RNA-GUIDED NUCLEIC ACID MODIFYING ENZYMES AND METHODS OF USE THEREOF

CROSS-REFERENCE

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 62/402,849, filed September 30, 2016, 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-343WO_SeqList_ST25.txt" created on September 28, 2017 and having a size of 244 KB. The contents of the text file are incorporated by reference herein in their entirety.

INTRODUCTION

[0003] The CRISPR-Cas system, an example of a pathway that was unknown to science prior to the DNA sequencing era, is now understood to confer bacteria and archaea with acquired immunity against phage and viruses. Intensive research over the past decade has uncovered the biochemistry of this system. CRISPR-Cas systems consist of Cas proteins, which are involved in acquisition, targeting and cleavage of foreign DNA or RNA, and a CRISPR array, which includes direct repeats flanking short spacer sequences that guide Cas proteins to their targets. Class 2 CRISPR-Cas are streamlined versions in which a single Cas protein bound to RNA is responsible for binding to and cleavage of a targeted sequence. The programmable nature of these minimal systems has enabled their use as a versatile technology that is revolutionizing the field of genome manipulation.

[0004] Current CRISPR-Cas technologies are based on systems from cultured bacteria, leaving untapped the vast majority of organisms that have not been isolated. To date, only a few Class 2 CRISPR/Cas systems have been discovered. There is a need in the art for additional Class 2 CRISPR/Cas systems (e.g., Cas protein plus guide RNA combinations).

SUMMARY

[0005] The present disclosure provides RNA-guided endonuclease polypeptides, referred to herein as "CasY" polypeptides (also referred to as "CasY proteins"); nucleic acids encoding the CasY polypeptides; and modified host cells comprising the CasY polypeptides and/or nucleic acids encoding same. CasY polypeptides are useful in a variety of applications, which are provided.

[0006] The present disclosure provides guide RNAs (referred to herein as "CasY guide RNAs") that bind to and provide sequence specificity to the CasY proteins; nucleic acids encoding the CasY guide RNAs; and modified host cells comprising the CasY guide RNAs and/or nucleic acids encoding same. CasY guide RNAs are useful in a variety of applications, which are provided.

[0007] The present disclosure provides methods of identifying a CRISPR RNA-guided endonuclease.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] **Figure 1** depicts examples of naturally occurring CasY protein sequences.

[0009] **Figure 2** depicts an alignment of naturally occurring CasY protein sequences.

[0010] **Figure 3 (panels a-b)** depicts a schematic domain representation for CasY. Also shown are results from various searches attempting to identify homologs of CasY. Also depicted are portions of the CasY-containing CRISPR loci there were identified.

[0011] **Figure 4** depicts a schematic diagram of CasY and C2c3 loci. Interference proteins are shown in green, acquisition proteins in red. Repeats folded using RNA structure are shown to the right revealing a strong hairpin at the 5' end, suggesting self processing of the CRISPR array by CasY.

[0012] **Figure 5 (panels a-d)** depicts experiments performed (PAM dependent plasmid interference by CasY) to determine a PAM sequence for CasY.

[0013] **Figure 6 (panels a-b)** presents 'repeat' sequences of naturally occurring CasY guide RNAs, and an example CasY guide RNA hybridizing to target DNA. (Top to bottom, SEQ ID NOs: 11-15, and 20)

[0014] **Figure 7 (panels a-b)** presents novel identified CRISPR-Cas systems from uncultivated organisms. **a,** Ratio of major lineages with and without isolated representatives in all bacteria and archaea, based on data of Hug et al.[32]. The results highlight the massive scale of as yet little investigated biology in these domains. Archaeal Cas9 and the novel CRISPR-CasY were found exclusively in lineages with no isolated representatives. **b,** Locus organization of the newly discovered CRISPR-Cas systems.

[0015] **Figure 8 (panels a-b)** presents ARMAN-1 CRISPR array diversity and identification of the ARMAN-1 Cas9 PAM sequence. **a,** CRISPR arrays reconstructed from 15 different AMD samples. White boxes indicate repeats and colored diamonds indicate spacers (identical spacers are similarly colored; unique spacers are in black). The conserved region of the array is highlighted (on the right). The diversity of recently acquired spacers (on the left) indicates the system is active. An analysis that also includes CRISPR fragments from the read data is presented in Figure 14. **b**, A single putative viral contig reconstructed from AMD metagenomic data contains 56 protospacers (red vertical bars) from the

ARMAN-1 CRISPR arrays. **c**, Sequence analysis revealed a conserved 'NGG' PAM motif downstream of the protospacers on the non-target strand.

[0016] **Figure 9 (panels a-d)** presents data showing that CasX mediates programmable DNA interference in *E. coli*. **a,** Diagram of CasX plasmid interference assays. *E. coli* expressing a minimal CasX locus is transformed with a plasmid containing a spacer matching the sequence in the CRISPR array (target) or plasmid containing a non-matching spacer (non-target). After being transformed, cultures are plated and colony forming units (cfu) quantified. **b,** Serial dilution of *E. coli* expressing the Planctomycetes CasX locus targeting spacer 1 (sX.1) and transformed with the specified target (sX1, CasX spacer 1; sX2, CasX spacer 2; NT, non-target). **c,** Plasmid interference by Deltaproteobacteria CasX. Experiments were conducted in triplicate and mean ± s.d. is shown. **d,** PAM depletion assays for the Planctomycetes CasX locus expressed in *E. coli*. PAM sequences depleted greater than 30-fold compared to a control library were used to generate the WebLogo.

[0017] **Figure 10 (panels a-c)** presents data showing CasX is a dual-guided CRISPR complex**. a,** Mapping of environmental RNA sequences (metatranscriptomic data) to the CasX CRISPR locus diagramed below (red arrow, putative tracrRNA; white boxes, repeat sequences; green diamonds, spacer sequences). Inset shows detailed view of the first repeat and spacer. **b,** Diagram of CasX double-stranded DNA interference. The site of RNA processing is indicated by black arrows. **c,** Results of plasmid interference assays with the putative tracrRNA knocked out of the CasX locus (T, target; NT, non-target). Experiments were conducted in triplicate and mean ± s.d. is shown.

[0018] **Figure 11 (panels a-c)** presents data showing expression of a CasY locus in *E. coli* is sufficient for DNA interference. **a,** Diagrams of CasY loci and neighboring proteins. **b,** WebLogo of 5′ PAM sequences depleted greater than 3-fold by CasY relative to a control library. **c,** Plasmid interference by *E. coli* expressing CasY.1 and transformed with targets containing the indicated PAM. Experiments were conducted in triplicate and mean ± s.d is shown.

[0019] **Figure 12 (panels a-b)** presents newly identified CRISPR-Cas in context of known systems. **a,** Simplified phylogenetic tree of the universal Cas1 protein. CRISPR types of known systems are noted on the wedges and branches; the newly described systems are in bold. Detailed Cas1 phylogeny is presented in Supplementary Data 2. **b,** Proposed evolutionary scenario that gave rise to the archaeal type II system as a result of a recombination between type II-B and type II-C loci.

[0020] **Figure 13** presents shows that archaeal Cas9 from ARMAN-4 is found on numerous contigs with a degenerate CRISPR array. Cas9 from ARMAN-4 is highlighted in dark red on 16 different contigs. Proteins with putative domains or functions are labeled whereas hypothetical proteins are unlabeled. Fifteen of the contigs contain two degenerate direct repeats (one bp mismatch) and a single,

conserved spacer. The remaining contig contains only one direct repeat. Unlike ARMAN-1, no additional Cas proteins are found adjacent to Cas9 in ARMAN-4.

[0021] **Figure 14** presents a full reconstruction of ARMAN-1 CRISPR arrays. Reconstruction of CRISPR arrays, that include reference assembled sequences, as well as array segments reconstructed from the short DNA reads. Green arrows indicate repeats and colored arrows indicate CRISPR spacers (identical spacers are colored the same whereas unique spacers are colored in black). In CRISPR systems, spacers are typically added unidirectionally, so the high variety of spacers on the left side is attributed to recent acquisition.

[0022] **Figure 15 (panels a-b)** shows that ARMAN-1 spacers map to genomes of archaeal community members. **a**, Protospacers (red arrows) from ARMAN-1 map to the genome of ARMAN-2, a nanoarchaeon from the same environment. Six protospacers map uniquely to a portion of the genome flanked by two long-terminal repeats (LTRs), and two additional protospacers match perfectly within the LTRs (blue and green). This region is likely a transposon, suggesting the CRISPR-Cas system of ARMAN-1 plays a role in suppressing mobilization of this element. **b**, Protospacers also map to a *Thermoplasmatales* archaeon (I-plasma), another member of the Richmond Mine ecosystem that is found in the same samples as ARMAN organisms. The protospacers cluster within a region of the genome encoding short, hypothetical proteins, suggesting this might also represent a mobile element.

[0023] **Figure 16 (panels a-e)** presents predicted secondary structure of ARMAN-1 crRNA and tracrRNA. **a,** The CRISPR repeat and tracrRNA anti-repeat are depicted in black whereas the spacer-derived sequence is shown as a series of green N's. No clear termination signal can be predicted from the locus, so three different tracrRNA lengths were tested based on their secondary structure – 69, 104, and 179 in red, blue, and pink, respectively. **b**, Engineered single-guide RNA corresponding to dual-guide in **a**. **c**, Dual-guide for ARMAN-4 Cas9 with two different hairpins on 3' end of tracrRNA (75 and 122). **d**, Engineered single-guide RNA corresponding to dual-guide in **c. e,** Conditions tested in *E. coli in vivo* targeting assay.

[0024] **Figure 17 (panels a-b)** presents purification schema for *in vitro* biochemistry studies. **a**, ARMAN-1 (AR1) and ARMAN-4 (AR4) Cas9 were expressed and purified under a variety of conditions as outlined in the Supplementary Materials. Proteins outlined in blue boxes were tested for cleavage activity *in vitro*. **b**, Fractions of AR1-Cas9 and AR4-Cas9 purifications were separated on a 10% SDS-PAGE gel.

[0025] **Figure 18** presents newly identified CRISPR-Cas systems compared to known proteins. Similarity of CasX and CasY to known proteins based on the following searches: (1) Blast search against the non-redundant (NR) protein database of NCBI, (2) Hidden markov model (HMM) search against an HMM database of all known proteins and (3) distant homology search using HHpred[30].

[0026] **Figure 19 (panels a-d)** presents data related to programed DNA interference by CasX. **a,** Plasmid interference assays for CasX2 (Planctomycetes) and CasX1 (Deltaproteobacteria), continued from Figure 9, panel c (sX1, CasX spacer 1; sX2, CasX spacer 2; NT, non-target). Experiments were conducted in triplicate and mean ± s.d. is shown. **b,** Serial dilution of *E. coli* expressing a CasX locus and transformed with the specified target, continued from Figure 9, panel b. **c,** PAM depletion assays for the Deltaproteobacteria CasX and **d,** Planctomycetes CasX expressed in *E. coli.* PAM sequences depleted greater than the indicated PAM depletion value threshold (PDVT) compared to a control library were used to generate the WebLogo.

[0027] **Figure 20** presents an evolutionary tree of Cas9 homologs. Maximum-likelihood phylogenic tree of Cas9 proteins, showing the previously described systems colored based on their type: II-A in blue, II-B in green and II-C in purple. The Archaeal Cas9, cluster with type II-C CRISPR-Cas systems, together with two newly described bacterial Cas9 from uncultivated bacteria.

[0028] **Figure 21** presents a table of cleavage conditions assayed for Cas9 from ARMAN-1 and ARMAN-4.

DEFINITIONS

[0029] "Heterologous," as used herein, means a nucleotide or polypeptide sequence that is not found in the native nucleic acid or protein, respectively. For example, relative to a CasY polypeptide, a heterologous polypeptide comprises an amino acid sequence from a protein other than the CasY polypeptide. In some cases, a portion of a CasY protein from one species is fused to a portion of a CasY protein from a different species. The CasY sequence from each species could therefor be condidered to be heterlogous relative to one another. As another example, a CasY protein (e.g., a dCasY protein) can be fused to an active domain from a non-CasY protein (e.g., a histone deacetylase), and the sequence of the active domain could be considered a heterologous polypeptide (it is heterologous to the CasY protein).

[0030] The terms "polynucleotide" and "nucleic acid," used interchangeably herein, refer to a polymeric form of nucleotides of any length, either ribonucleotides or deoxynucleotides. 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. The terms "polynucleotide" and "nucleic acid" should be understood to include, as applicable to the embodiment being described, single-stranded (such as sense or antisense) and double-stranded polynucleotides .

[0031] The terms "polypeptide," "peptide," and "protein", are used interchangeably herein, refer to a polymeric form of amino acids of any length, which can include genetically coded and non-genetically coded amino acids, chemically or biochemically modified or derivatized amino acids, and polypeptides having modified peptide backbones. The term includes fusion proteins, including, but not

limited to, fusion proteins with a heterologous amino acid sequence, fusions with heterologous and homologous leader sequences, with or without N-terminal methionine residues; immunologically tagged proteins; and the like.

[0032] The term "naturally-occurring" as used herein as applied to a nucleic acid, a protein, a cell, or an organism, refers to a nucleic acid, cell, protein, or organism that is found in nature.

[0033] As used herein the term "isolated" is meant to describe a polynucleotide, a polypeptide, or a cell that is in an environment different from that in which the polynucleotide, the polypeptide, or the cell naturally occurs. An isolated genetically modified host cell may be present in a mixed population of genetically modified host cells.

[0034] As used herein, the term "exogenous nucleic acid" refers to a nucleic acid that is not normally or naturally found in and/or produced by a given bacterium, organism, or cell in nature. As used herein, the term "endogenous nucleic acid" refers to a nucleic acid that is normally found in and/or produced by a given bacterium, organism, or cell in nature. An "endogenous nucleic acid" is also referred to as a "native nucleic acid" or a nucleic acid that is "native" to a given bacterium, organism, or cell.

[0035] "Recombinant," as used herein, means that a particular nucleic acid (DNA or RNA) is the product of various combinations of cloning, restriction, 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. Generally, DNA sequences encoding the structural coding sequence can be assembled from cDNA fragments and short oligonucleotide linkers, 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. Such sequences can be provided in the form of an open reading frame uninterrupted by internal non-translated sequences, or introns, which are typically present in eukaryotic genes. 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).

[0036] Thus, e.g., the term "recombinant" polynucleotide or "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 redundant codon encoding the same or a conservative amino acid, while typically introducing or removing a sequence recognition site. 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.

[0037] Similarly, the term "recombinant" polypeptide refers to a polypeptide which is not naturally occurring, e.g., is made by the artificial combination of two otherwise separated segments of amino sequence through human intervention. Thus, e.g., a polypeptide that comprises a heterologous amino acid sequence is recombinant.

[0038] By "construct" or "vector" is meant a recombinant nucleic acid, generally recombinant DNA, which has been generated for the purpose of the expression and/or propagation of a specific nucleotide sequence(s), or is to be used in the construction of other recombinant nucleotide sequences.

[0039] 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 expression of a coding sequence and/or production of an encoded polypeptide in a host cell.

[0040] The term "transformation" is used interchangeably herein with "genetic modification" and refers to a permanent or transient genetic change induced in a cell following introduction of new nucleic acid (e.g., DNA exogenous to the cell) into the cell. Genetic change ("modification") can be accomplished either by incorporation of the new nucleic acid into the genome of the host cell, or by transient or stable maintenance of the new nucleic acid as an episomal element. Where the cell is a eukaryotic cell, a permanent genetic change is generally achieved by introduction of new DNA into the genome of the cell. In prokaryotic cells, permanent changes can be introduced into the chromosome or via extrachromosomal elements such as plasmids and expression vectors, which may contain one or more selectable markers to aid in their maintenance in the recombinant host cell. Suitable methods of genetic modification include viral infection, transfection, conjugation, protoplast fusion, electroporation, particle gun technology, calcium phosphate precipitation, direct microinjection, and the like. The choice of method is generally dependent on the type of cell being transformed and the circumstances under which the transformation is taking place (i.e. *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.

[0041] "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. As used herein, the terms "heterologous promoter" and "heterologous control regions" refer to promoters and other control regions that are not normally associated with a particular nucleic acid in nature. For example, a

"transcriptional control region heterologous to a coding region" is a transcriptional control region that is not normally associated with the coding region in nature.

[0042] A "host cell," as used herein, denotes an *in vivo* or *in vitro* eukaryotic cell, a prokaryotic 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 (e.g., an expression vector), and include the progeny of the original cell which has been genetically modified 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 prokaryotic host cell is a genetically modified prokaryotic host cell (e.g., a bacterium), by virtue of introduction into a suitable prokaryotic host cell of a heterologous nucleic acid, e.g., an exogenous nucleic acid that is foreign to (not normally found in nature in) the prokaryotic host cell, or a recombinant nucleic acid that is not normally found in the prokaryotic host cell; and a subject eukaryotic host cell is a genetically modified eukaryotic host cell, by virtue of introduction into a suitable eukaryotic host cell of a heterologous nucleic acid, e.g., an exogenous nucleic acid that is foreign to the eukaryotic host cell, or a recombinant nucleic acid that is not normally found in the eukaryotic host cell.

[0043] 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 consists 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; and a group of amino acids having sulfur-containing side chains consists of cysteine and methionine. Exemplary conservative amino acid substitution groups are: valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, and asparagine-glutamine.

[0044] 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 similarity can be determined in a number of different manners. To determine sequence identity, sequences can be aligned using the methods and computer programs, including BLAST, available over the world wide web at ncbi.nlm.nih.gov/BLAST. See, e.g., Altschul *et al.* (1990), *J. Mol. Biol.* 215:403-10. Another alignment algorithm is FASTA, available in the Genetics Computing Group (GCG) package, from

Madison, Wisconsin, USA, a wholly owned subsidiary of Oxford Molecular Group, Inc. Other techniques for alignment are described in Methods in Enzymology, vol. 266: Computer Methods for Macromolecular Sequence Analysis (1996), ed. Doolittle, Academic Press, Inc., a division of Harcourt Brace & Co., San Diego, California, USA. Of particular interest are alignment programs that permit gaps in the sequence. The Smith-Waterman is one type of algorithm that permits gaps in sequence alignments. See *Meth. Mol. Biol.* 70: 173-187 (1997). Also, the GAP program using the Needleman and Wunsch alignment method can be utilized to align sequences. See *J. Mol. Biol.* 48: 443-453 (1970).

[0045] As used herein, the terms "treatment," "treating," and the like, refer to obtaining a desired pharmacologic and/or physiologic effect. The effect may be prophylactic in terms of completely or partially preventing a disease or symptom thereof and/or may be therapeutic in terms of a partial or complete cure for a disease and/or adverse effect attributable to the disease. "Treatment," as used herein, covers any treatment of a disease in a mammal, e.g., in a human, and includes: (a) preventing the disease from occurring in a subject which may be predisposed to the disease but has not yet been diagnosed as having it; (b) inhibiting the disease, i.e., arresting its development; and (c) relieving the disease, i.e., causing regression of the disease.

[0046] The terms "individual," "subject," "host," and "patient," used interchangeably herein, refer to an individual organism, e.g., a mammal, including, but not limited to, murines, simians, humans, mammalian farm animals, mammalian sport animals, and mammalian pets.

[0047] 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.

[0048] 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.

[0049] 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.

[0050] 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 CasY polypeptide" includes a plurality of such polypeptides and reference to "the guide RNA" includes reference to one or more guide RNAs 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.

[0051] 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.

[0052] 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

[0053] The present disclosure provides RNA-guided endonuclease polypeptides, referred to herein as "CasY" polypeptides (also referred to as "CasY proteins"); nucleic acids encoding the CasY polypeptides; and modified host cells comprising the CasY polypeptides and/or nucleic acids encoding same. CasY polypeptides are useful in a variety of applications, which are provided.

[0054] The present disclosure provides guide RNAs (referred to herein as "CasY guide RNAs") that bind to and provide sequence specificity to the CasY proteins; nucleic acids encoding the CasY guide

RNAs; and modified host cells comprising the CasY guide RNAs and/or nucleic acids encoding same. CasY guide RNAs are useful in a variety of applications, which are provided.

[0055] The present disclosure provides methods of identifying a CRISPR RNA-guided endonuclease.

COMPOSITIONS

CRISPR/CASY PROTEINS AND GUIDE RNAS

[0056] A CRISPR/Cas endonuclease (e.g., a CasY protein) interacts with (binds to) a corresponding guide RNA (e.g., a CasY guide RNA) to form a ribonucleoprotein (RNP) complex that is targeted to a particular site in a target nucleic acid via base pairing between the guide RNA and a target sequence within the target nucleic acid molecule. A guide RNA includes a nucleotide sequence (a guide sequence) that is complementary to a sequence (the target site) of a target nucleic acid. Thus, a CasY protein forms a complex with a CasY guide RNA and the guide RNA provides sequence specificity to the RNP complex via the guide sequence. The CasY protein of the complex provides the site-specific activity. In other words, the CasY protein is guided to a target site (e.g., stabilized at a target site) within a target nucleic acid sequence (e.g. a chromosomal sequence or an extrachromosomal sequence, e.g., an episomal sequence, a minicircle sequence, a mitochondrial sequence, a chloroplast sequence, etc.) by virtue of its association with the guide RNA.

[0057] The present disclosure provides compositions comprising a CasY polypeptide (and/or a nucleic acid encoding the CasY polypeptide) (e.g., where the CasY polypeptide can be a naturally existing protein, a nickase CasY protein, a dCasY protein, a chimeric CasY protein, etc.). The present disclosure provides compositions comprising a CasY guide RNA (and/or a nucleic acid encoding the CasY guide RNA). The present disclosure provides compositions comprising (a) a CasY polypeptide (and/or a nucleic acid encoding the CasY polypeptide) (e.g., where the CasY polypeptide can be a naturally existing protein, a nickase CasY protein, a dCasY protein, a chimeric CasY protein, etc.) and (b) a CasY guide RNA (and/or a nucleic acid encoding the CasY guide RNA). The present disclosure provides a nucleic acid/protein complex (RNP complex) comprising: (a) a CasY polypeptide of the present disclosure (e.g., where the CasY polypeptide can be a naturally existing protein, a nickase CasY protein, a dCasY protein, a chimeric CasY protein, etc.); and (b) a CasY guide RNA.

CasY protein

[0058] A CasY polypeptide (this term is used interchangeably with the term "CasY protein") can bind and/or modify (e.g., cleave, nick, methylate, demethylate, etc.) a target nucleic acid and/or a polypeptide associated with target nucleic acid (e.g., methylation or acetylation of a histone tail) (e.g., in some cases the CasY protein includes a fusion partner with an activity, and in some cases the CasY protein provides nuclease activity). In some cases, the CasY protein is a naturally-occurring protein (e.g.,

naturally occurs in prokaryotic cells). In other cases, the CasY protein is not a naturally-occurring polypeptide (e.g., the CasY protein is a variant CasY protein, a chimeric protein, and the like).

[0059] Assays to determine whether given protein interacts with a CasY guide RNA can be any convenient binding assay that tests for binding between a protein and a nucleic acid. Suitable binding assays (e.g., gel shift assays) will be known to one of ordinary skill in the art (e.g., assays that include adding a CasY guide RNA and a protein to a target nucleic acid). Assays to determine whether a protein has an activity (e.g., to determine if the protein has nuclease activity that cleaves a target nucleic acid and/or some heterologous activity) can be any convenient assay (e.g., any convenient nucleic acid cleavage assay that tests for nucleic acid cleavage). Suitable assays (e.g., cleavage assays) will be known to one of ordinary skill in the art.

[0060] A naturally occurring CasY protein functions as an endonuclease that catalyzes a double strand break at a specific sequence in a targeted double stranded DNA (dsDNA). The sequence specificity is provided by the associated guide RNA, which hybridizes to a target sequence within the target DNA. The naturally occurring CasY guide RNA is a crRNA, where the crRNA includes (i) a guide sequence that hybridizes to a target sequence in the target DNA and (ii) a protein binding segment which includes a stem-loop (hairpin – dsRNA duplex) that binds to the CasY protein.

