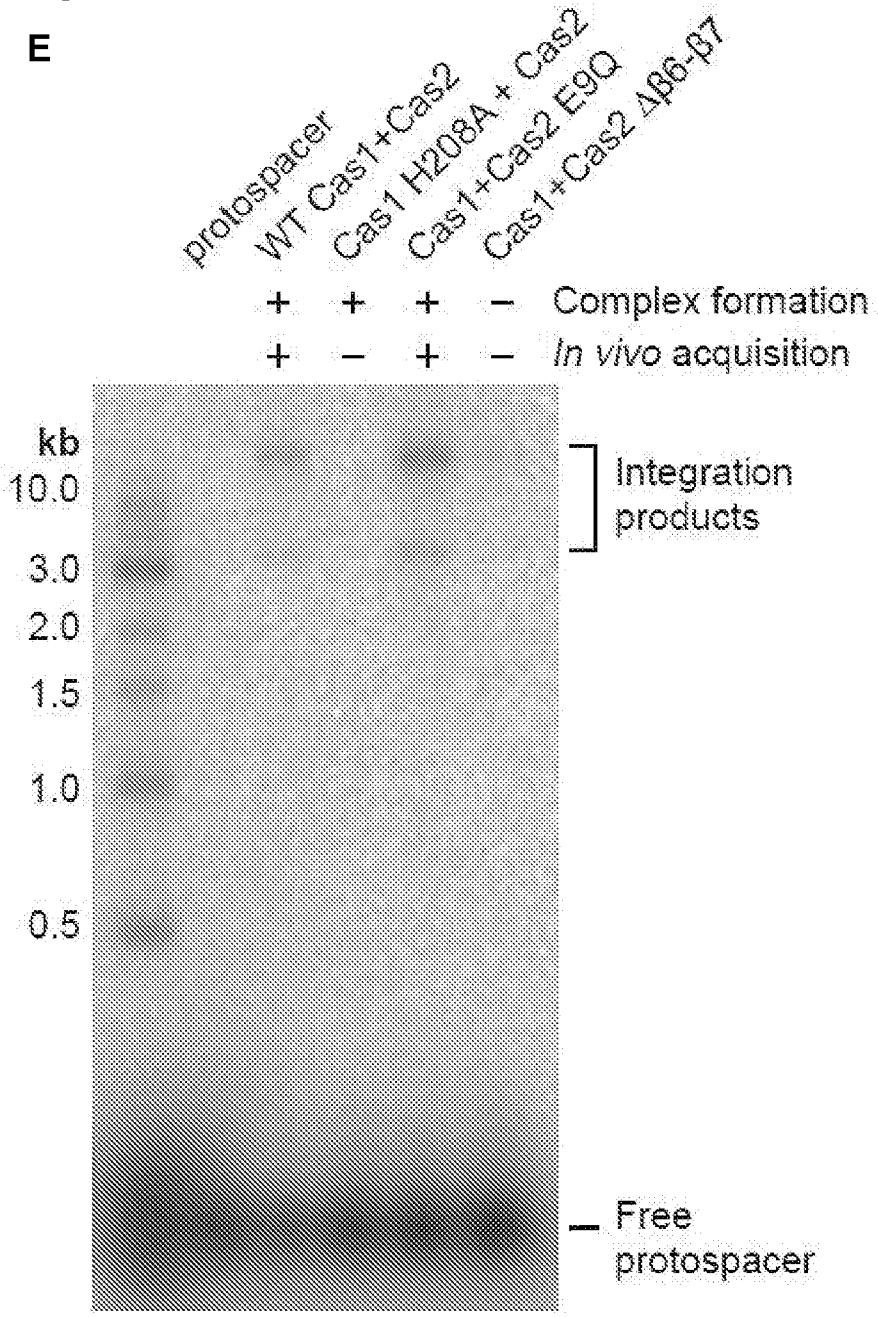

E
protospacer
WT Cas1+Cas2
Cas1 H208A + Cas2
Cas1+Cas2 E9Q
Cas1+Cas2 Δβ6-β7
+ + + – Complex formation
+ – + – In vivo acquisition
kb
10.0
3.0
2.0
1.5
1.0
0.5
Integration products
Free protospacer

Figure 2

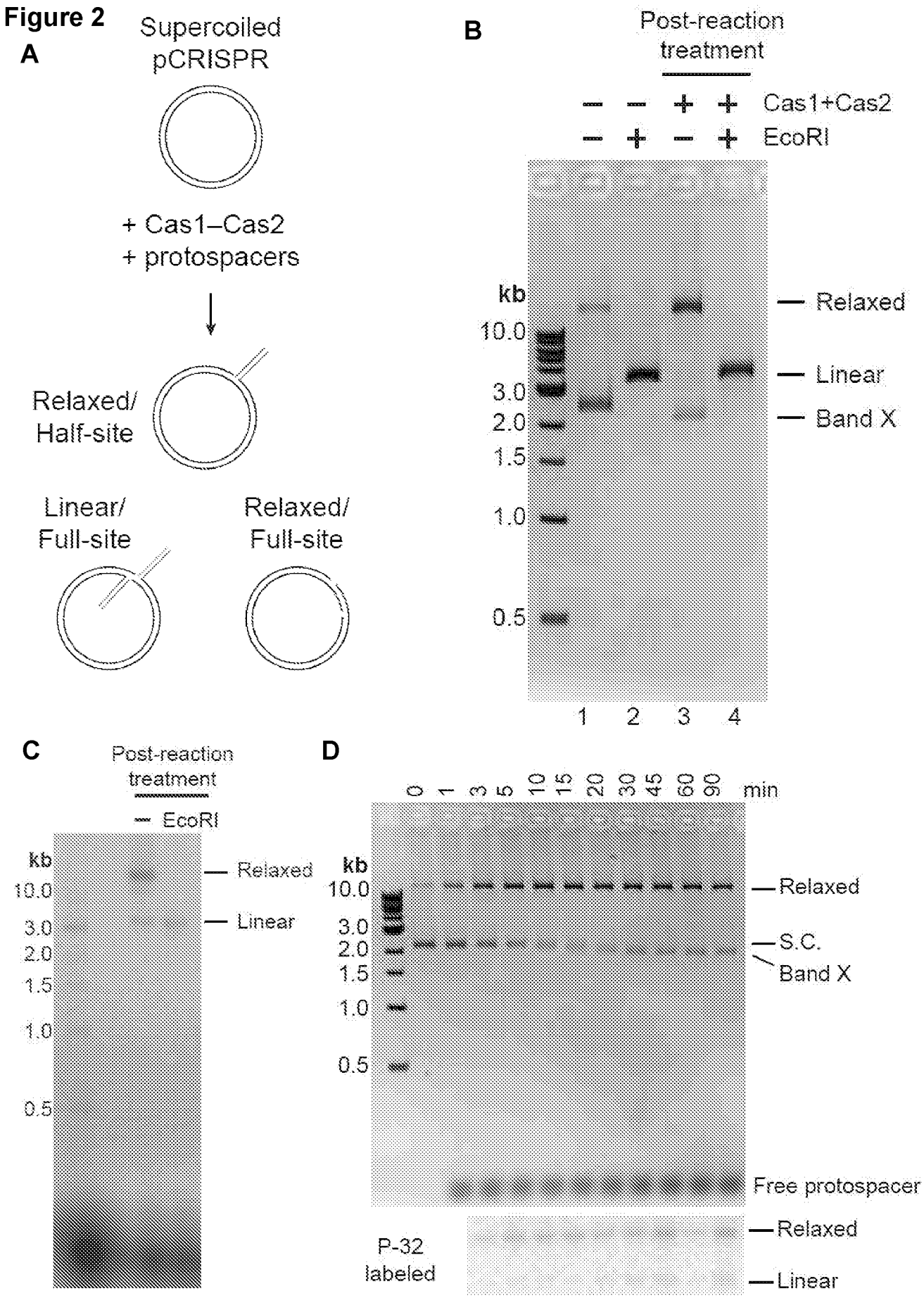

A
Supercoiled pCRISPR
+ Cas1–Cas2
+ protospacers
Relaxed/ Half-site
Linear/ Full-site
Relaxed/ Full-site
B
Post-reaction treatment
Cas1+Cas2
EcoRI
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
Band X
1 2 3 4
C
Post-reaction treatment
EcoRI
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
D
0 1 3 5 10 15 20 30 45 60 90 min
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
S.C.
Band X
Free protospacer
P-32 labeled
Relaxed
Linear

Figure 2

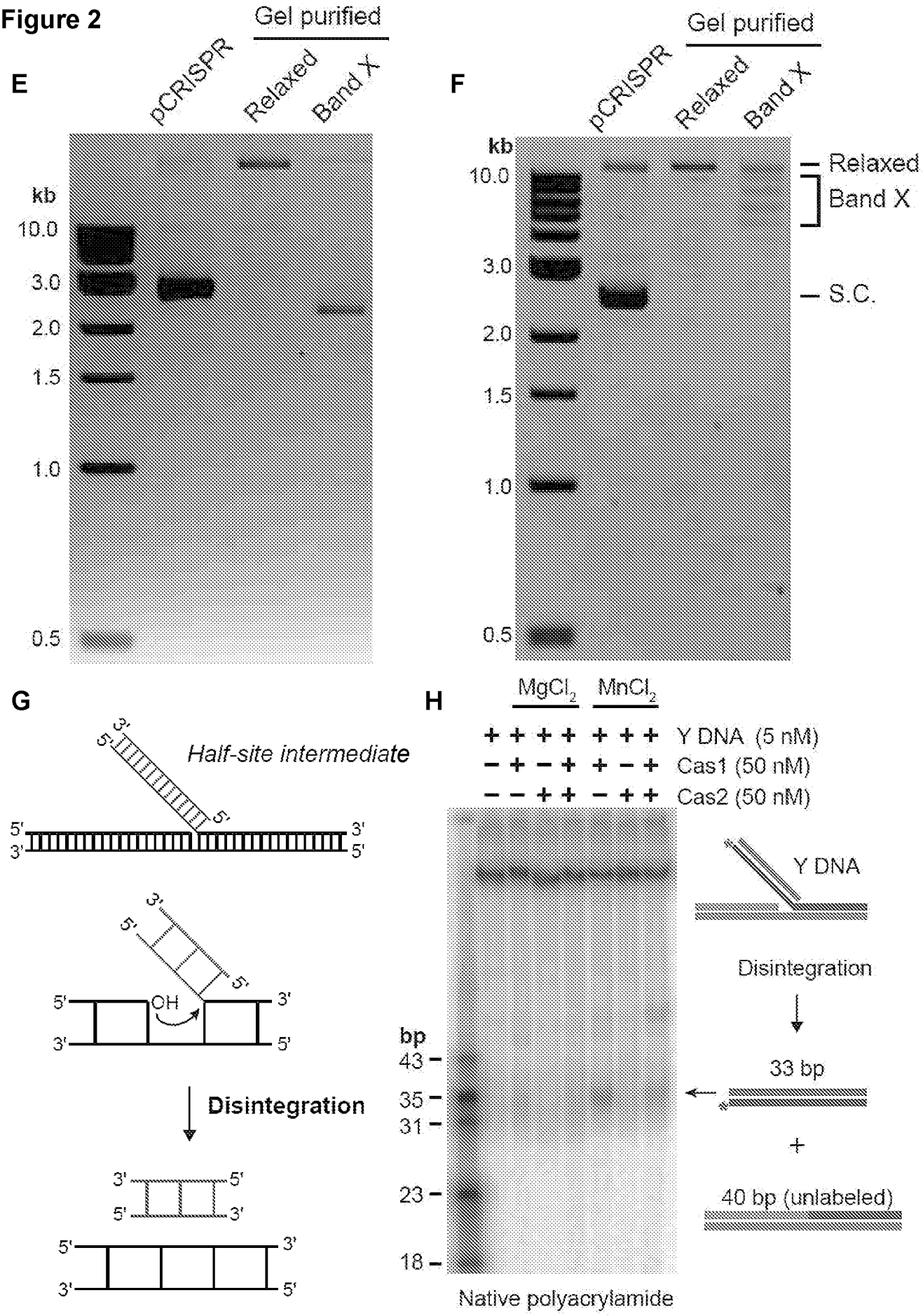

E
Gel purified
pCRISPR
Relaxed
Band X
kb
10.0
3.0
2.0
1.5
1.0
0.5
F
Gel purified
pCRISPR
Relaxed
Band X
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Band X
S.C.
G
Half-site intermediate
3'
5'
OH
Disintegration
H
MgCl2
MnCl2
+ + + + + + + Y DNA (5 nM)
− + − + + − + Cas1 (50 nM)
− − + + − + + Cas2 (50 nM)
bp
43
35
31
23
18
Native polyacrylamide
Y DNA
Disintegration
33 bp
+
40 bp (unlabeled)

Figure 3

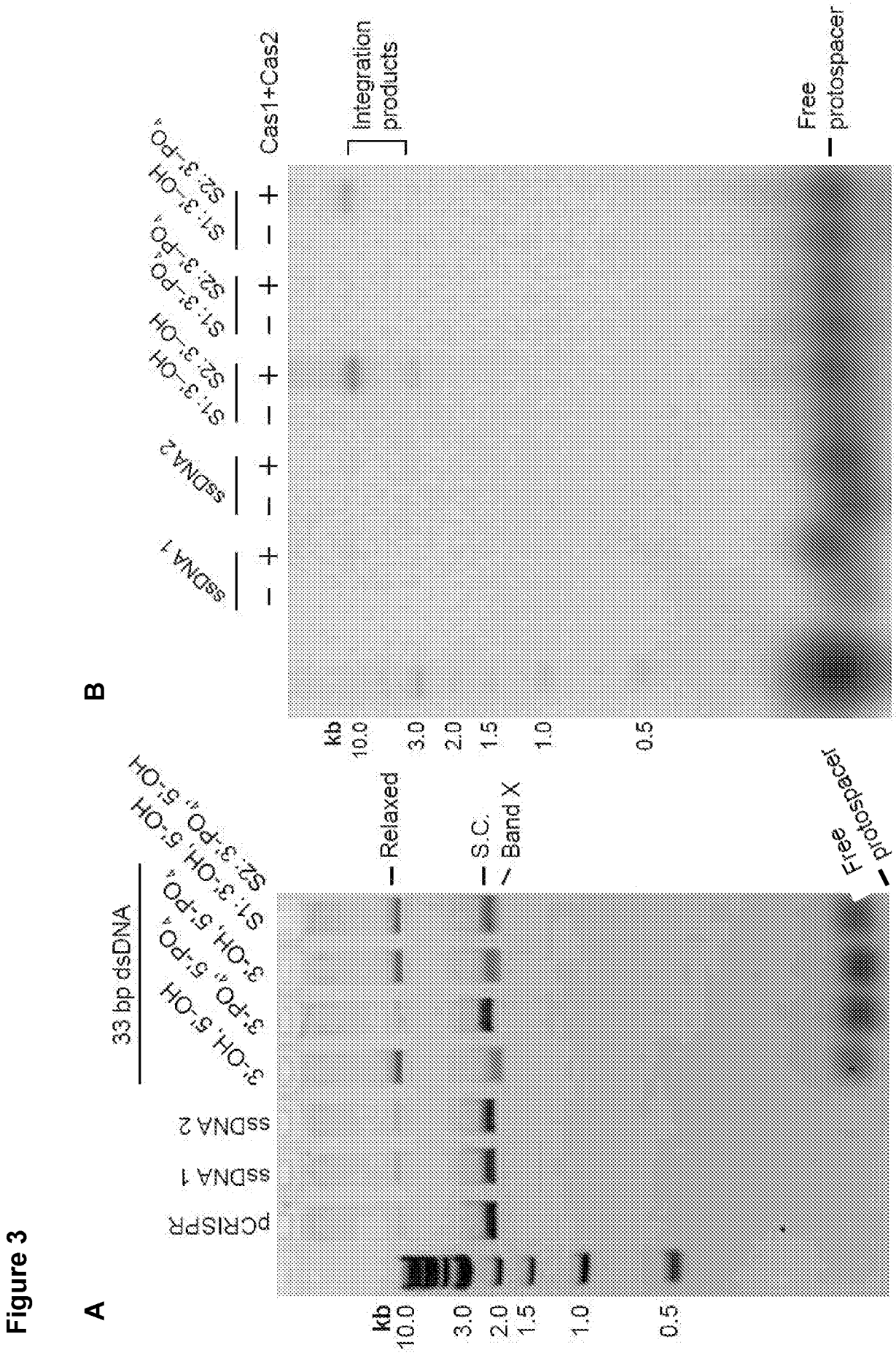

A
pCRISPR
ssDNA 1
ssDNA 2
33 bp dsDNA
3'-OH, 5'-OH
3'-PO4, 5'-PO4
3'-OH, 5'-PO4
S1: 3'-OH, 5'-OH
S2: 3'-PO4, 5'-OH
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
S.C.
Band X
Free protospacer
B
ssDNA 1
ssDNA 2
Cas1+Cas2
S1: 3'-OH
S2: 3'-OH
S1: 3'-PO4
S2: 3'-OH
S1: 3'-OH
S2: 3'-PO4
S1: 3'-PO4
S2: 3'-PO4
Integration products
Free protospacer

Figure 3

C

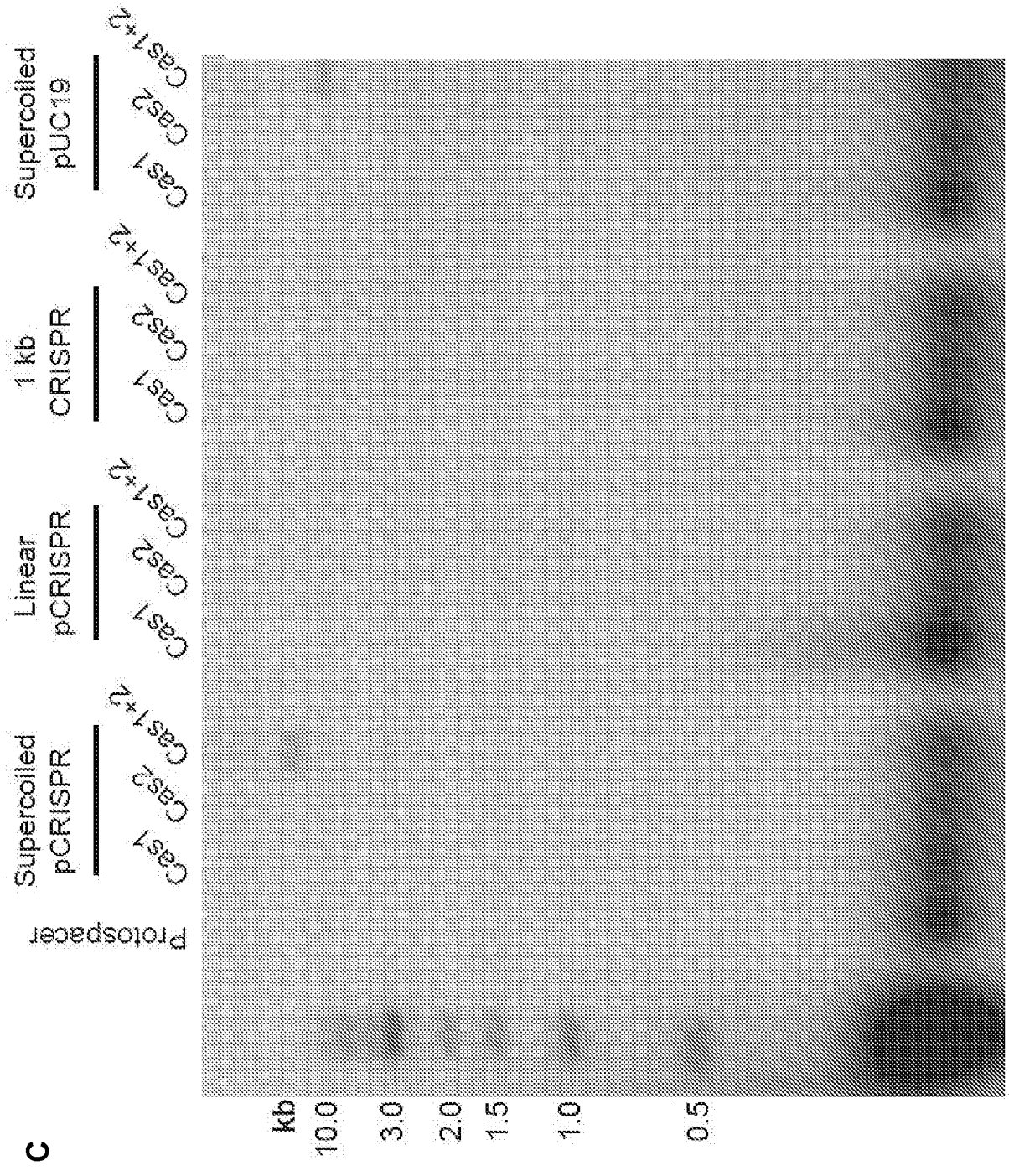

kb
10.0
3.0
2.0
1.5
1.0
0.5
Protospacer
Supercoiled pCRISPR
Linear pCRISPR
1 kb CRISPR
Supercoiled pUC19
Cas1
Cas2
Cas1+2

Figure 3

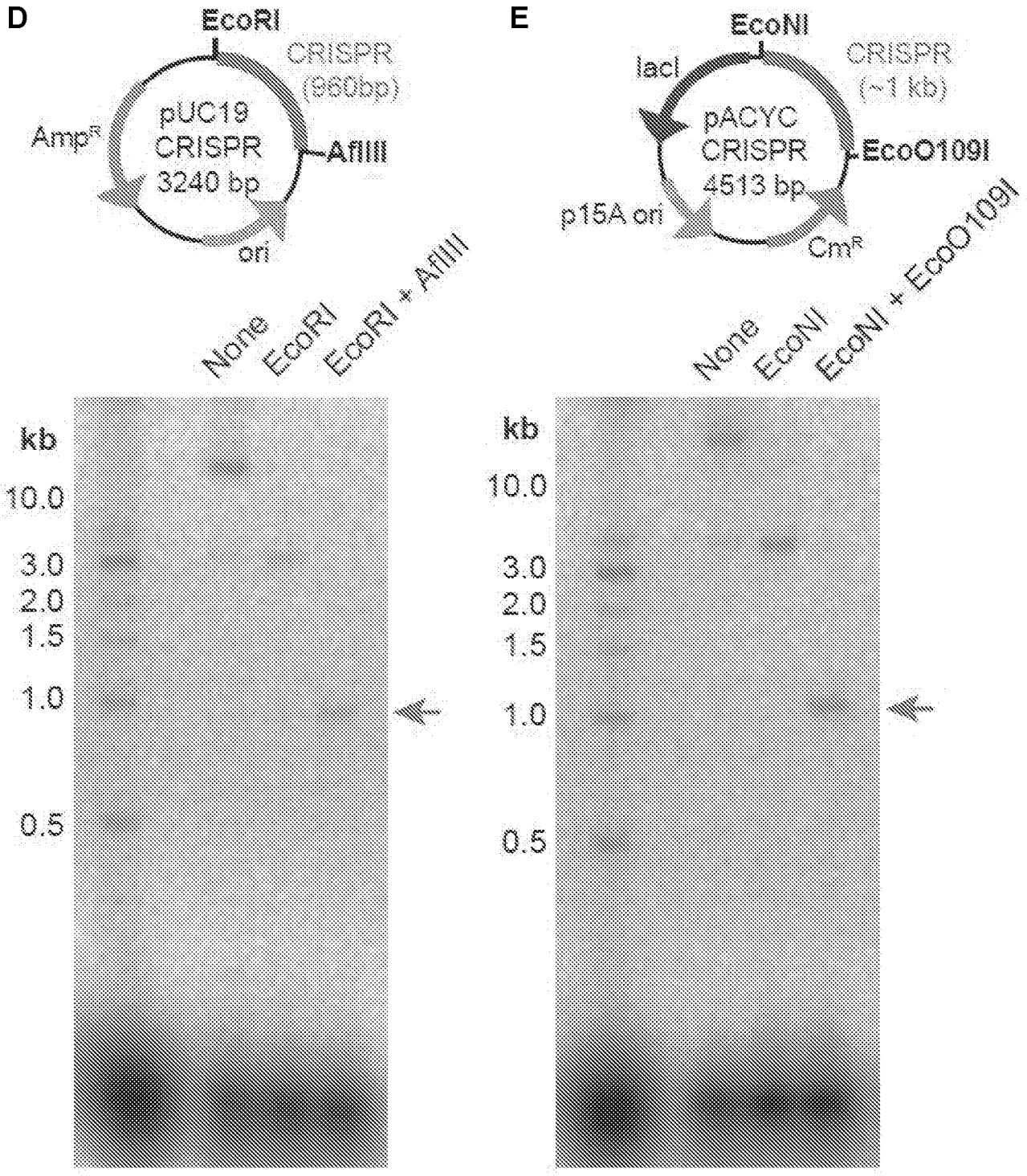

D
EcoRI
CRISPR
(960bp)
pUC19
CRISPR
3240 bp
AmpR
AflIII
ori
None
EcoRI
EcoRI + AflIII
kb
10.0
3.0
2.0
1.5
1.0
0.5
E
EcoNI
lacI
CRISPR
(~1 kb)
pACYC
CRISPR
4513 bp
EcoO109I
p15A ori
CmR
None
EcoNI
EcoNI + EcoO109I
kb
10.0
3.0
2.0
1.5
1.0
0.5