[0061] In some embodiments, the CasY protein of the subject methods and/or compositions is (or is derived from) a naturally occurring (wild type) protein. Examples of naturally occurring CasY proteins are depicted in Figure 1 and are set forth as SEQ ID NOs: 1-7. Examples of naturally occurring CasY proteins are depicted in Figure 1 and are set forth as SEQ ID NOs: 1-8. An alignment of example naturally occurring CasY proteins is presented in Figure 2 (the proteins are labeled as “Y1.”, “Y2.”, “Y3.” , etc.). Partial DNA scaffolds of 7 naturally occurring CasY CRISPR loci (assembled from sequencing data) are set forth as SEQ ID NOs: 21-27. It is important to note that this newly discovered protein (CasY) is short compared to previously identified CRISPR-Cas endonucleases, and thus use of this protein as an alternative provides the advantage that the nucleotide sequence encoding the protein is relatively short. This is useful, for example, in cases where a nucleic acid encoding the CasY protein is desirable, e.g., in situations that employ a viral vector (e.g., an AAV vector), for delivery to a cell such as a eukaryotic cell (e.g.., mammalian cell, human cell, mouse cell, in vitro, ex vivo, in vivo) for research and/or clinical applications. It is also noted herein that bacteria harboring CasY CRISPR loci were present in environmental samples that were collected at low temperature (e.g., 10-17°C). Thus, CasY is expected to be able to function well at low temperatures (e.g., 10-14°C, 10-17°C, 10-20°C) (e.g., better than other Cas endoconucleases discovered to date).

[0062] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or

more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 1. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 1, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0063] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 2. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 2. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 2. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 2. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 2. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 2, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0064] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 3. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 3. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 3. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 3. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 3. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 3, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0065] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 4. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 4. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 4. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 4. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 4. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 4, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0066] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 5. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 5. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 5. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 5. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 5. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 5, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0067] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 6. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 6. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 6. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 6. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 6. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 6, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0068] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 7. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 7. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 7. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 7. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 7. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 7, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0069] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 8. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 8. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 8. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 8. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 8. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 8, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0070] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 9. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 9. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 9. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the CasY protein sequence set forth as SEQ ID NO: 9. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 9. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth as SEQ ID NO: 9, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0071] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth in any one of SEQ ID NOs: 1-4. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth in any one of SEQ ID NOs: 1-4, with the exception that the sequence includes an amino acid

substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0072] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth in any one of SEQ ID NOs: 1-5. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth in any one of SEQ ID NOs: 1-5, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0073] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth in any one of SEQ ID NOs:

1-7. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth in any one of SEQ ID NOs: 1-7, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

[0074] In some cases, a CasY protein includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth in any one of SEQ ID NOs: 1-8. In some cases, a CasY protein includes an amino acid sequence having the CasY protein sequence set forth in any one of SEQ ID NOs: 1-8, with the exception that the sequence includes an amino acid substitution (e.g., 1, 2, or 3 amino acid substitutions) that reduces the naturally occurring catalytic activity of the protein (e.g., such as at amino acid positions described below).

CasY protein domains

[0075] The domains of a CasY protein are depicted in Figure 3. As can be seen in the schematic representation of Figure 3 (amino acids are numbered based on the CasY1 protein (SEQ ID NO: 1)), a CasY protein includes an N-terminal domain roughly 800-1000 amino acids in length (e.g., about 815 for CasY1 and about 980 for CasY5), and a C-terminal domain that includes 3 partial RuvC domains (RuvC-I, RuvC-II, and RuvC-III, also referred to herein as subdomains) that are not contiguous with respect to the primary amino acid sequence of the CasY protein, but form a RuvC domain once the protein is produced and folds. Thus, in some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence with an N-terminal domain (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids). In some cases, a

CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having a length (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids) that is N-termal to a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III).

[0076] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth as SEQ ID NO: 1. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having amino acids 1-812 of the CasY protein sequence set forth as SEQ ID NO: 1.

[0077] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-4. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel

a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-4. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-4. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-4. In some cases, a CasY protein includes a fragment of the amino acid sequence of any one of SEQ ID NOs: 1-4 that corresponds to amino acids 1-812 of the CasY protein sequence set forth as SEQ ID NO: 1.

[0078] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-5. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-5. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-5. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-5. In some cases, a CasY protein includes a fragment of the amino acid sequence of any one of SEQ ID NOs: 1-5 that corresponds to amino acids 1-812 of the CasY protein sequence set forth as SEQ ID NO: 1.

[0079] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more,

98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-7. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-7. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-7. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-7. In some cases, a CasY protein includes a fragment of the amino acid sequence of any one of SEQ ID NOs: 1-7 that corresponds to amino acids 1-812 of the CasY protein sequence set forth as SEQ ID NO: 1.

[0080] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-8. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-8. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of SEQ ID NOs: 1-8. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth in any one of

SEQ ID NOs: 1-8. In some cases, a CasY protein includes a fragment of the amino acid sequence of any one of SEQ ID NOs: 1-8 that corresponds to amino acids 1-812 of the CasY protein sequence set forth as SEQ ID NO: 1.

[0081] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). For example, in some cases, a CasY protein includes a first amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes a first amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes a first amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes an amino acid sequence corresponding to amino acids 1-812 of the CasY protein sequence set forth as SEQ ID NO: 1; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III).

In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more,

99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). For example, in some cases, a CasY protein includes a first amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes a first amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes a first amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes an amino acid sequence corresponding to amino acids 1-812 of the CasY protein sequence set forth as SEQ ID NO: 1; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III).

[0082] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). For example, in some cases, a CasY protein includes a first amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more,

90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes a first amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes a first amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes an amino acid sequence corresponding to amino acids 1-812 of the CasY protein sequence set forth as SEQ ID NO: 1; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III).

[0083] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). For example, in some cases, a CasY protein includes a first amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes a first amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or

more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes a first amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III). In some cases, a CasY protein includes an amino acid sequence corresponding to amino acids 1-812 of the CasY protein sequence set forth as SEQ ID NO: 1; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes a split Ruv C domain (e.g., 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III).

[0084] In some embodiments, the split RuvC domain of a CasY protein (of the subject compositions and/or methods) includes a region between the RuvC-II and RuvC-III subdomains that is larger than the RuvC-III subdomain. For example, in some cases, the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2). In some cases, the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1). In some cases, the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, or 1 and 1.2).

[0085] In some embodiments (for a CasY protein of the subject compositions and/or methods), the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less). For example, in some cases, the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less). In some embodiments, the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4).

[0086] In some cases (for a CasY protein of the subject compositions and/or methods), the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1. In some cases, the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.3 (e.g., 1 and 1.2).

[0087] In some cases (for a CasY protein of the subject compositions and/or methods), the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65, 68, or 70 amino acids in length). In some cases, the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids).

[0088] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0089] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 75% or more sequence identity (e.g., 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-

terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0090] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 85% or more sequence identity (e.g., 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the

ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0091] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains

over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0092] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 75% or more sequence identity (e.g., 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0093] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 85% or more sequence identity (e.g., 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0094] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2);

(ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0095] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 75% or more sequence identity (e.g., 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6,

1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0096] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 85% or more sequence identity (e.g., 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino

acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0097] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0098] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 75% or more sequence identity (e.g., 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and

RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[0099] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence having 85% or more sequence identity (e.g., 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the N-terminal domain (e.g., the domain depicted as amino acids 1-812 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8; and a second amino acid sequence, C-terminal to the first aminio acid sequence, that includes 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III – where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the

length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[00100] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence with an N-terminal domain (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence (C-terminal to the first) having a split Ruv C domain with 3 partial RuvC domains - RuvC-I, RuvC-II, and RuvC-III, where: (i) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is 1.1 or greater (e.g., 1.2); (ii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (iii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.4, 1 and 1.3, 1 and 1.2); (iv) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 2 or less (e.g., 1.8 or less, 1.7 or less, 1.6 or less, 1.5 or less, or 1.4 or less); (v) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is 1.5 or less (e.g., 1.4 or less); (vi) the ratio of the length of the RuvC-II subdomain over the length of the RuvC-III subdomain is in a range of from 1 to 2 (e.g., from 1.1 to 2, 1.2 to 2, 1 to 1.8, 1.1 to 1.8, 1.2 to 1.8, 1 to 1.6, 1.1 to 1.6, 1.2 to 1.6, 1 to 14, 1.1 to 1.4, or 1.2 to 1.4); (vii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1; (viii) the ratio of the length of the region between the RuvC-II and RuvC-III subdomains over the length of the RuvC-III subdomain is greater than 1and between 1 and 1.5 (e.g., 1 and 1.2); (ix) the region between the RuvC-II and RuvC-III subdomains is at least 60 amino acids in length (e.g., at least 65 or at least 70 amino acids in length); (x) the region between the RuvC-II and RuvC-III subdomains is at least 65 amino

acids in length; (xi) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 60-110 amino acids (e.g., a range of from 60-105, 60-100, 60-95, 60-90, 65-110, 65-105, 65-100, 65-95, or 65-90 amino acids); or (xii) the region between the RuvC-II and RuvC-III subdomains has a length in a range of from 65-95 amino acids.

[00101] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth as SEQ ID NO: 1. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of the CasY protein sequence set forth as SEQ ID NO: 1. In some cases, a CasY protein includes an amino acid sequence having amino acids 812-1125 of the CasY protein sequence set forth as SEQ ID NO: 1.

[00102] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or

more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes a fragment of the amino acid sequence of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4 that corresponds to amino acids 812-1125 of the CasY protein sequence set forth as SEQ ID NO: 1.

[00103] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes a fragment of the amino acid sequence of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5 that corresponds to amino acids 812-1125 of the CasY protein sequence set forth as SEQ ID NO: 1.

[00104] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein

sequences set forth as SEQ ID NOs: 1-7. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes a fragment of the amino acid sequence of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7 that corresponds to amino acids 812-1125 of the CasY protein sequence set forth as SEQ ID NO: 1.

[00105] In some cases, a CasY protein (of the subject compositions and/or methods) includes an amino acid sequence having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. For example, in some cases, a CasY protein includes an amino acid sequence having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes an amino acid sequence having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes an amino acid sequence having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes a fragment of the amino

acid sequence of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8 that corresponds to amino acids 812-1125 of the CasY protein sequence set forth as SEQ ID NO: 1.

[00106] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. For example, in some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 90% or

more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having a fragment of the amino acid sequence of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-4 that corresponds to amino acids 812-1125 of the CasY protein sequence set forth as SEQ ID NO: 1.

[00107] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. For example, in some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 80% or more sequence identity (e.g., 85%

or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having a fragment of the amino acid sequence of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-5 that corresponds to amino acids 812-1125 of the CasY protein sequence set forth as SEQ ID NO: 1.

[00108] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. For example, in some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95%

or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having a fragment of the amino acid sequence of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-7 that corresponds to amino acids 812-1125 of the CasY protein sequence set forth as SEQ ID NO: 1.

[00109] In some cases, a CasY protein (of the subject compositions and/or methods) includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 20% or more sequence identity (e.g., 30% or more, 40% or more, 50% or more, 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with

the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. For example, in some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 50% or more sequence identity (e.g., 60% or more, 70% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 80% or more sequence identity (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having 90% or more sequence identity (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100% sequence identity) with the C-terminal domain (e.g., the domain depicted as amino acids 812-1125 for CasY1 in Figure 3, panel a) of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8. In some cases, a CasY protein includes a first amino acid sequence (an N-terminal domain) (e.g., not including any fused heterologous sequence such as an NLS and/or a domain with a catalytic activity) having a length in a range of from 750 to 1050 amino acids (e.g, from 750 to 1025, 750 to 1000, 750 to 950, 775 to 1050, 775 to 1025, 775 to 1000, 775 to 950, 800 to 1050, 800 to 1025, 800 to 1000, or 800 to 950 amino acids); and a second amino acid sequence, positioned C-terminal to the first amino acid sequence, having a fragment of the amino acid sequence of any one of the CasY protein sequences set forth as SEQ ID NOs: 1-8 that corresponds to amino acids 812-1125 of the CasY protein sequence set forth as SEQ ID NO: 1.

CasY Variants

[00110] A variant CasY protein has an amino acid sequence that is different by at least one amino acid (e.g., has a deletion, insertion, substitution, fusion) when compared to the amino acid sequence of the corresponding wild type CasY protein. A CasY protein that cleaves one strand but not the other of a double stranded target nucleic acid is referred to herein as a "nickase" (e.g., a "nickase CasY"). A CasY protein that has substantially no nuclease activity is referred to herein as a dead CasY protein ("dCasY") (with the caveat that nuclease activity can be provided by a heterologous polypeptide – a fusion partner – in the case of a chimeric CasY protein, which is described in more detail below). For any of the CasY variant proteins described herein (e.g., nickase CasY, dCasY, chimeric CasY), the CasY variant can include a CasY protein sequence with the same parameters described above (e.g., domains that are present, percent identity, and the like).

Variants – catalytic activity

[00111] In some cases, the CasY protein is a variant CasY protein, e.g., mutated relative to the naturally occurring catalytically active sequence, and exhibits reduced cleavage activity (e.g., exhibits 90%, or less, 80% or less, 70% or less, 60% or less, 50% or less, 40% or less, or 30% or less cleavage activity) when compared to the corresponding naturally occurring sequence. In some cases, such a variant CasY protein is a catalytically 'dead' protein (has substantially no cleavage activity) and can be referred to as a 'dCasY.' In some cases, the variant CasY protein is a nickase (cleaves only one strand of a double stranded target nucleic acid, e.g., a double stranded target DNA). As described in more detail herein, in some cases, a CasY protein (in some case a CasY protein with wild type cleavage activity and in some cases a variant CasY with reduced cleavage activity, e.g., a dCasY or a nickase CasY) is fused (conjugated) to a heterologous polypeptide that has an activity of interest (e.g., a catalytic activity of interest) to form a fusion protein (a chimeric CasY protein).

[00112] Catalytic residues of CasY include D828, E914, D1074 when numbered according to CasY1 (SEQ ID NO: 1) (these residues are underlined in Figure 1 for SEQ ID NO: 1). (see, e.g., the alignments of Figure 2, panels a and b).

[00113] Thus, in some cases, the CasY protein has reduced activity and one or more of the above described amino acids (or one or more corresponding amino acids of any CasY protein) are mutated (e.g., substituted with an alanine). In some cases, the variant CasY protein is a catalytically 'dead' protein (is catalytically inactive) and is referred to as 'dCasY.' A dCasY protein can be fused to a fusion partner that provides an activity, and in some cases, the dCasY (e.g., one without a fusion partner that provides catalytic activity – but which can have an NLS when expressed in a eukaryotic cell) can bind to target DNA and can block RNA polymerase from translating from a target DNA. In some cases, the variant

CasY protein is a nickase (cleaves only one strand of a double stranded target nucleic acid, e.g., a double stranded target DNA).

Variants – chimeric CasY (i.e., fusion proteins)

[00114] As noted above, in some cases, a CasY protein (in some cases a CasY protein with wild type cleavage activity and in some cases a variant CasY with reduced cleavage activity, e.g., a dCasY or a nickase CasY) is fused (conjugated) to a heterologous polypeptide that has an activity of interest (e.g., a catalytic activity of interest) to form a fusion protein (a chimeric CasY protein). A heterologous polypeptide to which a CasY protein can be fused is referred to herein as a 'fusion partner.'

[00115] In some cases the fusion partner can modulate transcription (e.g., inhibit transcription, increase transcription) of a target DNA. For example, in some cases the fusion partner is a protein (or a domain from a protein) that inhibits transcription (e.g., a transcriptional repressor, a protein that functions via recruitment of transcription inhibitor proteins, modification of target DNA such as methylation, recruitment of a DNA modifier, modulation of histones associated with target DNA, recruitment of a histone modifier such as those that modify acetylation and/or methylation of histones, and the like). In some cases the fusion partner is a protein (or a domain from a protein) that increases transcription (e.g., a transcription activator, a protein that acts via recruitment of transcription activator proteins, modification of target DNA such as demethylation, recruitment of a DNA modifier, modulation of histones associated with target DNA, recruitment of a histone modifier such as those that modify acetylation and/or methylation of histones, and the like).

[00116] In some cases, a chimeric CasY protein includes a heterologous polypeptide that has enzymatic activity that modifies a target nucleic acid (e.g., nuclease activity, methyltransferase activity, demethylase activity, DNA repair activity, DNA damage activity, deamination activity, dismutase activity, alkylation activity, depurination activity, oxidation activity, pyrimidine dimer forming activity, integrase activity, transposase activity, recombinase activity, polymerase activity, ligase activity, helicase activity, photolyase activity or glycosylase activity).

[00117] In some cases, a chimeric CasY protein includes a heterologous polypeptide that has enzymatic activity that modifies a polypeptide (e.g., a histone) associated with a target nucleic acid (e.g., methyltransferase activity, demethylase activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity or demyristoylation activity).

[00118] Examples of proteins (or fragments thereof) that can be used in increase transcription include but are not limited to: transcriptional activators such as VP16, VP64, VP48, VP160, p65 subdomain (e.g., from NFkB), and activation domain of EDLL and/or TAL acitvation domain (e.g., for

activity in plants); histone lysine methyltransferases such as SET1A, SET1B, MLL1 to 5, ASH1, SYMD2, NSD1, and the like; histone lysine demethylases such as JHDM2a/b, UTX, JMJD3, and the like; histone acetyltransferases such as GCN5, PCAF, CBP, p300, TAF1, TIP60/PLIP, MOZ/MYST3, MORF/MYST4, SRC1, ACTR, P160, CLOCK, and the like; and DNA demethylases such as Ten-Eleven Translocation (TET) dioxygenase 1 (TET1CD), TET1, DME, DML1, DML2, ROS1, and the like.

[00119] Examples of proteins (or fragments thereof) that can be used in decrease transcription include but are not limited to: transcriptional repressors such as the Krüppel associated box (KRAB or SKD); KOX1 repression domain; the Mad mSIN3 interaction domain (SID); the ERF repressor domain (ERD), the SRDX repression domain (e.g., for repression in plants), and the like; histone lysine methyltransferases such as Pr-SET7/8, SUV4-20H1, RIZ1, and the like; histone lysine demethylases such as JMJD2A/JHDM3A, JMJD2B, JMJD2C/GASC1, JMJD2D, JARID1A/RBP2, JARID1B/PLU-1, JARID1C/SMCX, JARID1D/SMCY, and the like; histone lysine deacetylases such as HDAC1, HDAC2, HDAC3, HDAC8, HDAC4, HDAC5, HDAC7, HDAC9, SIRT1, SIRT2, HDAC11, and the like; DNA methylases such as HhaI DNA m5c-methyltransferase (M.HhaI), DNA methyltransferase 1 (DNMT1), DNA methyltransferase 3a (DNMT3a), DNA methyltransferase 3b (DNMT3b), METI, DRM3 (plants), ZMET2, CMT1, CMT2 (plants), and the like; and periphery recruitment elements such as Lamin A, Lamin B, and the like.

[00120] In some cases the fusion partner has enzymatic activity that modifies the target nucleic acid (e.g., ssRNA, dsRNA, ssDNA, dsDNA). Examples of enzymatic activity that can be provided by the fusion partner include but are not limited to: nuclease activity such as that provided by a restriction enzyme (e.g., FokI nuclease), methyltransferase activity such as that provided by a methyltransferase (e.g., HhaI DNA m5c-methyltransferase (M.HhaI), DNA methyltransferase 1 (DNMT1), DNA methyltransferase 3a (DNMT3a), DNA methyltransferase 3b (DNMT3b), METI, DRM3 (plants), ZMET2, CMT1, CMT2 (plants), and the like); demethylase activity such as that provided by a demethylase (e.g., Ten-Eleven Translocation (TET) dioxygenase 1 (TET1CD), TET1, DME, DML1, DML2, ROS1, and the like) , DNA repair activity, DNA damage activity, deamination activity such as that provided by a deaminase (e.g., a cytosine deaminase enzyme such as rat APOBEC1), dismutase activity, alkylation activity, depurination activity, oxidation activity, pyrimidine dimer forming activity, integrase activity such as that provided by an integrase and/or resolvase (e.g., Gin invertase such as the hyperactive mutant of the Gin invertase, GinH106Y; human immunodeficiency virus type 1 integrase (IN); Tn3 resolvase; and the like), transposase activity, recombinase activity such as that provided by a recombinase (e.g., catalytic domain of Gin recombinase), polymerase activity, ligase activity, helicase activity, photolyase activity, and glycosylase activity).

[00121] In some cases the fusion partner has enzymatic activity that modifies a protein associated with the target nucleic acid (e.g., ssRNA, dsRNA, ssDNA, dsDNA) (e.g., a histone, an RNA binding protein, a DNA binding protein, and the like). Examples of enzymatic activity (that modifyies a protein associated with a target nucleic acid) that can be provided by the fusion partner include but are not limited to: methyltransferase activity such as that provided by a histone methyltransferase (HMT) (e.g., suppressor of variegation 3-9 homolog 1 (SUV39H1, also known as KMT1A), euchromatic histone lysine methyltransferase 2 (G9A, also known as KMT1C and EHMT2), SUV39H2, ESET/SETDB1, and the like, SET1A, SET1B, MLL1 to 5, ASH1, SYMD2, NSD1, DOT1L, Pr-SET7/8, SUV4-20H1, EZH2, RIZ1), demethylase activity such as that provided by a histone demethylase (e.g., Lysine Demethylase 1A (KDM1A also known as LSD1), JHDM2a/b, JMJD2A/JHDM3A, JMJD2B, JMJD2C/GASC1, JMJD2D, JARID1A/RBP2, JARID1B/PLU-1, JARID1C/SMCX, JARID1D/SMCY, UTX, JMJD3, and the like), acetyltransferase activity such as that provided by a histone acetylase transferase (e.g., catalytic core/fragement of the human acetyltransferase p300, GCN5, PCAF, CBP, TAF1, TIP60/PLIP, MOZ/MYST3, MORF/MYST4, HBO1/MYST2, HMOF/MYST1, SRC1, ACTR, P160, CLOCK, and the like), deacetylase activity such as that provided by a histone deacetylase (e.g., HDAC1, HDAC2, HDAC3, HDAC8, HDAC4, HDAC5, HDAC7, HDAC9, SIRT1, SIRT2, HDAC11, and the like), kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity, and demyristoylation activity.

[00122] An additional examples of a suitable fusion partners are dihydrofolate reductase (DHFR) destabilization domain (e.g., to generate a chemically controllable chimeric CasY protein), and a chloroplast transit peptide. Suitable chloroplast transit peptides include, but are not limited to:

[00123] MASMISSSAVTTVSRASRGQSAAMAPFGGLKSMTGFPVRKVNTDITSITSNGGRV KCMQVWPPIGKKKFETLSYLPPLTRDSRA (SEQ ID NO:83);
MASMISSSAVTTVSRASRGQSAAMAPFGGLKSMTGFPVRKVNTDITSITSNGGRVKS (SEQ ID NO:84);
MASSMLSSATMVASPAQATMVAPFNGLKSSAAFPATRKANNDITSITSNGGRVNCMQVWPPIEK KKFETLSYLPDLTDSGGRVNC (SEQ ID NO:85);
MAQVSRICNGVQNPSLISNLSKSSQRKSPLSVSLKTQQHPRAYPISSSWGLKKSGMTLIGSELRPL KVMSSVSTAC (SEQ ID NO:86);
MAQVSRICNGVWNPSLISNLSKSSQRKSPLSVSLKTQQHPRAYPISSSWGLKKSGMTLIGSELRPL KVMSSVSTAC (SEQ ID NO:87);
MAQINNMAQGIQTLNPNSNFHKPQVPKSSSFLVFGSKKLKNSANSMLVLKKDSIFMQLFCSFRIS ASVATAC (SEQ ID NO:88);

MAALVTSQLATSGTVLSVTDRFRRPGFQGLRPRNPADAALGMRTVGASAAPKQSRKPHRFDRR CLSMVV (SEQ ID NO:89);
MAALTTSQLATSATGFGIADRSAPSSLLRHGFQGLKPRSPAGGDATSLSVTTSARATPKQQRSVQ RGSRRFPSVVVC (SEQ ID NO:90);
MASSVLSSAAVATRSNVAQANMVAPFTGLKSAASFPVSRKQNLDITSIASNGGRVQC (SEQ ID NO:91);
MESLAATSVFAPSRVAVPAARALVRAGTVVPTRRTSSTSGTSGVKCSAAVTPQASPVISRSAAAA (SEQ ID NO:92); and
MGAAATSMQSLKFSNRLVPPSRRLSPVPNNVTCNNLPKSAAPVRTVKCCASSWNSTINGAAATT NGASAASS (SEQ ID NO:93).

[00124] In some case, a CasY fusion polypeptide of the present disclosure comprises: a) a CasY polypeptide of the present disclosure; and b) a chloroplast transit peptide. Thus, for example, a CRISPR-CasY complex can be targeted to the chloroplast. In some cases, this targeting may be achieved by the presence of an N-terminal extension, called a chloroplast transit peptide (CTP) or plastid transit peptide. Chromosomal transgenes from bacterial sources must have a sequence encoding a CTP sequence fused to a sequence encoding an expressed polypeptide if the expressed polypeptide is to be compartmentalized in the plant plastid (e.g. chloroplast). Accordingly, localization of an exogenous polypeptide to a chloroplast is often 1 accomplished by means of operably linking a polynucleotide sequence encoding a CTP sequence to the 5' region of a polynucleotide encoding the exogenous polypeptide. The CTP is removed in a processing step during translocation into the plastid. Processing efficiency may, however, be affected by the amino acid sequence of the CTP and nearby sequences at the NH 2 terminus of the peptide. Other options for targeting to the chloroplast which have been described are the maize cab-m7 signal sequence (U.S. Pat. No. 7,022,896, WO 97/41228) a pea glutathione reductase signal sequence (WO 97/41228) and the CTP described in US2009029861.