Figure 4

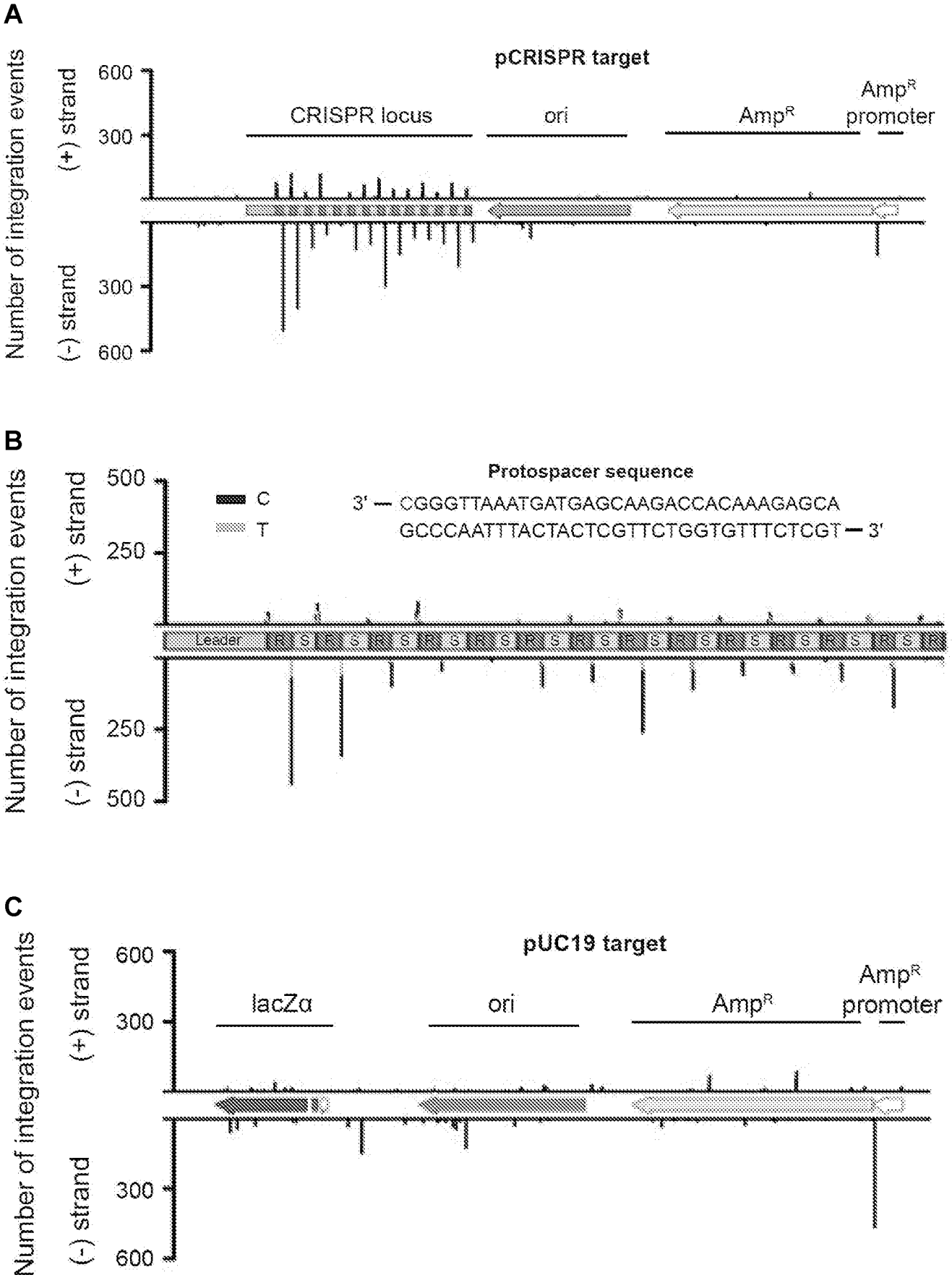
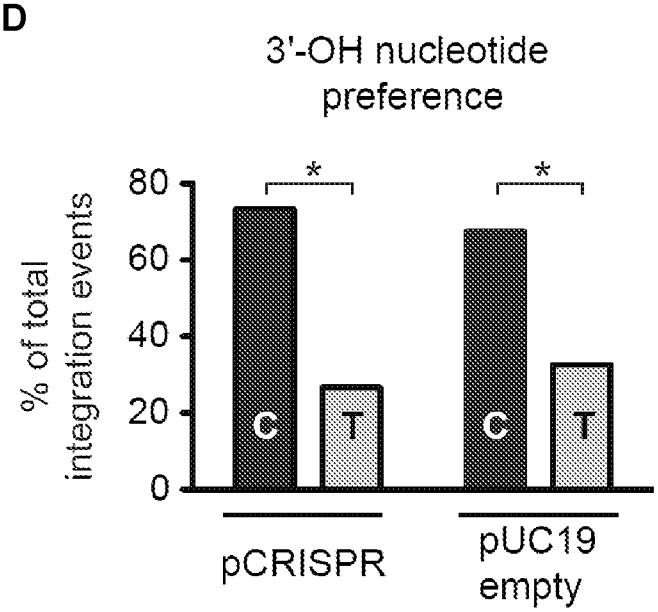

A
pCRISPR target
CRISPR locus
ori
Amp^R
Amp^R promoter
Number of integration events
(+) strand
(-) strand
600
300
300
600
B
Protospacer sequence
C
T
3' — CGGGTTAAATGATGAGCAAGACCACAAAGAGCA
GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT — 3'
Leader
Number of integration events
(+) strand
(-) strand
500
250
250
500
C
pUC19 target
lacZα
ori
Amp^R
Amp^R promoter
Number of integration events
(+) strand
(-) strand
600
300
300
600

Figure 4

D


3'-OH nucleotide preference
% of total integration events
80
60
40
20
0
*
*
C
T
C
T
pCRISPR
pUC19 empty

E

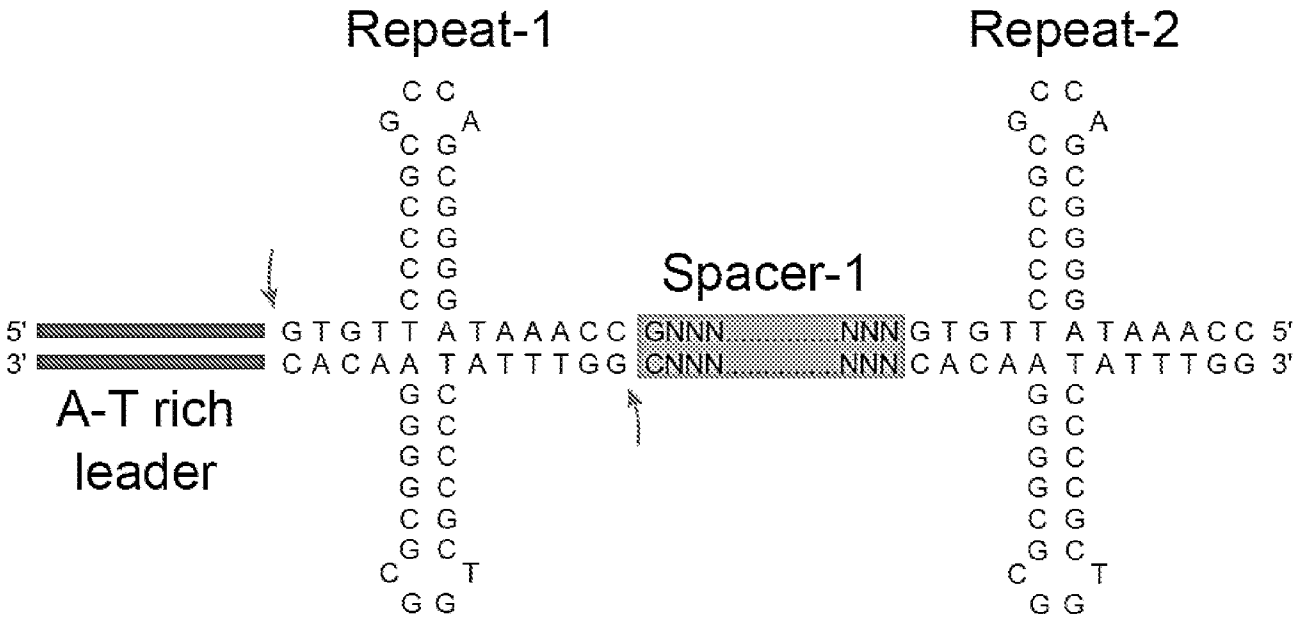

Repeat-1
Repeat-2
Spacer-1
5'
3'
GTGTTATAAACC
CACAATATTTGG
GNNN NNN
CNNN NNN
A-T rich leader

Figure 5

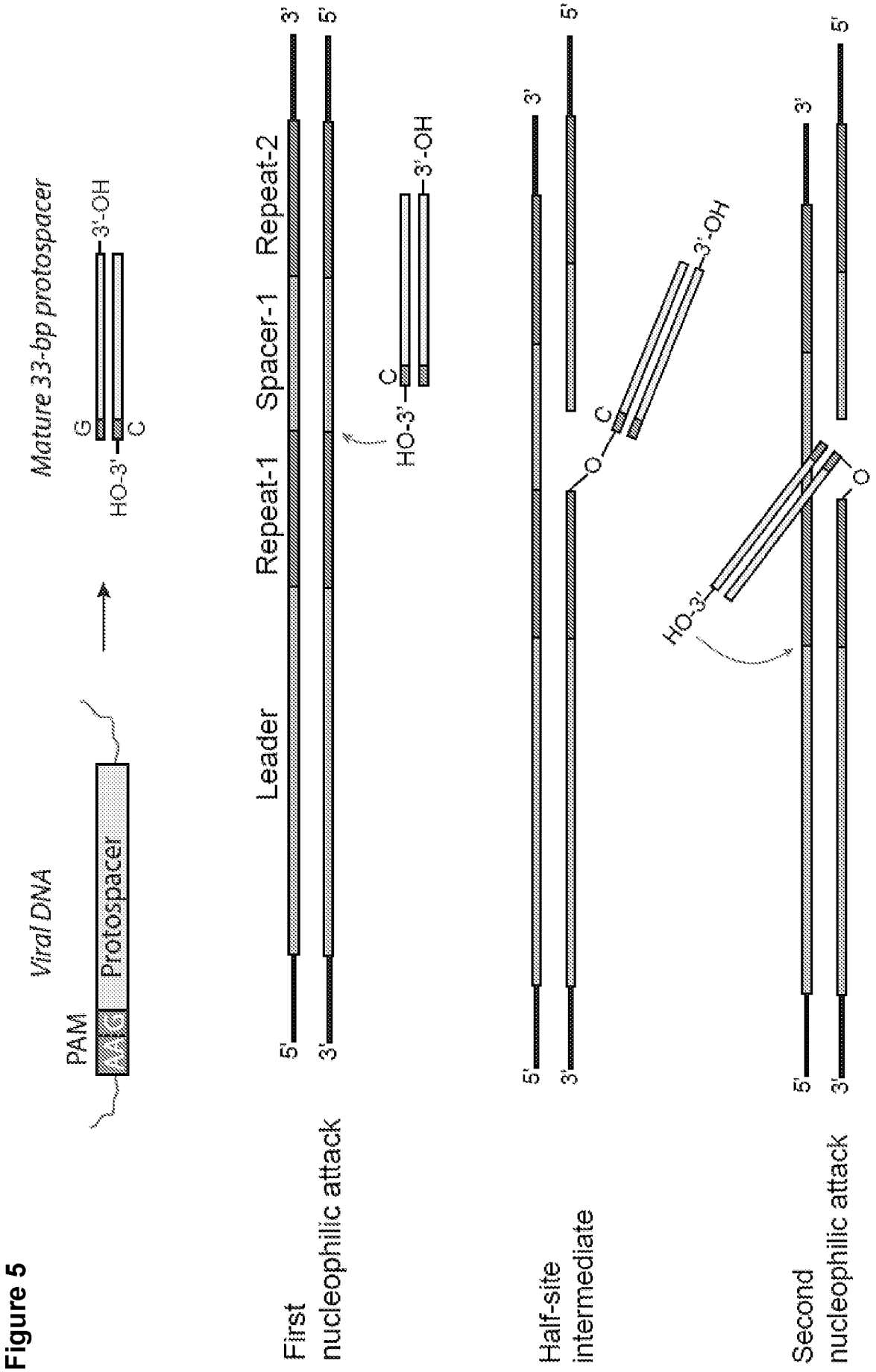

Viral DNA
PAM
Protospacer
Mature 33-bp protospacer
G
C
3'-OH
HO-3'
Leader
Repeat-1
Spacer-1
Repeat-2
5'
3'
First nucleophilic attack
Half-site intermediate
Second nucleophilic attack
O

Figure 5 (Cont.)

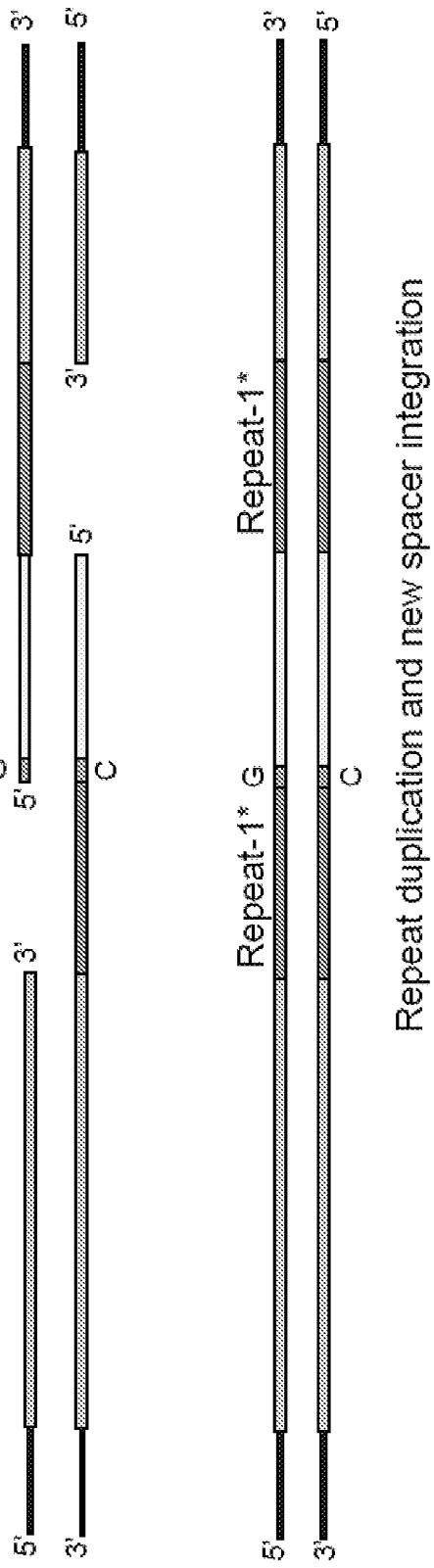


Figure 6

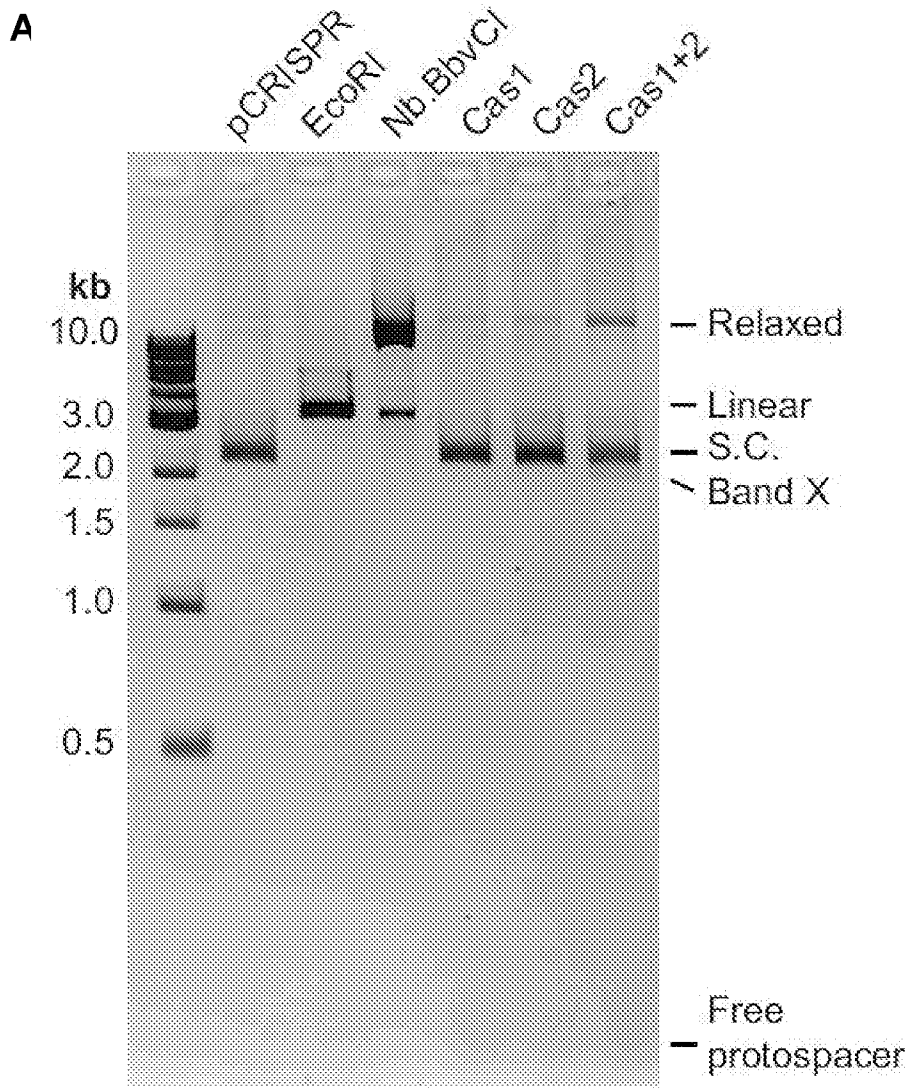

A
pCRISPR
EcoRI
Nb.BbvCI
Cas1
Cas2
Cas1+2
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
S.C.
Band X
Free
protospacer

Figure 6

B

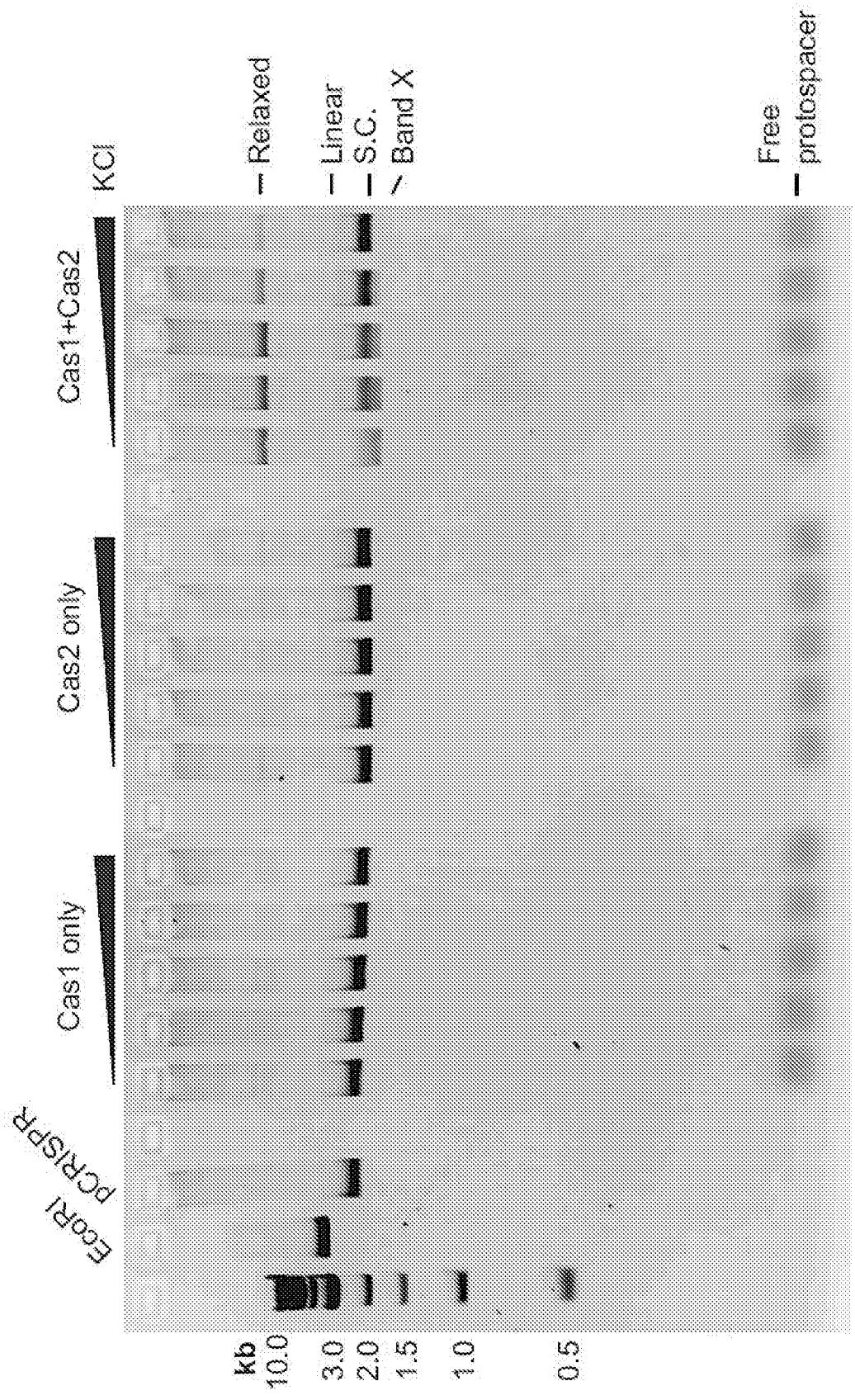

EcoRI PCRISPR
Cas1 only
Cas2 only
Cas1+Cas2
KCl
Relaxed
Linear
S.C.
Band X
Free protospacer
kb
10.0
3.0
2.0
1.5
1.0
0.5