[00125] In some cases, a CasY fusion polypeptide of the present disclosure can comprise: a) a CasY polypeptide of the present disclosure; and b) an endosomal escape peptide. In some cases, an endosomal escape polypeptide comprises the amino acid sequence GLFXALLXLLXSLWXLLLXA (SEQ ID NO:94), wherein each X is independently selected from lysine, histidine, and arginine. In some cases, an endosomal escape polypeptide comprises the amino acid sequence GLFHALLHLLHSLWHLLLHA (SEQ ID NO:95).

[00126] For examples of some of the above fusion partners (and more) used in the context of fusions with Cas9, Zinc Finger, and/or TALE proteins (for site specific target nucleic modification, modulation of transcription, and/or target protein modification, e.g., histone modification), see, e.g.: Nomura et al, J Am Chem Soc. 2007 Jul 18;129(28):8676-7; Rivenbark et al., Epigenetics. 2012

Apr;7(4):350-60; Nucleic Acids Res. 2016 Jul 8;44(12):5615-28; Gilbert et al., Cell. 2013 Jul 18;154(2):442-51; Kearns et al., Nat Methods. 2015 May;12(5):401-3; Mendenhall et al., Nat Biotechnol. 2013 Dec;31(12):1133-6; Hilton et al., Nat Biotechnol. 2015 May;33(5):510-7; Gordley et al., Proc Natl Acad Sci U S A. 2009 Mar 31;106(13):5053-8; Akopian et al., Proc Natl Acad Sci U S A. 2003 Jul 22;100(15):8688-91; Tan et., al., J Virol. 2006 Feb;80(4):1939-48; Tan et al., Proc Natl Acad Sci U S A. 2003 Oct 14;100(21):11997-2002; Papworth et al., Proc Natl Acad Sci U S A. 2003 Feb 18;100(4):1621-6; Sanjana et al., Nat Protoc. 2012 Jan 5;7(1):171-92; Beerli et al., Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14628-33; Snowden et al., Curr Biol. 2002 Dec 23;12(24):2159-66; Xu et.al., Xu et al., Cell Discov. 2016 May 3;2:16009; Komor et al., Nature. 2016 Apr 20;533(7603):420-4; Chaikind et al., Nucleic Acids Res. 2016 Aug 11; Choudhury at. al., Oncotarget. 2016 Jun 23; Du et al., Cold Spring Harb Protoc. 2016 Jan 4; Pham et al., Methods Mol Biol. 2016;1358:43-57; Balboa et al., Stem Cell Reports. 2015 Sep 8;5(3):448-59; Hara et al., Sci Rep. 2015 Jun 9;5:11221; Piatek et al., Plant Biotechnol J. 2015 May;13(4):578-89; Hu et al., Nucleic Acids Res. 2014 Apr;42(7):4375-90; Cheng et al., Cell Res. 2013 Oct;23(10):1163-71; and Maeder et al., Nat Methods. 2013 Oct;10(10):977-9.

[00127] Additional suitable heterologous polypeptide include, but are not limited to, a polypeptide that directly and/or indirectly provides for increased transcription and/or translation of a target nucleic acid (e.g., a transcription activator or a fragment thereof, a protein or fragment thereof that recruits a transcription activator, a small molecule/drug-responsive transcription and/or translation regulator, a translation-regulating protein, etc.). Non-limiting examples of heterologous polypeptides to accomplish increased or decreased transcription include transcription activator and transcription repressor domains. In some such cases, a chimeric CasY polypeptide is targeted by the guide nucleic acid (guide RNA) to a specific location (i.e., sequence) in the target nucleic acid and exerts locus-specific regulation such as blocking RNA polymerase binding to a promoter (which selectively inhibits transcription activator function), and/or modifying the local chromatin status (e.g., when a fusion sequence is used that modifies the target nucleic acid or modifies a polypeptide associated with the target nucleic acid). In some cases, the changes are transient (e.g., transcription repression or activation). In some cases, the changes are inheritable (e.g., when epigenetic modifications are made to the target nucleic acid or to proteins associated with the target nucleic acid, e.g., nucleosomal histones).

[00128] Non-limiting examples of heterologous polypeptides for use when targeting ssRNA target nucleic acids include (but are not limited to): splicing factors (e.g., RS domains); protein translation components (e.g., translation initiation, elongation, and/or release factors; e.g., eIF4G); RNA methylases; RNA editing enzymes (e.g., RNA deaminases, e.g., adenosine deaminase acting on RNA (ADAR), including A to I and/or C to U editing enzymes); helicases; RNA-binding proteins; and the like. It is

understood that a heterologous polypeptide can include the entire protein or in some cases can include a fragment of the protein (e.g., a functional domain).

[00129] The heterologous polypeptide of a subject chimeric CasY polypeptide can be any domain capable of interacting with ssRNA (which, for the purposes of this disclosure, includes intramolecular and/or intermolecular secondary structures, e.g., double-stranded RNA duplexes such as hairpins, stem-loops, etc.), whether transiently or irreversibly, directly or indirectly, including but not limited to an effector domain selected from the group comprising; Endonucleases (for example RNase III, the CRR22 DYW domain, Dicer, and PIN (PilT N-terminus) domains from proteins such as SMG5 and SMG6); proteins and protein domains responsible for stimulating RNA cleavage (for example CPSF, CstF, CFIm and CFIIm); Exonucleases (for example XRN-1 or Exonuclease T) ; Deadenylases (for example HNT3); proteins and protein domains responsible for nonsense mediated RNA decay (for example UPF1, UPF2, UPF3, UPF3b, RNP S1, Y14, DEK, REF2, and SRm160); proteins and protein domains responsible for stabilizing RNA (for example PABP) ; proteins and protein domains responsible for repressing translation (for example Ago2 and Ago4); proteins and protein domains responsible for stimulating translation (for example Staufen); proteins and protein domains responsible for (e.g., capable of) modulating translation (e.g., translation factors such as initiation factors, elongation factors, release factors, etc., e.g., eIF4G); proteins and protein domains responsible for polyadenylation of RNA (for example PAP1, GLD-2, and Star- PAP) ; proteins and protein domains responsible for polyuridinylation of RNA (for example CI D1 and terminal uridylate transferase) ; proteins and protein domains responsible for RNA localization (for example from IMP1, ZBP1, She2p, She3p, and Bicaudal-D); proteins and protein domains responsible for nuclear retention of RNA (for example Rrp6); proteins and protein domains responsible for nuclear export of RNA (for example TAP, NXF1, THO, TREX, REF, and Aly) ; proteins and protein domains responsible for repression of RNA splicing (for example PTB, Sam68, and hnRNP A1) ; proteins and protein domains responsible for stimulation of RNA splicing (for example Serine/Arginine-rich (SR) domains) ; proteins and protein domains responsible for reducing the efficiency of transcription (for example FUS (TLS)); and proteins and protein domains responsible for stimulating transcription (for example CDK7 and HIV Tat). Alternatively, the effector domain may be selected from the group comprising Endonucleases; proteins and protein domains capable of stimulating RNA cleavage; Exonucleases; Deadenylases; proteins and protein domains having nonsense mediated RNA decay activity; proteins and protein domains capable of stabilizing RNA; proteins and protein domains capable of repressing translation; proteins and protein domains capable of stimulating translation; proteins and protein domains capable of modulating translation (e.g., translation factors such as initiation factors, elongation factors, release factors, etc., e.g., eIF4G); proteins and protein domains capable of polyadenylation of RNA; proteins and protein domains capable of polyuridinylation of RNA; proteins and

protein domains having RNA localization activity; proteins and protein domains capable of nuclear retention of RNA; proteins and protein domains having RNA nuclear export activity; proteins and protein domains capable of repression of RNA splicing; proteins and protein domains capable of stimulation of RNA splicing; proteins and protein domains capable of reducing the efficiency of transcription ; and proteins and protein domains capable of stimulating transcription. Another suitable heterologous polypeptide is a PUF RNA-binding domain, which is described in more detail in WO2012068627, which is hereby incorporated by reference in its entirety.

[00130] Some RNA splicing factors that can be used (in whole or as fragments thereof) as heterologous polypeptides for a chimeric CasY polypeptide have modular organization, with separate sequence-specific RNA binding modules and splicing effector domains. For example, members of the Serine/ Arginine-rich (SR) protein family contain N-terminal RNA recognition motifs (RRMs) that bind to exonic splicing enhancers (ESEs) in pre-mRNAs and C-terminal RS domains that promote exon inclusion. As another example, the hnRNP protein hnRNP Al binds to exonic splicing silencers (ESSs) through its RRM domains and inhibits exon inclusion through a C-terminal Glycine-rich domain. Some splicing factors can regulate alternative use of splice site (ss) by binding to regulatory sequences between the two alternative sites. For example, ASF/SF2 can recognize ESEs and promote the use of intron proximal sites, whereas hnRNP Al can bind to ESSs and shift splicing towards the use of intron distal sites. One application for such factors is to generate ESFs that modulate alternative splicing of endogenous genes, particularly disease associated genes. For example, Bcl-x pre-mRNA produces two splicing isoforms with two alternative 5' splice sites to encode proteins of opposite functions. The long splicing isoform Bcl-xL is a potent apoptosis inhibitor expressed in long-lived postmitotic cells and is up-regulated in many cancer cells, protecting cells against apoptotic signals. The short isoform Bcl-xS is a pro-apoptotic isoform and expressed at high levels in cells with a high turnover rate (e.g., developing lymphocytes). The ratio of the two Bcl-x splicing isoforms is regulated by multiple cώ-elements that are located in either the core exon region or the exon extension region (i.e., between the two alternative 5' splice sites). For more examples, see WO2010075303, which is hereby incorporated by reference in its entirety.

[00131] Further suitable fusion partners include, but are not limited to proteins (or fragments thereof) that are boundary elements (e.g., CTCF), proteins and fragments thereof that provide periphery recruitment (e.g., Lamin A, Lamin B, etc.), protein docking elements (e.g., FKBP/FRB, Pil1/Aby1, etc.).

[00132] Examples of various additional suitable heterologous polypeptide (or fragments thereof) for a subject chimeric CasY polypeptide include, but are not limited to those described in the following applications (which publications are related to other CRISPR endonucleases such as Cas9, but the described fusion partners can also be used with CasY instead): PCT patent applications:

WO2010075303, WO2012068627, and WO2013155555, and can be found, for example, in U.S. patents and patent applications: 8,906,616; 8,895,308; 8,889,418; 8,889,356; 8,871,445; 8,865,406; 8,795,965; 8,771,945; 8,697,359; 20140068797; 20140170753; 20140179006; 20140179770; 20140186843; 20140186919; 20140186958; 20140189896; 20140227787; 20140234972; 20140242664; 20140242699; 20140242700; 20140242702; 20140248702; 20140256046; 20140273037; 20140273226; 20140273230; 20140273231; 20140273232; 20140273233; 20140273234; 20140273235; 20140287938; 20140295556; 20140295557; 20140298547; 20140304853; 20140309487; 20140310828; 20140310830; 20140315985; 20140335063; 20140335620; 20140342456; 20140342457; 20140342458; 20140349400; 20140349405; 20140356867; 20140356956; 20140356958; 20140356959; 20140357523; 20140357530; 20140364333; and 20140377868; all of which are hereby incorporated by reference in their entirety.

[00133] In some cases, a heterologous polypeptide (a fusion partner) provides for subcellular localization, i.e., the heterologous polypeptide contains a subcellular localization sequence (e.g., a nuclear localization signal (NLS) for targeting to the nucleus, a sequence to keep the fusion protein out of the nucleus, e.g., a nuclear export sequence (NES), a sequence to keep the fusion protein retained in the cytoplasm, 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 embodiments, a CasY fusion polypeptide does not include a NLS so that the protein is not targeted to the nucleus (which can be advantageous, e.g., when the target nucleic acid is an RNA that is present in the cyosol). In some embodiments, the heterologous polypeptide can provide a tag (i.e., the heterologous polypeptide is a detectable label) for ease of tracking and/or purification (e.g., a fluorescent protein, e.g., green fluorescent protein (GFP), YFP, RFP, CFP, mCherry, tdTomato, and the like; a histidine tag, e.g., a 6XHis tag; a hemagglutinin (HA) tag; a FLAG tag; a Myc tag; and the like).

[00134] In some cases a CasY protein (e.g., a wild type CasY protein, a variant CasY protein, a chimeric CasY protein, a dCasY protein, a chimeric CasY protein where the CasY portion has reduced nuclease activity - such as a dCasY protein fused to a fusion partner, and the like) includes (is fused to) a nuclear localization signal (NLS) (e.g, in some cases 2 or more, 3 or more, 4 or more, or 5 or more NLSs). Thus, in some cases, a CasY polypeptide includes one or more NLSs (e.g., 2 or more, 3 or more, 4 or more, or 5 or more NLSs). In some cases, one or more NLSs (2 or more, 3 or more, 4 or more, or 5 or more NLSs) are positioned at or near (e.g., within 50 amino acids of) the N-terminus and/or the C-terminus. In some cases, one or more NLSs (2 or more, 3 or more, 4 or more, or 5 or more NLSs) are positioned at or near (e.g., within 50 amino acids of) the N-terminus. In some cases, one or more NLSs (2 or more, 3 or more, 4 or more, or 5 or more NLSs) are positioned at or near (e.g., within 50 amino acids of) the C-terminus. In some cases, one or more NLSs (3 or more, 4 or more, or 5 or more NLSs) are

positioned at or near (e.g., within 50 amino acids of) both the N-terminus and the C-terminus. In some cases, an NLS is positioned at the N-terminus and an NLS is positioned at the C-terminus.

[00135] In some cases a CasY protein (e.g., a wild type CasY protein, a variant CasY protein, a chimeric CasY protein, a dCasY protein, a chimeric CasY protein where the CasY portion has reduced nuclease activity - such as a dCasY protein fused to a fusion partner, and the like) includes (is fused to) between 1 and 10 NLSs (e.g., 1-9, 1-8, 1-7, 1-6, 1-5, 2-10, 2-9, 2-8, 2-7, 2-6, or 2-5 NLSs). In some cases a CasY protein (e.g., a wild type CasY protein, a variant CasY protein, a chimeric CasY protein, a dCasY protein, a chimeric CasY protein where the CasY portion has reduced nuclease activity - such as a dCasY protein fused to a fusion partner, and the like) includes (is fused to) between 2 and 5 NLSs (e.g., 2-4, or 2-3 NLSs).

[00136] Non-limiting examples of NLSs include an NLS sequence derived from: the NLS of the SV40 virus large T-antigen, having the amino acid sequence PKKKRKV (SEQ ID NO: 96); the NLS from nucleoplasmin (e.g., the nucleoplasmin bipartite NLS with the sequence KRPAATKKAGQAKKKK (SEQ ID NO: 97)); the c-myc NLS having the amino acid sequence PAAKRVKLD (SEQ ID NO: 98) or RQRRNELKRSP (SEQ ID NO: 99); the hRNPA1 M9 NLS having the sequence NQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGY (SEQ ID NO: 100); the sequence RMRIZFKNKGKDTAELRRRRVEVSVELRKAKKDEQILKRRNV (SEQ ID NO: 101) of the IBB domain from importin-alpha; the sequences VSRKRPRP (SEQ ID NO: 102) and PPKKARED (SEQ ID NO: 103) of the myoma T protein; the sequence PQPKKKPL (SEQ ID NO: 104) of human p53; the sequence SALIKKKKKMAP (SEQ ID NO: 105) of mouse c-abl IV; the sequences DRLRR (SEQ ID NO: 106) and PKQKKRK (SEQ ID NO: 107) of the influenza virus NS1; the sequence RKLKKKIKKL (SEQ ID NO: 108) of the Hepatitis virus delta antigen; the sequence REKKKFLKRR (SEQ ID NO: 109) of the mouse Mx1 protein; the sequence KRKGDEVDGVDEVAKKKSKK (SEQ ID NO: 110) of the human poly(ADP-ribose) polymerase; and the sequence RKCLQAGMNLEARKTKK (SEQ ID NO: 111) of the steroid hormone receptors (human) glucocorticoid. In general, NLS (or multiple NLSs) are of sufficient strength to drive accumulation of the CasY protein in a detectable amount in the nucleus of a eukaryotic cell. Detection of accumulation in the nucleus may be performed by any suitable technique. For example, a detectable marker may be fused to the CasY protein such that location within a cell may be visualized. Cell nuclei may also be isolated from cells, the contents of which may then be analyzed by any suitable process for detecting protein, such as immunohistochemistry, Western blot, or enzyme activity assay. Accumulation in the nucleus may also be determined indirectly.

[00137] In some cases, a CasY fusion polypeptide includes a "Protein Transduction Domain" or PTD (also known as a CPP – cell penetrating peptide), which refers 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 a polypeptide (e.g., linked to a wild type CasY to generate a fusino protein, or linked to a variant CasY protein such as a dCasY, nickase CasY, or chimeric CasY protein to generate a fusion protein). In some embodiments, a PTD is covalently linked to the carboxyl terminus of a polypeptide (e.g., linked to a wild type CasY to generate a fusino protein, or linked to a variant CasY protein such as a dCasY, nickase CasY, or chimeric CasY protein to generate a fusion protein). In some cases, the PTD is inserted interally in the CasY fusion polypeptide (i.e., is not at the N- or C-terminus of the CasY fusion polypeptide) at a suitable insertion site. In some cases, a subject CasY fusion polypeptide includes (is conjugated to, is fused to) one or more PTDs (e.g., two or more, three or more, four or more PTDs). 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). Thus, in some cases, a CasY fusion polypeptide includes one or more NLSs (e.g., 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 CasY guide nucleic acid, a polynucleotide encoding a CasY guide nucleic acid, a polynucleotide encoding a CasY fusion polypeptide, a donor polynucleotide, etc.). Examples of 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:112); 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:113); Transportan GWTLNSAGYLLGKINLKALAALAKKIL (SEQ ID NO:114); KALAWEAKLAKALAKALAKHLAKALAKALKCEA (SEQ ID NO:115); and RQIKIWFQNRRMKWKK (SEQ ID NO:116). Exemplary PTDs include but are not limited to, YGRKKRRQRRR (SEQ ID NO:117), RKKRRQRRR (SEQ ID NO:118); 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:119); RKKRRQRR (SEQ ID NO:120); YARAAARQARA (SEQ ID NO:121); THRLPRRRRRR (SEQ ID NO:122); and GGRRARRRRRR (SEQ ID NO:123). 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.

Linkers (e.g., for fusion partners)

[00138] In some embodiments, a subject CasY protein can fused to a fusion partner via a linker polypeptide (e.g., one or more linker polypeptides). The linker polypeptide may have any of a variety of amino acid sequences. Proteins can be joined by a spacer peptide, generally of a flexible nature, although other chemical linkages are not excluded. Suitable linkers include polypeptides of between 4 amino acids and 40 amino acids in length, or between 4 amino acids and 25 amino acids in length. These linkers can be produced by using synthetic, linker-encoding oligonucleotides to couple the proteins, or can be encoded by a nucleic acid sequence encoding the fusion protein. Peptide linkers with a degree of flexibility can be used. The linking peptides may have virtually any amino acid sequence, bearing in mind that the preferred linkers will have a sequence that results in a generally flexible peptide. The use of small amino acids, such as glycine and alanine, are of use in creating a flexible peptide. The creation of such sequences is routine to those of skill in the art. A variety of different linkers are commercially available and are considered suitable for use.

[00139] Examples of linker polypeptides include glycine polymers $(G)_n$, glycine-serine polymers (including, for example, $(GS)_n$, $GSGGS_n$ (SEQ ID NO: 124), $GGSGGS_n$ (SEQ ID NO: 125), and $GGGS_n$ (SEQ ID NO: 126), where n is an integer of at least one), glycine-alanine polymers, alanine-serine polymers. Exemplary linkers can comprise amino acid sequences including, but not limited to, GGSG (SEQ ID NO: 127), GGSGG (SEQ ID NO: 128), GSGSG (SEQ ID NO: 129), GSGGG (SEQ ID NO: 130), GGGSG (SEQ ID NO: 131), GSSSG (SEQ ID NO: 132), and the like. The ordinarily skilled artisan will recognize that design of a peptide conjugated to any desired element can include linkers that are all or partially flexible, such that the linker can include a flexible linker as well as one or more portions that confer less flexible structure.

Detectable labels

[00140] In some cases, a CasY polypeptide of the present disclosure comprises a detectable label. Suitable detectable labels and/or moieties that can provide a detectable signal can include, but are not limited to, an enzyme, a radioisotope, a member of a specific binding pair; a fluorophore; a fluorescent protein; a quantum dot; and the like.

[00141] Suitable fluorescent proteins include, but are not limited to, green fluorescent protein (GFP) or variants thereof, blue fluorescent variant of GFP (BFP), cyan fluorescent variant of GFP (CFP), yellow fluorescent variant of GFP (YFP), enhanced GFP (EGFP), enhanced CFP (ECFP), enhanced YFP (EYFP), GFPS65T, Emerald, Topaz (TYFP), Venus, Citrine, mCitrine, GFPuv, destabilised EGFP

(dEGFP), destabilised ECFP (dECFP), destabilised EYFP (dEYFP), mCFPm, Cerulean, T-Sapphire, CyPet, YPet, mKO, HcRed, t-HcRed, DsRed, DsRed2, DsRed-monomer, J-Red, dimer2, t-dimer2(12), mRFP1, pocilloporin, Renilla GFP, Monster GFP, paGFP, Kaede protein and kindling protein, Phycobiliproteins and Phycobiliprotein conjugates including B-Phycoerythrin, R-Phycoerythrin and Allophycocyanin. Other examples of fluorescent proteins include mHoneydew, mBanana, mOrange, dTomato, tdTomato, mTangerine, mStrawberry, mCherry, mGrape1, mRaspberry, mGrape2, mPlum (Shaner et al. (2005) *Nat. Methods* 2:905-909), and the like. Any of a variety of fluorescent and colored proteins from Anthozoan species, as described in, e.g., Matz et al. (1999) *Nature Biotechnol.* 17:969-973, is suitable for use.

[00142] Suitable enzymes include, but are not limited to, horse radish peroxidase (HRP), alkaline phosphatase (AP), beta-galactosidase (GAL), glucose-6-phosphate dehydrogenase, beta-N-acetylglucosaminidase, β-glucuronidase, invertase, Xanthine Oxidase, firefly luciferase, glucose oxidase (GO), and the like.

Protospacer Adjacent Motif (PAM)

[00143] A CasY protein binds to target DNA at a target sequence defined by the region of complementarity between the DNA-targeting RNA and the target DNA. As is the case for many CRISPR endonucleases, site-specific binding (and/or cleavage) of a double stranded target DNA occurs at locations determined by both (i) base-pairing complementarity between the guide RNA and the target DNA; and (ii) a short motif [referred to as the protospacer adjacent motif (PAM)] in the target DNA.

[00144] In some embodiments, the PAM for a CasY protein is immediately 5' of the target sequence of the non-complementary strand of the target DNA (the complementary strand hybridizes to the guide sequence of the guide RNA while the non-complementary strand does not directly hybridize with the guide RNA and is the reverse complement of the non-complementary strand). In some embodiments (e.g., when CasY1 as described herein is used), the PAM sequence of the non-complementary strand is 5'-TA-3' (and in some cases XTA, where X is C, A, or T). As an example, see Figure 5 and Figure 7 (in which the PAM is TA, or CTA if you consider the PAM to be XTA where X is C, A, or T). In some embodiments (e.g., when CasY1 as described herein is used), the PAM sequence of the non-complementary strand is 5'-TA-3' (and in some cases HTA, where H is C, A, or T). As an example, see Figure 5 and Figure 7 (in which the PAM is TA, or CTA if the PAM is considered to be HTA where H is C, A, or T). In some cases (e.g., when CasY2 as described herein is used), the PAM sequence of the non-complementary strand is a 5'-YR-3' flanking sequence 5' of the target (where Y is a T or C and R is an A or G). In some cases (e.g., when CasY2 as described herein is used), the PAM sequence of the non-complementary strand is 5'-TR-3' (e.g., 5'-DTR-3') (where R is an A or G and D is an A, G, or T). As an example, see Figure 5d.

[00145] In some cases, differefnt CasY proteins (i.e., CasY proteins from various species) may be advantageous to use in the various provided methods in order to capitalize on various enzymatic characteristics of the different CasY proteins (e.g., for different PAM sequence preferences; for increased or decreased enzymatic activity; for an increased or decreased level of cellular toxicity; to change the balance between NHEJ, homology-directed repair, single strand breaks, double strand breaks, etc.; to take advantage of a short total sequence; and the like). CasY proteins from different species may require different PAM sequences in the target DNA. Thus, for a particular CasY protein of choice, the PAM sequence requirement may be different than the 5'-TA-3' (or XTA, HTA) sequence described above. Various methods (including in silico and/or wet lab methods) for identification of the appropriate PAM sequence are known in the art and are routine, and any convenient method can be used. The TA (XTA, HTA) PAM sequence described herein was identified using a PAM depletion assay (e.g., see Figure 5 of the working examples below).

CasY Guide RNA

[00146] A nucleic acid molecule that binds to a CasY protein, forming a ribonucleoprotein complex (RNP), and targets the complex to a specific location within a target nucleic acid (e.g., a target DNA) is referred to herein as a "CasY guide RNA" or simply as a "guide RNA." It is to be understood that in some cases, a hybrid DNA/RNA can be made such that a CasY guide RNA includes DNA bases in addition to RNA bases, but the term "CasY guide RNA" is still used to encompass such a molecule herein.

[00147] A CasY guide RNA can be said to include two segments, a targeting segment and a protein-binding segment. The targeting segment of a CasY guide RNA includes a nucleotide sequence (a guide sequence) that is complementary to (and therefore hybridizes with) a specific sequence (a target site) within a target nucleic acid (e.g., a target ssRNA, a target ssDNA, the complementary strand of a double stranded target DNA, etc.). The protein-binding segment (or "protein-binding sequence") interacts with (binds to) a CasY polypeptide. The protein-binding segment of a subject CasY guide RNA includes two complementary stretches of nucleotides that hybridize to one another to form a double stranded RNA duplex (dsRNA duplex). Site-specific binding and/or cleavage of a target nucleic acid (e.g., genomic DNA) can occur at locations (e.g., target sequence of a target locus) determined by base-pairing complementarity between the CasY guide RNA (the guide sequence of the CasY guide RNA) and the target nucleic acid.

[00148] A CasY guide RNA and a CasY protein, e.g., a fusion CasY polypeptide, form a complex (e.g., bind via non-covalent interactions). The CasY guide RNA provides target specificity to the complex by including a targeting segment, which includes a guide sequence (a nucleotide sequence that is complementary to a sequence of a target nucleic acid). The CasY protein of the complex provides the site-

specific activity (e.g., cleavage activity provided by the CasY protein and/or an activity provided by the fusion partner in the case of a chimeric CasY protein). In other words, the CasY protein is guided to a target nucleic acid sequence (e.g. a target sequence) by virtue of its association with the CasY guide RNA.

[00149] The "guide sequence" also referred to as the "targeting sequence" of a CasY guide RNA can be modified so that the CasY guide RNA can target a CasY protein (e.g., a naturally occurring CasY protein, a fusion CasY polypeptide (chimeric CasY), and the like) to any desired sequence of any desired target nucleic acid, with the exception (e.g., as described herein) that the PAM sequence can be taken into account. Thus, for example, a CasY guide RNA can have a guide sequence with complementarity to (e.g., can hybridize to) a sequence in a nucleic acid in a eukaryotic cell, e.g., a viral nucleic acid, a eukaryotic nucleic acid (e.g., a eukaryotic chromosome, chromosomal sequence, a eukaryotic RNA, etc.), and the like.

Guide sequence of a CasY guide RNA

[00150] A subject CasY guide RNA includes a guide sequence (i.e., a targeting sequence), which is a nucleotide sequence that is complementary to a sequence (a target site) in a target nucleic acid. In other words, the guide sequence of a CasY guide RNA can interact with a target nucleic acid (e.g., double stranded DNA (dsDNA), single stranded DNA (ssDNA), single stranded RNA (ssRNA), or double stranded RNA (dsRNA)) in a sequence-specific manner via hybridization (i.e., base pairing). The guide sequence of a CasY guide RNA can be modified (e.g., by genetic engineering)/designed to hybridize to any desired target sequence (e.g., while taking the PAM into account, e.g., when targeting a dsDNA target) within a target nucleic acid (e.g., a eukaryotic target nucleic acid such as genomic DNA).

[00151] In some embodiments, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 60% or more (e.g., 65% or more, 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%). In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 80% or more (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%). In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 90% or more (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100%). In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 100%.

[00152] In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 100% over the seven contiguous 3'-most nucleotides of the target site of the target nucleic acid.

[00153] In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 60% or more (e.g., 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 17 or more (e.g., 18 or more, 19 or more, 20 or more, 21 or more, 22 or more) contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 80% or more (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 17 or more (e.g., 18 or more, 19 or more, 20 or more, 21 or more, 22 or more) contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 90% or more (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 17 or more (e.g., 18 or more, 19 or more, 20 or more, 21 or more, 22 or more) contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 100% over 17 or more (e.g., 18 or more, 19 or more, 20 or more, 21 or more, 22 or more) contiguous nucleotides.

[00154] In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 60% or more (e.g., 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 19 or more (e.g., 20 or more, 21 or more, 22 or more) contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 80% or more (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 19 or more (e.g., 20 or more, 21 or more, 22 or more) contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 90% or more (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 19 or more (e.g., 20 or more, 21 or more, 22 or more) contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 100% over 19 or more (e.g., 20 or more, 21 or more, 22 or more) contiguous nucleotides.

[00155] In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 60% or more (e.g., 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 17-25 contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 80% or more (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 17-25 contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 90% or more (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 17-25 contiguous

nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 100% over 17-25 contiguous nucleotides.

[00156] In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 60% or more (e.g., 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 19-25 contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 80% or more (e.g., 85% or more, 90% or more, 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 19-25 contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 90% or more (e.g., 95% or more, 97% or more, 98% or more, 99% or more, or 100%) over 19-25 contiguous nucleotides. In some cases, the percent complementarity between the guide sequence and the target site of the target nucleic acid is 100% over 19-25 contiguous nucleotides.

[00157] In some cases, the guide sequence has a length in a range of from 17-30 nucleotides (nt) (e.g., from 17-25, 17-22, 17-20, 19-30, 19-25, 19-22, 19-20, 20-30, 20-25, or 20-22 nt). In some cases, the guide sequence has a length in a range of from 17-25 nucleotides (nt) (e.g., from 17-22, 17-20, 19-25, 19-22, 19-20, 20-25, or 20-22 nt). In some cases, the guide sequence has a length of 17 or more nt (e.g., 18 or more, 19 or more, 20 or more, 21 or more, or 22 or more nt; 19 nt, 20 nt, 21 nt, 22 nt, 23 nt, 24 nt, 25 nt, etc.). In some cases, the guide sequence has a length of 19 or more nt (e.g., 20 or more, 21 or more, or 22 or more nt; 19 nt, 20 nt, 21 nt, 22 nt, 23 nt, 24 nt, 25 nt, etc.). In some cases the guide sequence has a length of 17 nt. In some cases the guide sequence has a length of 18 nt. In some cases the guide sequence has a length of 19 nt. In some cases the guide sequence has a length of 20 nt. In some cases the guide sequence has a length of 21 nt. In some cases the guide sequence has a length of 22 nt. In some cases the guide sequence has a length of 23 nt.

Protein-binding segment of a CasY guide RNA

[00158] The protein-binding segment of a subject CasY guide RNA interacts with a CasY protein. The CasY guide RNA guides the bound CasY protein to a specific nucleotide sequence within target nucleic acid via the above mentioned guide sequence. The protein-binding segment of a CasY guide RNA comprises two stretches of nucleotides that are complementary to one another and hybridize to form a double stranded RNA duplex (dsRNA duplex). Thus, the protein-binding segment includes a dsRNA duplex.

[00159] In some cases, the dsRNA duplex region includes a range of from 5-25 base pairs (bp) (e.g., from 5-22, 5-20, 5-18, 5-15, 5-12, 5-10, 5-8, 8-25, 8-22, 8-18, 8-15, 8-12, 12-25, 12-22, 12-18, 12-15, 13-25, 13-22, 13-18, 13-15, 14-25, 14-22, 14-18, 14-15, 15-25, 15-22, 15-18, 17-25, 17-22, or 17-18 bp, e.g., 5 bp, 6 bp, 7 bp, 8 bp, 9 bp, 10 bp, etc.). In some cases, the dsRNA duplex region includes a

range of from 6-15 base pairs (bp) (e.g., from 6-12, 6-10, or 6-8 bp, e.g., 6 bp, 7 bp, 8 bp, 9 bp, 10 bp, etc.). In some cases, the duplex region includes 5 or more bp (e.g., 6 or more, 7 or more, or 8 or more bp). In some cases, the duplex region includes 6 or more bp (e.g., 7 or more, or 8 or more bp). In some cases, not all nucleotides of the duplex region are paired, and therefore the duplex forming region can include a bulge. The term "bulge" herein is used to mean a stretch of nucleotides (which can be one nucleotide) that do not contribute to a double stranded duplex, but which are surround 5' and 3' by nucleotides that do contribute, and as such a bulge is considered part of the duplex region. In some cases, the dsRNA includes 1 or more bulges (e.g., 2 or more, 3 or more, 4 or more bulges). In some cases, the dsRNA duplex includes 2 or more bulges (e.g., 3 or more, 4 or more bulges). In some cases, the dsRNA duplex includes 1-5 bulges (e.g., 1-4, 1-3, 2-5, 2-4, or 2-3 bulges).

[00160] Thus, in some cases, the stretches of nucleotides that hybridize to one another to form the dsRNA duplex have 70%-100% complementarity (e.g., 75%-100%, 80%-10%, 85%-100%, 90%-100%, 95%-100% complementarity) with one another. In some cases, the stretches of nucleotides that hybridize to one another to form the dsRNA duplex have 70%-100% complementarity (e.g., 75%-100%, 80%-10%, 85%-100%, 90%-100%, 95%-100% complementarity) with one another. In some cases, the stretches of nucleotides that hybridize to one another to form the dsRNA duplex have 85%-100% complementarity (e.g., 90%-100%, 95%-100% complementarity) with one another. In some cases, the stretches of nucleotides that hybridize to one another to form the dsRNA duplex have 70%-95% complementarity (e.g., 75%-95%, 80%-95%, 85%-95%, 90%-95% complementarity) with one another.

[00161] In other words, in some embodiments, the dsRNA duplex includes two stretches of nucleotides that have 70%-100% complementarity (e.g., 75%-100%, 80%-10%, 85%-100%, 90%-100%, 95%-100% complementarity) with one another. In some cases, the dsRNA duplex includes two stretches of nucleotides that have 85%-100% complementarity (e.g., 90%-100%, 95%-100% complementarity) with one another. In some cases, the dsRNA duplex includes two stretches of nucleotides that have 70%-95% complementarity (e.g., 75%-95%, 80%-95%, 85%-95%, 90%-95% complementarity) with one another.

[00162] The duplex region of a subject CasY guide RNA can include one or more (1, 2, 3, 4, 5, etc) mutations relative to a naturally occurring duplex region. For example, in some cases a base pair can be maintained while the nucleotides contributing to the base pair from each segment can be different. In some cases, the duplex region of a subject CasY guide RNA includes more paired bases, less paired bases, a smaller bulge, a larger bulge, fewer bulges, more bulges, or any convenient combination thereof, as compared to a naturally occurring duplex region (of a naturally occurring CasY guide RNA).

[00163] Examples of various Cas9 guide RNAs can be found in the art, and in some cases variations similar to those introduced into Cas9 guide RNAs can also be introduced into CasY guide

RNAs of the present disclosure (e.g., mutations to the dsRNA duplex region, extension of the 5' or 3' end for added stability for to provide for interaction with another protein, and the like). For example, see Jinek et al., Science. 2012 Aug 17;337(6096):816-21; Chylinski et al., RNA Biol. 2013 May;10(5):726-37; Ma et al., Biomed Res Int. 2013;2013:270805; Hou et al., Proc Natl Acad Sci U S A. 2013 Sep 24;110(39):15644-9; Jinek et al., Elife. 2013;2:e00471; Pattanayak et al., Nat Biotechnol. 2013 Sep;31(9):839-43; Qi et al, Cell. 2013 Feb 28;152(5):1173-83; Wang et al., Cell. 2013 May 9;153(4):910-8; Auer et al., Genome Res. 2013 Oct 31; Chen et al., Nucleic Acids Res. 2013 Nov 1;41(20):e19; Cheng et al., Cell Res. 2013 Oct;23(10):1163-71; Cho et al., Genetics. 2013 Nov;195(3):1177-80; DiCarlo et al., Nucleic Acids Res. 2013 Apr;41(7):4336-43; Dickinson et al., Nat Methods. 2013 Oct;10(10):1028-34; Ebina et al., Sci Rep. 2013;3:2510; Fujii et. al, Nucleic Acids Res. 2013 Nov 1;41(20):e187; Hu et al., Cell Res. 2013 Nov;23(11):1322-5; Jiang et al., Nucleic Acids Res. 2013 Nov 1;41(20):e188; Larson et al., Nat Protoc. 2013 Nov;8(11):2180-96; Mali et. at., Nat Methods. 2013 Oct;10(10):957-63; Nakayama et al., Genesis. 2013 Dec;51(12):835-43; Ran et al., Nat Protoc. 2013 Nov;8(11):2281-308; Ran et al., Cell. 2013 Sep 12;154(6):1380-9; Upadhyay et al., G3 (Bethesda). 2013 Dec 9;3(12):2233-8; Walsh et al., Proc Natl Acad Sci U S A. 2013 Sep 24;110(39):15514-5; Xie et al., Mol Plant. 2013 Oct 9; Yang et al., Cell. 2013 Sep 12;154(6):1370-9; Briner et al., Mol Cell. 2014 Oct 23;56(2):333-9; and U.S. patents and patent applications: 8,906,616; 8,895,308; 8,889,418; 8,889,356; 8,871,445; 8,865,406; 8,795,965; 8,771,945; 8,697,359; 20140068797; 20140170753; 20140179006; 20140179770; 20140186843; 20140186919; 20140186958; 20140189896; 20140227787; 20140234972; 20140242664; 20140242699; 20140242700; 20140242702; 20140248702; 20140256046; 20140273037; 20140273226; 20140273230; 20140273231; 20140273232; 20140273233; 20140273234; 20140273235; 20140287938; 20140295556; 20140295557; 20140298547; 20140304853; 20140309487; 20140310828; 20140310830; 20140315985; 20140335063; 20140335620; 20140342456; 20140342457; 20140342458; 20140349400; 20140349405; 20140356867; 20140356956; 20140356958; 20140356959; 20140357523; 20140357530; 20140364333; and 20140377868; all of which are hereby incorporated by reference in their entirety.

[00164] A CasY guide RNA comprises both the guide sequence and two stretches ("duplex-forming segments") of nucleotides that hybridize to form the dsRNA duplex of the protein-binding segment. The particular sequence of a given CasY guide RNA can be characteristic of the species in which the a crRNA is found. Examples of suitable CasY guide RNAs are provided herein.

Example Guide RNA sequences

[00165] The repeat seqences (non-guide sequence portion of example CasY guide RNAs) depicted in Figure 6 (panels a and b) are from the natural locus for CasY1-Y5. In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) the crRNA sequence CTCCGAAAGTATCGGGGATAAAGGC (SEQ ID NO: 31) [RNA is

CUCCGAAAGUAUCGGGGAUAAAGGC (SEQ ID NO: 11)] (e.g., see Figure 6). In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 80% or more identity (e.g., 85% or more, 90% or more, 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CTCCGAAAGTATCGGGGATAAAGGC (SEQ ID NO: 31) [RNA is CUCCGAAAGUAUCGGGGAUAAAGGC (SEQ ID NO: 11)]. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 90% or more identity (e.g., 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CTCCGAAAGTATCGGGGATAAAGGC (SEQ ID NO: 31) [RNA is CUCCGAAAGUAUCGGGGAUAAAGGC (SEQ ID NO: 11)].

[00166] In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) the crRNA sequence CACCGAAATTTGGAGAGGATAAGGC (SEQ ID NO: 32) [RNA is CACCGAAAUUUGGAGAGGAUAAGGC (SEQ ID NO: 12)] (e.g., see Figure 6). In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 80% or more identity (e.g., 85% or more, 90% or more, 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CACCGAAATTTGGAGAGGATAAGGC (SEQ ID NO: 32) [RNA is CACCGAAAUUUGGAGAGGAUAAGGC (SEQ ID NO: 12)]. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 90% or more identity (e.g., 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CACCGAAATTTGGAGAGGATAAGGC (SEQ ID NO: 32) [RNA is CACCGAAAUUUGGAGAGGAUAAGGC (SEQ ID NO: 12)].

[00167] In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) the crRNA sequence CTCCGAATTATCGGGAGGATAAGGC (SEQ ID NO: 33) [RNA is CUCCGAAUUAUCGGGAGGAUAAGGC (SEQ ID NO: 13)] (e.g., see Figure 6). In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 80% or more identity (e.g., 85% or more, 90% or more, 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CTCCGAATTATCGGGAGGATAAGGC (SEQ ID NO: 33) [RNA is CUCCGAAUUAUCGGGAGGAUAAGGC (SEQ ID NO: 13)]. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 90% or more identity (e.g., 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CTCCGAATTATCGGGAGGATAAGGC (SEQ ID NO: 33) [RNA is CUCCGAAUUAUCGGGAGGAUAAGGC (SEQ ID NO: 13)].

[00168] In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) the crRNA sequence CCCCGAATATAGGGGACAAAAAGGC (SEQ ID NO: 34) [RNA is CCCCGAAUAUAGGGGACAAAAAGGC (SEQ ID NO: 14)] (e.g., see Figure 6). In some cases, a

subject CasY guide RNA comprises a nucleotide sequence having 80% or more identity (e.g., 85% or more, 90% or more, 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CCCCGAATATAGGGGACAAAAAGGC (SEQ ID NO: 34) [RNA is CCCCGAAUAUAGGGGACAAAAAGGC (SEQ ID NO: 14)]. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 90% or more identity (e.g., 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CCCCGAATATAGGGGACAAAAAGGC (SEQ ID NO: 34) [RNA is CCCCGAAUAUAGGGGACAAAAAGGC (SEQ ID NO: 14)].

[00169] In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) the crRNA sequence GTCTAGACATACAGGTGGAAAGGTGAGAGTAAAGAC (SEQ ID NO: 35) [RNA is GUCUAGACAUACAGGUGGAAAGGUGAGAGUAAAGAC (SEQ ID NO: 15)] (e.g., see Figure 6). In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 80% or more identity (e.g., 85% or more, 90% or more, 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence GTCTAGACATACAGGTGGAAAGGTGAGAGTAAAGAC (SEQ ID NO: 35) [RNA is GUCUAGACAUACAGGUGGAAAGGUGAGAGUAAAGAC (SEQ ID NO: 15)]. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 90% or more identity (e.g., 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence GTCTAGACATACAGGTGGAAAGGTGAGAGTAAAGAC (SEQ ID NO: 35) [RNA is GUCUAGACAUACAGGUGGAAAGGUGAGAGUAAAGAC (SEQ ID NO: 15)].

[00170] In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) the crRNA sequence set forth in any one of SEQ ID NOs: 11-15. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 80% or more identity (e.g., 85% or more, 90% or more, 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence set forth in any one of SEQ ID NOs: 11-15. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 90% or more identity (e.g., 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence set forth in any one of SEQ ID NOs: 11-15.

[00171] In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) the crRNA sequence set forth in any one of SEQ ID NOs: 11-14. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 80% or more identity (e.g., 85% or more, 90% or more, 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence set forth in any one of SEQ ID NOs: 11-14. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 90% or more identity (e.g., 93% or more, 95% or more, 97% or

more, 98% or more, or 100% identity) with the crRNA sequence set forth in any one of SEQ ID NOs: 11-14.

[00172] The repeat seqence (non-guide sequence portion of an example CasY guide RNA) from the natural locus for CasY18 is CTCCGTGAATACGTGGGGTAAAGGC (SEQ ID NO: 36) [RNA is CUCCGUGAAUACGUGGGGUAAAGGC (SEQ ID NO: 16)]. In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) the crRNA sequence CTCCGTGAATACGTGGGGTAAAGGC (SEQ ID NO: 36) [RNA is CUCCGUGAAUACGUGGGGUAAAGGC (SEQ ID NO: 16)]. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 80% or more identity (e.g., 85% or more, 90% or more, 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CTCCGTGAATACGTGGGGTAAAGGC (SEQ ID NO: 36) [RNA is CUCCGUGAAUACGUGGGGUAAAGGC (SEQ ID NO: 16)]. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 90% or more identity (e.g., 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence CTCCGTGAATACGTGGGGTAAAGGC (SEQ ID NO: 36) [RNA is CUCCGUGAAUACGUGGGGUAAAGGC (SEQ ID NO: 16)].

[00173] In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) the crRNA sequence set forth in any one of SEQ ID NOs: 11-16. In some cases, a subject CasY guide RNA comprises (e.g., in addition to a guide sequence) a nucleotide sequence having 80% or more identity (e.g., 85% or more, 90% or more, 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence set forth in any one of SEQ ID NOs: 11-16. In some cases, a subject CasY guide RNA comprises a nucleotide sequence having 90% or more identity (e.g., 93% or more, 95% or more, 97% or more, 98% or more, or 100% identity) with the crRNA sequence set forth in any one of SEQ ID NOs: 11-16.

CASY SYSTEMS

[00174] The present disclosure provides a CasY system. A CasY system of the present disclosure can comprise: a) a CasY polypeptide of the present disclosure and a CasY guide RNA; b) a CasY polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; c) a CasY fusion polypeptide of the present disclosure and a CasY guide RNA; d) a CasY fusion polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; e) an mRNA encoding a CasY polypeptide of the present disclosure; and a CasY guide RNA; f) an mRNA encoding a CasY polypeptide of the present disclosure, a CasY guide RNA, and a donor templat nucleic acid; g) an mRNA encoding a CasY fusion polypeptide of the present disclosure; and a CasY guide RNA; h) an mRNA encoding a CasY fusion polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid;

i) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure and a nucleotide sequence encoding a CasY guide RNA; j) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, a nucleotide sequence encoding a CasY guide RNA, and a nucleotide sequence encoding a donor template nucleic acid; k) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure and a nucleotide sequence encoding a CasY guide RNA; l) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, a nucleotide sequence encoding a CasY guide RNA, and a nucleotide sequence encoding a donor template nucleic acid; m) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; n) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; and a donor template nucleic acid; o) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; p) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; and a donor template nucleic acid; q) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, a nucleotide sequence encoding a first CasY guide RNA, and a nucleotide sequence encoding a second CasY guide RNA; or r) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, a nucleotide sequence encoding a first CasY guide RNA, and a nucleotide sequence encoding a second CasY guide RNA; or some variation of one of (a) through (r).

NUCLEIC ACIDS

[00175] The present disclosure provides one ore more nucleic acids comprising one or more of: a donor polynucleotide sequence, a nucleotide sequence encoding a CasY polypeptide (e.g., a wild type CasY protein, a nickase CasY protein, a dCasY protein, chimeric CasY protein, and the like), a CasY guide RNA, and a nucleotide sequence encoding a CasY guide RNA. The present disclosure provides a nucleic acid comprising a nucleotide sequence encoding a CasY fusion polypeptide. The present disclosure provides a recombinant expression vector that comprises a nucleotide sequence encoding a CasY polypeptide. The present disclosure provides a recombinant expression vector that comprises a nucleotide sequence encoding a CasY fusion polypeptide. The present disclosure provides a recombinant expression vector that comprises: a) a nucleotide sequence encoding a CasY polypeptide; and b) a

nucleotide sequence encoding a CasY guide RNA(s). The present disclosure provides a recombinant expression vector that comprises: a) a nucleotide sequence encoding a CasY fusion polypeptide; and b) a nucleotide sequence encoding a CasY guide RNA(s). In some cases, the nucleotide sequence encoding the CasY protein and/or the nucleotide sequence encoding the CasY guide RNA is operably linked to a promoter that is operable in a cell type of choice (e.g., a prokarytoic cell, a eukaryotic cell, a plant cell, an animal cell, a mammalian cell, a primate cell, a rodent cell, a human cell, etc.).