Figure 6

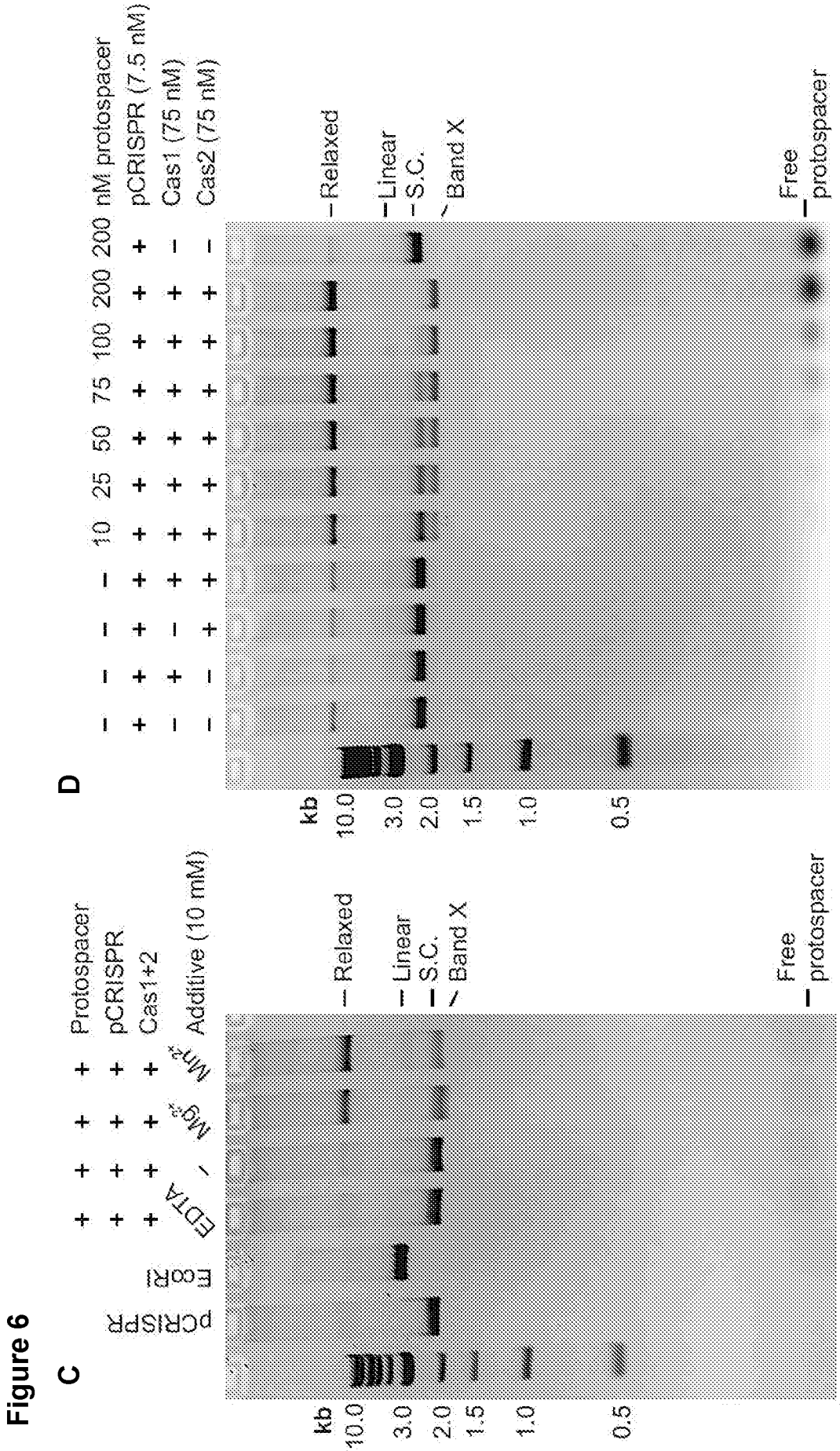

C
pCRISPR
EcoRI
EDTA + + + +
- + + +
Mg2+ + + + +
Mn2+ + + + +
Additive (10 mM) + + + +
Protospacer + + + +
pCRISPR + + + +
Cas1+2 + + + +
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
S.C.
Band X
Free
protospacer
D
nM protospacer 10 25 50 75 100 200 200
pCRISPR (7.5 nM) - + + + + + + +
Cas1 (75 nM) - + + + + + - -
Cas2 (75 nM) - + + + + + + -
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
S.C.
Band X
Free
protospacer

Figure 6

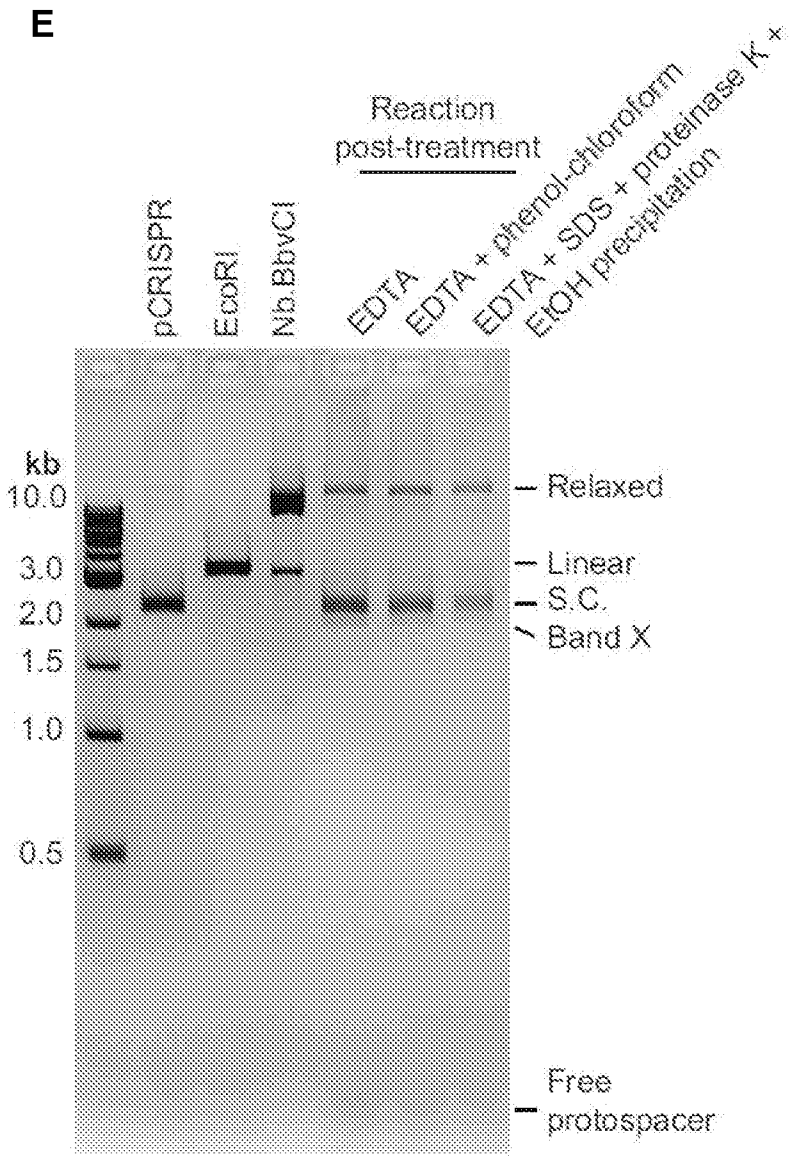

E
Reaction post-treatment
pCRISPR
EcoRI
Nb.BbvCI
EDTA
EDTA + phenol-chloroform
EDTA + SDS + proteinase K +
EtOH precipitation
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
S.C.
Band X
Free protospacer

Figure 7

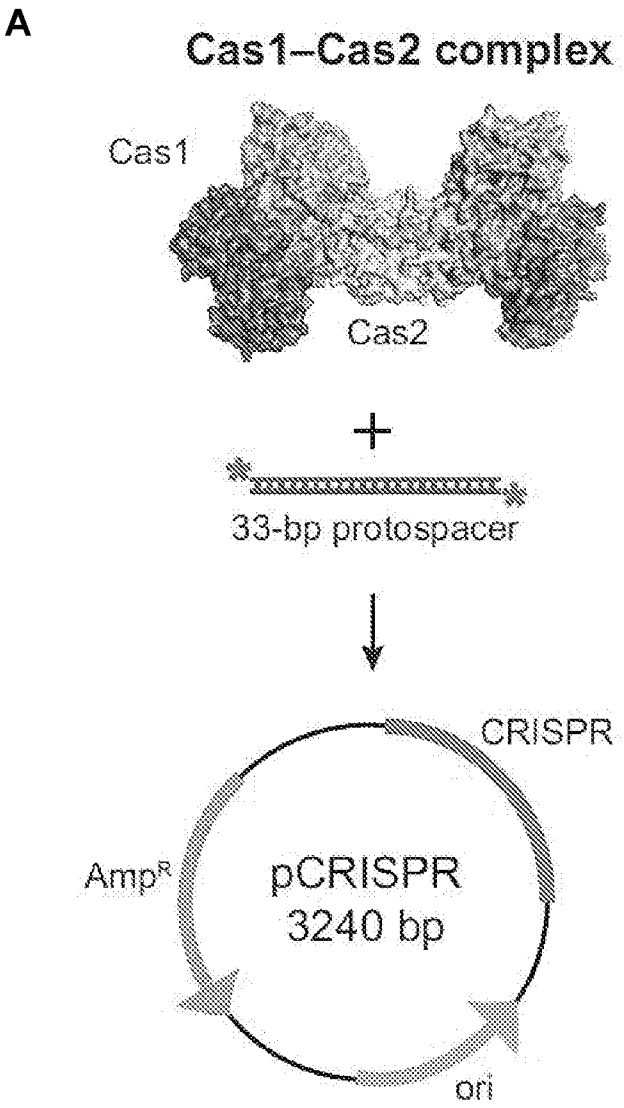

A
Cas1–Cas2 complex
Cas1
Cas2
+
33-bp protospacer
CRISPR
Amp^R
pCRISPR
3240 bp
ori

Figure 7

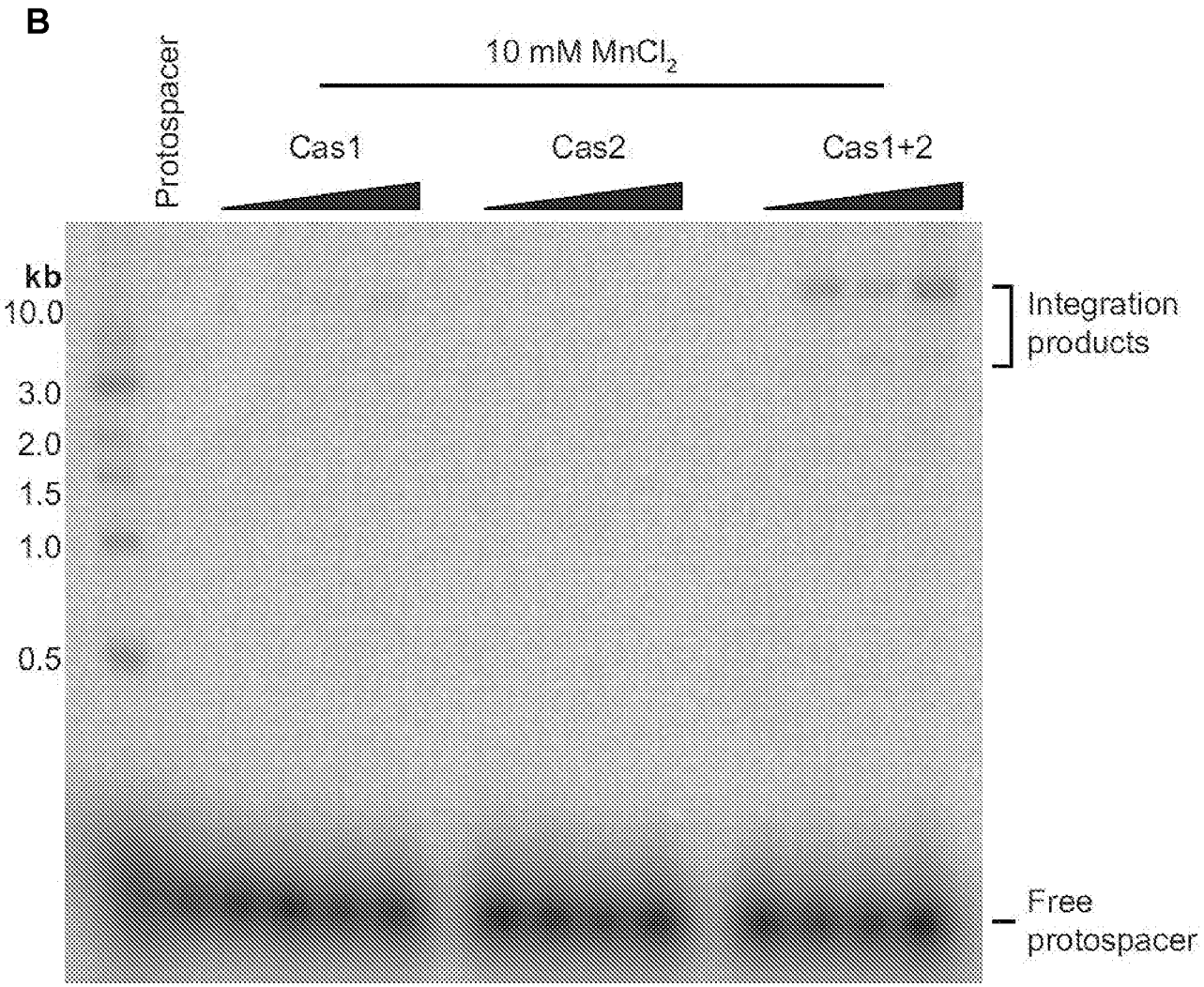

B
Protospacer
10 mM MnCl2
Cas1
Cas2
Cas1+2
kb
10.0
3.0
2.0
1.5
1.0
0.5
Integration products
Free protospacer

Figure 7

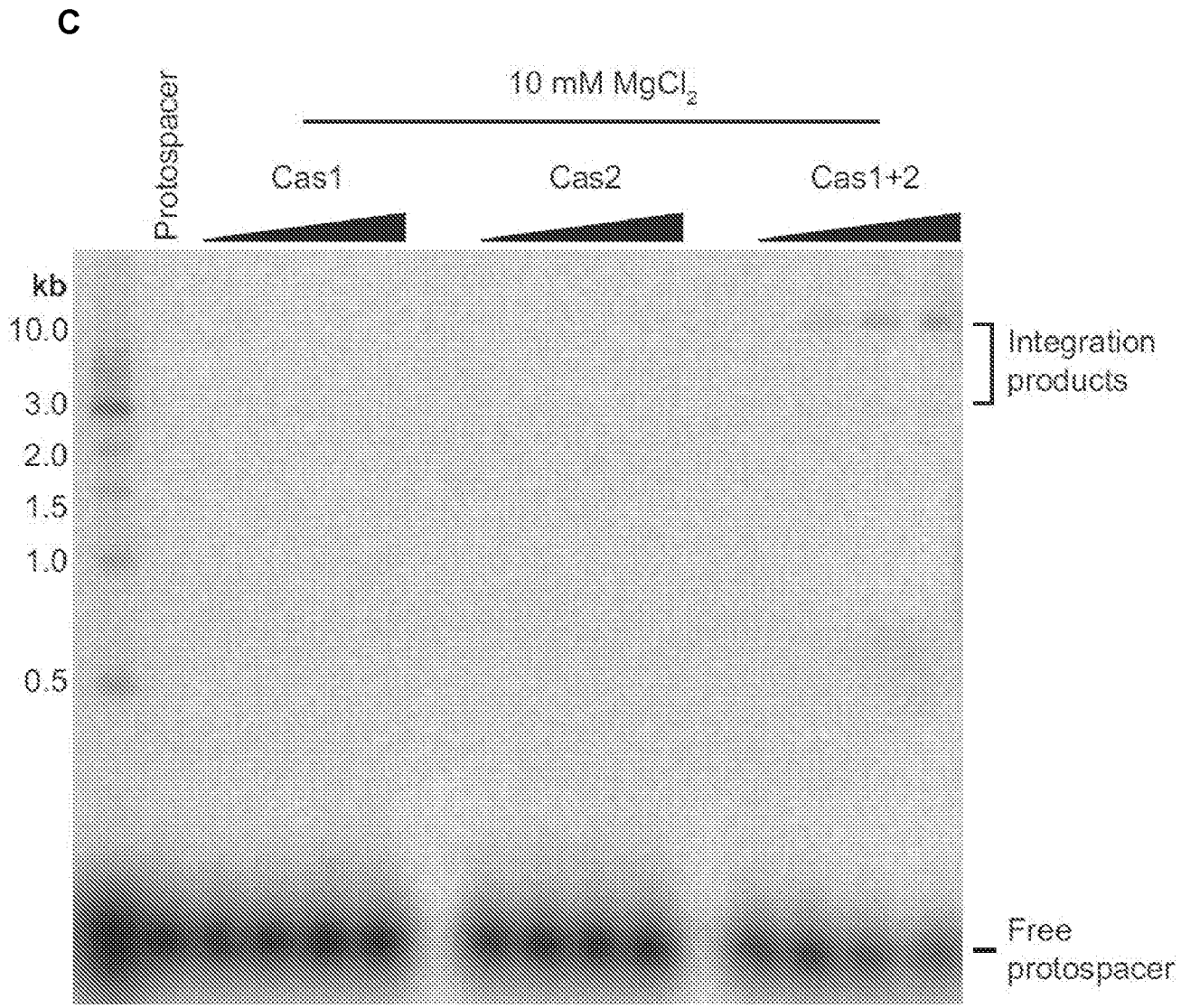
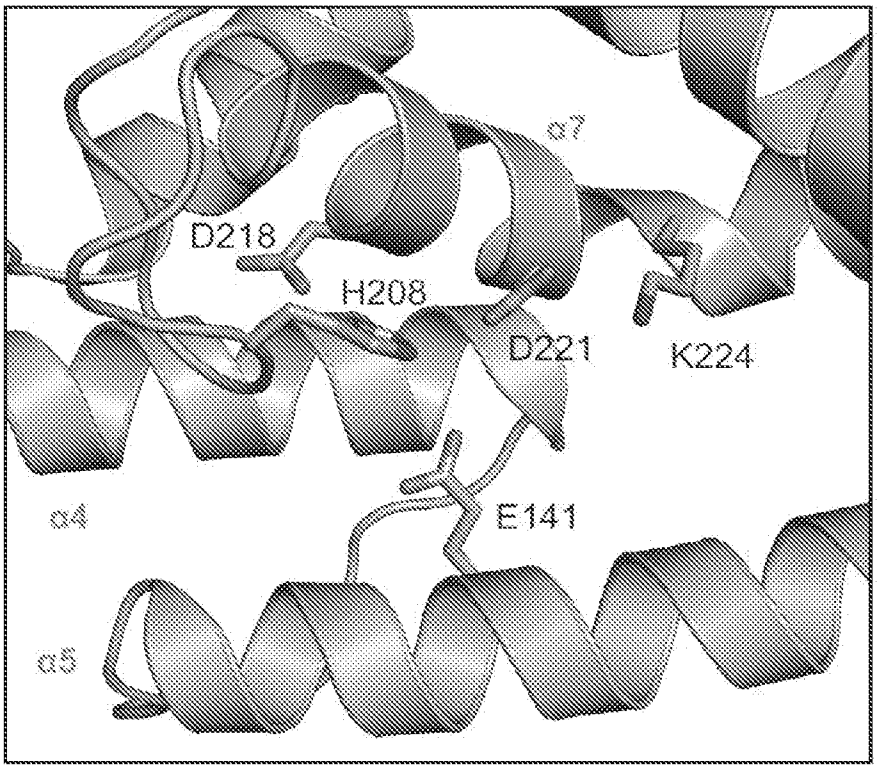

C
10 mM MgCl2
Protospacer
Cas1
Cas2
Cas1+2
kb
10.0
3.0
2.0
1.5
1.0
0.5
Integration products
Free protospacer

Figure 8

A



Cas1 active site

Figure 8

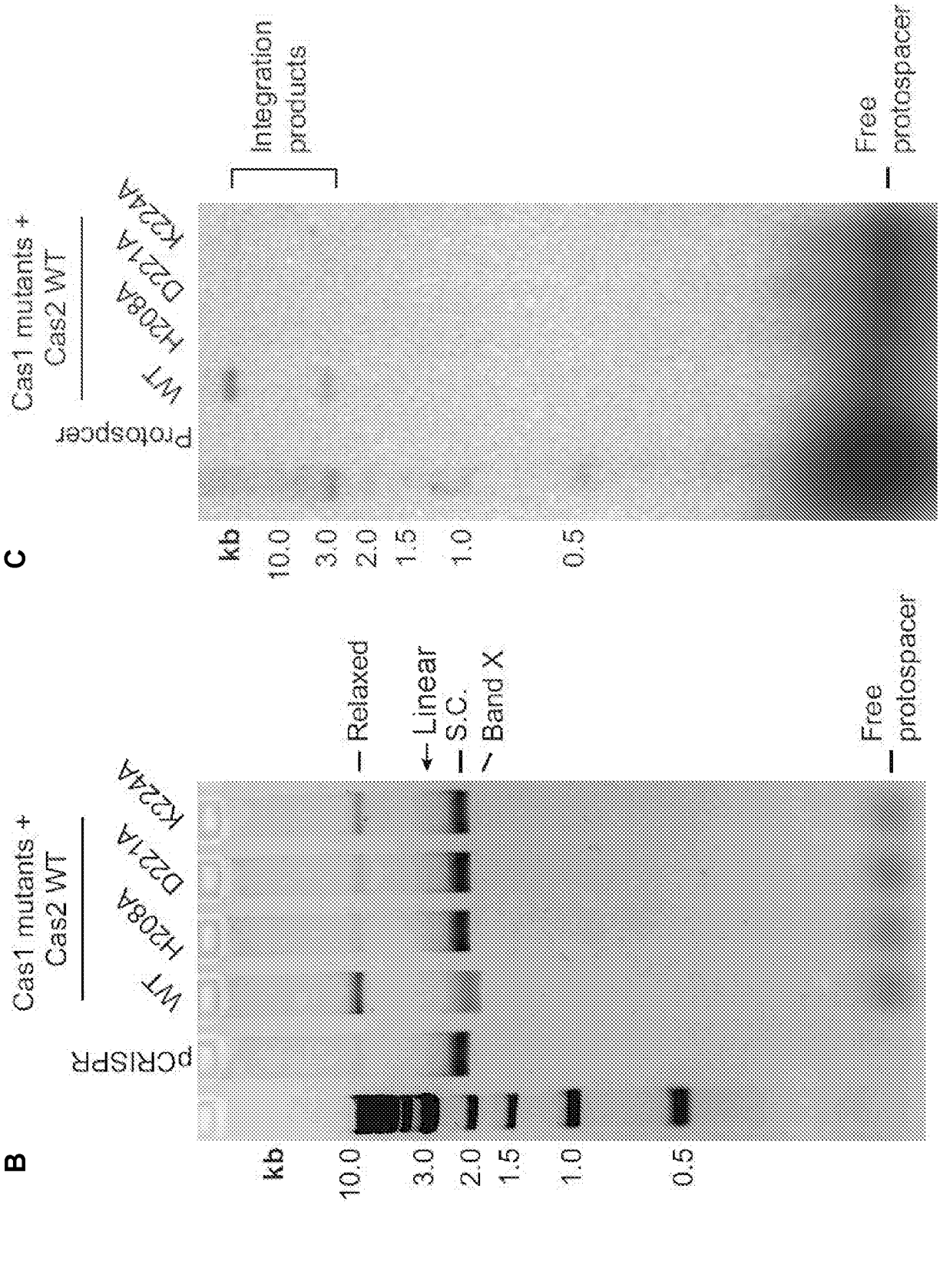

B
pCRISPR
Cas1 mutants +
Cas2 WT
WT
H208A
D221A
K224A
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
S.C.
Band X
Free
protospacer
C
Protospacer
Cas1 mutants +
Cas2 WT
WT
H208A
D221A
K224A
kb
10.0
3.0
2.0
1.5
1.0
0.5
Integration
products
Free
protospacer