[00176] In some cases, a nucleotide sequence encoding a CasY polypeptide of the present disclosure is codon optimized. This type of optimization can entail a mutation of a CasY-encoding nucleotide sequence to mimic the codon preferences of the intended host organism or cell while encoding the same protein. Thus, the codons can be changed, but the encoded protein remains unchanged. For example, if the intended target cell was a human cell, a human codon-optimized CasY-encoding nucleotide sequence could be used. As another non-limiting example, if the intended host cell were a mouse cell, then a mouse codon-optimized CasY-encoding nucleotide sequence could be generated. As another non-limiting example, if the intended host cell were a plant cell, then a plant codon-optimized CasY-encoding nucleotide sequence could be generated. As another non-limiting example, if the intended host cell were an insect cell, then an insect codon-optimized CasY-encoding nucleotide sequence could be generated.

[00177] The present disclosure provides one or more recombinant expression vectors that include (in differnet recombinant expression vectors in some cases, and in the same recombinant expression vector in some cases): (i) a nucleotide seqeunce of a donor template nucleic acid (where the donor template comprises a nucleotide sequence having homology to a target sequence of a target nucleic acid (e.g., a target genome)); (ii) a nucleotide sequence that encodes a CasY guide RNA that hybridizes to a target sequence of the target locus of the targeted genome (e.g., operably linked to a promoter that is operable in a target cell such as a eukaryotic cell); and (iii) a nucleotide sequence encoding a CasY protein (e.g., operably linked to a promoter that is operable in a target cell such as a eukaryotic cell). The present disclosure provides one or more recombinant expression vectors that include (in differnet recombinant expression vectors in some cases, and in the same recombinant expression vector in some cases): (i) a nucleotide seqeunce of a donor template nucleic acid (where the donor template comprises a nucleotide sequence having homology to a target sequence of a target nucleic acid (e.g., a target genome)); and (ii) a nucleotide sequence that encodes a CasY guide RNA that hybridizes to a target sequence of the target locus of the targeted genome (e.g., operably linked to a promoter that is operable in a target cell such as a eukaryotic cell). The present disclosure provides one or more recombinant expression vectors that include (in differnet recombinant expression vectors in some cases, and in the same recombinant expression vector in some cases): (i) a nucleotide sequence that encodes a CasY guide

RNA that hybridizes to a target sequence of the target locus of the targeted genome (e.g., operably linked to a promoter that is operable in a target cell such as a eukaryotic cell); and (ii) a nucleotide sequence encoding a CasY protein (e.g., operably linked to a promoter that is operable in a target cell such as a eukaryotic cell).

[00178] Suitable expression vectors include viral expression 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 (AAV) (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. In some cases, a recombinant expression vector of the present disclosure is a recombinant adeno-associated virus (AAV) vector. In some cases, a recombinant expression vector of the present disclosure is a recombinant lentivirus vector. In some cases, a recombinant expression vector of the present disclosure is a recombinant retroviral vector.

[00179] 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.

[00180] In some embodiments, a nucleotide sequence encoding a CasY guide RNA is operably linked to a control element, e.g., a transcriptional control element, such as a promoter. In some embodiments, a nucleotide sequence encoding a CasY protein or a CasY fusion polypeptide is operably linked to a control element, e.g., a transcriptional control element, such as a promoter.

[00181] The transcriptional control element can be a promoter. In some cases, the promoter is a constitutively active promoter. In some cases, the promoter is a regulatable promoter. In some cases, the promoter is an inducible promoter. In some cases, the promoter is a tissue-specific promoter. In some cases, the promoter is a cell type-specific promoter. In some cases, the transcriptional control element (e.g., the promoter) is functional in a targeted cell type or targeted cell population. For example, in some

cases, the transcriptional control element can be functional in eukaryotic cells, e.g., hematopoietic stem cells (e.g., mobilized peripheral blood (mPB) CD34(+) cell, bone marrow (BM) CD34(+) cell, etc.).

[00182] Non-limiting examples of eukaryotic promoters (promoters functional in a eukaryotic cell) include EF1α, those from cytomegalovirus (CMV) immediate early, herpes simplex virus (HSV) thymidine kinase, early and late SV40, long terminal repeats (LTRs) from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art. The expression vector may also contain a ribosome binding site for translation initiation and a transcription terminator. The expression vector may also include appropriate sequences for amplifying expression. The expression vector may also include nucleotide sequences encoding protein tags (e.g., 6xHis tag, hemagglutinin tag, fluorescent protein, etc.) that can be fused to the CasY protein, thus resulting in a chimeric CasY polypeptide.

[00183] In some embodiments, a nucleotide sequence encoding a CasY guide RNA and/or a CasY fusion polypeptide is operably linked to an inducible promoter. In some embodiments, a nucleotide sequence encoding a CasY guide RNA and/or a CasY fusion protein is operably linked to a constitutive promoter.

[00184] A promoter can be a constitutively active promoter (i.e., a promoter that is constitutively in an active/"ON" state), it may be an inducible promoter (i.e., a promoter whose state, active/"ON" or inactive/"OFF", is controlled by an external stimulus, e.g., the presence of a particular temperature, compound, or protein.), it may be a spatially restricted promoter (i.e., transcriptional control element, enhancer, etc.)(e.g., tissue specific promoter, cell type specific promoter, etc.), and it may be a temporally restricted promoter (i.e., the promoter is in the "ON" state or "OFF" state during specific stages of embryonic development or during specific stages of a biological process, e.g., hair follicle cycle in mice).

[00185] Suitable promoters can be derived from viruses and can therefore be referred to as viral promoters, or they can be derived from any organism, including prokaryotic or eukaryotic organisms. Suitable promoters can be used to drive expression by any RNA polymerase (e.g., pol I, pol II, pol III). Exemplary promoters include, but are not limited to the SV40 early promoter, mouse mammary tumor virus long terminal repeat (LTR) promoter; adenovirus major late promoter (Ad MLP); a herpes simplex virus (HSV) promoter, a cytomegalovirus (CMV) promoter such as the CMV immediate early promoter region (CMVIE), a rous sarcoma virus (RSV) promoter, a human U6 small nuclear promoter (U6) (Miyagishi et al., Nature Biotechnology 20, 497 - 500 (2002)), an enhanced U6 promoter (e.g., Xia et al., Nucleic Acids Res. 2003 Sep 1;31(17)), a human H1 promoter (H1), and the like.

[00186] In some cases, a nucleotide sequence encoding a CasY guide RNA is operably linked to (under the control of) a promoter operable in a eukaryotic cell (e.g., a U6 promoter, an enhanced U6 promoter, an H1 promoter, and the like). As would be understood by one of ordinary skill in the art, when

expressing an RNA (e.g., a guide RNA) from a nucleic acid (e.g., an expression vector) using a U6 promoter (e.g., in a eukaryotic cell), or another PolIII promoter, the RNA may need to be mutated if there are several Ts in a row (coding for Us in the RNA). This is because a string of Ts (e.g., 5 Ts) in DNA can act as a terminator for polymerase III (PolIII). Thus, in order to ensure transcription of a guide RNA in a eukaryotic cell it may sometimes be necessary to modify the sequence encoding the guide RNA to eliminate runs of Ts. In some cases, a nucleotide sequence encoding a CasY protein (e.g., a wild type CasY protein, a nickase CasY protein, a dCasY protein, a chimeric CasY protein and the like) is operably linked to a promoter operable in a eukaryotic cell (e.g., a CMV promoter, an EF1α promoter, an estrogen receptor-regulated promoter, and the like).

[00187] Examples of inducible promoters include, but are not limited toT7 RNA polymerase promoter, T3 RNA polymerase promoter, Isopropyl-beta-D-thiogalactopyranoside (IPTG)-regulated promoter, lactose induced promoter, heat shock promoter, Tetracycline-regulated promoter, Steroid-regulated promoter, Metal-regulated promoter, estrogen receptor-regulated promoter, etc. Inducible promoters can therefore be regulated by molecules including, but not limited to, doxycycline; estrogen and/or an estrogen analog; IPTG; etc.

[00188] Inducible promoters suitable for use include any inducible promoter described herein or known to one of ordinary skill in the art. Examples of inducible promoters include, without limitation, chemically/biochemically-regulated and physically-regulated promoters such as alcohol-regulated promoters, tetracycline-regulated promoters (e.g., anhydrotetracycline (aTc)-responsive promoters and other tetracycline-responsive promoter systems, which include a tetracycline repressor protein (tetR), a tetracycline operator sequence (tetO) and a tetracycline transactivator fusion protein (tTA)), steroid-regulated promoters (e.g., promoters based on the rat glucocorticoid receptor, human estrogen receptor, moth ecdysone receptors, and promoters from the steroid/retinoid/thyroid receptor superfamily), metal-regulated promoters (e.g., promoters derived from metallothionein (proteins that bind and sequester metal ions) genes from yeast, mouse and human), pathogenesis-regulated promoters (e.g., induced by salicylic acid, ethylene or benzothiadiazole (BTH)), temperature/heat-inducible promoters (e.g., heat shock promoters), and light-regulated promoters (e.g., light responsive promoters from plant cells).

[00189] In some cases, the promoter is a spatially restricted promoter (i.e., cell type specific promoter, tissue specific promoter, etc.) such that in a multi-cellular organism, the promoter is active (i.e., "ON") in a subset of specific cells. Spatially restricted promoters may also be referred to as enhancers, transcriptional control elements, control sequences, etc. Any convenient spatially restricted promoter may be used as long as the promoter is functional in the targeted host cell (e.g., eukaryotic cell; prokaryotic cell).

[00190] In some cases, the promoter is a reversible promoter. Suitable reversible promoters, including reversible inducible promoters are known in the art. Such reversible promoters may be isolated and derived from many organisms, e.g., eukaryotes and prokaryotes. Modification of reversible promoters derived from a first organism for use in a second organism, e.g., a first prokaryote and a second a eukaryote, a first eukaryote and a second a prokaryote, etc., is well known in the art. Such reversible promoters, and systems based on such reversible promoters but also comprising additional control proteins, include, but are not limited to, alcohol regulated promoters (e.g., alcohol dehydrogenase I (alcA) gene promoter, promoters responsive to alcohol transactivator proteins (AlcR), etc.), tetracycline regulated promoters, (e.g., promoter systems including TetActivators, TetON, TetOFF, etc.), steroid regulated promoters (e.g., rat glucocorticoid receptor promoter systems, human estrogen receptor promoter systems, retinoid promoter systems, thyroid promoter systems, ecdysone promoter systems, mifepristone promoter systems, etc.), metal regulated promoters (e.g., metallothionein promoter systems, etc.), pathogenesis-related regulated promoters (e.g., salicylic acid regulated promoters, ethylene regulated promoters, benzothiadiazole regulated promoters, etc.), temperature regulated promoters (e.g., heat shock inducible promoters (e.g., HSP-70, HSP-90, soybean heat shock promoter, etc.), light regulated promoters, synthetic inducible promoters, and the like.

[00191] Methods of introducing a nucleic acid (e.g., a nucleic acid comprising a donor polynucleotide sequence, one or more nucleic acids encoding a CasY protein and/or a CasY guide RNA, and the like) into a host cell are known in the art, and any convenient method can be used to introduce a nucleic acid (e.g., an expression construct) into a cell. Suitable methods include e.g., viral infection, transfection, lipofection, electroporation, calcium phosphate precipitation, polyethyleneimine (PEI)-mediated transfection, DEAE-dextran mediated transfection, liposome-mediated transfection, particle gun technology, calcium phosphate precipitation, direct microinjection, nanoparticle-mediated nucleic acid delivery, and the like.

[00192] Introducing the recombinant expression vector into cells can occur in any culture media and under any culture conditions that promote the survival of the cells. Introducing the recombinant expression vector into a target cell can be carried out *in vivo* or *ex vivo*. Introducing the recombinant expression vector into a target cell can be carried out *in vitro*.

[00193] In some embodiments, a CasY protein can be provided as RNA. The RNA can be provided by direct chemical synthesis or may be transcribed in vitro from a DNA (e.g., encoding the CasY protein). Once synthesized, the RNA may be introduced into a cell by any of the well-known techniques for introducing nucleic acids into cells (e.g., microinjection, electroporation, transfection, etc.).

[00194] Nucleic acids may be provided to the cells using well-developed transfection techniques; see, e.g. Angel and Yanik (2010) PLoS ONE 5(7): e11756, and the commercially available

TransMessenger® reagents from Qiagen, Stemfect™ RNA Transfection Kit from Stemgent, and TransIT®-mRNA Transfection Kit from Mirus Bio LLC. See also Beumer et al. (2008) PNAS 105(50):19821-19826.

[00195] Vectors may be provided directly to a target host cell. In other words, the cells are contacted with vectors comprising the subject nucleic acids (e.g., recombinant expression vectors having the donor template sequence and encoding the CasY guide RNA; recombinant expression vectors encoding the CasY protein; etc.) such that the vectors are taken up by the cells. Methods for contacting cells with nucleic acid vectors that are plasmids, include electroporation, calcium chloride transfection, microinjection, and lipofection are well known in the art. For viral vector delivery, cells can be contacted with viral particles comprising the subject viral expression vectors.

[00196] Retroviruses, for example, lentiviruses, are suitable for use in methods of the present disclosure. Commonly used retroviral vectors are "defective", i.e. unable to produce viral proteins required for productive infection. Rather, replication of the vector requires growth in a packaging cell line. To generate viral particles comprising nucleic acids of interest, the retroviral nucleic acids comprising the nucleic acid are packaged into viral capsids by a packaging cell line. Different packaging cell lines provide a different envelope protein (ecotropic, amphotropic or xenotropic) to be incorporated into the capsid, this envelope protein determining the specificity of the viral particle for the cells (ecotropic for murine and rat; amphotropic for most mammalian cell types including human, dog and mouse; and xenotropic for most mammalian cell types except murine cells). The appropriate packaging cell line may be used to ensure that the cells are targeted by the packaged viral particles. Methods of introducing subject vector expression vectors into packaging cell lines and of collecting the viral particles that are generated by the packaging lines are well known in the art. Nucleic acids can also introduced by direct micro-injection (e.g., injection of RNA).

[00197] Vectors used for providing the nucleic acids encoding CasY guide RNA and/or a CasY polypeptide to a target host cell can include suitable promoters for driving the expression, that is, transcriptional activation, of the nucleic acid of interest. In other words, in some cases, the nucleic acid of interest will be operably linked to a promoter. This may include ubiquitously acting promoters, for example, the CMV-β-actin promoter, or inducible promoters, such as promoters that are active in particular cell populations or that respond to the presence of drugs such as tetracycline. By transcriptional activation, it is intended that transcription will be increased above basal levels in the target cell by 10 fold, by 100 fold, more usually by 1000 fold. In addition, vectors used for providing a nucleic acid encoding a CasY guide RNA and/or a CasY protein to a cell may include nucleic acid sequences that encode for selectable markers in the target cells, so as to identify cells that have taken up the CasY guide RNA and/or CasY protein.

[00198] A nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide, or a CasY fusion polypeptide, is in some cases an RNA. Thus, a CasY fusion protein can be introduced into cells as RNA. Methods of introducing RNA into cells are known in the art and may include, for example, direct injection, transfection, or any other method used for the introduction of DNA. A CasY protein may instead be provided to cells as a polypeptide. Such a polypeptide may optionally be fused to a polypeptide domain that increases solubility of the product. The domain may be linked to the polypeptide through a defined protease cleavage site, e.g. a TEV sequence, which is cleaved by TEV protease. The linker may also include one or more flexible sequences, e.g. from 1 to 10 glycine residues. In some embodiments, the cleavage of the fusion protein is performed in a buffer that maintains solubility of the product, e.g. in the presence of from 0.5 to 2 M urea, in the presence of polypeptides and/or polynucleotides that increase solubility, and the like. Domains of interest include endosomolytic domains, e.g. influenza HA domain; and other polypeptides that aid in production, e.g. IF2 domain, GST domain, GRPE domain, and the like. The polypeptide may be formulated for improved stability. For example, the peptides may be PEGylated, where the polyethyleneoxy group provides for enhanced lifetime in the blood stream.

[00199] Additionally or alternatively, a CasY polypeptide of the present disclosure may be fused to a polypeptide permeant domain to promote uptake by the cell. A number of permeant domains are known in the art and may be used in the non-integrating polypeptides of the present disclosure, including peptides, peptidomimetics, and non-peptide carriers. For example, a permeant peptide may be derived from the third alpha helix of *Drosophila melanogaster transcription factor Antennapaedia, referred to as penetratin, which comprises the* amino acid sequence RQIKIWFQNRRMKWKK (SEQ ID NO: 133). As another example, the permeant peptide comprises the HIV-1 tat basic region amino acid sequence, which may include, for example, amino acids 49-57 of naturally-occurring tat protein. Other permeant domains include poly-arginine motifs, for example, the region of amino acids 34-56 of HIV-1 rev protein, nona-arginine, octa-arginine, and the like. (See, for example, Futaki et al. (2003) Curr Protein Pept Sci. 2003 Apr; 4(2): 87-9 and 446; and Wender et al. (2000) Proc. Natl. Acad. Sci. U.S.A 2000 Nov. 21; 97(24):13003-8; published U.S. Patent applications 20030220334; 20030083256; 20030032593; and 20030022831, herein specifically incorporated by reference for the teachings of translocation peptides and peptoids). The nona-arginine (R9) sequence is one of the more efficient PTDs that have been characterized (Wender et al. 2000; Uemura et al. 2002). The site at which the fusion is made may be selected in order to optimize the biological activity, secretion or binding characteristics of the polypeptide. The optimal site will be determined by routine experimentation.

[00200] A CasY polypeptide of the present disclosure may be produced *in vitro* or by eukaryotic cells or by prokaryotic cells, and it may be further processed by unfolding, e.g. heat denaturation, dithiothreitol reduction, etc. and may be further refolded, using methods known in the art.

[00201] Modifications of interest that do not alter primary sequence include chemical derivatization of polypeptides, e.g., acylation, acetylation, carboxylation, amidation, etc. Also included are modifications of glycosylation, e.g. those made by modifying the glycosylation patterns of a polypeptide during its synthesis and processing or in further processing steps; e.g. by exposing the polypeptide to enzymes which affect glycosylation, such as mammalian glycosylating or deglycosylating enzymes. Also embraced are sequences that have phosphorylated amino acid residues, e.g. phosphotyrosine, phosphoserine, or phosphothreonine.

[00202] Also suitable for inclusion in embodiments of the present disclosure are nucleic acids (e.g., encoding a CasY guide RNA, encoding a CasY fusion protein, etc.) and proteins (e.g., a CasY fusion protein derived from a wild type protein or a variant protein) that have been modified using ordinary molecular biological techniques and synthetic chemistry so as to improve their resistance to proteolytic degradation, to change the target sequence specificity, to optimize solubility properties, to alter protein activity (e.g., transcription modulatory activity, enzymatic activity, etc.) or to render them more suitable. Analogs of such polypeptides include those containing residues other than naturally occurring L-amino acids, e.g. D-amino acids or non-naturally occurring synthetic amino acids. D-amino acids may be substituted for some or all of the amino acid residues.

[00203] A CasY polypeptide of the present disclosure may be prepared by *in vitro* synthesis, using conventional methods as known in the art. Various commercial synthetic apparatuses are available, for example, automated synthesizers by Applied Biosystems, Inc., Beckman, etc. By using synthesizers, naturally occurring amino acids may be substituted with unnatural amino acids. The particular sequence and the manner of preparation will be determined by convenience, economics, purity required, and the like.

[00204] If desired, various groups may be introduced into the peptide during synthesis or during expression, which allow for linking to other molecules or to a surface. Thus cysteines can be used to make thioethers, histidines for linking to a metal ion complex, carboxyl groups for forming amides or esters, amino groups for forming amides, and the like.

[00205] A CasY polypeptide of the present disclosure may also be isolated and purified in accordance with conventional methods of recombinant synthesis. A lysate may be prepared of the expression host and the lysate purified using high performance liquid chromatography (HPLC), exclusion chromatography, gel electrophoresis, affinity chromatography, or other purification technique. For the most part, the compositions which are used will comprise 20% or more by weight of the desired product, more usually 75% or more by weight, preferably 95% or more by weight, and for therapeutic purposes, usually 99.5% or more by weight, in relation to contaminants related to the method of preparation of the product and its purification. Usually, the percentages will be based upon total protein. Thus, in some

cases, a CasY polypeptide, or a CasY fusion polypeptide, of the present disclosure is at least 80% pure, at least 85% pure, at least 90% pure, at least 95% pure, at least 98% pure, or at least 99% pure (e.g., free of contaminants, non-CasY proteins or other macromolecules, etc.).

[00206] To induce cleavage or any desired modification to a target nucleic acid (e.g., genomic DNA), or any desired modification to a polypeptide associated with target nucleic acid, the CasY guide RNA and/or the CasY polypeptide of the present disclosure and/or the donor template sequence, whether they be introduced as nucleic acids or polypeptides, are provided to the cells for about 30 minutes to about 24 hours, e.g., 1 hour, 1.5 hours, 2 hours, 2.5 hours, 3 hours, 3.5 hours 4 hours, 5 hours, 6 hours, 7 hours, 8 hours, 12 hours, 16 hours, 18 hours, 20 hours, or any other period from about 30 minutes to about 24 hours, which may be repeated with a frequency of about every day to about every 4 days, e.g., every 1.5 days, every 2 days, every 3 days, or any other frequency from about every day to about every four days. The agent(s) may be provided to the subject cells one or more times, e.g. one time, twice, three times, or more than three times, and the cells allowed to incubate with the agent(s) for some amount of time following each contacting event e.g. 16-24 hours, after which time the media is replaced with fresh media and the cells are cultured further.

[00207] In cases in which two or more different targeting complexes are provided to the cell (e.g., two different CasY guide RNAs that are complementary to different sequences within the same or different target nucleic acid), the complexes may be provided simultaneously (e.g. as two polypeptides and/or nucleic acids), or delivered simultaneously. Alternatively, they may be provided consecutively, e.g. the targeting complex being provided first, followed by the second targeting complex, etc. or vice versa.

[00208] To improve the delivery of a DNA vector into a target cell, the DNA can be protected from damage and its entry into the cell facilitated, for example, by using lipoplexes and polyplexes. Thus, in some cases, a nucleic acid of the present disclosure (e.g., a recombinant expression vector of the present disclosure) can be covered with lipids in an organized structure like a micelle or a liposome. When the organized structure is complexed with DNA it is called a lipoplex. There are three types of lipids, anionic (negatively-charged), neutral, or cationic (positively-charged). Lipoplexes that utilize cationic lipids have proven utility for gene transfer. Cationic lipids, due to their positive charge, naturally complex with the negatively charged DNA. Also as a result of their charge, they interact with the cell membrane. Endocytosis of the lipoplex then occurs, and the DNA is released into the cytoplasm. The cationic lipids also protect against degradation of the DNA by the cell.

[00209] Complexes of polymers with DNA are called polyplexes. Most polyplexes consist of cationic polymers and their production is regulated by ionic interactions. One large difference between the methods of action of polyplexes and lipoplexes is that polyplexes cannot release their DNA load into the

cytoplasm, so to this end, co-transfection with endosome-lytic agents (to lyse the endosome that is made during endocytosis) such as inactivated adenovirus must occur. However, this is not always the case; polymers such as polyethylenimine have their own method of endosome disruption as does chitosan and trimethylchitosan.

[00210] Dendrimers, a highly branched macromolecule with a spherical shape, may be also be used to genetically modify stem cells. The surface of the dendrimer particle may be functionalized to alter its properties. In particular, it is possible to construct a cationic dendrimer (i.e., one with a positive surface charge). When in the presence of genetic material such as a DNA plasmid, charge complementarity leads to a temporary association of the nucleic acid with the cationic dendrimer. On reaching its destination, the dendrimer-nucleic acid complex can be taken up into a cell by endocytosis.

[00211] In some cases, a nucleic acid of the disclosure (e.g., an expression vector) includes an insertion site for a guide sequence of interest. For example, a nucleic acid can include an insertion site for a guide sequence of interest, where the insertion site is immediately adjacent to a nucleotide sequence encoding the portion of a CasY guide RNA that does not change when the guide sequence is changed to hybrized to a desired target sequence (e.g., sequences that contribute to the CasY binding aspect of the guide RNA, e.g, the sequences that contribute to the dsRNA duplex(es) of the CasY guide RNA – this portion of the guide RNA can also be referred to as the 'scaffold' or 'constant region' of the guide RNA). Thus, in some cases, a subject nucleic acid (e.g., an expression vector) includes a nucleotide sequence encoding a CasY guide RNA, except that the portion encoding the guide sequence portion of the guide RNA is an insertion sequence (an insertion site). An insertion site is any nucleotide sequence used for the insertion of a the desired sequence. "Insertion sites" for use with various technologies are known to those of ordinary skill in the art and any convenient insertion site can be used. An insertion site can be for any method for manipulating nucleic acid sequences. For example, in some cases the insertion site is a multiple cloning site (MCS) (e.g., a site including one or more restriction enzyme recognition sequences), a site for ligation independent cloning, a site for recombination based cloning (e.g., recombination based on *att* sites), a nucleotide sequence recognized by a CRISPR/Cas (e.g. Cas9) based technology, and the like.