Figure 9

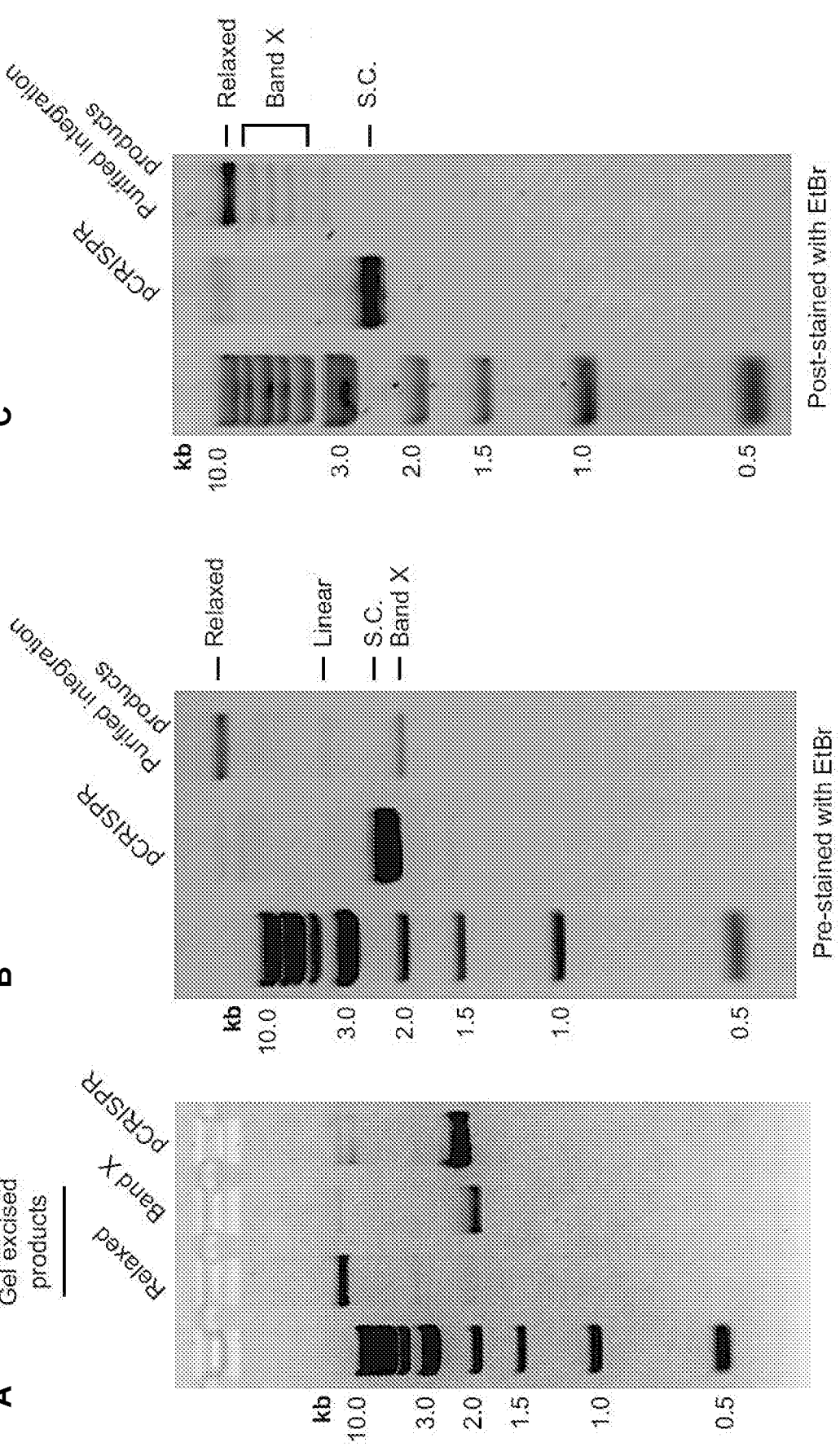

A
Gel excised products
Relaxed
Band X
PCRISPR
kb
10.0
3.0
2.0
1.5
1.0
0.5
B
PCRISPR
Purified integration products
Relaxed
Linear
S.C.
Band X
Pre-stained with EtBr
C
PCRISPR
Purified integration products
Relaxed
Band X
S.C.
Post-stained with EtBr

Figure 9

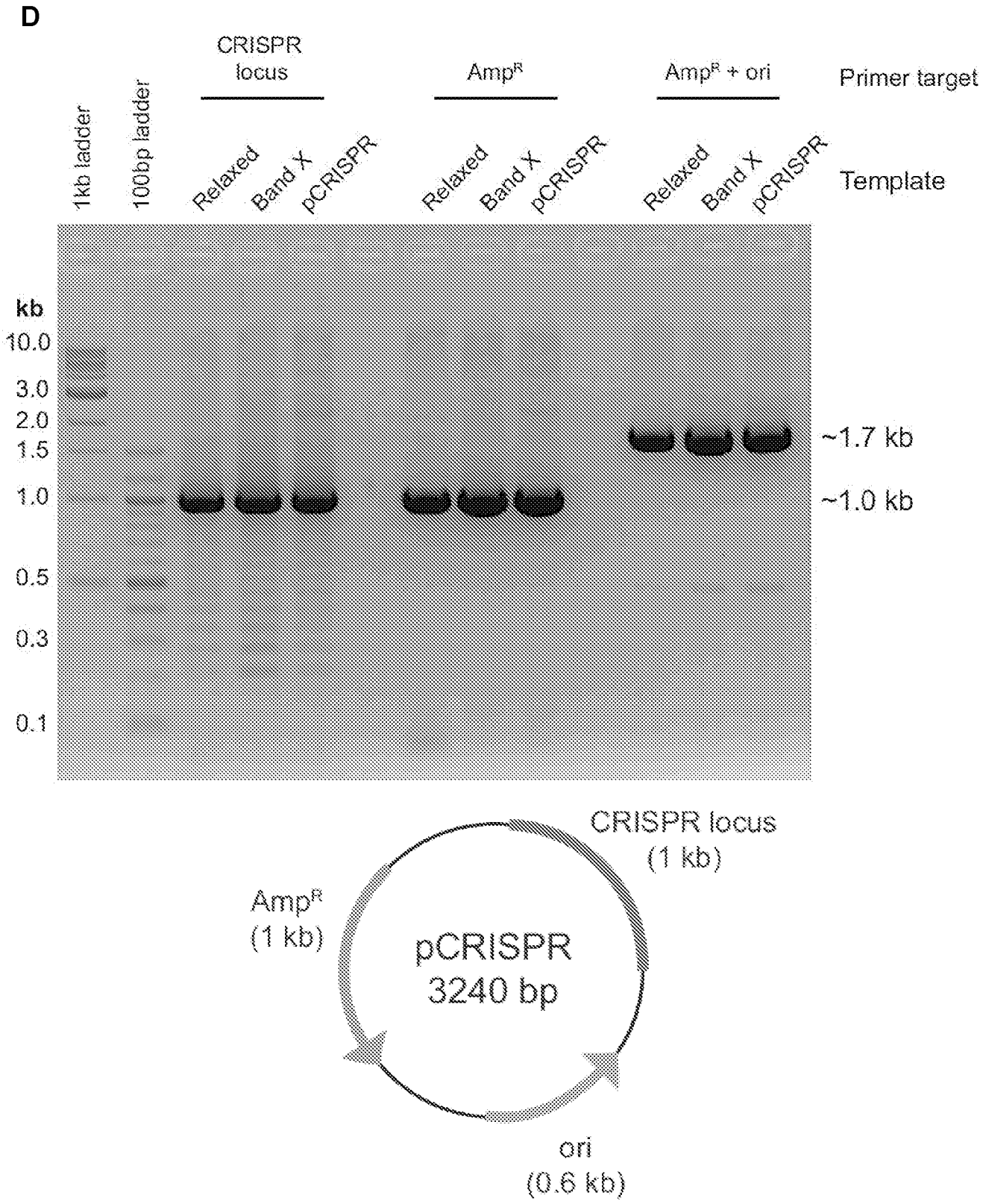

D
CRISPR locus
Amp^R
Amp^R + ori
Primer target
1kb ladder
100bp ladder
Relaxed
Band X
pCRISPR
Relaxed
Band X
pCRISPR
Relaxed
Band X
pCRISPR
Template
kb
10.0
3.0
2.0
1.5
1.0
0.5
0.3
0.1
~1.7 kb
~1.0 kb
CRISPR locus
(1 kb)
Amp^R
(1 kb)
pCRISPR
3240 bp
ori
(0.6 kb)

Figure 10

A

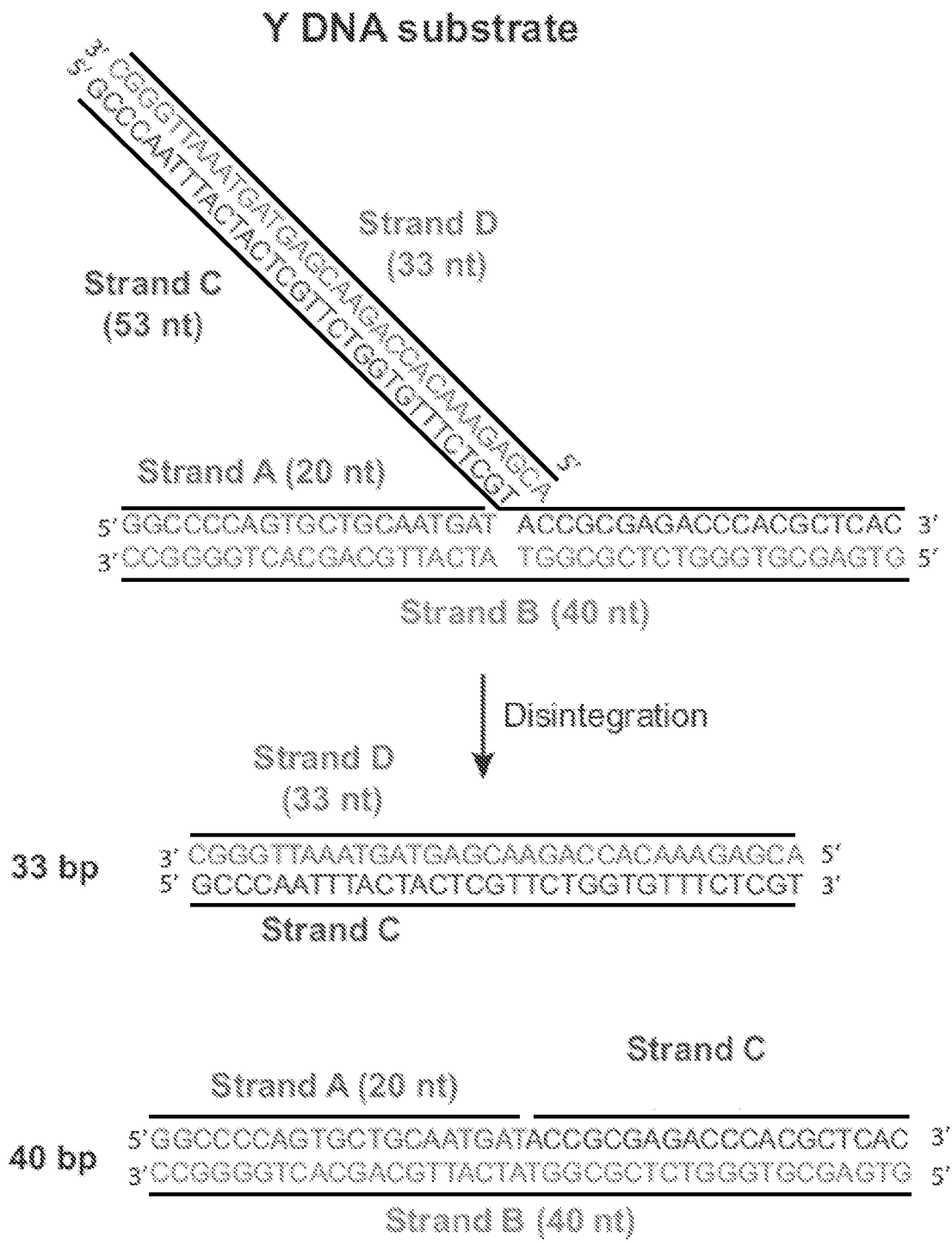

Y DNA substrate
3' CGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
5' GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT
Strand D
(33 nt)
Strand C
(53 nt)
Strand A (20 nt)
5' GGCCCCAGTGCTGCAATGAT ACCGCGAGACCCACGCTCAC 3'
3' CCGGGGTCACGACGTTACTA TGGCGCTCTGGGTGCGAGTG 5'
Strand B (40 nt)
Disintegration
Strand D
(33 nt)
33 bp
3' CGGGTTAAATGATGAGCAAGACCACAAAGAGCA 5'
5' GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT 3'
Strand C
Strand C
Strand A (20 nt)
40 bp
5' GGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCAC 3'
3' CCGGGGTCACGACGTTACTATGGCGCTCTGGGTGCGAGTG 5'
Strand B (40 nt)

Figure 10

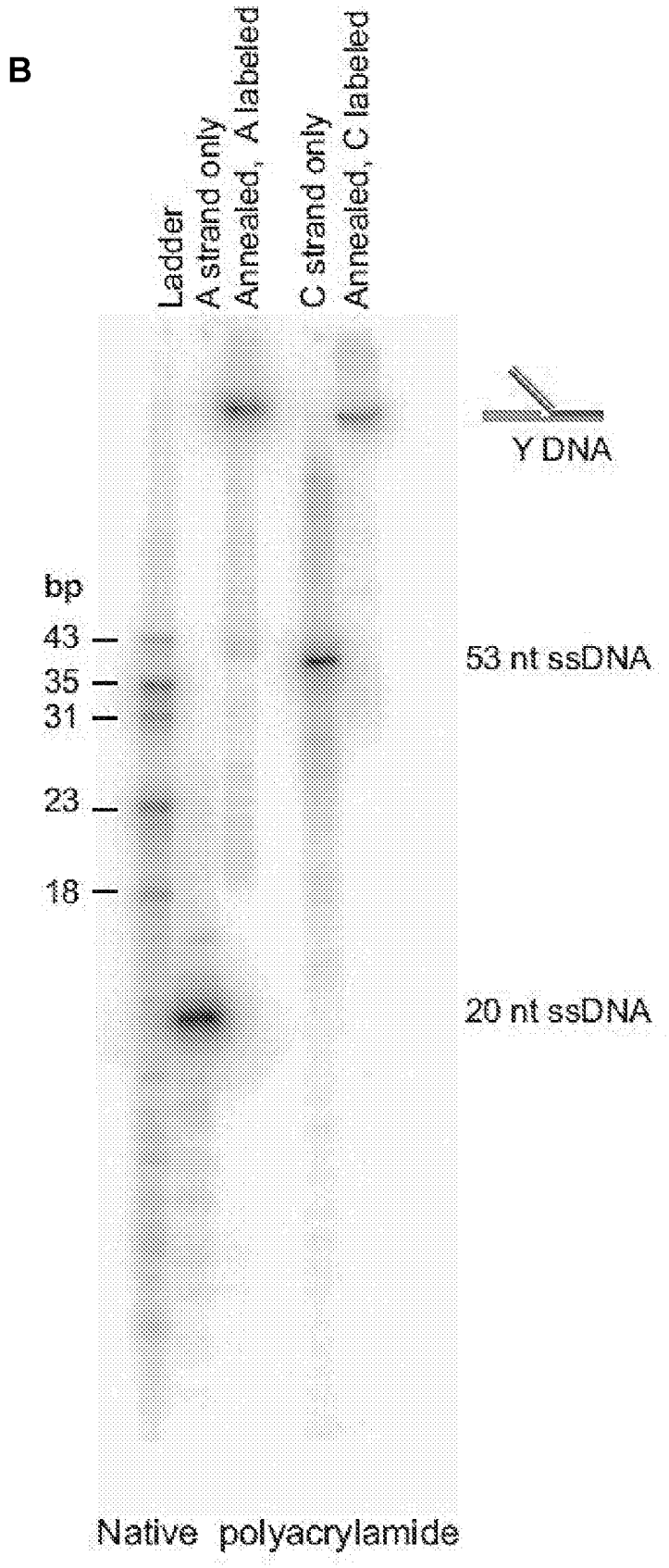

B
Ladder
A strand only
Annealed, A labeled
C strand only
Annealed, C labeled
Y DNA
bp
43
35
31
23
18
53 nt ssDNA
20 nt ssDNA
Native polyacrylamide

Figure 10

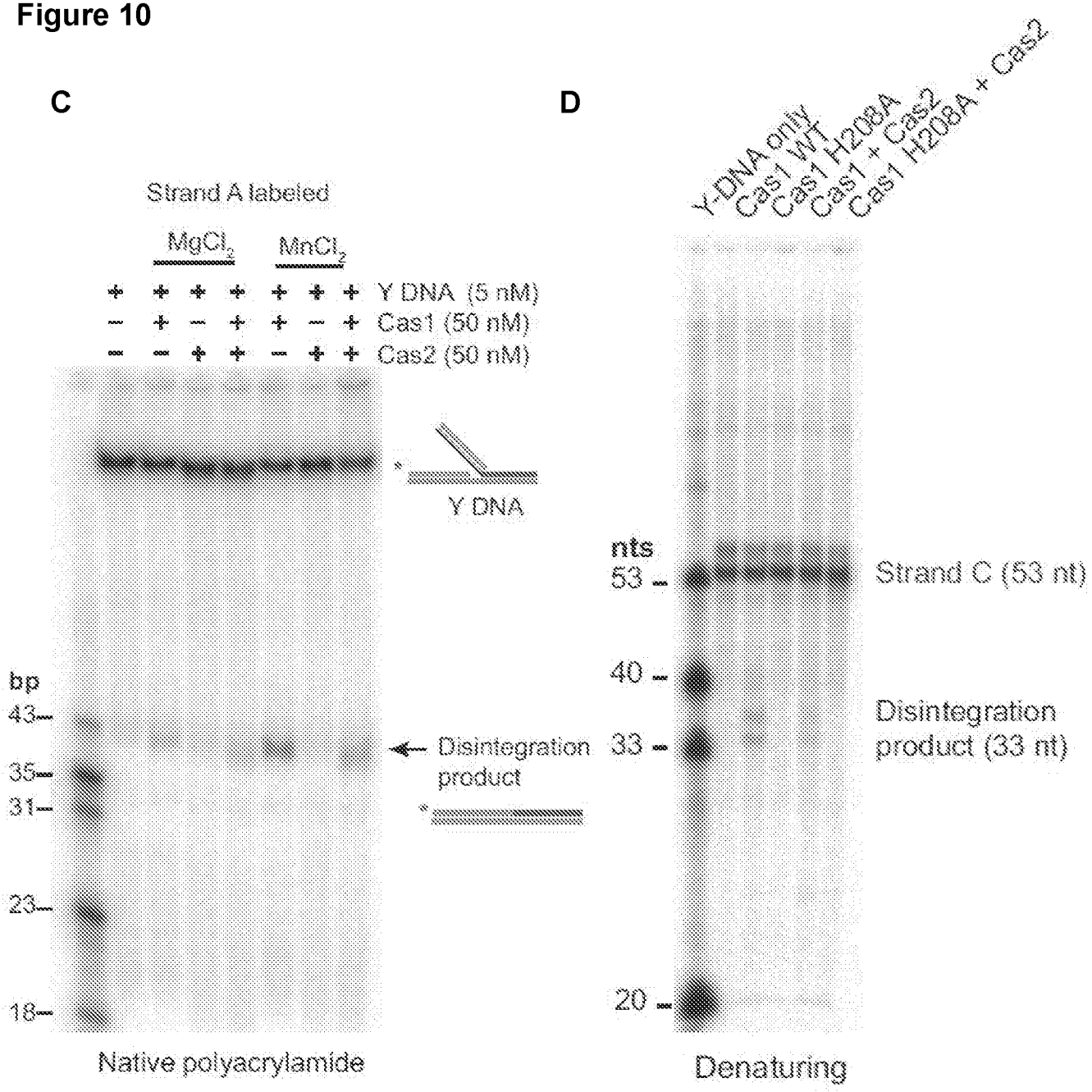

C
Strand A labeled
MgCl2
MnCl2
+ + + + + + + Y DNA (5 nM)
– + – + + – + Cas1 (50 nM)
– – + + – + + Cas2 (50 nM)
Y DNA
bp
43
35
31
23
18
Disintegration product
Native polyacrylamide
D
Y-DNA only
Cas1 WT
Cas1 H208A
Cas1 + Cas2
Cas1 H208A + Cas2
nts
53
40
33
20
Strand C (53 nt)
Disintegration product (33 nt)
Denaturing

Figure 11

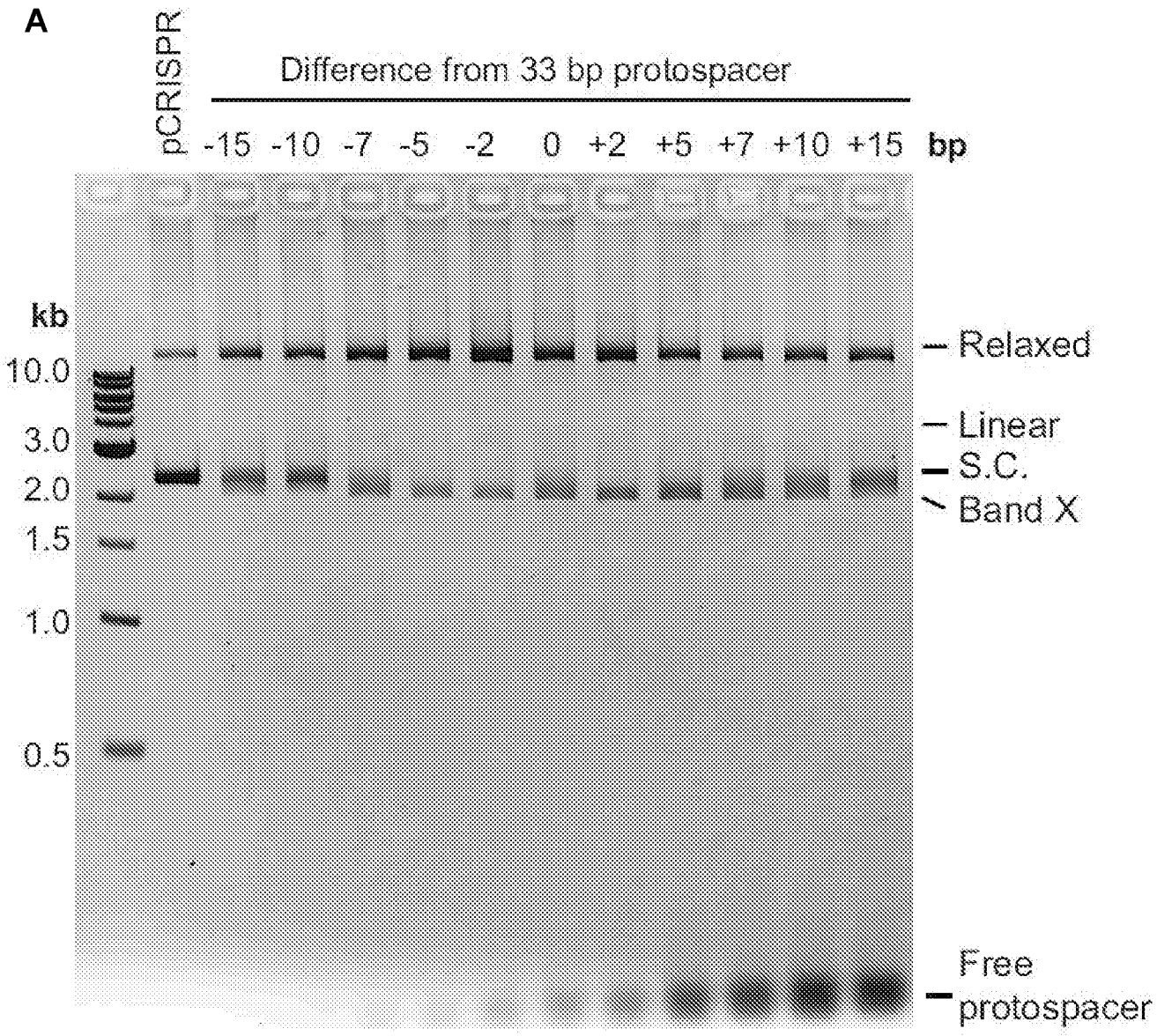

A
pCRISPR
Difference from 33 bp protospacer
-15 -10 -7 -5 -2 0 +2 +5 +7 +10 +15 bp
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
S.C.
Band X
Free protospacer

Figure 11

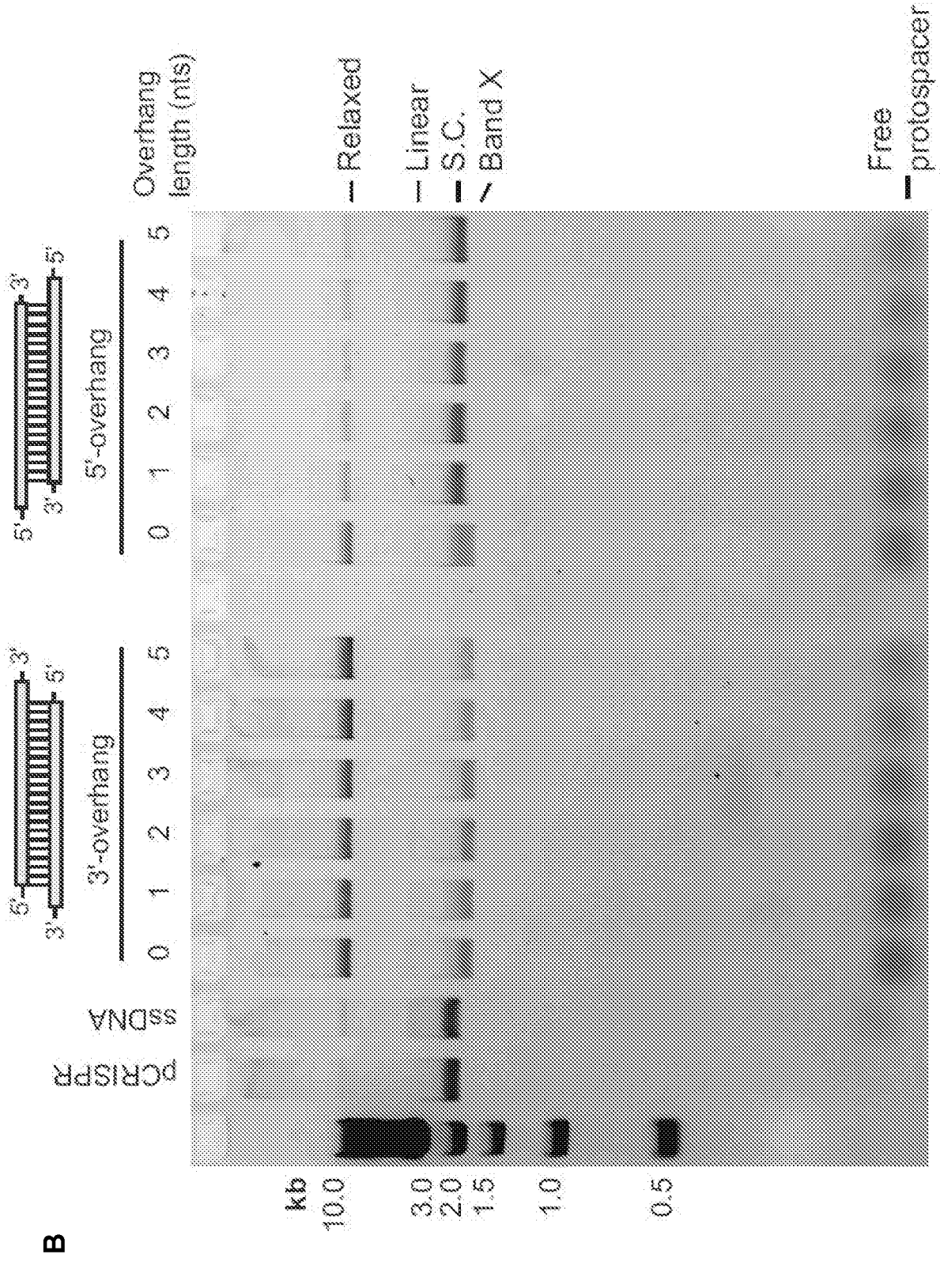

B
pCRISPR
ssDNA
3'-overhang
0 1 2 3 4 5
5'-overhang
0 1 2 3 4 5
Overhang length (nts)
Relaxed
Linear
S.C.
Band X
Free protospacer
kb
10.0
3.0
2.0
1.5
1.0
0.5

Figure 11

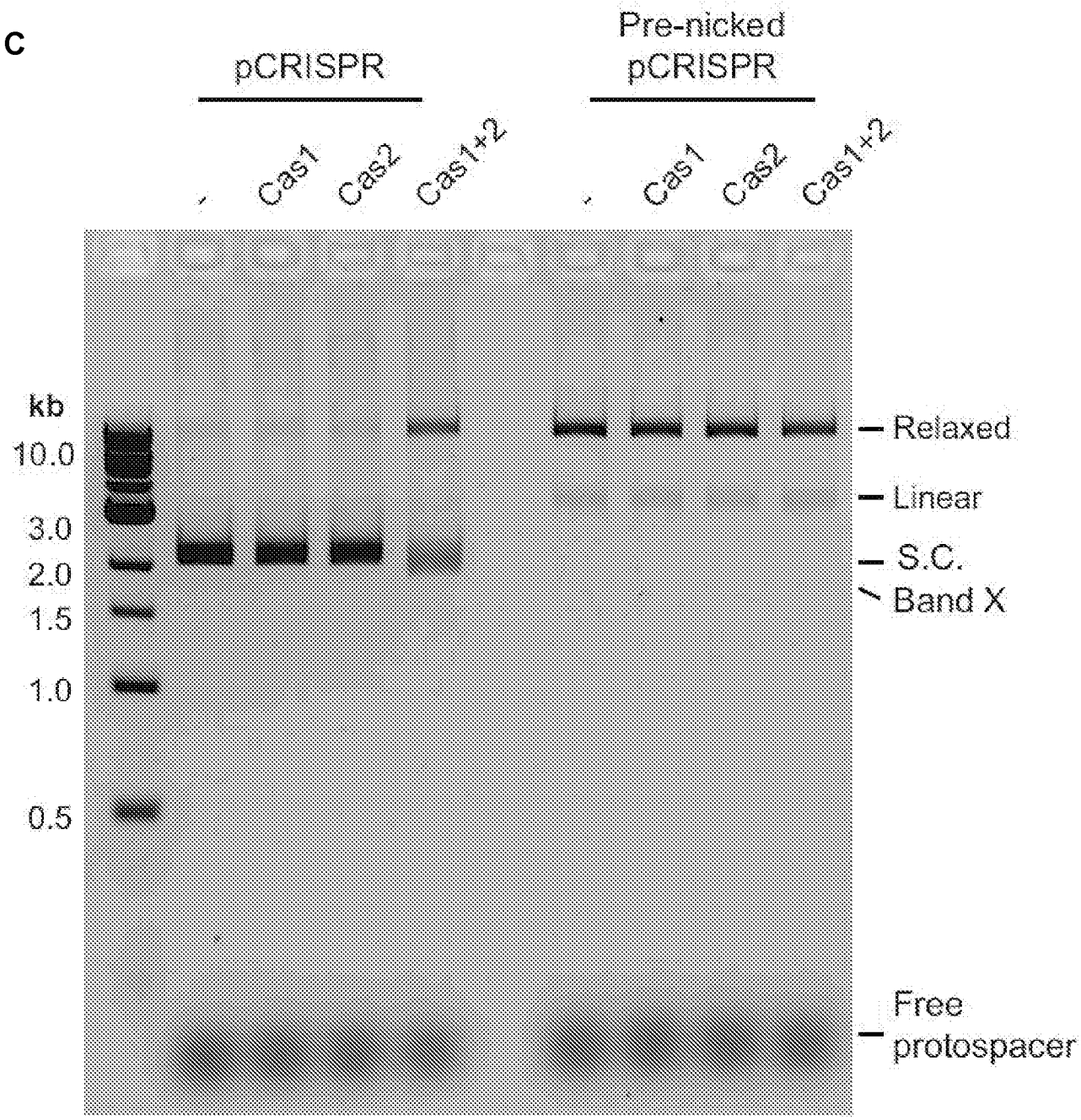

C
pCRISPR
Pre-nicked pCRISPR
-
Cas1
Cas2
Cas1+2
-
Cas1
Cas2
Cas1+2
kb
10.0
3.0
2.0
1.5
1.0
0.5
Relaxed
Linear
S.C.
Band X
Free protospacer

Figure 11

D

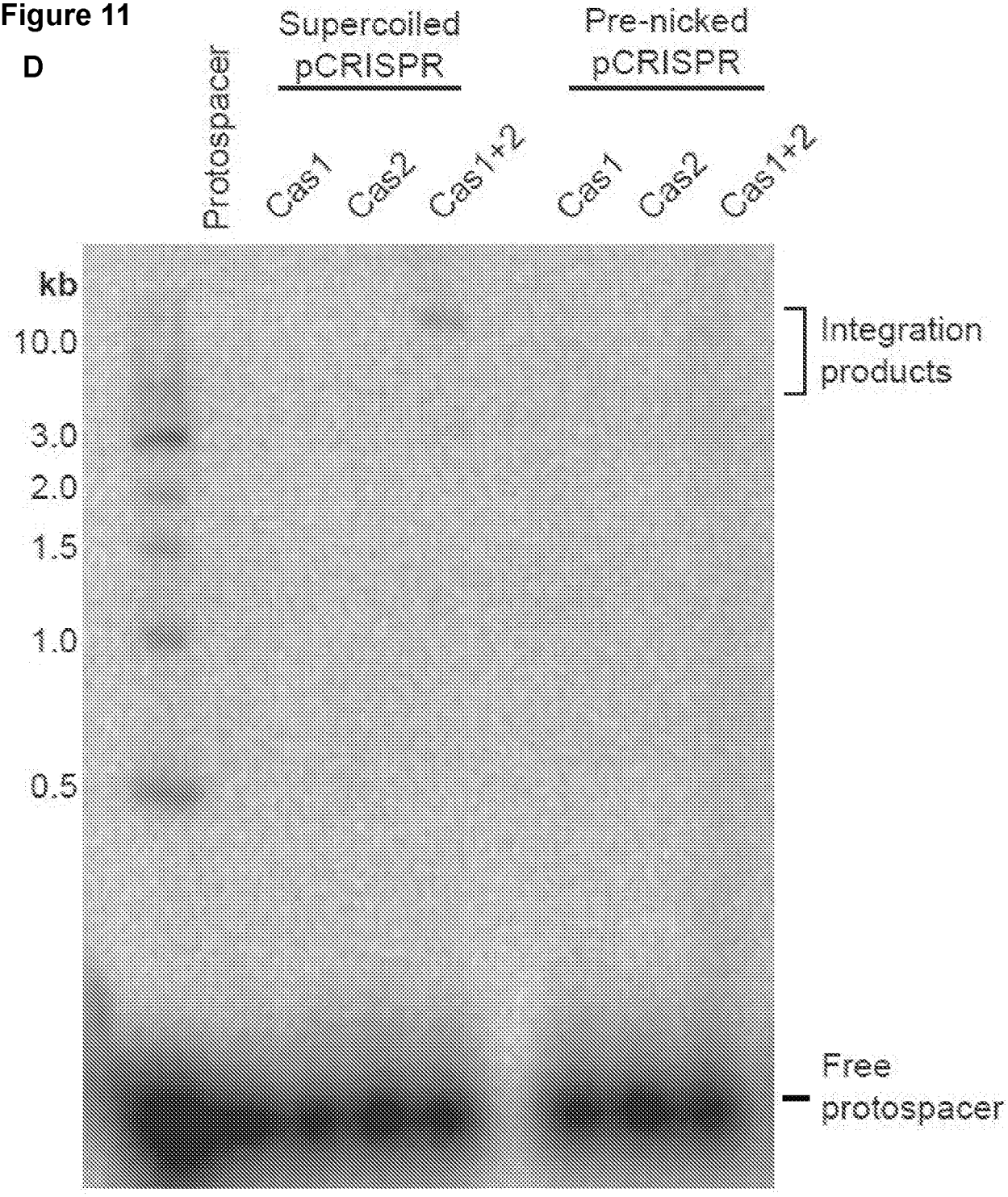

Supercoiled pCRISPR
Pre-nicked pCRISPR
Protospacer
Cas1
Cas2
Cas1+2
Cas1
Cas2
Cas1+2
kb
10.0
3.0
2.0
1.5
1.0
0.5
Integration products
Free protospacer

Figure 11

E

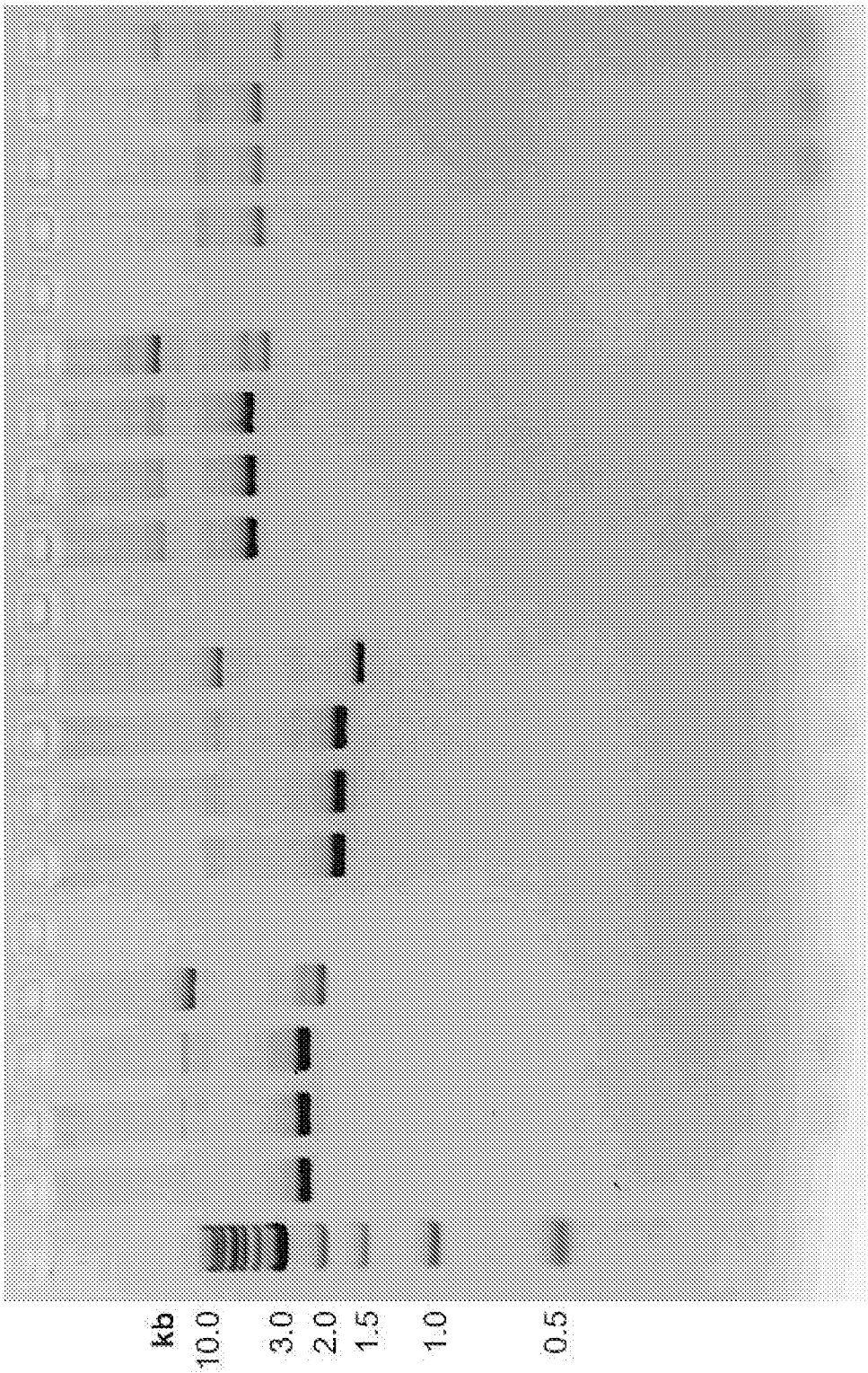

pUC19 CRISPR
pUC19 empty vector
pACYC Duet CRISPR
pACYC Duet empty vector
plasmid
Cas1
Cas2
Cas1+2
kb
10.0
3.0
2.0
1.5
1.0
0.5

Figure 12

A

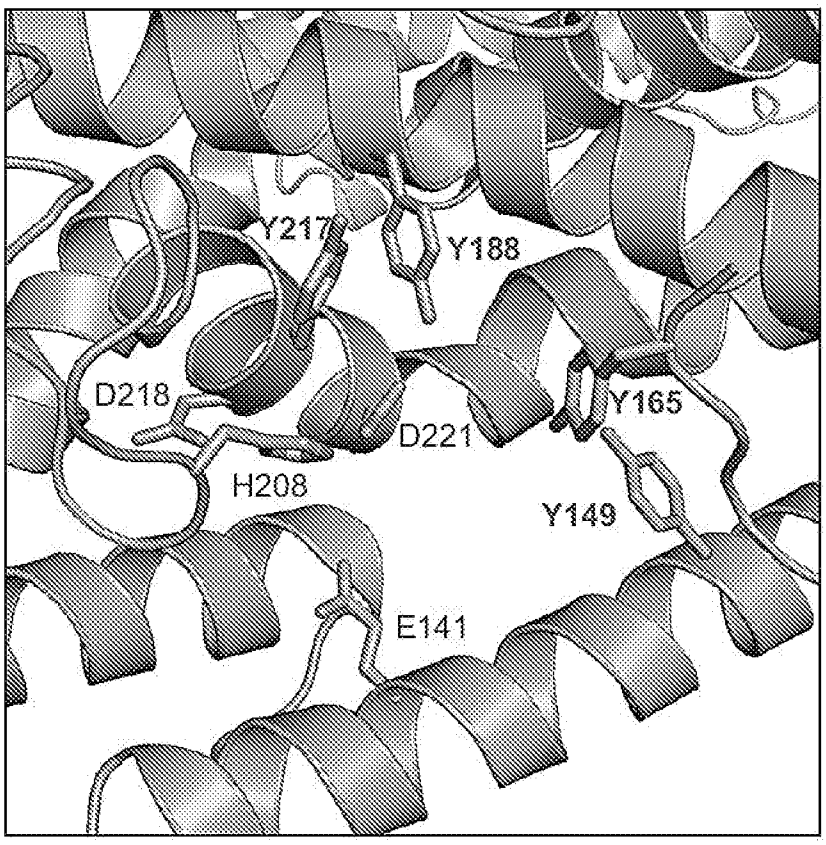


Figure 12

B

α5
α6
α7
Y163
Y168
Y168
Y217
E. coli
T. maritima
P. horikoshii
P. aeolicus
P. aeruginosa
180
186
186
178
226
240
261
261
253
287

Figure 12

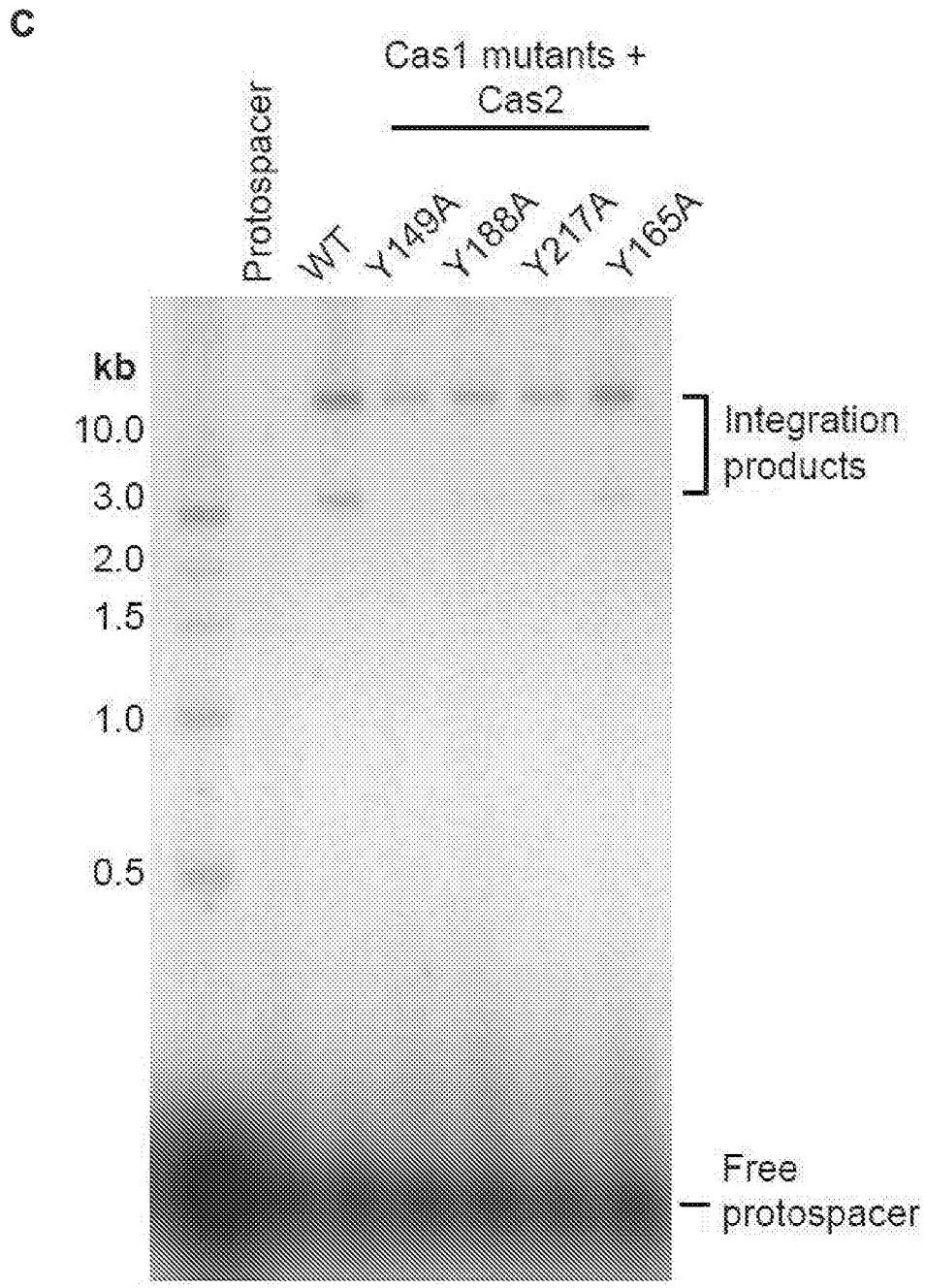

C
Cas1 mutants + Cas2
Protospacer
WT
Y149A
Y188A
Y217A
Y165A
kb
10.0
3.0
2.0
1.5
1.0
0.5
Integration products
Free protospacer