[00212] An insertion site can be any desirable length, and can depend on the type of insertion site (e.g., can depend on whether (and how many) the site includes one or more restriction enzyme recognition sequences, whether the site includes a target site for a CRISPR/Cas protein, etc.). In some cases, an insertion site of a subject nucleic acid is 3 or more nucleotides (nt) in length (e.g., 5 or more, 8 or more, 10 or more, 15 or more, 17 or more, 18 or more, 19 or more, 20 or more or 25 or more, or 30 or more nt in length). In some cases, the length of an insertion site of a subject nucleic acid has a length in a range of from 2 to 50 nucleotides (nt) (e.g., from 2 to 40 nt, from 2 to 30 nt, from 2 to 25 nt, from 2 to 20

nt, from 5 to 50 nt, from 5 to 40 nt, from 5 to 30 nt, from 5 to 25 nt, from 5 to 20 nt, from 10 to 50 nt, from 10 to 40 nt, from 10 to 30 nt, from 10 to 25 nt, from 10 to 20 nt, from 17 to 50 nt, from 17 to 40 nt, from 17 to 30 nt, from 17 to 25 nt). In some cases, the length of an insertion site of a subject nucleic acid has a length in a range of from 5 to 40 nt.

Nucleic acid modifications

[00213] In some embodiments, a subject nucleic acid (e.g., a CasY guide RNA) has 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). 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 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.

[00214] Suitable nucleic acid modifications include, but are not limited to: 2'Omethyl modified nucleotides, 2' Fluoro modified nucleotides, locked nucleic acid (LNA) modified nucleotides, peptide nucleic acid (PNA) modified nucleotides, nucleotides with phosphorothioate linkages, and a 5' cap (e.g., a 7-methylguanylate cap (m7G)). Additional details and additional modifications are described below.

[00215] A 2'-O-Methyl modified nucleotide (also referred to as 2'-O-Methyl RNA) is a naturally occurring modification of RNA found in tRNA and other small RNAs that arises as a post-transcriptional modification. Oligonucleotides can be directly synthesized that contain 2'-O-Methyl RNA. This modification increases Tm of RNA:RNA duplexes but results in only small changes in RNA:DNA stability. It is stabile with respect to attack by single-stranded ribonucleases and is typically 5 to 10-fold less susceptible to DNases than DNA. It is commonly used in antisense oligos as a means to increase stability and binding affinity to the target message.

[00216] 2' Fluoro modified nucleotides (e.g., 2' Fluoro bases) have a fluorine modified ribose which increases binding affinity (Tm) and also confers some relative nuclease resistance when compared to

native RNA. These modifications are commonly employed in ribozymes and siRNAs to improve stability in serum or other biological fluids.

[00217] LNA bases have a modification to the ribose backbone that locks the base in the C3'-endo position, which favors RNA A-type helix duplex geometry. This modification significantly increases Tm and is also very nuclease resistant. Multiple LNA insertions can be placed in an oligo at any position except the 3'-end. Applications have been described ranging from antisense oligos to hybridization probes to SNP detection and allele specific PCR. Due to the large increase in Tm conferred by LNAs, they also can cause an increase in primer dimer formation as well as self-hairpin formation. In some cases, the number of LNAs incorporated into a single oligo is 10 bases or less.

[00218] The phosphorothioate (PS) bond (i.e., a phosphorothioate linkage) substitutes a sulfur atom for a non-bridging oxygen in the phosphate backbone of a nucleic acid (e.g., an oligo). This modification renders the internucleotide linkage resistant to nuclease degradation. Phosphorothioate bonds can be introduced between the last 3-5 nucleotides at the 5'- or 3'-end of the oligo to inhibit exonuclease degradation. Including phosphorothioate bonds within the oligo (e.g., throughout the entire oligo) can help reduce attack by endonucleases as well.

[00219] In some embodiments, a subject nucleic acid has one or more nucleotides that are 2'-O-Methyl modified nucleotides. In some embodiments, a subject nucleic acid (e.g., a dsRNA, a siNA, etc.) has one or more 2' Fluoro modified nucleotides. In some embodiments, a subject nucleic acid (e.g., a dsRNA, a siNA, etc.) has one or more LNA bases. In some embodiments, a subject nucleic acid (e.g., a dsRNA, a siNA, etc.) has one or more nucleotides that are linked by a phosphorothioate bond (i.e., the subject nucleic acid has one or more phosphorothioate linkages). In some embodiments, a subject nucleic acid (e.g., a dsRNA, a siNA, etc.) has a 5' cap (e.g., a 7-methylguanylate cap (m7G)). In some embodiments, a subject nucleic acid (e.g., a dsRNA, a siNA, etc.) has a combination of modified nucleotides. For example, a subject nucleic acid (e.g., a dsRNA, a siNA, etc.) can have a 5' cap (e.g., a 7-methylguanylate cap (m7G)) in addition to having one or more nucleotides with other modifications (e.g., a 2'-O-Methyl nucleotide and/or a 2' Fluoro modified nucleotide and/or a LNA base and/or a phosphorothioate linkage).

Modified backbones and modified internucleoside linkages

[00220] Examples of suitable nucleic acids (e.g., a CasY guide RNA) 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.

[00221] 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.

[00222] 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, the disclosure of which is incorporated herein by reference in its entirety. Suitable amide internucleoside linkages are disclosed in U.S. Pat. No. 5,602,240, the disclosure of which is incorporated herein by reference in its entirety.

[00223] 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.

[00224] 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

[00225] 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.

[00226] 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, the disclosures of which are incorporated herein by reference in their entirety.

[00227] 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, the disclosure of which is incorporated herein by reference in its entirety. A variety of compounds within the morpholino class of polynucleotides have been prepared, having a variety of different linking groups joining the monomeric subunits.

[00228] 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, the disclosure of which is incorporated herein by reference in its entirety). 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.

[00229] 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, the disclosure of which is incorporated herein by reference in its entirety). 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 (e.g., Wahlestedt et al., *Proc. Natl. Acad. Sci.* U.S.A., 2000, 97, 5633-5638, the disclosure of which is incorporated herein by reference in its entirety).

[00230] 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 (e.g., Koshkin et al., *Tetrahedron*, 1998, 54, 3607-3630, the disclosure of which is incorporated herein by reference in its entirety). LNAs and preparation thereof are also described in WO 98/39352 and WO 99/14226, as well as U.S. applications 20120165514, 20100216983, 20090041809, 20060117410, 20040014959, 20020094555, and 20020086998, the disclosures of which are incorporated herein by reference in their entirety.

Modified sugar moieties

[00231] 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, the disclosure of which is incorporated herein by reference in its entirety) 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-$N(CH_3)_2$.

[00232] 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

[00233] 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).

[00234] 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; the disclosures of which are incorporated herein by reference in their entirety. 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; the disclosure of which is incorporated herein by reference in its entirety) and are suitable base substitutions, e.g., when combined with 2'-O-methoxyethyl sugar modifications.

Conjugates

[00235] 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.

[00236] 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).

[00237] 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 (e.g., the nucleus). In some embodiments, a PTD is covalently linked to the 3' end of an exogenous polynucleotide. In some embodiments, a PTD is covalently linked to the 5' end of an exogenous polynucleotide. 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:112); 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:113); Transportan GWTLNSAGYLLGKINLKALAALAKKIL SEQ ID NO:114); KALAWEAKLAKALAKALAKHLAKALAKALKCEA SEQ ID NO:115); and RQIKIWFQNRRMKWKK SEQ ID NO:116). Exemplary PTDs include but are not limited to, YGRKKRRQRRR SEQ ID NO:117), RKKRRQRRR SEQ ID NO:118); 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:119); RKKRRQRR SEQ ID NO:120); YARAAARQARA SEQ ID NO:121); THRLPRRRRRR SEQ ID NO:122); and GGRRARRRRRR SEQ ID NO:123). 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.

Introducing components into a target cell

[00238] A CasY guide RNA (or a nucleic acid comprising a nucleotide sequence encoding same) and/or a CasY polypeptide of the present disclosure (or a nucleic acid comprising a nucleotide sequence encoding same) and/or a CasY fusion polypeptide of the present disclosure (or a nucleic acid that includes a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure) and/or a donor polynucleotide (donor template) can be introduced into a host cell by any of a variety of well-known methods.

[00239] Any of a variety of compounds and methods can be used to deliver to a target cell a CasY system of the present disclosure (e.g., where a CasY system comprises: a) a CasY polypeptide of the present disclosure and a CasY guide RNA; b) a CasY polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; c) a CasY fusion polypeptide of the present disclosure and a CasY guide RNA; d) a CasY fusion polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; e) an mRNA encoding a CasY polypeptide of the present disclosure; and a CasY guide RNA; f) an mRNA encoding a CasY polypeptide of the present disclosure, a CasY guide RNA, and a donor templat nucleic acid; g) an mRNA encoding a CasY fusion polypeptide of the present disclosure; and a CasY guide RNA; h) an mRNA encoding a CasY fusion polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; i) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure and a nucleotide sequence encoding a CasY guide RNA; j) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, a nucleotide sequence encoding a CasY guide RNA, and a nucleotide sequence encoding a donor template nucleic acid; k) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure and a nucleotide sequence encoding a CasY guide RNA; l) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, a nucleotide sequence encoding a CasY guide RNA, and a nucleotide sequence encoding a donor template nucleic acid; m) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; n) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; and a donor template nucleic acid; o) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, and a second recombinant expression vector comprising a

nucleotide sequence encoding a CasY guide RNA; p) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; and a donor template nucleic acid; q) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, a nucleotide sequence encoding a first CasY guide RNA, and a nucleotide sequence encoding a second CasY guide RNA; or r) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, a nucleotide sequence encoding a first CasY guide RNA, and a nucleotide sequence encoding a second CasY guide RNA; or some variation of one of (a) through (r). As a non-limiting example, a CasY system of the present disclosure can be combined with a lipid. As another non-limiting example, a CasY system of the present disclosure can be combined with a particle, or formulated into a particle.

[00240] Methods of introducing a nucleic acid into a host cell are known in the art, and any convenient method can be used to introduce a subject nucleic acid (e.g., an expression construct/vector) into a target cell (e.g., prokaryotic cell, eukaryotic cell, plant cell, animal cell, mammalian cell, human cell, and the like). Suitable methods include, e.g., viral 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.

[00241] In some cases, a CasY polypeptide of the present disclosure is provided as a nucleic acid (e.g., an mRNA, a DNA, a plasmid, an expression vector, a viral vector, etc.) that encodes the CasY polypeptide. In some cases, the CasY polypeptide of the present disclosure is provided directly as a protein (e.g., without an associated guide RNA or with an associate guide RNA, i.e., as a ribonucleoprotein complex). A CasY polypeptide of the present disclosure can be introduced into a cell (provided to the cell) by any convenient method; such methods are known to those of ordinary skill in the art. As an illustrative example, a CasY polypeptide of the present disclosure can be injected directly into a cell (e.g., with or without a CasY guide RNA or nucleic acid encoding a CasY guide RNA, and with or without a donor polynucleotide). As another example, a preformed complex of a CasY polypeptide of the present disclosure and a CasY guide RNA (an RNP) can be introduced into a cell (e.g, eukaryotic cell) (e.g., via injection, via nucleofection; via a protein transduction domain (PTD) conjugated to one or more components, e.g., conjugated to the CasY protein, conjugated to a guide RNA, conjugated to a CasY polypeptide of the present disclosure and a guide RNA; etc.).

[00242] In some cases, a CasY fusion polypeptide (e.g., dCasY fused to a fusion partner, nickase CasY fused to a fusion partner, etc.) of the present disclosure is provided as a nucleic acid (e.g., an mRNA, a DNA, a plasmid, an expression vector, a viral vector, etc.) that encodes the CasY fusion polypeptide. In some cases, the CasY fusion polypeptide of the present disclosure is provided directly as a protein (e.g., without an associated guide RNA or with an associate guide RNA, i.e., as a ribonucleoprotein complex). A CasY fusion polypeptide of the present disclosure can be introduced into a cell (provided to the cell) by any convenient method; such methods are known to those of ordinary skill in the art. As an illustrative example, a CasY fusion polypeptide of the present disclosure can be injected directly into a cell (e.g., with or without nucleic acid encoding a CasY guide RNA and with or without a donor polynucleotide). As another example, a preformed complex of a CasY fusion polypeptide of the present disclosure and a CasY guide RNA (an RNP) can be introduced into a cell (e.g., via injection, via nucleofection; via a protein transduction domain (PTD) conjugated to one or more components, e.g., conjugated to the CasY fusion protein, conjugated to a guide RNA, conjugated to a CasY fusion polypeptide of the present disclosure and a guide RNA; etc.).

[00243] In some cases, a nucleic acid (e.g., a CasY guide RNA; a nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure; etc.) is delivered to a cell (e.g., a target host cell) and/or a polypeptide (e.g., a CasY polypeptide; a CasY fusion polypeptide) in a particle, or associated with a particle. In some cases, a CasY system of the present disclosure is delivered to a cell in a particle, or associated with a particle. The terms "particle" and nanoparticle" can be used interchangeable, as appropriate. A recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure and/or a CasY guide RNA, an mRNA comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, and guide RNA may be delivered simultaneously using particles or lipid envelopes; for instance, a CasY polypeptide and a CasY guide RNA, e.g., as a complex (e.g., a ribonucleoprotein (RNP) complex), can be delivered via a particle, e.g., a delivery particle comprising lipid or lipidoid and hydrophilic polymer, e.g., a cationic lipid and a hydrophilic polymer, for instance wherein the cationic lipid comprises 1,2-dioleoyl-3-trimethylammonium-propane (DOTAP) or 1,2-ditetradecanoyl-sn-glycero-3-phosphocholine (DMPC) and/or wherein the hydrophilic polymer comprises ethylene glycol or polyethylene glycol (PEG); and/or wherein the particle further comprises cholesterol (e.g., particle from formulation 1=DOTAP 100, DMPC 0, PEG 0, Cholesterol 0; formulation number 2=DOTAP 90, DMPC 0, PEG 10, Cholesterol 0; formulation number 3=DOTAP 90, DMPC 0, PEG 5, Cholesterol 5). For example, a particle can be formed using a multistep process in which a CasY polypepide and a CasY guideRNA are mixed together, e.g., at a 1:1 molar ratio, e.g., at room temperature, e.g., for 30 minutes, e.g., in sterile, nuclease free 1 x phosphate-buffered saline (PBS); and separately, DOTAP, DMPC, PEG, and cholesterol as applicable for

the formulation are dissolved in alcohol, e.g., 100% ethanol; and, the two solutions are mixed together to form particles containing the complexes).

[00244] A CasY polypeptide of the present disclosure (or an mRNA comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure; or a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure) and/or CasY guide RNA (or a nucleic acid such as one or more expression vectors encoding the CasY guide RNA) may be delivered simultaneously using particles or lipid envelopes. For example, a biodegradable core-shell structured nanoparticle with a poly (β-amino ester) (PBAE) core enveloped by a phospholipid bilayer shell can be used. In some cases, particles/nanoparticles based on self assembling bioadhesive polymers are used; such particles/nanoparticles may be applied to oral delivery of peptides, intravenous delivery of peptides and nasal delivery of peptides, e.g., to the brain. Other embodiments, such as oral absorption and ocular delivery of hydrophobic drugs are also contemplated. A molecular envelope technology, which involves an engineered polymer envelope which is protected and delivered to the site of the disease, can be used. Doses of about 5 mg/kg can be used, with single or multiple doses, depending on various factors, e.g., the target tissue.

[00245] Lipidoid compounds (e.g., as described in US patent application 20110293703) are also useful in the administration of polynucleotides, and can be used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure (e.g., where a CasY system comprises: a) a CasY polypeptide of the present disclosure and a CasY guide RNA; b) a CasY polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; c) a CasY fusion polypeptide of the present disclosure and a CasY guide RNA; d) a CasY fusion polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; e) an mRNA encoding a CasY polypeptide of the present disclosure; and a CasY guide RNA; f) an mRNA encoding a CasY polypeptide of the present disclosure, a CasY guide RNA, and a donor templat nucleic acid; g) an mRNA encoding a CasY fusion polypeptide of the present disclosure; and a CasY guide RNA; h) an mRNA encoding a CasY fusion polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; i) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure and a nucleotide sequence encoding a CasY guide RNA; j) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, a nucleotide sequence encoding a CasY guide RNA, and a nucleotide sequence encoding a donor template nucleic acid; k) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure and a nucleotide sequence encoding a CasY guide RNA; l) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of

the present disclosure, a nucleotide sequence encoding a CasY guide RNA, and a nucleotide sequence encoding a donor template nucleic acid; m) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; n) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; and a donor template nucleic acid; o) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; p) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; and a donor template nucleic acid; q) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, a nucleotide sequence encoding a first CasY guide RNA, and a nucleotide sequence encoding a second CasY guide RNA; or r) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, a nucleotide sequence encoding a first CasY guide RNA, and a nucleotide sequence encoding a second CasY guide RNA; or some variation of one of (a) through (r). In one aspect, the aminoalcohol lipidoid compounds are combined with an agent to be delivered to a cell or a subject to form microparticles, nanoparticles, liposomes, or micelles. The aminoalcohol lipidoid compounds may be combined with other aminoalcohol lipidoid compounds, polymers (synthetic or natural), surfactants, cholesterol, carbohydrates, proteins, lipids, etc. to form the particles. These particles may then optionally be combined with a pharmaceutical excipient to form a pharmaceutical composition.

[00246] A poly(beta-amino alcohol) (PBAA) can be used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell. US Patent Publication No. 20130302401 relates to a class of poly(beta-amino alcohols) (PBAAs) that has been prepared using combinatorial polymerization.

[00247] Sugar-based particles may be used, for example GalNAc, as described with reference to WO2014118272 (incorporated herein by reference) and Nair, J K et al., 2014, Journal of the American Chemical Society 136 (49), 16958-16961) can be used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell.

[00248] In some cases, lipid nanoparticles (LNPs) are used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure,

a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell. Negatively charged polymers such as RNA may be loaded into LNPs at low pH values (e.g., pH 4) where the ionizable lipids display a positive charge. However, at physiological pH values, the LNPs exhibit a low surface charge compatible with longer circulation times. Four species of ionizable cationic lipids have been focused upon, namely 1,2-dilineoyl-3-dimethylammonium-propane (DLinDAP), 1,2-dilinoleyloxy-3-N,N-dimethylaminopropane (DLinDMA), 1,2-dilinoleyloxy-keto-N,N-dimethyl-3-aminopropane (DLinKDMA), and 1,2-dilinoleyl-4-(2-dimethylaminoethyl)-[1,3]-dioxolane (DLinKC2-DMA). Preparation of LNPs and is described in, e.g., Rosin et al. (2011) Molecular Therapy 19:1286-2200). The cationic lipids 1,2-dilineoyl-3-dimethylammonium-propane (DLinDAP), 1,2-dilinoleyloxy-3-N,N-dimethylaminopropane (DLinDMA), 1,2-dilinoleyloxyketo-N,N-dimethyl-3-aminopropane (DLinK-DMA), 1,2-dilinoleyl-4-(2-dimethylaminoethyl)-[1,3]-dioxolane (DLinKC2-DMA), (3-o-[2"-(methoxypolyethyleneglycol 2000) succinoyl]-1,2-dimyristoyl-sn-glycol (PEG-S-DMG), and R-3-[(.omega.-methoxy-poly(ethylene glycol)2000) carbamoyl]-1,2-dimyristyloxlpropyl-3-amine (PEG-C-DOMG) may be used. A nucleic acid (e.g., a CasY guide RNA; a nucleic acid of the present disclosure; etc.) may be encapsulated in LNPs containing DLinDAP, DLinDMA, DLinK-DMA, and DLinKC2-DMA (cationic lipid:DSPC:CHOL: PEGS-DMG or PEG-C-DOMG at 40:10:40:10 molar ratios). In some cases, 0.2% SP-DiOC18 is incorporated.

[00249] Spherical Nucleic Acid (SNA™) constructs and other nanoparticles (particularly gold nanoparticles) can be used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell.. See, e.g., Cutler et al., J. Am. Chem. Soc. 2011 133:9254-9257, Hao et al., Small. 2011 7:3158-3162, Zhang et al., ACS Nano. 2011 5:6962-6970, Cutler et al., J. Am. Chem. Soc. 2012 134:1376-1391, Young et al., Nano Lett. 2012 12:3867-71, Zheng et al., Proc. Natl. Acad. Sci. USA. 2012 109:11975-80, Mirkin, Nanomedicine 2012 7:635-638 Zhang et al., J. Am. Chem. Soc. 2012 134:16488-1691, Weintraub, Nature 2013 495:S14-S16, Choi et al., Proc. Natl. Acad. Sci. USA. 2013 110(19): 7625-7630, Jensen et al., Sci. Transl. Med. 5, 209ra152 (2013) and Mirkin, et al., Small, 10:186-192.

[00250] Self-assembling nanoparticles with RNA may be constructed with polyethyleneimine (PEI) that is PEGylated with an Arg-Gly-Asp (RGD) peptide ligand attached at the distal end of the polyethylene glycol (PEG).

[00251] In general, a "nanoparticle" refers to any particle having a diameter of less than 1000 nm. In some cases, nanoparticles suitable for use in delivering a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell have a diameter of 500 nm

or less, e.g., from 25 nm to 35 nm, from 35 nm to 50 nm, from 50 nm to 75 nm, from 75 nm to 100 nm, from 100 nm to 150 nm, from 150 nm to 200 nm, from 200 nm to 300 nm, from 300 nm to 400 nm, or from 400 nm to 500 nm. In some cases, nanoparticles suitable for use in delivering a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell have a diameter of from 25 nm to 200 nm. In some cases, nanoparticles suitable for use in delivering a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell have a diameter of 100 nm or less In some cases, nanoparticles suitable for use in delivering a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell have a diameter of from 35 nm to 60 nm.

[00252] Nanoparticles suitable for use in delivering a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell may be provided in different forms, e.g., as solid nanoparticles (e.g., metal such as silver, gold, iron, titanium), non-metal, lipid-based solids, polymers), suspensions of nanoparticles, or combinations thereof. Metal, dielectric, and semiconductor nanoparticles may be prepared, as well as hybrid structures (e.g., core-shell nanoparticles). Nanoparticles made of semiconducting material may also be labeled quantum dots if they are small enough (typically below 10 nm) that quantization of electronic energy levels occurs. Such nanoscale particles are used in biomedical applications as drug carriers or imaging agents and may be adapted for similar purposes in the present disclosure.

[00253] Semi-solid and soft nanoparticles are also suitable for use in delivering a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell. A prototype nanoparticle of semi-solid nature is the liposome.

[00254] In some cases, an exosome is used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell. Exosomes are endogenous nano-vesicles that transport RNAs and proteins, and which can deliver RNA to the brain and other target organs.

[00255] In some cases, a liposome is used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell. Liposomes are

spherical vesicle structures composed of a uni- or multilamellar lipid bilayer surrounding internal aqueous compartments and a relatively impermeable outer lipophilic phospholipid bilayer. Liposomes can be made from several different types of lipids; however, phospholipids are most commonly used to generate liposomes. Although liposome formation is spontaneous when a lipid film is mixed with an aqueous solution, it can also be expedited by applying force in the form of shaking by using a homogenizer, sonicator, or an extrusion apparatus. Several other additives may be added to liposomes in order to modify their structure and properties. For instance, either cholesterol or sphingomyelin may be added to the liposomal mixture in order to help stabilize the liposomal structure and to prevent the leakage of the liposomal inner cargo. A liposome formulation may be mainly comprised of natural phospholipids and lipids such as 1,2-distearoryl-sn-glycero-3-phosphatidyl choline (DSPC), sphingomyelin, egg phosphatidylcholines and monosialoganglioside.