Figure 13

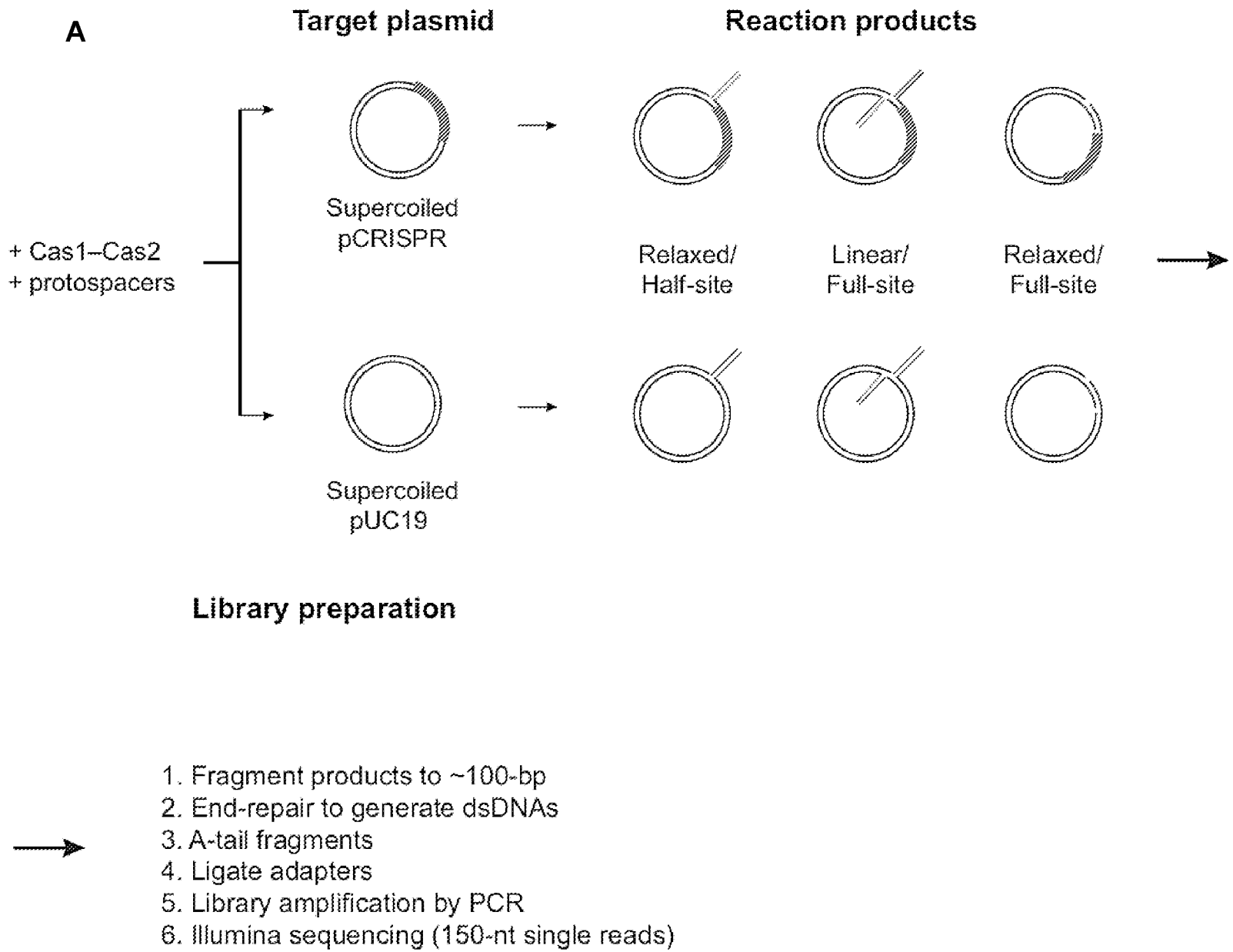

A
Target plasmid
Reaction products
+ Cas1–Cas2
+ protospacers
Supercoiled
pCRISPR
Relaxed/
Half-site
Linear/
Full-site
Relaxed/
Full-site
Supercoiled
pUC19
Library preparation
1. Fragment products to ~100-bp
2. End-repair to generate dsDNAs
3. A-tail fragments
4. Ligate adapters
5. Library amplification by PCR
6. Illumina sequencing (150-nt single reads)

Figure 13

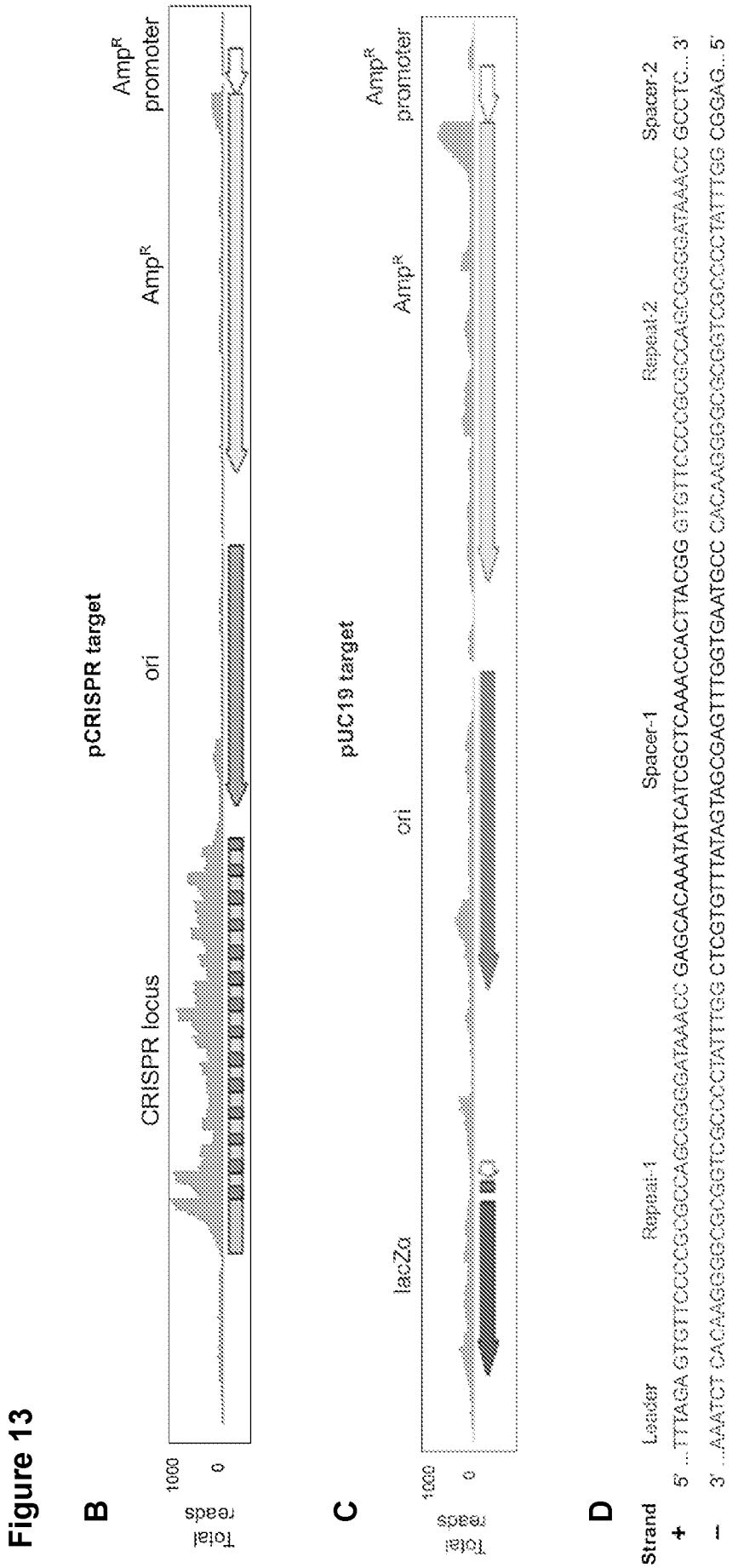
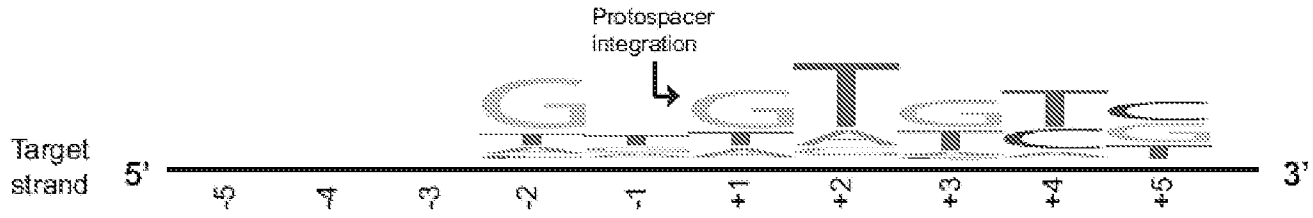

B
pCRISPR target
CRISPR locus
ori
Amp^R
Amp^R promoter
Total reads
1000
0
C
pUC19 target
lacZα
ori
Amp^R
Amp^R promoter
Total reads
1000
0
D
Strand
Leader
Repeat-1
Spacer-1
Repeat-2
Spacer-2
+ 5' ...TTTAGA GTGTTCCCCGCGCCAGCGGGGATAAACC GAGCACAAATATCATCGCTCAAACCACTTACGG GTGTTCCCCGCGCCAGCGGGGATAAACC GCCTC... 3'
– 3' ...AAATCT CACAAGGGGCGCGGTCGCGCCCTATTTGG CTCGTGTTTATAGTAGCGAGTTTGGTGAATGCC CACAAGGGGCGCGGTCGCGCCCTATTTGG CGGAG... 5'

Figure 13

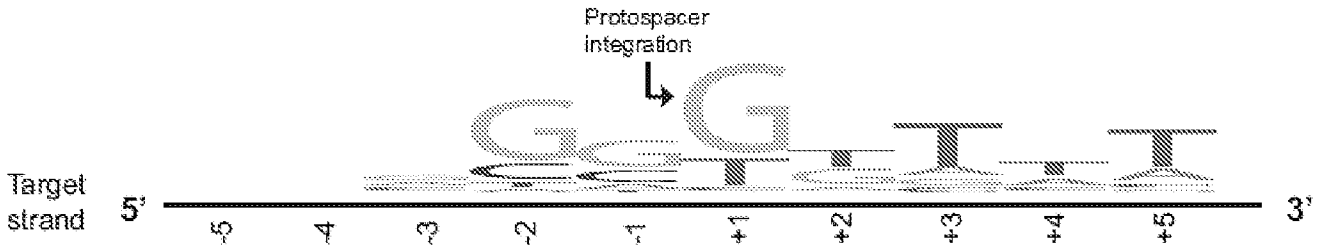
E

pCRISPR plus strand integration site



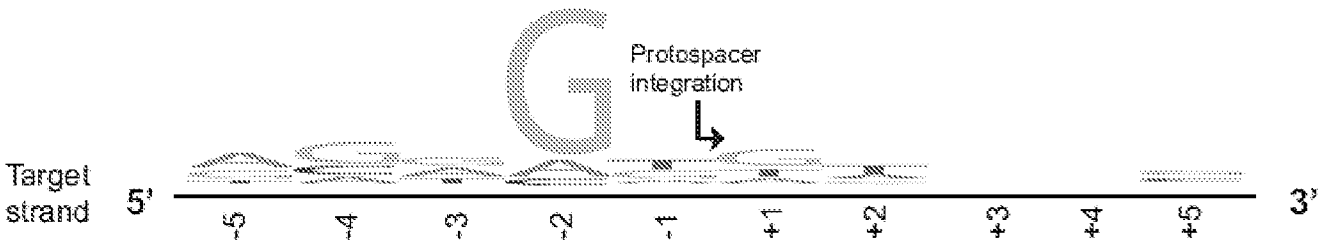
F

pCRISPR minus strand integration site



G

pUC19 plus strand integration site



H

pUC19 minus strand integration site

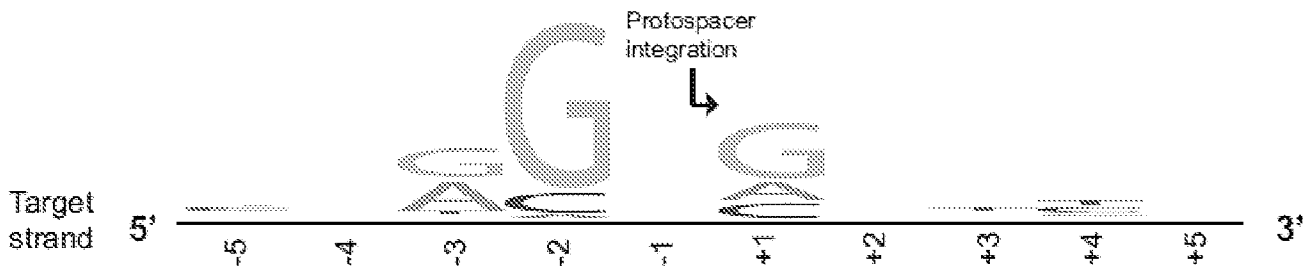
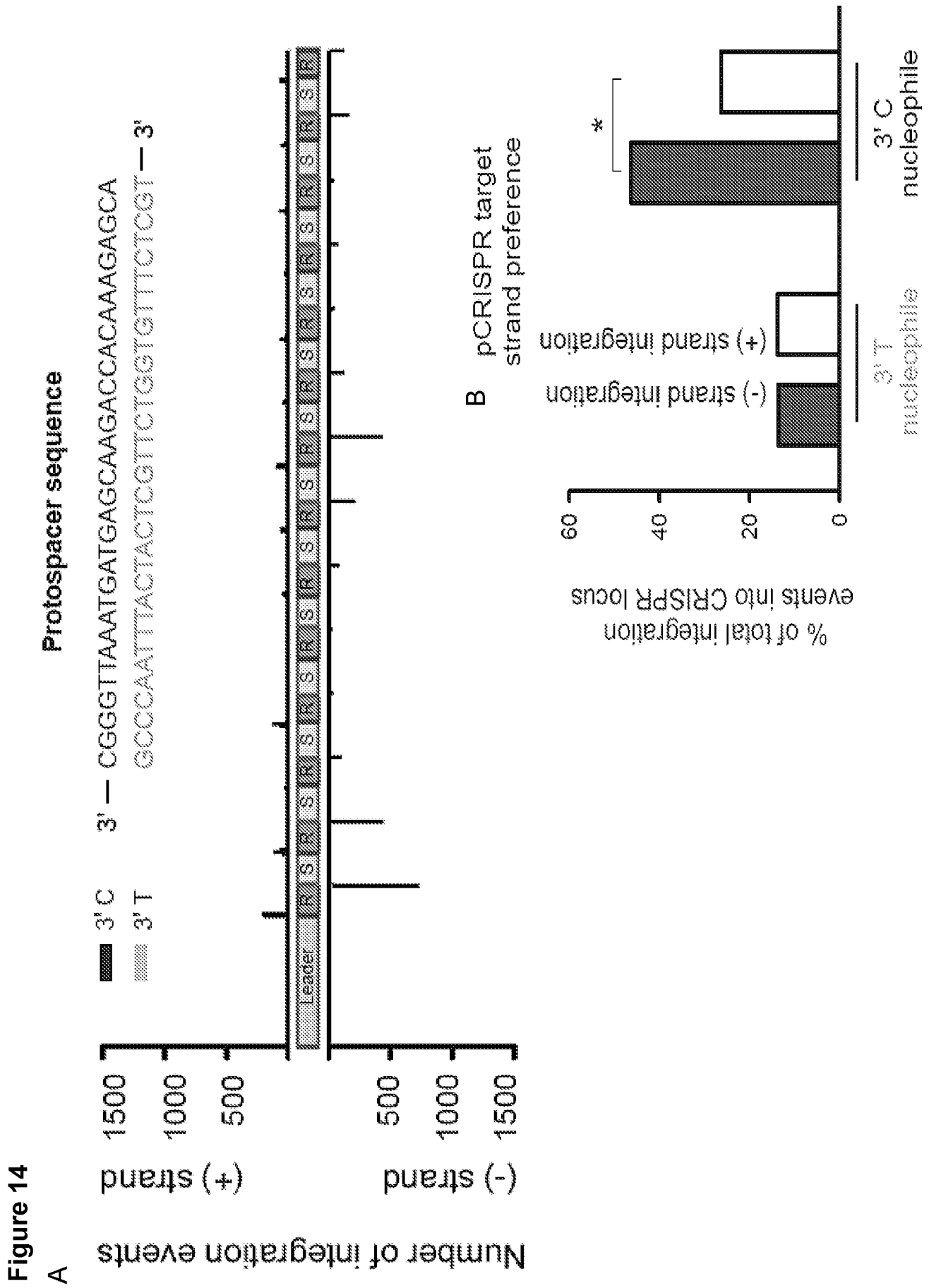


Figure 14


A
Number of integration events
(+) strand
(-) strand
1500
1000
500
Protospacer sequence
3' C
3' T
3' — CGGGTTAAATGATGAGCAAGACCACACAAAGAGCA
GCCCAATTTACTACTCGTTCTGGTGTTTCTCGT — 3'
Leader
R S R S R S R S R S R S R S R S R S R S R S R S R S R
B pCRISPR target strand preference
% of total integration events into CRISPR locus
60
40
20
0
(-) strand integration
(+) strand integration
*
3' T nucleophile
3' C nucleophile

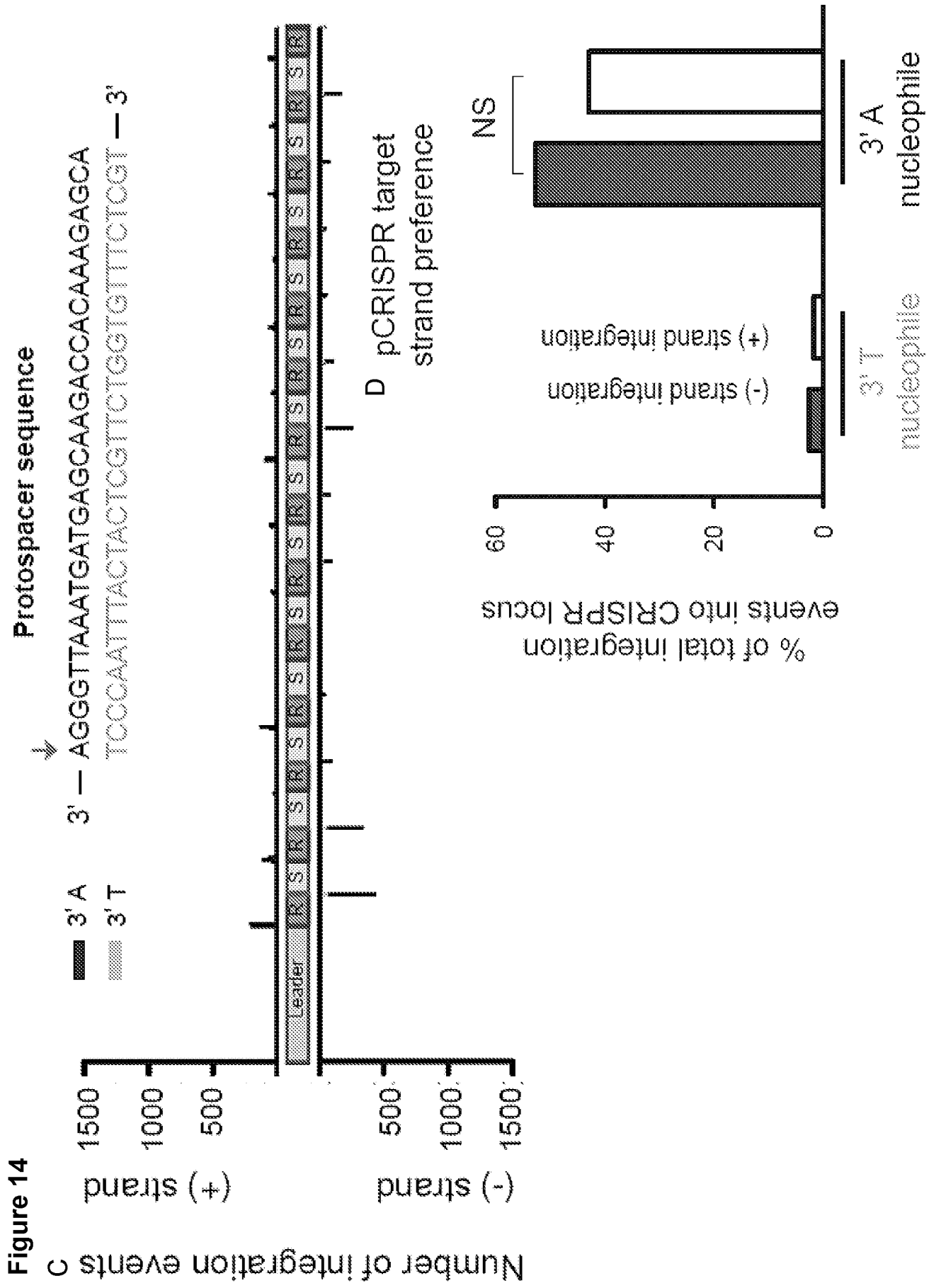

Protospacer sequence
3' A
3' T
3' — AGGGTTAAATGATGAGGCAAGACCACACAAAGAGCA
TCCCAATTTACTACTCCGTTCTGGTGTGTTTCTCGT — 3'
Leader
R S R S R S R S R S R S R S R S R S R S R S R
C Number of integration events
(+) strand
(-) strand
1500
1000
500
D pCRISPR target strand preference
% of total integration events into CRISPR locus
(-) strand integration
(+) strand integration
NS
3' T nucleophile
3' A nucleophile
60
40
20
0

Figure 14

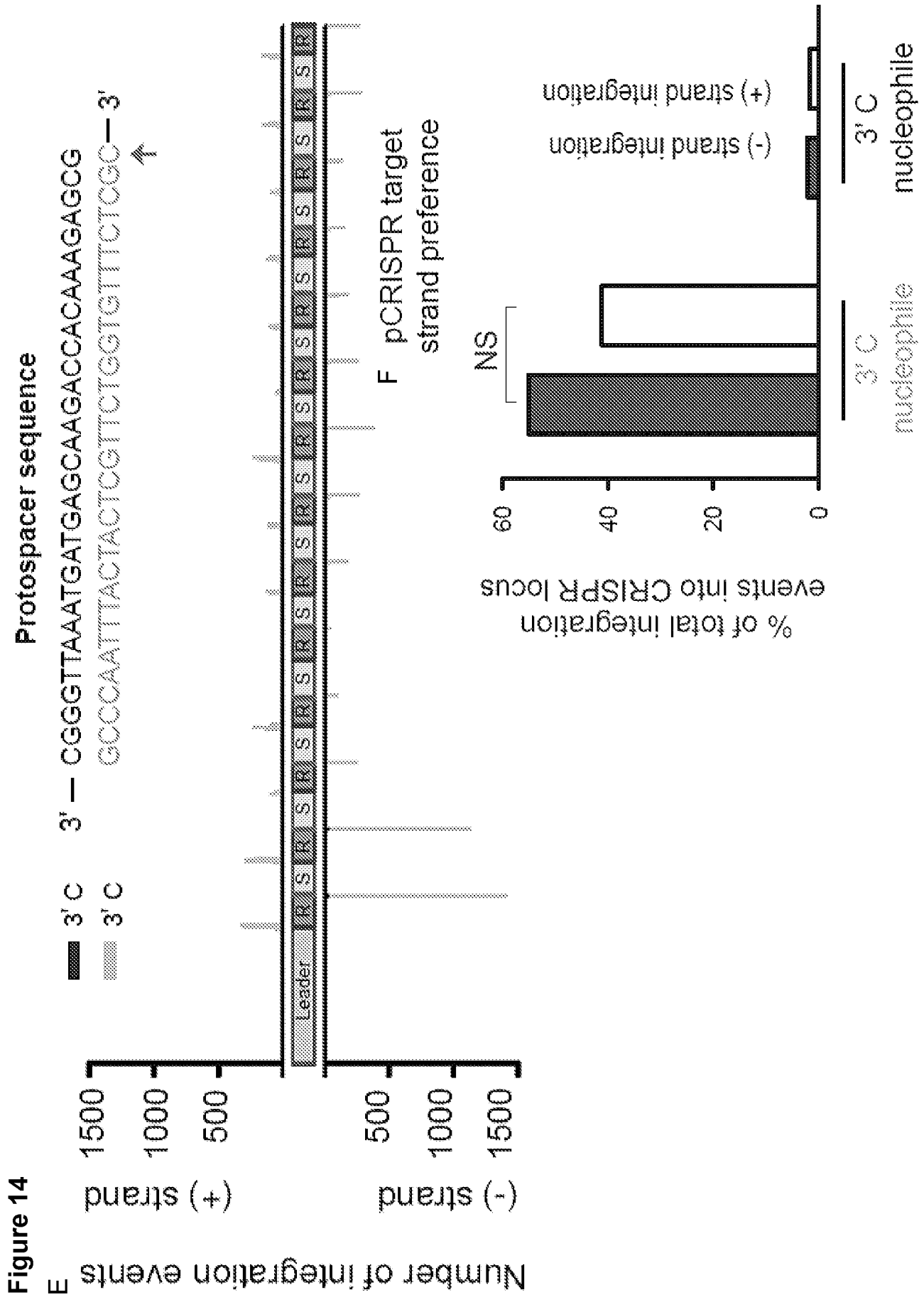

Protospacer sequence
3' C
3' — CGGGTTAAATGATGAGGCAAGACCACACAAAGAGCG
3' C
GCCCAATTTACTACTCCGTTCTGGTGTTTCTCGC — 3'
Leader
R S R S R S R S R S R S R S R S R S R S R S R S R
Number of integration events
(+) strand
(-) strand
1500
1000
500
500
1000
1500
E
F pCRISPR target strand preference
% of total integration events into CRISPR locus
60
40
20
0
NS
3' C nucleophile
3' C nucleophile
(-) strand integration
(+) strand integration

Figure 14

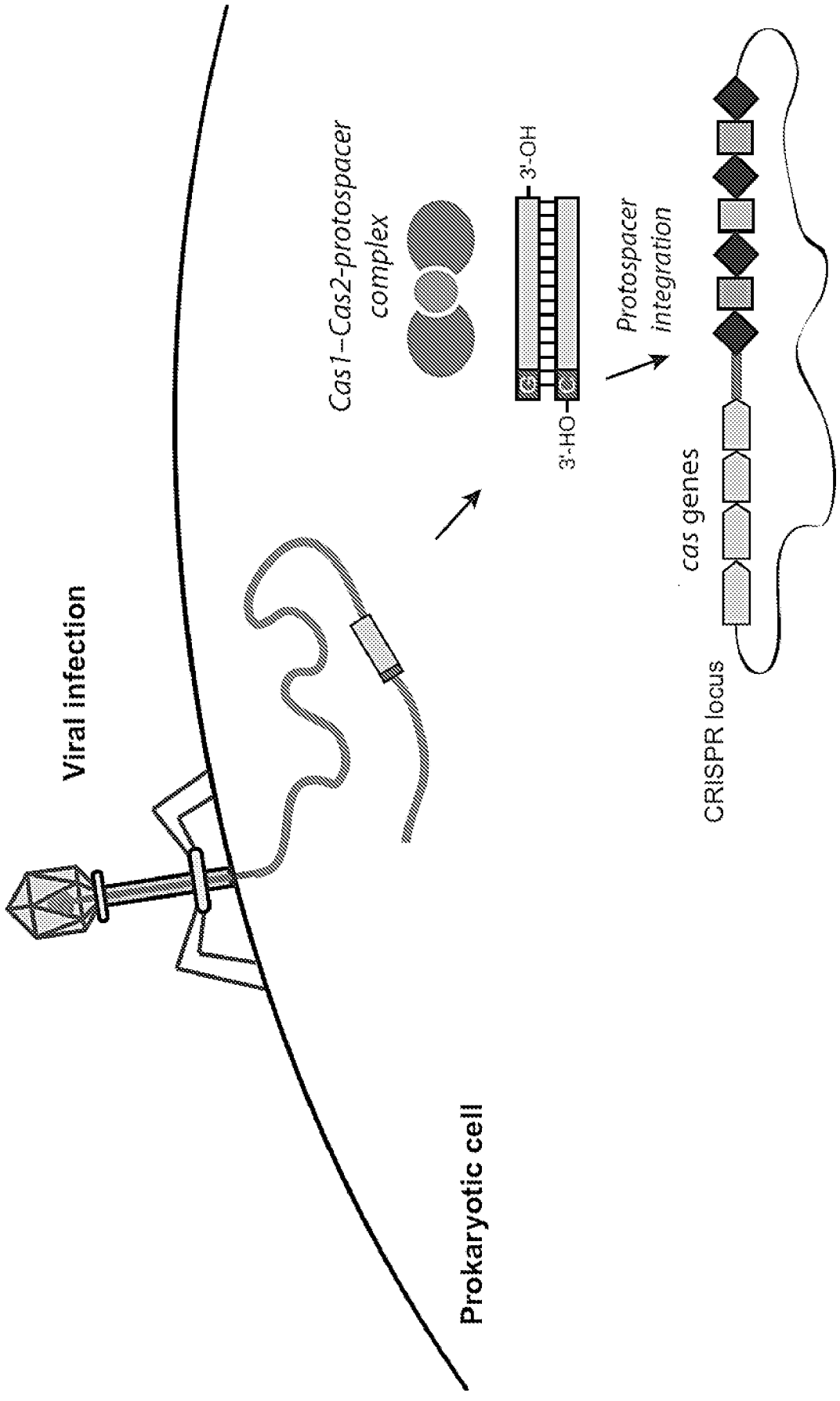

Viral infection
Prokaryotic cell
Cas1–Cas2-protospacer complex
3'-OH
3'-HO
Protospacer integration
cas genes
CRISPR locus

Figure 15

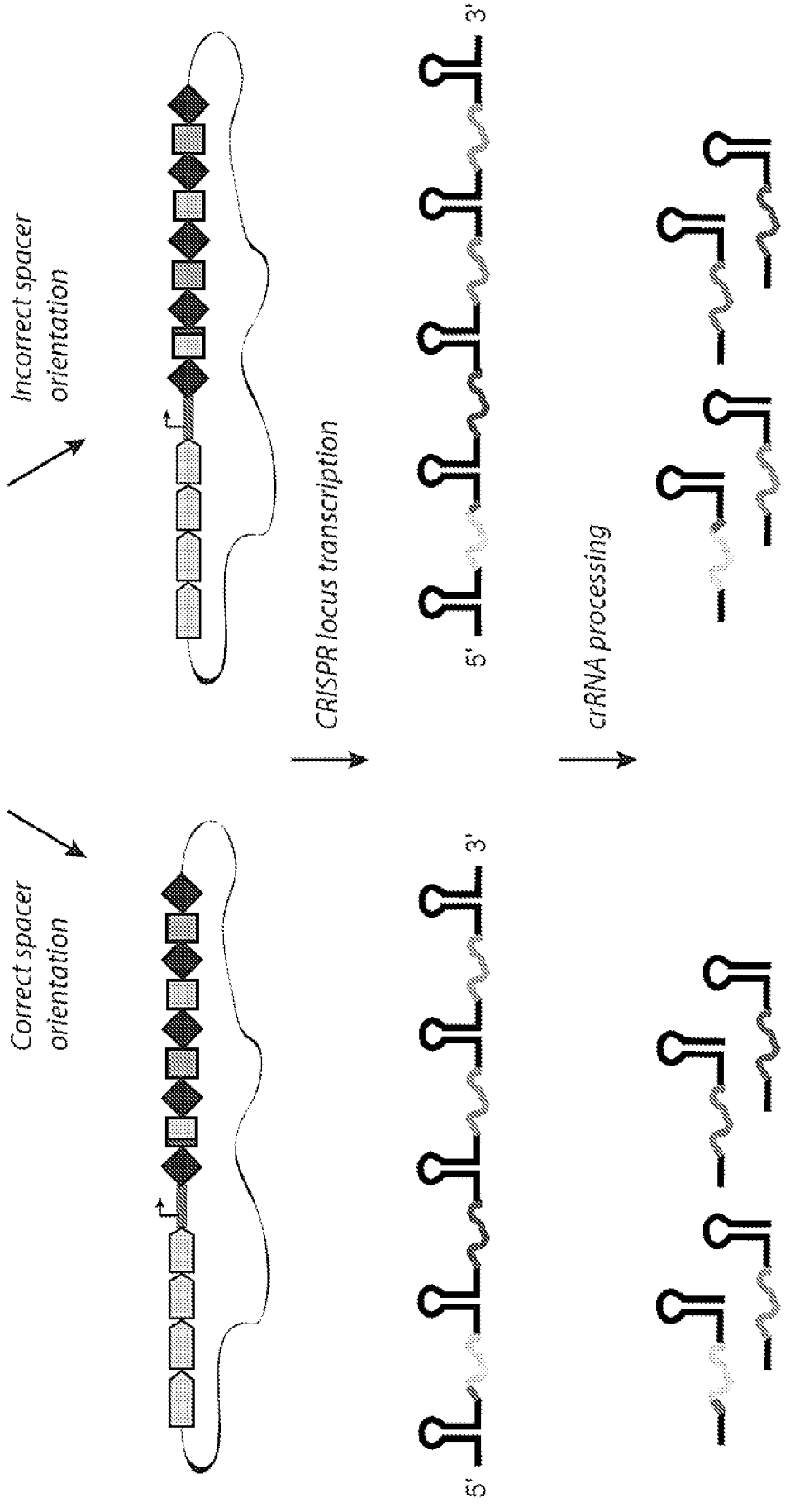


Figure 15 (Cont. 1)

Figure 15 (Cont. 2)

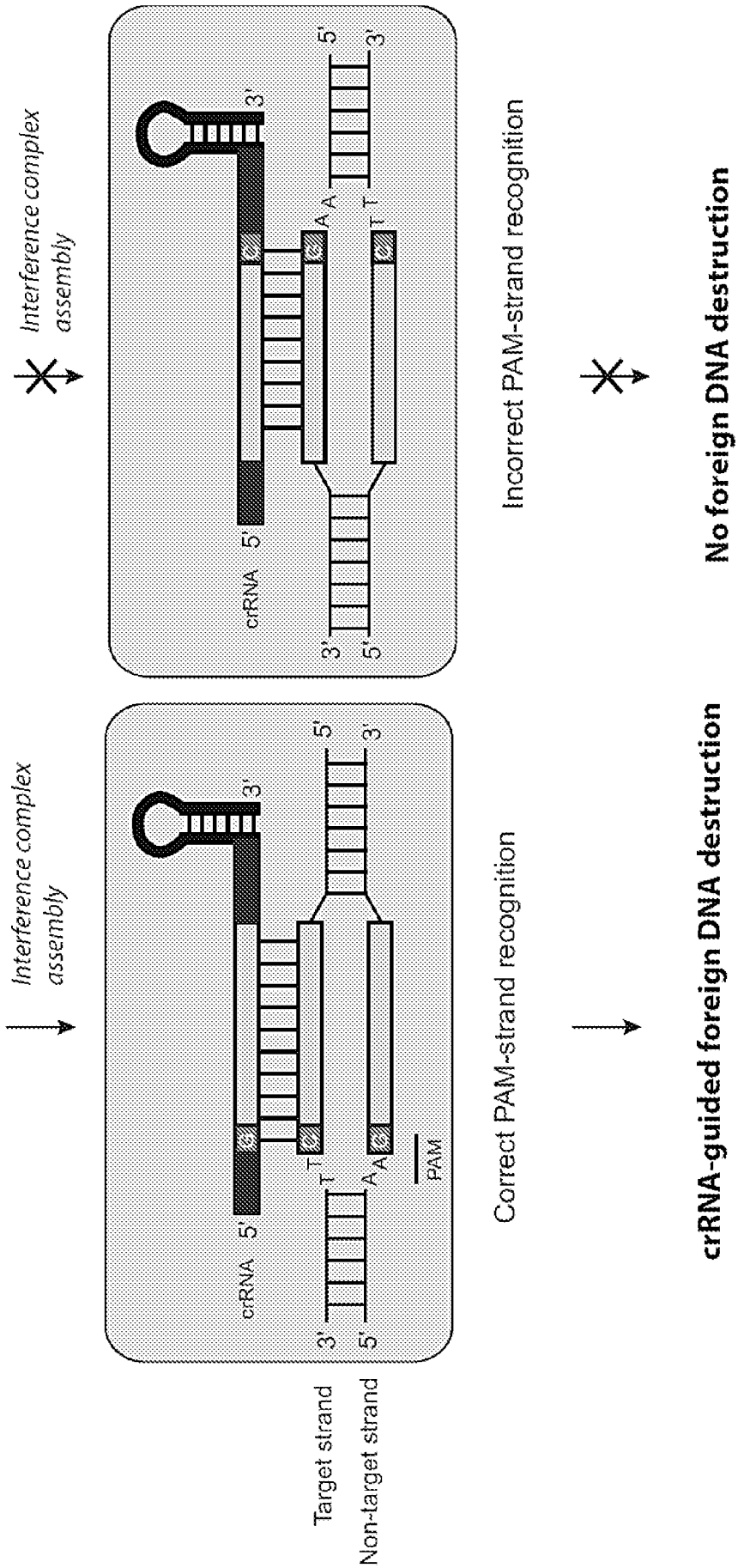


Figure 16

Type I-A

Thermoproteus tenax (Cas1–Cas2 fusion)

>gi|503893142|ref|WP_014127136.1| CRISPR-associated protein Cas1 and Cas2 [Thermoproteus tenax]
MDEVLLLTGGISITTRALRALLATGATVAVFSPRGEPLGIFMRPVGDATGAKRRCQYKAAEDGRGLQYAK
SWVFKKILGQRDNIKAWRRRLRGYSQYAESLAKALPGAGLHGAMETPRRRRRGQDGGQAGVRGRPTHPPV
PPGAGRRSPGGAPRGQEASLRRDPQRGQSSGALHMYVIVVYDITENDVRAKVADILRAYGLARIQRSAYV
GRLPPALVKELAERLARAVRGANADIAIFKVDKRTIDTSLRIPPRPPAGHVA [SEQ ID NO: 162]

Pyrobaculum neutrophilum Cas1

>gi|501318854|ref|WP_012350489.1| CRISPR-associated protein Cas1 [Pyrobaculum neutrophilum]
MAAYGARIRARKGLLLVETKEGAREYPLHEVDEVLLLTGGISITTRALRALLAAGATVAVFSPRGEPLGI
FMKPIGDATGAKRRCQYKAAEDGRGLQYAKSWVFKKMLGQRDNIKAWRRRLRGYSQYAESLAKALQALRD
AASPHAVLEAEAAAAEAYWAAYREVTGFPGRDQEGRDPVNAGLNYGYGILKALVYKSLILAGLDPYVGFL
HVDKSGRPSLALDFMEQWRPRVDAVVAKMADKLESEGGLLTRRSRLELAAAVLEELHAAKRPLSAEIHRE
ARALARSICT [SEQ ID NO: 68]

Pyrobaculum neutrophilum Cas2

>gi|501318855|ref|WP_012350490.1| CRISPR-associated protein Cas2 [Pyrobaculum neutrophilum]
MYIIVVYDITENDVRAKVADILRAYGLARIQRSAYVGRLPPALVKELAERLARAVKGANADIAIFKVDKR
AIETALRIPPRPPAGHAALH [SEQ ID NO: 137]

Type I-B

Haloarcula hispanica Cas1

>gi|344209644|ref|YP_004785821.1| CRISPR-associated protein Cas1 [Haloarcula hispanica ATCC 33960]
MDRNYHVFSDGRLERSDDTLRLVPDDDGDKQYIPIENAEAFFLHGQIDFNTRLMSFLNDRTVALHIFGWE
DYYAGSVMPKRGQTSGKTVVNQVRAYEDSQHRRQLAAAIVKGSIHNMRTNVVYYNGRDHDLDSVIEDLEA
AATRVDESLPIDELMGIEATAKKAYYRSFNQILPSEFQLSQREYNPPPNEINSLISFGNSMVYANCVSAI
RATALDPTISYLHEPGERRYSLSLDIADLFKPVLADRVLFRLVNRNQISESDFETELGSCLLNEAGRKTY
TKAFEEMLERTVEHPTLNRKVSYQYLMRLEAYKLKKHLLTGEAYDSFQRWW [SEQ ID NO: 42]

Haloarcula hispanica Cas2

>gi|344209643|ref|YP_004785820.1| CRISPR-associated protein Cas2 [Haloarcula hispanica ATCC 33960]
MYVVMVYDLEADRTQKALKIGRRYLTHVQNSVLEGEISEGDLSNLKNEIDDLLKPGESTIIYELSSDTLL
NRTVYGDDPTEDQRFL [SEQ ID NO: 111]

Figure 16 (cont. 1)

Type I-C

Desulfovibrio vulgaris Cas1

>gi|387134024|ref|YP_005704014.1| CRISPR-associated protein Cas1 [Desulfovibrio vulgaris RCH1]
MKKLLNTLYVTTQGTYLAKEGECIVVRVGDEVRLRVPVHSLGGVVCFGQVSCSPFLMGFAAERGLGFSFL
TEHGRFLARVQGPVSGNVLLRREQYRRADSPEASAEVARSIVSAKVVNARGVLQRAMRDHGDKVDGVALE
AEVLHLRACLMRLQQPAGLDAVRGIEGEAAKGYFSVFDNLILTREAAFRFEGRSRRPPLDRVNCLLSFIY
TLLGHDVRSALEGVGLDSAVGFLHRDRPGRHGLALDVMEEFRAVVADRLALSLINLGKLKKSDFEIQETG
AVRMTDDARKALLVAYQKRKQDEIVHPFLNERIPLGLVFHVQAMLMARWLRGDLDGYPPFVWK [SEQ ID NO: 36]

Desulfovibrio vulgaris Cas2

>gi|387134025|ref|YP_005704015.1| CRISPR-associated protein Cas2 [Desulfovibrio vulgaris RCH1]
MLVLISYDVSFEDPGGQRRLRRIAKACQDYGQRVQYSVFECVVDPAQWAKLKHRLLSEMDKEKDCLRFYY
LGANWRNKVEHVGAKPAYDPEGPLIL [SEQ ID NO: 101]

Type I-D

Thermofilum pendens Cas1

>gi|500076970|ref|WP_011752983.1| CRISPR-associated protein Cas1 [Thermofilum pendens]
MRLLVLRGVDEVTVSSRSTVVIKSGNRVFERALRDVDAVLVVGSGIKISSSLPPVLALHGIPLSILAKGH
VAVLLNPVGTKYNNYRALQYTLPKNKALAIALEYLKSRVRGMASIIRNRGGRLPALPEPPDPALYEDPAR
LESDIRSWEAAASNTLWDEVFKLLDPSAARELRERYGFAGRKPGHPDPLNKAISAMYAVLYTLSTKALVA
AGLDPTYGFLHRTQYSVPLAFDYAEAFKPLAVEAALDLVNEEGLPTLSEDGDLDKDSLNKAMKRLYRYLS
AKHRETGKTPYQQIHLKAFCLAKHLEGKCSREKLAFTWDKKRYIIHE [SEQ ID NO: 82]

Thermofilum pendens Cas2

>gi|500076971|ref|WP_011752984.1| CRISPR-associated Cas2 family protein [Thermofilum pendens]
MIVIVAYDISDEDRRGRLRRYLRRLGLARVNRSVYAGPGTATTAELVAERAKEIVEEGDSVFVIVVREDE
YQRAHVFDGRDYYIVSERKYEVY [SEQ ID NO: 158]

Figure 16 (cont. 2)

Type I-E

Escherichia coli K-12 Cas1

>gi|16130662|ref|NP_417235.1| multifunctional endonuclease Cas1, CRISPR adaptation protein; DNA repair enzyme [Escherichia coli str. K-12 substr. MG1655]
MTWLPLNPIPLKDRVSMIFLQYGQIDVIDGAFVLIDKTGIRTHIPVGSVACIMLEPGTRVSHAAVRLAAQ
VGTLLVWVGEAGVRVYASGQPGGARSDKLLYQAKLALDEDLRLKVVRKMFELRFGEPAPARRSVEQLRGI
EGSRVRATYALLAKQYGVTWNGRRYDPKDWEKGDTINQCISAATSCLYGVTEAAILAAGYAPAIGFVHTG
KPLSFVYDIADIIKFDTVVPKAFEIARRNPGEPDREVRLACRDIFRSSKTLAKLIPLIEDVLAAGEIQPP
APPEDAQPVAIPLPVSLGDAGHRSS [SEQ ID NO: 38]

Escherichia coli K-12 Cas2

>gi|90111482|ref|NP_417234.2| CRISPR adaptation ssRNA endonuclease [Escherichia coli str. K-12 substr. MG1655]
MSMLVVVTENVPPRLRGRLAIWLLEVRAGVYVGDVSAKIREMIWEQIAGLAEEGNVVMAWATNTETGFEF
QTFGLNRRTPVDLDGLRLVSFLPV [SEQ ID NO: 104]

Type I-F

Pseudomonas aeruginosa M18 Cas1

>gi|386058681|ref|YP_005975203.1| CRISPR-associated protein [Pseudomonas aeruginosa M18]
MDDISPSELKTILHSKRANLYYLQHCRVLVNGGRVEYVTDEGRHSHYWNIPIANTTSLLLGTGTSITQAA
MRELARAGVLVGFCGGGGTPLFSANEVDVEVSWLTPQSEYRPTEYLQLWVGFWFDEEKRLEAARHFQRVR
LERIRHSWLEDRVLRDAGFAVDATALAVAVEDSARALEQAPNHEHLLTEEARLSKRLFKLAAQATRYGEF
VRAKRGSGGDPANRFLDHGNYLAYGLAATATWVLGIPHGLAVLHGKTRRGGLVFDVADLIKDSLILPQAF
LSAMRGDEEQDFRQACLDNLSRAQALDFMIDTLKDVAQRSTVSA [SEQ ID NO: 65]

Figure 16 (cont. 3)

Type II-A

Campylobacter jejuni RM1221 Cas1

>gi|57238545|ref|YP_179676.1| CRISPR-associated Cas1 family protein [Campylobacter jejuni RM1221]
MSYDEAFKTLLISSNAKLNLELNHLVIKQDENIAKLFLKDINIIVLESLQVSISSALFNAFARHKIILLT
CDETHSINGVFTPFLGHFQSAKIAKEQMNVSAQKKAILWQKIIKNKILNQAFILKKYNKIEQSNELINLA
KKVSLNDSKNIEAVAAALYFKTLFGTSFSRDELCFENSALNYGYAIIRACIIRAVCISGLLPWLGIKHDN
IYNSFALCDDLIEVFRASVDDCVLKLKGESEFLSKDDKRALIGNLQSKINFDGQNYPLNRAINHYVANFK
NALLYEDELKIVKFDD [SEQ ID NO: 31]

Campylobacter jejuni RM1221 Cas2

>gi|57238544|ref|YP_179675.1| CRISPR-associated Cas2 family protein [Campylobacter jejuni RM1221]
MIEDKFMRVLLMFDVPTKSKKEQKLASKFRNNLIKLGYFMLQFSVYMRICKGLSSAKSSIENVKKILPPY
GNVRALIITEKQFDKMELLLGGIVFNEKVNNETNLTLFDIDSHGEFKYKNSNNEEIQLNKKQEKYHQQNL
FEF [SEQ ID NO: 93]