[00256] A stable nucleic-acid-lipid particle (SNALP) can be used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell. The SNALP formulation may contain the lipids 3-N-[(methoxypoly(ethylene glycol) 2000) carbamoyl]-1,2-dimyristyloxy-propylamine (PEG-C-DMA), 1,2-dilinoleyloxy-N,N-dimethyl-3-aminopropane (DLinDMA), 1,2-distearoyl-sn-glycero-3-phosphocholine (DSPC) and cholesterol, in a 2:40:10:48 molar percent ratio. The SNALP liposomes may be prepared by formulating D-Lin-DMA and PEG-C-DMA with distearoylphosphatidylcholine (DSPC), Cholesterol and siRNA using a 25:1 lipid/siRNA ratio and a 48/40/10/2 molar ratio of Cholesterol/D-Lin-DMA/DSPC/PEG-C-DMA. The resulting SNALP liposomes can be about 80-100 nm in size. A SNALP may comprise synthetic cholesterol (Sigma-Aldrich, St Louis, Mo., USA), dipalmitoylphosphatidylcholine (Avanti Polar Lipids, Alabaster, Ala., USA), 3-N-[(w-methoxy poly(ethylene glycol)2000)carbamoyl]-1,2-dimyrestyloxypropylamine, and cationic 1,2-dilinoleyloxy-3-N,Ndimethylaminopropane. A SNALP may comprise synthetic cholesterol (Sigma-Aldrich), 1,2-distearoyl-sn-glycero-3-phosphocholine (DSPC; Avanti Polar Lipids Inc.), PEG-cDMA, and 1,2-dilinoleyloxy-3-(N;N-dimethyl)aminopropane (DLinDMA).

[00257] Other cationic lipids, such as amino lipid 2,2-dilinoleyl-4-dimethylaminoethyl-[1,3]-dioxolane (DLin-KC2-DMA) can be used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell. A preformed vesicle with the following lipid composition may be contemplated: amino lipid, distearoylphosphatidylcholine (DSPC), cholesterol and (R)-2,3-bis(octadecyloxy) propyl-1-(methoxy poly(ethylene glycol)2000)propylcarbamate (PEG-lipid) in the molar ratio 40/10/40/10, respectively, and a FVII

siRNA/total lipid ratio of approximately 0.05 (w/w). To ensure a narrow particle size distribution in the range of 70-90 nm and a low polydispersity index of 0.11.+-.0.04 (n=56), the particles may be extruded up to three times through 80 nm membranes prior to adding the guide RNA. Particles containing the highly potent amino lipid 16 may be used, in which the molar ratio of the four lipid components 16, DSPC, cholesterol and PEG-lipid (50/10/38.5/1.5) which may be further optimized to enhance in vivo activity.

[00258] Lipids may be formulated with a CasY system of the present disclosure or component(s) thereof or nucleic acids encoding the same to form lipid nanoparticles (LNPs). Suitable lipids include, but are not limited to, DLin-KC2-DMA4, C12-200 and colipids disteroylphosphatidyl choline, cholesterol, and PEG-DMG may be formulated with a CasY system, or component thereof, of the present disclosure, using a spontaneous vesicle formation procedure. The component molar ratio may be about 50/10/38.5/1.5 (DLin-KC2-DMA or C12-200/disteroylphosphatidyl choline/cholesterol/PEG-DMG).

[00259] A CasY system of the present disclosure, or a component thereof, may be delivered encapsulated in PLGA microspheres such as that further described in US published applications 20130252281 and 20130245107 and 20130244279.

[00260] Supercharged proteins can be used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell. Supercharged proteins are a class of engineered or naturally occurring proteins with unusually high positive or negative net theoretical charge. Both supernegatively and superpositively charged proteins exhibit the ability to withstand thermally or chemically induced aggregation. Superpositively charged proteins are also able to penetrate mammalian cells. Associating cargo with these proteins, such as plasmid DNA, RNA, or other proteins, can enable the functional delivery of these macromolecules into mammalian cells both in vitro and in vivo.

[00261] Cell Penetrating Peptides (CPPs) can be used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell. CPPs typically have an amino acid composition that either contains a high relative abundance of positively charged amino acids such as lysine or arginine or has sequences that contain an alternating pattern of polar/charged amino acids and non-polar, hydrophobic amino acids.

[00262] An implantable device can be used to deliver a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure (e.g., a CasY guide RNA, a nucleic acid encoding a CasY guide RNA, a nucleic acid encoding CasY polypeptide, a donor template, and the like), or a CasY system of the

present disclosure, to a target cell (e.g., a target cell *in vivo*, where the target cell is a target cell in circulation, a target cell in a tissue, a target cell in an organ, etc.). An implantable device suitable for use in delivering a CasY polypeptide of the present disclosure, a CasY fusion polypeptide of the present disclosure, an RNP of the present disclosure, a nucleic acid of the present disclosure, or a CasY system of the present disclosure, to a target cell (e.g., a target cell *in vivo*, where the target cell is a target cell in circulation, a target cell in a tissue, a target cell in an organ, etc.) can include a container (e.g., a reservoir, a matrix, etc.) that comprises the CasY polypeptide, the CasY fusion polypeptide, the RNP, or the CasY system (or component thereof, e.g., a nucleic acid of the present disclosure).

[00263] A suitable implantable device can comprise a polymeric substrate, such as a matrix for example, that is used as the device body, and in some cases additional scaffolding materials, such as metals or additional polymers, and materials to enhance visibility and imaging. An implantable delivery device can be advantageous in providing release locally and over a prolonged period, where the polypeptide and/or nucleic acid to be delivered is released directly to a target site, e.g., the extracellular matrix (ECM), the vasculature surrounding a tumor, a diseased tissue, etc. Suitable implantable delivery devices include devices suitable for use in delivering to a cavity such as the abdominal cavity and/or any other type of administration in which the drug delivery system is not anchored or attached, comprising a biostable and/or degradable and/or bioabsorbable polymeric substrate, which may for example optionally be a matrix. In some cases, a suitable implantable drug delivery device comprises degradable polymers, wherein the main release mechanism is bulk erosion. In some cases, a suitable implantable drug delivery device comprises non degradable, or slowly degraded polymers, wherein the main release mechanism is diffusion rather than bulk erosion, so that the outer part functions as membrane, and its internal part functions as a drug reservoir, which practically is not affected by the surroundings for an extended period (for example from about a week to about a few months). Combinations of different polymers with different release mechanisms may also optionally be used. The concentration gradient at the can be maintained effectively constant during a significant period of the total releasing period, and therefore the diffusion rate is effectively constant (termed "zero mode" diffusion). By the term "constant" it is meant a diffusion rate that is maintained above the lower threshold of therapeutic effectiveness, but which may still optionally feature an initial burst and/or may fluctuate, for example increasing and decreasing to a certain degree. The diffusion rate can be so maintained for a prolonged period, and it can be considered constant to a certain level to optimize the therapeutically effective period, for example the effective silencing period.

[00264] In some cases, the implantable delivery system is designed to shield the nucleotide based therapeutic agent from degradation, whether chemical in nature or due to attack from enzymes and other factors in the body of the subject.

[00265] The site for implantation of the device, or target site, can be selected for maximum therapeutic efficacy. For example, a delivery device can be implanted within or in the proximity of a tumor environment, or the blood supply associated with a tumor. The target location can be, e.g.: 1) the brain at degenerative sites like in Parkinson or Alzheimer disease at the basal ganglia, white and gray matter; 2) the spine, as in the case of amyotrophic lateral sclerosis (ALS); 3) uterine cervix; 4) active and chronic inflammatory joints; 5) dermis as in the case of psoriasis; 7) sympathetic and sensoric nervous sites for analgesic effect; 7) a bone; 8) a site of acute or chronic infection; 9) Intra vaginal; 10) Inner ear--auditory system, labyrinth of the inner ear, vestibular system; 11) Intra tracheal; 12) Intra-cardiac; coronary, epicardiac; 13) urinary tract or bladder; 14) biliary system; 15) parenchymal tissue including and not limited to the kidney, liver, spleen; 16) lymph nodes; 17) salivary glands; 18) dental gums; 19) Intra-articular (into joints); 20) Intra-ocular; 21) Brain tissue; 22) Brain ventricles; 23) Cavities, including abdominal cavity (for example but without limitation, for ovary cancer); 24) Intra esophageal; and 25) Intra rectal; and 26) into the vasculature.

[00266] The method of insertion, such as implantation, may optionally already be used for other types of tissue implantation and/or for insertions and/or for sampling tissues, optionally without modifications, or alternatively optionally only with non-major modifications in such methods. Such methods optionally include but are not limited to brachytherapy methods, biopsy, endoscopy with and/or without ultrasound, such as stereotactic methods into the brain tissue, laparoscopy, including implantation with a laparoscope into joints, abdominal organs, the bladder wall and body cavities.

MODIFIED HOST CELLS

[00267] The present disclosure provides a modified cell comprising a CasY polypeptide of the present disclosure and/or a nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure. The present disclosure provides a modified cell comprising a CasY polypeptide of the present disclosure, where the modified cell is a cell that does not normally comprise a CasY polypeptide of the present disclosure. The present disclosure provides a modified cell (e.g., a genetically modified cell) comprising nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure. The present disclosure provides a genetically modified cell that is genetically modified with an mRNA comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure. The present disclosure provides a genetically modified cell that is genetically modified with a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure. The present disclosure provides a genetically modified cell that is genetically modified with a recombinant expression vector comprising: a) a nucleotide sequence encoding a CasY polypeptide of the present disclosure; and b) a nucleotide sequence encoding a CasY guide RNA of the present disclosure. The present disclosure provides a genetically modified cell that is genetically modified with a

recombinant expression vector comprising: a) a nucleotide sequence encoding a CasY polypeptide of the present disclosure; b) a nucleotide sequence encoding a CasY guide RNA of the present disclosure; and c) a nucleotide sequence encoding a donor template.

[00268] A cell that serves as a recipient for a CasY polypeptide of the present disclosure and/or a nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure and/or a CasY guide RNA of the present disclosure, can be any of a variety of cells, including, e.g., *in vitro* cells; *in vivo* cells; *ex vivo* cells; primary cells; cancer cells; animal cells; plant cells; algal cells; fungal cells; etc. A cell that serves as a recipient for a CasY polypeptide of the present disclosure and/or a nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure and/or a CasY guide RNA of the present disclosure is referred to as a "host cell" or a "target cell." A host cell or a target cell can be a recipient of a CasY system of the present disclosure. A host cell or a target cell can be a recipient of a CasY RNP of the present disclosure. A host cell or a target cell can be a recipient of a single component of a CasY system of the present disclosure.

[00269] Non-limiting examples of cells (target cells) include: a prokaryotic cell, eukaryotic cell, a bacterial cell, an archaeal cell, a cell of a single-cell eukaryotic organism, a protozoa cell, a cell from a plant (e.g., cells from plant crops, fruits, vegetables, grains, soy bean, corn, maize, wheat, seeds, tomatos, rice, cassava, sugarcane, pumpkin, hay, potatos, cotton, cannabis, tobacco, flowering plants, conifers, gymnosperms, angiosperms, ferns, clubmosses, hornworts, liverworts, mosses, dicotyledons, monocotyledons, etc.), an algal cell, (e.g., *Botryococcus braunii*, *Chlamydomonas reinhardtii*, *Nannochloropsis gaditana*, *Chlorella pyrenoidosa*, *Sargassum patens*, *C. agardh*, and the like), seaweeds (e.g. kelp) a fungal cell (e.g., a yeast cell, a cell from a mushroom), 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 (e.g., an ungulate (e.g., a pig, a cow, a goat, a sheep); a rodent (e.g., a rat, a mouse); a non-human primate; a human; a feline (e.g., a cat); a canine (e.g., a dog); etc.), and the like. In some cases, the cell is a cell that does not originate from a natural organism (e.g., the cell can be a synthetically made cell; also referred to as an artificial cell).

[00270] A cell can be an *in vitro* cell (e.g., established cultured cell line). A cell can be an *ex vivo* cell (cultured cell from an individual). A cell can be and *in vivo* cell (e.g., a cell in an individual). A cell can be an isolated cell. A cell can be a cell inside of an organism. A cell can be an organism. A cell can be a cell in a cell culture (e.g., *in vitro* cell culture). A cell can be one of a collection of cells. A cell can be a prokaryotic cell or derived from a prokaryotic cell. A cell can be a bacterial cell or can be derived from a bacterial cell. A cell can be an archaeal cell or derived from an archaeal cell. A cell can be a eukaryotic cell or derived from a eukaryotic cell. A cell can be a plant cell or derived from a plant cell. A cell can be an animal cell or derived from an animal cell. A cell can be an invertebrate cell or derived from an

invertebrate cell. A cell can be a vertebrate cell or derived from a vertebrate cell. A cell can be a mammalian cell or derived from a mammalian cell. A cell can be a rodent cell or derived from a rodent cell. A cell can be a human cell or derived from a human cell. A cell can be a microbe cell or derived from a microbe cell. A cell can be a fungi cell or derived from a fungi cell. A cell can be an insect cell. A cell can be an arthropod cell. A cell can be a protozoan cell. A cell can be a helminth cell.

[00271] Suitable cells include a stem cell (e.g. an embryonic stem (ES) cell, an induced pluripotent stem (iPS) cell; a germ cell (e.g., an oocyte, a sperm, an oogonia, a spermatogonia, etc.); a somatic cell, e.g. a fibroblast, an oligodendrocyte, a glial cell, a hematopoietic cell, a neuron, a muscle cell, a bone cell, a hepatocyte, a pancreatic cell, etc.

[00272] Suitable cells include human embryonic stem cells, fetal cardiomyocytes, myofibroblasts, mesenchymal stem cells, autotransplated expanded cardiomyocytes, adipocytes, totipotent cells, pluripotent cells, blood stem cells, myoblasts, adult stem cells, bone marrow cells, mesenchymal cells, embryonic stem cells, parenchymal cells, epithelial cells, endothelial cells, mesothelial cells, fibroblasts, osteoblasts, chondrocytes, exogenous cells, endogenous cells, stem cells, hematopoietic stem cells, bone-marrow derived progenitor cells, myocardial cells, skeletal cells, fetal cells, undifferentiated cells, multi-potent progenitor cells, unipotent progenitor cells, monocytes, cardiac myoblasts, skeletal myoblasts, macrophages, capillary endothelial cells, xenogenic cells, allogenic cells, and post-natal stem cells.

[00273] In some cases, the cell is an immune cell, a neuron, an epithelial cell, and endothelial cell, or a stem cell. In some cases, the immune cell is a T cell, a B cell, a monocyte, a natural killer cell, a dendritic cell, or a macrophage. In some cases, the immune cell is a cytotoxic T cell. In some cases, the immune cell is a helper T cell. In some cases, the immune cell is a regulatory T cell (Treg).

[00274] In some cases, the cell is a stem cell. Stem cells include adult stem cells. Adult stem cells are also referred to as somatic stem cells.

[00275] Adult stem cells are resident in differentiated tissue, but retain the properties of self-renewal and ability to give rise to multiple cell types, usually cell types typical of the tissue in which the stem cells are found. Numerous examples of somatic stem cells are known to those of skill in the art, including muscle stem cells; hematopoietic stem cells; epithelial stem cells; neural stem cells; mesenchymal stem cells; mammary stem cells; intestinal stem cells; mesodermal stem cells; endothelial stem cells; olfactory stem cells; neural crest stem cells; and the like.

[00276] Stem cells of interest include mammalian stem cells, where the term "mammalian" refers to any animal classified as a mammal, including humans; non-human primates; domestic and farm animals; and zoo, laboratory, sports, or pet animals, such as dogs, horses, cats, cows, mice, rats, rabbits, etc. In some cases, the stem cell is a human stem cell. In some cases, the stem cell is a rodent (e.g., a mouse; a rat) stem cell. In some cases, the stem cell is a non-human primate stem cell.

[00277] Stem cells can express one or more stem cell markers, e.g., SOX9, KRT19, KRT7, LGR5, CA9, FXYD2, CDH6, CLDN18, TSPAN8, BPIFB1, OLFM4, CDH17, and PPARGC1A.

[00278] In some embodiments, the stem cell is a hematopoietic stem cell (HSC). HSCs are mesoderm-derived cells that can be isolated from bone marrow, blood, cord blood, fetal liver and yolk sac. HSCs are characterized as $CD34^+$ and $CD3^-$. HSCs can repopulate the erythroid, neutrophil-macrophage, megakaryocyte and lymphoid hematopoietic cell lineages *in vivo*. *In vitro*, HSCs can be induced to undergo at least some self-renewing cell divisions and can be induced to differentiate to the same lineages as is seen *in vivo*. As such, HSCs can be induced to differentiate into one or more of erythroid cells, megakaryocytes, neutrophils, macrophages, and lymphoid cells.

[00279] In other embodiments, the stem cell is a neural stem cell (NSC). Neural stem cells (NSCs) are capable of differentiating into neurons, and glia (including oligodendrocytes, and astrocytes). A neural stem cell is a multipotent stem cell which is capable of multiple divisions, and under specific conditions can produce daughter cells which are neural stem cells, or neural progenitor cells that can be neuroblasts or glioblasts, e.g., cells committed to become one or more types of neurons and glial cells respectively. Methods of obtaining NSCs are known in the art.

[00280] In other embodiments, the stem cell is a mesenchymal stem cell (MSC). MSCs originally derived from the embryonal mesoderm and isolated from adult bone marrow, can differentiate to form muscle, bone, cartilage, fat, marrow stroma, and tendon. Methods of isolating MSC are known in the art; and any known method can be used to obtain MSC. See, e.g., U.S. Pat. No. 5,736,396, which describes isolation of human MSC.

[00281] A cell is in some cases a plant cell. A plant cell can be a cell of a monocotyledon. A cell can be a cell of a dicotyledon.

[00282] In some cases, the cell is a plant cell. For example, the cell can be a cell of a major agricultural plant, e.g., Barley, Beans (Dry Edible), Canola, Corn, Cotton (Pima), Cotton (Upland), Flaxseed, Hay (Alfalfa), Hay (Non-Alfalfa), Oats, Peanuts, Rice, Sorghum, Soybeans, Sugarbeets, Sugarcane, Sunflowers (Oil), Sunflowers (Non-Oil), Sweet Potatoes , Tobacco (Burley), Tobacco (Flue-cured), Tomatoes, Wheat (Durum), Wheat (Spring), Wheat (Winter), and the like. As another example, the cell is a cell of a vegetable crops which include but are not limited to, e.g., alfalfa sprouts, aloe leaves, arrow root, arrowhead, artichokes, asparagus, bamboo shoots, banana flowers, bean sprouts, beans, beet tops, beets, bittermelon, bok choy, broccoli, broccoli rabe (rappini), brussels sprouts, cabbage, cabbage sprouts, cactus leaf (nopales), calabaza, cardoon, carrots, cauliflower, celery, chayote, chinese artichoke (crosnes), chinese cabbage, chinese celery, chinese chives, choy sum, chrysanthemum leaves (tung ho), collard greens, corn stalks, corn-sweet, cucumbers, daikon, dandelion greens, dasheen, dau mue (pea tips), donqua (winter melon), eggplant, endive, escarole, fiddle head ferns, field cress, frisee, gai choy

(chinese mustard), gailon, galanga (siam, thai ginger), garlic, ginger root, gobo, greens, hanover salad greens, huauzontle, jerusalem artichokes, jicama, kale greens, kohlrabi, lamb's quarters (quilete), lettuce (bibb), lettuce (boston), lettuce (boston red), lettuce (green leaf), lettuce (iceberg), lettuce (lolla rossa), lettuce (oak leaf - green), lettuce (oak leaf - red), lettuce (processed), lettuce (red leaf), lettuce (romaine), lettuce (ruby romaine), lettuce (russian red mustard), linkok, lo bok, long beans, lotus root, mache, maguey (agave) leaves, malanga, mesculin mix, mizuna, moap (smooth luffa), moo, moqua (fuzzy squash), mushrooms, mustard, nagaimo, okra, ong choy, onions green, opo (long squash), ornamental corn, ornamental gourds, parsley, parsnips, peas, peppers (bell type), peppers, pumpkins, radicchio, radish sprouts, radishes, rape greens, rape greens, rhubarb, romaine (baby red), rutabagas, salicornia (sea bean), sinqua (angled/ridged luffa), spinach, squash, straw bales, sugarcane, sweet potatoes, swiss chard, tamarindo, taro, taro leaf, taro shoots, tatsoi, tepeguaje (guaje), tindora, tomatillos, tomatoes, tomatoes (cherry), tomatoes (grape type), tomatoes (plum type), tumeric, turnip tops greens, turnips, water chestnuts, yampi, yams (names), yu choy, yuca (cassava), and the like.

[00283] A cell is in some cases an arthropod cell. For example, the cell can be a cell of a sub-order, a family, a sub-family, a group, a sub-group, or a species of, e.g., *Chelicerata*, *Myriapodia*, *Hexipodia*, *Arachnida*, *Insecta*, *Archaeognatha*, *Thysanura*, *Palaeoptera*, *Ephemeroptera*, *Odonata*, *Anisoptera*, *Zygoptera*, *Neoptera*, *Exopterygota*, *Plecoptera* , *Embioptera* , *Orthoptera*, *Zoraptera* , *Dermaptera*, *Dictyoptera*, *Notoptera*, *Grylloblattidae*, *Mantophasmatidae*, *Phasmatodea* , *Blattaria*, *Isoptera*, *Mantodea*, *Parapneuroptera*, *Psocoptera*, *Thysanoptera*, *Phthiraptera*, *Hemiptera*, *Endopterygota or Holometabola* , *Hymenoptera* , *Coleoptera*, *Strepsiptera*, *Raphidioptera*, *Megaloptera*, *Neuroptera* , *Mecoptera* , *Siphonaptera*, *Diptera*, *Trichoptera*, or *Lepidoptera.*

[00284] A cell is in some cases an insect cell. For example, in some cases, the cell is a cell of a mosquito, a grasshopper, a true bug, a fly, a flea, a bee, a wasp, an ant, a louse, a moth, or a beetle.

KITS

[00285] The present disclosure provides a kit comprising a CasY system of the present disclosure, or a component of a CasY system of the present disclosure.

[00286] A kit of the present disclosure can comprise: a) a CasY polypeptide of the present disclosure and a CasY guide RNA; b) a CasY polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; c) a CasY fusion polypeptide of the present disclosure and a CasY guide RNA; d) a CasY fusion polypeptide of the present disclosure, a CasY guide RNA, and a donor template nucleic acid; e) an mRNA encoding a CasY polypeptide of the present disclosure; and a CasY guide RNA; f) an mRNA encoding a CasY polypeptide of the present disclosure, a CasY guide RNA, and a donor templat nucleic acid; g) an mRNA encoding a CasY fusion polypeptide of the present disclosure; and a CasY guide RNA; h) an mRNA encoding a CasY fusion polypeptide of the present disclosure, a

CasY guide RNA, and a donor template nucleic acid; i) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure and a nucleotide sequence encoding a CasY guide RNA; j) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, a nucleotide sequence encoding a CasY guide RNA, and a nucleotide sequence encoding a donor template nucleic acid; k) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure and a nucleotide sequence encoding a CasY guide RNA; l) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, a nucleotide sequence encoding a CasY guide RNA, and a nucleotide sequence encoding a donor template nucleic acid; m) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; n) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; and a donor template nucleic acid; o) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; p) a first recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, and a second recombinant expression vector comprising a nucleotide sequence encoding a CasY guide RNA; and a donor template nucleic acid; q) a recombinant expression vector comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure, a nucleotide sequence encoding a first CasY guide RNA, and a nucleotide sequence encoding a second CasY guide RNA; or r) a recombinant expression vector comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure, a nucleotide sequence encoding a first CasY guide RNA, and a nucleotide sequence encoding a second CasY guide RNA; or some variation of one of (a) through (r).

[00287] A kit of the present disclosure can comprise: a) a component, as described above, of a CasY system of the present disclosure, or can comprise a CasY system of the present disclosure; and b) one or more additional reagents, e.g., i) a buffer; ii) a protease inhibitor; iii) a nuclease inhibitor; iv) a reagent required to develop or visualize a detectable label; v) a positive and/or negative control target DNA; vi) a positive and/or negative control CasY guide RNA; and the like. A kit of the present disclosure can comprise: a) a component, as described above, of a CasY system of the present disclosure, or can comprise a CasY system of the present disclosure; and b) a therapeutic agent.

[00288] A kit of the present disclosure can comprise a recombinant expression vector comprising: a) an insertion site for inserting a nucleic acid comprising a nucleotide sequence encoding a portion of a

CasY guide RNA that hybridizes to a target nucleotide sequence in a target nucleic acid; and b) a nucleotide sequence encoding the CasY-binding portion of a CasY guide RNA. A kit of the present disclosure can comprise a recombinant expression vector comprising: a) an insertion site for inserting a nucleic acid comprising a nucleotide sequence encoding a portion of a CasY guide RNA that hybridizes to a target nucleotide sequence in a target nucleic acid; b) a nucleotide sequence encoding the CasY-binding portion of a CasY guide RNA; and c) a nucleotide sequence encoding a CasY polypeptide of the present disclosure.