Streptococcus pyogenes MGAS1882 Cas1

>gi|383493862|ref|YP_005411538.1| CRISPR-associated protein Cas1 [Streptococcus pyogenes MGAS1882]
MAGWRTVVVNTHSKLSYKNNHLIFKDAYKTELIHLSEIDILLLETTDIVLSTMLVKRLVDENVLVIFCDD
KRLPTAMLMPFYGRHDSSLQLGKQMSWSETVKSQVWTTIIAQKILNQSCYLGACSYFEKSQSIMDLYHGL
ENFDPSNREGHAARIYFNTLFGNDFSRDLEHPTNAGLDYGYTLLLSMFAREVVVSGCMTQFGLKHANQFN
QFNFASDIMEPFRPLVDKIVYENRNQPFPKIKRELFTLFSDTFSYNGKEMYLTNIISDYTKKVVKALNNE
GKGVPEFRI [SEQ ID NO: 74]

Streptococcus pyogenes MGAS1882 Cas2

>gi|383493863|ref|YP_005411539.1| CRISPR-associated protein Cas2 [Streptococcus pyogenes MGAS1882]
MSYRYMRMILMFDMPTDTAEERKAYRKFRKFLLSEGFIMHQFSIYSKLLLNNTANNAMIGRLREHNPNKG
NITLLTVTEKQFARMIYLHGERNNCIANSDERLVFLGEAFDES [SEQ ID NO: 147]

Type II-B

Francisella novicida 3523 Cas1

>gi|387824702|ref|YP_005824173.1| CRISPR-associated protein Cas1 [Francisella cf. tularensis subsp. novicida 3523]
MLVGANQHAVVQARINQYRYIHENALALSTKLIIAKIKNQRATLSYFNKHHKSVNLLNAIEELKRIAQLI
KNAKTLNDVLGYEGYAANIYFSSLARDKFLSESFANREGRGSQEIANSMLNFGYAILSSYILNAVTNAGL
EPYLGFLHQKRPGKMSLVLDLMEEYRAWVVDRVVIKLREQYKNKKSIDPKLKSALISEIQATIAKKYIYN
GKKLKLEHIIQRQVYRLSGEFAGEHNYKPYIFKW [SEQ ID NO: 39]

Francisella novicida 3523 Cas2

>gi|383493863|ref|YP_005411539.1| CRISPR-associated protein Cas2 [Streptococcus pyogenes MGAS1882]
MSYRYMRMILMFDMPTDTAEERKAYRKFRKFLLSEGFIMHQFSIYSKLLLNNTANNAMIGRLREHNPNKG
NITLLTVTEKQFARMIYLHGERNNCIANSDERLVFLGEAFDES [SEQ ID NO: 107]

Figure 16 (cont. 4)

Type II-C

Neisseria meningitides 8013 Cas1

>gi|677989519|sp|C9X1G6.1|CAS1_NEIM8 RecName: Full=CRISPR-associated endonuclease Cas1
MTWRSLLIQNGGKLSLQRRQLLIQQNGESHTVPLEDIAVIIIENRETLITAPLLSALAEHGATLLTCDEQ
FLPCGQWLPYAQYHRQLKILKLQLNISEPLKKQLWQHIVRQKILNQAFVADETGNDLAAKRLRTLASEVR
SGDTGNREAQAAALYFQALFGEKFTRNDNNAVNAALNYTYAVLRAAVARTLTLYGWLPALGLFHRSELNP
FNLADDFIEPLRPLADLTVIHLYEQGRLKAELTPGIKQHLIKTLHYQISIERQHFSTLAAIDKMVSSFQA
GVTDKNAKQLKLPEILPLKEYQYE [SEQ ID NO: 61]

Neisseria meningitides 8013 Cas2

>gi|677990551|sp|C9X1G7.1|CAS2_NEIM8 RecName: Full=CRISPR-associated endoribonuclease Cas2
MSEAKFMRIIVFFDLPVITAAKRKAANQFRQFLLKDGYQMLQLSVYSRIVKGRDSLQKHHNRLCANLPQE
GSIRCLEITEKQYAAMKLLLGELKTQEKKVNSDQLLLF [SEQ ID NO: 132]

Type III-A

Staphylococcus epidermidis RP62A Cas1

>gi|57865886|ref|YP_190006.1| CRISPR-associated Cas1 family protein [Staphylococcus epidermidis RP62A]
MKDVIYVENHYFVTAKENSIKFRNVIDKSEKFYLFEEIEAIIFDHYKSYFSHKLVIKCIENDIAIIFCDK
KHSPVTQLISSYGMVNRLKRIQSQFQLSGRTKDRIWKKIVINKIFNQTRCLENNLHNENVKLMLGFAKEV
SSGDKSNKEAHATRIYFKDLFGKQFKRGRYNDVINSGLNYGYSILRSFINKELAIHGFEMSLGIKHQSKE
NPFNLADDIIEVFRPFVDNIVYEIVFKKNIDTFDINEKKLLLNVLYERCIIDKKVVRLLDSVKIVIQSLI
RCYEENTPTYLLLPKMIEVGN [SEQ ID NO: 73]

Staphylococcus epidermidis RP62A Cas2

>gi|57865885|ref|YP_190005.1| CRISPR-associated Cas2 family protein [Staphylococcus epidermidis RP62A]
MYLLVSFDLPRDTKLERRIASKYRLRLIELGFSMKQFSLYERYVSNVQKKDKILEILKQEIPDTGSITLY
ILPDEVNSNQITILGKEVKLAVNKEPKLIFI [SEQ ID NO: 143]

Type III-B

Sulfolobus islandicus M.16.4 Cas1

>gi|238619345|ref|YP_002914170.1| CRISPR-associated protein Cas1 [Sulfolobus islandicus M.16.4]
MRTLVISEYGAYIYVKKNMLVIKKGDNKVEISPSEVDEILITASCSISTSALSLALTHGISVMFLNSRDT
PWGILLPSVITETVKTKKAQYETIVAKKDIRYGEEIISSKIYNQSVHLKYWTRLTGTRNDYKELLGKDEP
TAARIYWRNISQLLPKDIGFDGRDVDGVDQFNMALNYSYAILYNTIFKYLVIAGLDPYLGFIHKDRPGNE
SLVYDFSEMFKPYIDFLLVRALRSGFRLKVKDGLIEENSRGDLAKLIRKGMEEKVKEESDHNPKTLIQAI
RAHAVKLASSIREGKEYKGFKLVM [SEQ ID NO: 76]

Sulfolobus islandicus M.16.4 Cas2

>gi|238619346|ref|YP_002914171.1| CRISPR-associated protein Cas2 [Sulfolobus islandicus M.16.4]
MKLLVVYDVSDDSKRSKLANNLKKLGLERIQRSAFEGDIDSQRVKDLIRVVRLIVDTSTDIVHIIPLGVR
DWERRIVIGKEGLEEWLV [SEQ ID NO: 149]

Figure 17

Type I-A
Thermoproteus tenax – 7 CRISPR loci, only one Cas1–Cas2 (fusion protein)

Leader sequences

>CRISPR I
CAAAACTGCTATCATGCCACCTCAAGAGTCCGCCGAGATCAACGTCAGTCTCGTGAGAAAA
AAGCTTAAAAGCTCGTAGAAACAAAGACAACAATACCCG [SEQ ID NO: 163]

>CRISPR II
AGAGGAATATATGAATGTAAATGGGCGTTTGAAGGCGACCGCCGGGCGGGCGGAGGATC
TGGGCAAAATTTAAATAGCCGGGGCCGCAATCGACGTGGGC [SEQ ID NO: 164]

>CRISPR III
GGAGAATGTGTGAATGTAAAGGGCAGCGGAAAAGCGGCCGCCGGGCGGGCGGAGGATCT
GGGCAAAATTTAAATACCCGAGGCCGCAGTCAACGTGGGCG [SEQ ID NO: 165]

>CRISPR IV
AAAGCCCTCCCCGGGGCAACACAAGGTCGCGCCGGGACTAACACAAAGACCCACAGAGG
AAAACTTAAAAATCGGCAAAAACTAAAGACCAGAAGGCCCG [SEQ ID NO: 166]

>CRISPR V
AAAGCCCTCCCCGGGGCAACACAAGGTCGCGCCGGGACCAAGACAAAGACCCGCAGAGG
AAAACTTAAAAATCGGCAGAAACTAAAGACCAGGAGGCCCG [SEQ ID NO: 167]

>CRISPR VI
GCAGTCCGGGGAGAGCTATCTCAAGAACACGCTGCAATCAACGCAAAACCCACCAGAGAA
AAACTTAAAAACAGCCAAAAACCAAAACCCAGAAGGCCCG [SEQ ID NO: 168]

>CRISPR VII
AACCCCACCCCTTAGCCAACACGAGAACACGCCAGAACTAACACAAAACCCACCAGAGAA
AAACTTAAAAACAGCCAAAAGCCAAAACCCAGAAGGCCCG [SEQ ID NO: 169]

Repeat sequences

> Repeat I
GAATCTCAGATAGAGATTTGAAGG [SEQ ID NO: 189]

>Repeat II
AGTGGAAATCAAAAGATAGTAGAAAC [SEQ ID NO: 190]

>Repeat III
GTGGAAATCAAAAGATAGTAGAAAG [SEQ ID NO: 191]

>Repeat IV
GAATCTCAAAGAGAGGATTGAAAG [SEQ ID NO: 192]

Figure 17 (cont. 1)

>Repeat V
GAATCTCAAAGAGAGGATTGAAAG [SEQ ID NO: 193]

>Repeat VI
GAATCTCAAAAAGAGGATTGAAAG [SEQ ID NO: 194]

>Repeat VII
GAATCTCAAAGAGAGGATTGAAAG [SEQ ID NO: 195]

Type I-B
Haloarcula hispanica N601

Leader sequence
AGGTGAATCGACAATTATCTACGAGCTGTCTTCTGATACACTGCTCAACCGGACAGTCTAC
GGCGACGATCCAACTGAGGATCAGCGGTTTCTATAGTCG [SEQ ID NO: 170]

Repeat sequence
GTTTCAGACGAACCCTCGTGGGGTTGAAGC [SEQ ID NO: 196]

Type I-C
Desulfovibrio vulgaris RCH1

Leader sequence
AGAAGGCCCCCTGATTCTGTAGCGCGAACCCCAAGCGACCCTGTTTTTCCCGGCAGGTTC
GCGAACATGACAATGACCTGTTTTCATTGAATTGCGTAACCTTTAATGCAGGCTGGTCACA
CATCTTGGCGGTGCTGCTGGCCCGGTTCGCGGAACCCTCGTCGCAAGGTCAATACTGCCA
ACGTGTTTGATGGCCGACA [SEQ ID NO: 171]

Repeat sequence
GTCGCCCCCCACGCGGGGGCGTGGATTGAAAC [SEQ ID NO: 197]

Type I-D
Thermofilum pendens Hrk 5

Leader sequence
CCGTAGGTGTCTGCGCGTAGGCTGGAGAAACGGGCGCCGAGAGGTTCGGGAACCGCTCC
TCGCATCGTCAGTCTAGGATAAGGTTGAGGCAGTTTTAGCGGAAAGATTGGGCTCTAAAAA
TTGATTGAATTGATCGTTTTTCTAAACTTTTCACGATTTTTCGAACAAGAATATTAGAGAATG
CAACCTCTTCTGTTACC [SEQ ID NO: 172]

Repeat sequence
CTCTTAGTCTTGCTGATTTTGAAC [SEQ ID NO: 198]

Figure 17 (cont. 2)

Type I-E
Escherichia coli BL21(DE3)
Leader sequence
ACTCTTTAACATAATGGATGTGTTGTTTGTGTGATACTATAAAGTTGGTAGATTGTGACTGG
CTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGA [SEQ ID NO: 173]

Repeat sequence
GTGTTCCCCGCGCCAGCGGGGATAAACCG [SEQ ID NO: 199]

Escherichia coli K-12 (MG1655)
Leader sequence
AAGAATTAGCTGATCTTTAATAATAAGGAAATGTTACATTAAGGTTGGTGGGTTGTTTTTATG
GGAAAAAATGCTTTAAGAACAAATGTATACTTTTAGA [SEQ ID NO: 174]

Repeat sequence
GAGTTCCCCGCGCCAGCGGGGATAAACCG [SEQ ID NO: 200]

Type I-F
Pseudomonas aeruginosa M18 – 3 CRISPR loci

Leader sequences

>CRISPR I
TATAAATCAGCAAGTTACGAGACCTCGAAAAAAGAGGGTTTCTGGCAGGAAAAACTCGGTA
TTTCCTTTTCCTTCAAATGGTTATAGGTTTTCGGGGCTA [SEQ ID NO: 175]

>CRISPR II
TTTTAGATCAAAGGGTTAGAGATCGCTGCAAAAAGAGGGTTTTTCCGGGCTTTGGCGCTGG
AGCCCTTGGAGCTTGGAAGGTTGATGGTTTTTGGGTCTA [SEQ ID NO: 176]

>CRISPR III
TAGCTCCGAAAACCTATAACCATTTGAAGGAAAAAGAAATACCGAGTTTTTCCCGCCAGAAA
CCCTCTTTTTTCGAGGTCTCGTAACTTGCTGATTTATA [SEQ ID NO: 177]

Repeat sequences

>CRISPR I
GTTCACTGCCGTATAGGCAGCTAAGAAA [SEQ ID NO: 201]

>CRISPR II
GTTCACTGCCGTATAGGCAGCTAAGAAA [SEQ ID NO: 202]

>CRISPR III
TTTCTTAGCTGCCTACACGGCAGTGAAC [SEQ ID NO: 203]

Figure 17 (cont. 3)

Type II-A
Campylobacter jejuni RM1221
Leader sequences
(Left flanking the CRISPR locus)
AGATATTTACCAGATAATGAAAATTTCGGGGTTTTTTCATGAAAAATAGCAAAAATTATGCTA
TAATCTCATAAGAAATTTAAAAAGGGACTAAAATAAA [SEQ ID NO: 178]

(Right flanking the CRISPR locus)
AGATATTTACCAGATAATGAAAATTTCGGGGTTTTTTCATGAAAAATAGCAAAAATTATGCTA
TAATCTCATAAGAAATTTAAAAAGGGACTAAAATAAA [SEQ ID NO: 179]

Repeat sequence
GTTTTAGTCCCTTTTTAAATTTCTTTATGGTAAAAT [SEQ ID NO: 204]

Streptococcus pyogenes MGAS1882
Leader sequences
(Left flanking the CRISPR locus)
CTCTTAATAAATGCAGTAATACAGGGGCTTTTCAAGACTGAAGTCTAGCTGAGACAAATAGT
GCGATTACGAAATTTTTTAGACAAAAATAGTCTACGAG [SEQ ID NO: 180]

(Right flanking the CRISPR locus)
AACATTGCCGATGATAACTTGAGAAAGAGGGTTCATACCAGCAGTCGGATACCTTTCTATT
CTTCCTGTTAAAACGTTTTCATGTTATAATAGGCAAAAG [SEQ ID NO: 181]

Repeat sequence
GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC [SEQ ID NO: 205]

Type II-B
Francisella novicida 3523
Leader sequences
(Left flanking the CRISPR locus)
TTGAATTGATATTTTGCTATATAACTATTTTAGCGGTTATATTATCTGATATACTTTACCCAGA
TAGAAACTTCTAAATATTATTCGTAAAATTAATTAA [SEQ ID NO: 182]

(Right flanking the CRISPR locus)
AGCCAAGACAAACCTAGAGTTTGCAAGGCTTTGAGAGAATTATTTGCTGGTGTTTTTTCGCA
AACATCCTACTTGTAAAAAAATGTAGACATTTTGCGAA [SEQ ID NO: 183]

Repeat sequence
GTTTCAGTAGCTAGATTATTTGATATACTGCTGTTAG [SEQ ID NO: 206]

Figure 17 (cont. 4)

Type II-C
Neisseria meningitides 8013

Leader sequence
AAAAAAGGTCAATTCAGACCAATTATTGTTATTTTAAGCCCATTTTTTCATAACAAATAAAAC
GGGAAACCCTTATGAAATAAGGATTTCCCGTCGAAGT [SEQ ID NO: 184]

Repeat sequence
ATTGTAGCACTGCGAAATGAGAAAGGGAGCTACAAC [SEQ ID NO: 207]

Type III-A
Staphylococcus epidermidis RP62A

Leader sequences
(Left flanking the CRISPR locus)
TTTCTACATAAATAACATCTTTCATTTTTCCATCCCCTAGAAATTAATCAATGCGTATTTTATT
CAAAATCTACAATTTTTTAGAGTGTTGTTAGATTTA [SEQ ID NO: 185]

(Right flanking the CRISPR locus)
GATACTTTAACAAATGCCATCACAACTATATTTCAAGCATCATTTTTGCTGTCAAAAAATATG
ACAATCACTAGTACAAGATTATATGATATGTCACTTT [SEQ ID NO: 186]

Repeat sequence
GTTCTCGTCCCCTTTTCTTCGGGGTGGGTATCGATCC [SEQ ID NO: 208]

Type III-B
Sulfolobus islandicus M.16.4

Leader sequences
(Left flanking the CRISPR locus)
TCTTATTCGATATAAACGTACTTATATCTATTAGGACTTCGTCTTTTCCCATACGGCTTCCCT
AGATTTAGATTTCAAACAAGTCATAGAATATAGTATA [SEQ ID NO: 187]

(Right flanking the CRISPR locus)
TGAGGGTTTATTAACTTTTATCTAATATCTTGTTCTCTCTTCGTTATTTATAAGCTTTACTCTG
TATTATCTTTTCAATTTTTCTCCTCATCCTTTCACT [SEQ ID NO: 188]

Repeat sequence
CTTTCAATTCTATAGTAGATTATC [SEQ ID NO: 209]

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 15/67161

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - A01N 63/00; C12N 1/21, 15/00, 9/14; C07H 21/04 (2016.01)
CPC - C12N 15/00, 15/63, 9/16
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8): A01N 63/00; C12N 1/21, 15/00, 9/14; C07H 21/04 (2016.01)
CPC: C12N 15/00, 15/63, 9/16

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
CPC: C12N 15/00, 15/63, 9/16 (text search)
USPC: 424/93.1, 93.21; 435/196, 471, 320.1, 5; 523/23.1 (text search)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Electronic data bases: PatBase; Google Patents; Google Scholar
Search terms: Cas1, Cas 2, donor, target, integration, eukaryotic cell, in vitro, CRISPR array, spacer integration, integrase, adaptive immunity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ARSLAN et al. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. Nucl Acids Res ePub 11 June 2014 Vol 42 No 12 Pages 7884-7893. Especially pg 7882 col 2 para 2, pg 7885 fig 1A, C.	1-5
Y	NUNEZ et al. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat Struct Mol Biol ePub 14 May 2014 Vol 21 No 6 Pages 528-534. Especailly pg 532 col 1 para 2, pg 532 fig 4E.	1-5

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(ε) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
18 April 2016	06 MAY 2016
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 15/67161

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 6-36, 41-44, 48-53, 61
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
--------Go to Extra Sheet for continuation----------

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
Claims 1-5

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 15/67161

--------continuation of Box III (Lack of Unity of Invention)---------

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Group I: Claims 1-5, drawn to a method of nucleic acid integration comprising contacting a target DNA with a donor DNA molecule and a Cas1 protein.

Group II: Claims 37-40, 45-47, drawn to a composition or kit comprising a Cas1 protein or nucleic acid encoding Cas1 protein, and at least one of a donor DNA and a Cas2 protein or nucleic acid encoding Cas2 protein.

Group III: Claims 54-57, drawn to a composition comprising a mutant Cas1 protein [i.e. amino acid alteration], or a nucleic acid encoding a mutant Cas1 protein.

Group IV: Claims 58-60, drawn to a composition comprising a mutant Cas2 protein [i.e. amino acid alteration], or a nucleic acid encoding a mutant Cas2 protein.

The inventions listed as Groups I-IV do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

Special Technical Features:

Group I has the special technical feature of the specific method step of contacting a target DNA molecule with a donor DNA molecule and a Cas1 protein, not required by Groups II-IV.

Group II has the special technical feature of a kit in one or more containers, not required by Groups I, III or IV.

Group II has the special technical feature of a composition or kit including Cas2 protein or a nucleic acid encoding Cas2 protein, not required by Groups I, III or IV.

Group III has the special technical feature of a mutant Cas1 protein [i.e. amino acid alteration], or a nucleic acid encoding a mutant Cas1 protein, not required by Groups I, II or IV.

Group IV has the special technical feature of a mutant Cas2 protein [i.e. amino acid alteration], or a nucleic acid encoding a mutant Cas2 protein, not required by Groups I-III.

Common Technical Feature:

1. Groups I-III share the common technical feature of a Cas1 protein, or nucleic acid encoding Cas1.

2. Groups II-IV share the common technical feature of a Cas2 protein, or a nucleic acid encoding Cas2.

3. Groups I and II share the common technical feature of a Cas1 protein and a linear, donor DNA.

4. Group III and IV share the common technical feature of a mutant Cas protein [i.e. amino acid alteration], or a nucleic acid encoding a mutant Cas protein.

However, said common technical features do not represent a contribution over the prior art, and is obvious over the publication titled "Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system" by ARSLAN et al. (hereinafter "Arslan") [published July 2014 in Nucl Acids Res Vol 42 No 12 Pages 7884-7893] in view of the publication titled "Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity" by NUNEZ et al. (hereinafter "Nunez") [published June 2014 in Nat Struct Mol Biol Vol 21 No 6 Pages 528-534].

As to common technical features #1 and #2, Arslan teaches plasmids encoding Cas1 and Cas2 and induction of expression of said plasmid to generate Cas1 and Cas2 proteins (pg 7885 fig 1A; pCR001 encoding Cas1 and Cas2).

As to common technical feature #3, Arslan teaches Cas1 (pg 7885 fig 1A), donor DNA (pg 7885 fig 1A; gDNA including CRISPR array) and donor DNA (pg 7885 fig 1A; extended CRISPR array; pg 7886 col 2 para 3; " The integrated spacers [i.e. donor DNA molecule] were derived either from the genomic DNA or from plasmid DNA, and insertion of up to three spacers into a single CRISPR array could be observed (Supplementary Table S1)"). Arslan does not teach a specific donor DNA molecule, per se, and also does not specifically teaches that it is linear. However, Nunez teaches specific donor DNA interaction with Cas1-Cas2 complex in vitro (pg 532 col 1 para 2; "Our discovery that Cas1 and Cas2 form an essential complex that is required for CRISPR spacer acquisition in vivo led us to test for CRISPR DNA binding by Cas1 and Cas2") and indicates that the donor DNA molecules are linear (pg 532 col 1 para 2; "We first tested the ability of Cas1 and/or Cas2 to bind a 186-bp 5'-biotinylated dsDNA containing two CRISPR repeats, two spacers and the minimal 60-bp leader sequence shown to be required for spacer acquisition in vivo"; pg 532 fig 4E; biotinylated CRISPR).

----continued on next sheet----

-----continued from previous sheet------

As to common technical feature #4, Nunez teaches a mutant Cas1 protein and also a mutant Cas2 protein (pg 528 col 2 para 2; "We combined an in vivo spacer acquisition assay with mutagenesis and immunoprecipitation experiments to show that physical disruption of complex formation abrogates spacer acquisition. Although active site mutations in Cas1 inhibit spacer acquisition, the catalytic activity of Cas2 is not required for either Cas1-Cas2 complex formation or new spacer acquisition").

As the common technical feature was known in the art at the time of the invention, this cannot be considered a common special technical feature that would otherwise unify the groups. The inventions lack unity with one another.

Therefore, Groups I-IV lack unity of invention under PCT Rule 13 because they do not share a same or corresponding special technical feature.

Note concerning item 4: Claims 6-36, 41-44, 48-53, 61 are multiple dependent claims, and are not drafted according to the second and third sentences of PCT Rule 6.4(a).