UTILITY

[00289] A CasY polypeptide of the present disclosure, or a CasY fusion polypeptide of the present disclosure, finds use in a variety of methods (e.g., in combination with a CasY guide RNA and in some cases further in combination with a donor template). For example, a CasY polypeptide of the present disclosure can be used to (i) modify (e.g., cleave, e.g., nick; methylate; etc.) target nucleic acid (DNA or RNA; single stranded or double stranded); (ii) modulate transcription of a target nucleic acid; (iii) label a target nucleic acid; (iv) bind a target nucleic acid (e.g., for purposes of isolation, labeling, imaging, tracking, etc.); (v) modify a polypeptide (e.g., a histone) associated with a target nucleic acid; and the like. Thus, the present disclosure provides a method of modifying a target nucleic acid. In some cases, a method of the present disclosure for modifying a target nucleic acid comprises contacting the target nucleic acid with: a) a CasY polypeptide of the present disclosure; and b) one or more (e.g., two) CasY guide RNAs. In some cases, a method of the present disclosure for modifying a target nucleic acid comprises contacting the target nucleic acid with: a) a CasY polypeptide of the present disclosure; b) a CasY guide RNA; and c) a donor nucleic acid (e.g, a donor template). In some cases, the contacting step is carried out in a cell *in vitro*. In some cases, the contacting step is carried out in a cell *in vivo*. In some cases, the contacting step is carried out in a cell *ex vivo*.

[00290] Because a method that uses a CasY polypeptide includes binding of the CasY polypeptide to a particular region in a target nucleic acid (by virtue of being targeted there by an associated CasY guide RNA), the methods are generally referred to herein as methods of binding (e.g., a method of binding a target nucleic acid). However, it is to be understood that in some cases, while a method of binding may result in nothing more than binding of the target nucleic acid, in other cases, the method can have different final results (e.g., the method can result in modification of the target nucleic acid, e.g., cleavage/methylation/etc., modulation of transcription from the target nucleic acid; modulation of translation of the target nucleic acid; genome editing; modulation of a protein associated with the target nucleic acid; isolation of the target nucleic acid; etc.).

[00291] For examples of suitable methods, see, for example, Jinek et al., Science. 2012 Aug 17;337(6096):816-21; Chylinski et al., RNA Biol. 2013 May;10(5):726-37; Ma et al., Biomed Res Int.

2013;2013:270805; Hou et al., Proc Natl Acad Sci U S A. 2013 Sep 24;110(39):15644-9; Jinek et al., Elife. 2013;2:e00471; Pattanayak et al., Nat Biotechnol. 2013 Sep;31(9):839-43; Qi et al, Cell. 2013 Feb 28;152(5):1173-83; Wang et al., Cell. 2013 May 9;153(4):910-8; Auer et al., Genome Res. 2013 Oct 31; Chen et al., Nucleic Acids Res. 2013 Nov 1;41(20):e19; Cheng et al., Cell Res. 2013 Oct;23(10):1163-71; Cho et al., Genetics. 2013 Nov;195(3):1177-80; DiCarlo et al., Nucleic Acids Res. 2013 Apr;41(7):4336-43; Dickinson et al., Nat Methods. 2013 Oct;10(10):1028-34; Ebina et al., Sci Rep. 2013;3:2510; Fujii et al, Nucleic Acids Res. 2013 Nov 1;41(20):e187; Hu et al., Cell Res. 2013 Nov;23(11):1322-5; Jiang et al., Nucleic Acids Res. 2013 Nov 1;41(20):e188; Larson et al., Nat Protoc. 2013 Nov;8(11):2180-96; Mali et. at., Nat Methods. 2013 Oct;10(10):957-63; Nakayama et al., Genesis. 2013 Dec;51(12):835-43; Ran et al., Nat Protoc. 2013 Nov;8(11):2281-308; Ran et al., Cell. 2013 Sep 12;154(6):1380-9; Upadhyay et al., G3 (Bethesda). 2013 Dec 9;3(12):2233-8; Walsh et al., Proc Natl Acad Sci U S A. 2013 Sep 24;110(39):15514-5; Xie et al., Mol Plant. 2013 Oct 9; Yang et al., Cell. 2013 Sep 12;154(6):1370-9; and U.S. patents and patent applications: 8,906,616; 8,895,308; 8,889,418; 8,889,356; 8,871,445; 8,865,406; 8,795,965; 8,771,945; 8,697,359; 20140068797; 20140170753; 20140179006; 20140179770; 20140186843; 20140186919; 20140186958; 20140189896; 20140227787; 20140234972; 20140242664; 20140242699; 20140242700; 20140242702; 20140248702; 20140256046; 20140273037; 20140273226; 20140273230; 20140273231; 20140273232; 20140273233; 20140273234; 20140273235; 20140287938; 20140295556; 20140295557; 20140298547; 20140304853; 20140309487; 20140310828; 20140310830; 20140315985; 20140335063; 20140335620; 20140342456; 20140342457; 20140342458; 20140349400; 20140349405; 20140356867; 20140356956; 20140356958; 20140356959; 20140357523; 20140357530; 20140364333; and 20140377868; each of which is hereby incorporated by reference in its entirety.

[00292] For example, the present disclosure provides (but is not limited to) methods of cleaving a target nucleic acid; methods of editing a target nucleic acid; methods of modulating transcription from a target nucleic acid; methods of isolating a target nucleic acid, methods of binding a target nucleic acid, methods of imaging a target nucleic acid, methods of modifying a target nucleic acid, and the like.

[00293] As used herein, the terms/phrases "contact a target nucleic acid" and "contacting a target nucleic acid", for example, with a CasY polypeptide or with a CasY fusion polypeptide, etc., encompass all methods for contacting the target nucleic acid. For example, a CasY polypeptide can be provided to a cell as protein, RNA (encoding the CasY polypeptide), or DNA (encoding the CasY polypeptide); while a CasY guide RNA can be provided as a guide RNA or as a nucleic acid encoding the guide RNA. As such, when, for example, performing a method in a cell (e.g., inside of a cell *in vitro*, inside of a cell *in vivo*, inside of a cell *ex vivo*), a method that includes contacting the target nucleic acid encompasses the introduction into the cell of any or all of the components in their active/final state (e.g., in the form of a protein(s) for CasY polypeptide; in the form of a protein for a CasY fusion polypeptide; in the form of an

RNA in some cases for the guide RNA), and also encompasses the introduction into the cell of one or more nucleic acids encoding one or more of the components (e.g., nucleic acid(s) comprising nucleotide sequence(s) encoding a CasY polypeptide or a CasY fusion polypeptide, nucleic acid(s) comprising nucleotide sequence(s) encoding guide RNA(s), nucleic acid comprising a nucleotide sequence encoding a donor template, and the like). Because the methods can also be performed *in vitro* outside of a cell, a method that includes contacting a target nucleic acid, (unless otherwise specified) encompasses contacting outside of a cell *in vitro*, inside of a cell *in vitro*, inside of a cell *in vivo*, inside of a cell *ex vivo*, etc.

[00294] In some cases, a method of the present disclosure for modifying a target nucleic acid comprises introducing into a target cell a CasY locus, e.g., a nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide as well as nucleotide sequences of about 1 kilobase (kb) to 5 kb in length surrounding the CasY-encoding nucleotide sequence from a cell (e.g., in some cases a cell that in its natural state (the state in which it occurs in nature) comprises a CasY locus) comprising a CasY locus, where the target cell does not normally (in its natural state) comprise a CasY locus. However, one or more spacer sequences, encoding guide sequences for the encoded crRNA(s), can be modified such that one or more target sequences of interest are targeted. Thus, for example, in some cases, a method of the present disclosure for modifying a target nucleic acid comprises introducing into a target cell a CasY locus, e.g., a nucleic acid obtained from a source cell (e.g., in some cases a cell that in its natural state (the state in which it occurs in nature) comprises a CasY locus), where the nucleic acid has a length of from 100 nucleotides (nt) to 5 kb in length (e.g., from 100 nt to 500 nt, from 500 nt to 1 kb, from 1 kb to 1.5 kb, from 1.5 kb to 2 kb, from 2 kb to 2.5 kb, from 2.5 kb to 3 kb, from 3 kb to 3.5 kb, from 3.5 kb to 4 kb, or from 4 kb to 5 kb in length) and comprises a nucleotide sequence encoding a CasY polypeptide. As noted above, in some such cases, one or more spacer sequences, encoding guide sequences for the encoded crRNA(s), can be modified such that one or more target sequences of interest are targeted. In some cases, the method comprises introducing into a target cell: i) a CasY locus; and ii) a donor DNA template. In some cases, the target nucleic acid is in a cell-free composition *in vitro*. In some cases, the target nucleic acid is present in a target cell. In some cases, the target nucleic acid is present in a target cell, where the target cell is a prokaryotic cell. In some cases, the target nucleic acid is present in a target cell, where the target cell is a eukaryotic cell. In some cases, the target nucleic acid is present in a target cell, where the target cell is a mammalian cell. In some cases, the target nucleic acid is present in a target cell, where the target cell is a plant cell.

[00295] In some cases, a method of the present disclosure for modifying a target nucleic acid comprises contacting a target nucleic acid with a CasY polypeptide of the present disclosure, or with a CasY fusion polypeptide of the present disclosure. In some cases, abmethod of the present disclosure for

modifying a target nucleic acid comprises contacting a target nucleic acid with a CasY polypeptide and a CasY guide RNA. In some cases, abmethod of the present disclosure for modifying a target nucleic acid comprises contacting a target nucleic acid with a CasY polypeptide, a first CasY guide RNA, and a second CasY guide RNA In some cases, a method of the present disclosure for modifying a target nucleic acid comprises contacting a target nucleic acid with a CasY polypeptide of the present disclosure and a CasY guide RNA and a donor DNA template.

Target nucleic acids and target cells of interest

[00296] A CasY polypeptide of the present disclosure, or a CasY fusion polypeptide of the present disclosure, when bound to a CasY guide RNA, can bind to a target nucleic acid, and in some cases, can bind to and modify a target nucleic acid. A target nucleic acid can be any nucleic acid (e.g., DNA, RNA), can be double stranded or single stranded, can be any type of nucleic acid (e.g., a chromosome (genomic DNA), derived from a chromosome, chromosomal DNA, plasmid, viral, extracellular, intracellular, mitochondrial, chloroplast, linear, circular, etc.) and can be from any organism (e.g., as long as the CasY guide RNA comprises a nucleotide sequence that hybridizes to a target sequence in a target nucleic acid, such that the target nucleic acid can be targeted).

[00297] A target nucleic acid can be DNA or RNA. A target nucleic acid can be double stranded (e.g., dsDNA, dsRNA) or single stranded (e.g., ssRNA, ssDNA). In some cases, a target nucleic acid is single stranded. In some cases, a target nucleic acid is a single stranded RNA (ssRNA). In some cases, a target ssRNA (e.g., a target cell ssRNA, a viral ssRNA, etc.) is selected from: mRNA, rRNA, tRNA, non-coding RNA (ncRNA), long non-coding RNA (lncRNA), and microRNA (miRNA). In some cases, a target nucleic acid is a single stranded DNA (ssDNA) (e.g., a viral DNA). As noted above, in some cases, a target nucleic acid is single stranded.

[00298] A target nucleic acid can be located anywhere, for example, outside of a cell *in vitro*, inside of a cell *in vitro*, inside of a cell *in vivo*, inside of a cell *ex vivo*. Suitable target cells (which can comprise target nucleic acids such as genomic DNA) include, but are not limited to: a bacterial cell; an archaeal cell; 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, a cnidarian, an echinoderm, a nematode, etc.); a cell of an insect (e.g., a mosquito; a bee; an agricultural pest; etc.); a cell of an arachnid (e.g., a spider; a tick; etc.); a cell from a vertebrate animal (e.g., a fish, an amphibian, a reptile, a bird, a mammal); a cell from a mammal (e.g., a cell from a rodent; a cell from a human; a cell of a non-human mammal; a cell of a rodent (e.g., a mouse, a rat); a cell of a lagomorph (e.g., a rabbit); a cell of an ungulate (e.g., a cow, a horse, a camel, a llama, a vicuña, a sheep, a goat, etc.); a cell of a marine mammal (e.g., a whale, a seal, an elephant seal, a dolphin, a sea

lion; etc.) and the like. 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 (e.g., an oocyte, a sperm, an oogonia, a spermatogonia, etc.), an adult stem 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.).

[00299] 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 are maintained for fewer than 10 passages *in vitro*. Target cells can be unicellular organisms and/or can be grown in culture. 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. can be conveniently harvested by biopsy.

[00300] In some of the above applications, the subject methods may be employed to induce target nucleic acid cleavage, target nucleic acid modification, and/or to bind target nucleic acids (e.g., for visualization, for collecting and/or analyzing, etc.) in mitotic or post-mitotic cells *in vivo* and/or *ex vivo* and/or *in vitro* (e.g., to disrupt production of a protein encoded by a targeted mRNA, to cleave or otherwise modify target DNA, to geneically modify a target cell, and the like). Because the guide RNA provides specificity by hybridizing to target nucleic acid, a mitotic and/or post-mitotic cell of interest in the disclosed methods may include a cell from any organism (e.g. a bacterial cell, an archaeal cell, 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.). In some cases, a subject CasY protein (and/or nucleic acid encoding the protein such as DNA and/or RNA), and/or CasY guide RNA (and/or a DNA encoding the guide RNA), and/or donor template, and/or RNP can be intrduced into an individual (i.e., the target cell can be in vivo) (e.g., a mammal, a rat, a mouse, a pig, a primate, a non-human primate, a human, etc.). In some case, such an administration can be for the purpose of treating and/or preventing a disease, e.g., by editing the genome of targeted cells.

[00301] Plant cells include cells of a monocotyledon, and cells of a dicotyledon. The cells can be root cells, leaf cells, cells of the xylem, cells of the phloem, cells of the cambium, apical meristem cells, parenchyma cells, collenchyma cells, sclerenchyma cells, and the like. Plant cells include cells of agricultural crops such as wheat, corn, rice, sorghum, millet, soybean, etc. Plant cells include cells of agricultural fruit and nut plants, e.g., plant that produce apricots, oranges, lemons, apples, plums, pears, almonds, etc.

[00302] Additional examples of target cells are listed above in the section titled "Modified cells." Non-limiting examples of cells (target cells) include: a prokaryotic cell, eukaryotic cell, a bacterial cell, an archaeal cell, a cell of a single-cell eukaryotic organism, a protozoa cell, a cell from a plant (e.g., cells from plant crops, fruits, vegetables, grains, soy bean, corn, maize, wheat, seeds, tomatos, rice, cassava, sugarcane, pumpkin, hay, potatos, cotton, cannabis, tobacco, flowering plants, conifers, gymnosperms, angiosperms, ferns, clubmosses, hornworts, liverworts, mosses, dicotyledons, monocotyledons, etc.), an algal cell, (e.g., *Botryococcus braunii*, *Chlamydomonas reinhardtii*, *Nannochloropsis gaditana*, *Chlorella pyrenoidosa*, *Sargassum patens*, *C. agardh*, and the like), seaweeds (e.g. kelp) a fungal cell (e.g., a yeast cell, a cell from a mushroom), 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 (e.g., an ungulate (e.g., a pig, a cow, a goat, a sheep); a rodent (e.g., a rat, a mouse); a non-human primate; a human; a feline (e.g., a cat); a canine (e.g., a dog); etc.), and the like. In some cases, the cell is a cell that does not originate from a natural organism (e.g., the cell can be a synthetically made cell; also referred to as an artificial cell).

[00303] A cell can be an *in vitro* cell (e.g., established cultured cell line). A cell can be an *ex vivo* cell (cultured cell from an individual). A cell can be and *in vivo* cell (e.g., a cell in an individual). A cell can be an isolated cell. A cell can be a cell inside of an organism. A cell can be an organism. A cell can be a cell in a cell culture (e.g., *in vitro* cell culture). A cell can be one of a collection of cells. A cell can be a prokaryotic cell or derived from a prokaryotic cell. A cell can be a bacterial cell or can be derived from a bacterial cell. A cell can be an archaeal cell or derived from an archaeal cell. A cell can be a eukaryotic cell or derived from a eukaryotic cell. A cell can be a plant cell or derived from a plant cell. A cell can be an animal cell or derived from an animal cell. A cell can be an invertebrate cell or derived from an invertebrate cell. A cell can be a vertebrate cell or derived from a vertebrate cell. A cell can be a mammalian cell or derived from a mammalian cell. A cell can be a rodent cell or derived from a rodent cell. A cell can be a human cell or derived from a human cell. A cell can be a microbe cell or derived from a microbe cell. A cell can be a fungi cell or derived from a fungi cell. A cell can be an insect cell. A cell can be an arthropod cell. A cell can be a protozoan cell. A cell can be a helminth cell.

[00304] Suitable cells include a stem cell (e.g. an embryonic stem (ES) cell, an induced pluripotent stem (iPS) cell; a germ cell (e.g., an oocyte, a sperm, an oogonia, a spermatogonia, etc.); a somatic cell, e.g. a fibroblast, an oligodendrocyte, a glial cell, a hematopoietic cell, a neuron, a muscle cell, a bone cell, a hepatocyte, a pancreatic cell, etc.

[00305] Suitable cells include human embryonic stem cells, fetal cardiomyocytes, myofibroblasts, mesenchymal stem cells, autotransplated expanded cardiomyocytes, adipocytes, totipotent cells, pluripotent cells, blood stem cells, myoblasts, adult stem cells, bone marrow cells, mesenchymal cells, embryonic stem cells, parenchymal cells, epithelial cells, endothelial cells, mesothelial cells, fibroblasts, osteoblasts, chondrocytes, exogenous cells, endogenous cells, stem cells, hematopoietic stem cells, bone-marrow derived progenitor cells, myocardial cells, skeletal cells, fetal cells, undifferentiated cells, multi-potent progenitor cells, unipotent progenitor cells, monocytes, cardiac myoblasts, skeletal myoblasts, macrophages, capillary endothelial cells, xenogenic cells, allogenic cells, and post-natal stem cells.

[00306] In some cases, the cell is an immune cell, a neuron, an epithelial cell, and endothelial cell, or a stem cell. In some cases, the immune cell is a T cell, a B cell, a monocyte, a natural killer cell, a dendritic cell, or a macrophage. In some cases, the immune cell is a cytotoxic T cell. In some cases, the immune cell is a helper T cell. In some cases, the immune cell is a regulatory T cell (Treg).

[00307] In some cases, the cell is a stem cell. Stem cells include adult stem cells. Adult stem cells are also referred to as somatic stem cells.

[00308] Adult stem cells are resident in differentiated tissue, but retain the properties of self-renewal and ability to give rise to multiple cell types, usually cell types typical of the tissue in which the stem cells are found. Numerous examples of somatic stem cells are known to those of skill in the art, including muscle stem cells; hematopoietic stem cells; epithelial stem cells; neural stem cells; mesenchymal stem cells; mammary stem cells; intestinal stem cells; mesodermal stem cells; endothelial stem cells; olfactory stem cells; neural crest stem cells; and the like.

[00309] Stem cells of interest include mammalian stem cells, where the term "mammalian" refers to any animal classified as a mammal, including humans; non-human primates; domestic and farm animals; and zoo, laboratory, sports, or pet animals, such as dogs, horses, cats, cows, mice, rats, rabbits, etc. In some cases, the stem cell is a human stem cell. In some cases, the stem cell is a rodent (e.g., a mouse; a rat) stem cell. In some cases, the stem cell is a non-human primate stem cell.

[00310] Stem cells can express one or more stem cell markers, e.g., SOX9, KRT19, KRT7, LGR5, CA9, FXYD2, CDH6, CLDN18, TSPAN8, BPIFB1, OLFM4, CDH17, and PPARGC1A.

[00311] In some embodiments, the stem cell is a hematopoietic stem cell (HSC). HSCs are mesoderm-derived cells that can be isolated from bone marrow, blood, cord blood, fetal liver and yolk sac. HSCs are characterized as $CD34^+$ and $CD3^-$. HSCs can repopulate the erythroid, neutrophil-

macrophage, megakaryocyte and lymphoid hematopoietic cell lineages *in vivo*. *In vitro*, HSCs can be induced to undergo at least some self-renewing cell divisions and can be induced to differentiate to the same lineages as is seen *in vivo*. As such, HSCs can be induced to differentiate into one or more of erythroid cells, megakaryocytes, neutrophils, macrophages, and lymphoid cells.

[00312] In other embodiments, the stem cell is a neural stem cell (NSC). Neural stem cells (NSCs) are capable of differentiating into neurons, and glia (including oligodendrocytes, and astrocytes). A neural stem cell is a multipotent stem cell which is capable of multiple divisions, and under specific conditions can produce daughter cells which are neural stem cells, or neural progenitor cells that can be neuroblasts or glioblasts, e.g., cells committed to become one or more types of neurons and glial cells respectively. Methods of obtaining NSCs are known in the art.

[00313] In other embodiments, the stem cell is a mesenchymal stem cell (MSC). MSCs originally derived from the embryonal mesoderm and isolated from adult bone marrow, can differentiate to form muscle, bone, cartilage, fat, marrow stroma, and tendon. Methods of isolating MSC are known in the art; and any known method can be used to obtain MSC. See, e.g., U.S. Pat. No. 5,736,396, which describes isolation of human MSC.

[00314] A cell is in some cases a plant cell. A plant cell can be a cell of a monocotyledon. A cell can be a cell of a dicotyledon.

[00315] In some cases, the cell is a plant cell. For example, the cell can be a cell of a major agricultural plant, e.g., Barley, Beans (Dry Edible), Canola, Corn, Cotton (Pima), Cotton (Upland), Flaxseed, Hay (Alfalfa), Hay (Non-Alfalfa), Oats, Peanuts, Rice, Sorghum, Soybeans, Sugarbeets, Sugarcane, Sunflowers (Oil), Sunflowers (Non-Oil), Sweet Potatoes , Tobacco (Burley), Tobacco (Flue-cured), Tomatoes, Wheat (Durum), Wheat (Spring), Wheat (Winter), and the like. As another example, the cell is a cell of a vegetable crops which include but are not limited to, e.g., alfalfa sprouts, aloe leaves, arrow root, arrowhead, artichokes, asparagus, bamboo shoots, banana flowers, bean sprouts, beans, beet tops, beets, bittermelon, bok choy, broccoli, broccoli rabe (rappini), brussels sprouts, cabbage, cabbage sprouts, cactus leaf (nopales), calabaza, cardoon, carrots, cauliflower, celery, chayote, chinese artichoke (crosnes), chinese cabbage, chinese celery, chinese chives, choy sum, chrysanthemum leaves (tung ho), collard greens, corn stalks, corn-sweet, cucumbers, daikon, dandelion greens, dasheen, dau mue (pea tips), donqua (winter melon), eggplant, endive, escarole, fiddle head ferns, field cress, frisee, gai choy (chinese mustard), gailon, galanga (siam, thai ginger), garlic, ginger root, gobo, greens, hanover salad greens, huauzontle, jerusalem artichokes, jicama, kale greens, kohlrabi, lamb's quarters (quilete), lettuce (bibb), lettuce (boston), lettuce (boston red), lettuce (green leaf), lettuce (iceberg), lettuce (lolla rossa), lettuce (oak leaf - green), lettuce (oak leaf - red), lettuce (processed), lettuce (red leaf), lettuce (romaine), lettuce (ruby romaine), lettuce (russian red mustard), linkok, lo bok, long beans, lotus root, mache,

maguey (agave) leaves, malanga, mesculin mix, mizuna, moap (smooth luffa), moo, moqua (fuzzy squash), mushrooms, mustard, nagaimo, okra, ong choy, onions green, opo (long squash), ornamental corn, ornamental gourds, parsley, parsnips, peas, peppers (bell type), peppers, pumpkins, radicchio, radish sprouts, radishes, rape greens, rape greens, rhubarb, romaine (baby red), rutabagas, salicornia (sea bean), sinqua (angled/ridged luffa), spinach, squash, straw bales, sugarcane, sweet potatoes, swiss chard, tamarindo, taro, taro leaf, taro shoots, tatsoi, tepeguaje (guaje), tindora, tomatillos, tomatoes, tomatoes (cherry), tomatoes (grape type), tomatoes (plum type), tumeric, turnip tops greens, turnips, water chestnuts, yampi, yams (names), yu choy, yuca (cassava), and the like.

[00316] A cell is in some cases an arthropod cell. For example, the cell can be a cell of a sub-order, a family, a sub-family, a group, a sub-group, or a species of, e.g., *Chelicerata*, *Myriapodia*, *Hexipodia*, *Arachnida*, *Insecta*, *Archaeognatha*, *Thysanura*, *Palaeoptera*, *Ephemeroptera*, *Odonata*, *Anisoptera*, *Zygoptera*, *Neoptera*, *Exopterygota*, *Plecoptera* , *Embioptera* , *Orthoptera*, *Zoraptera* , *Dermaptera*, *Dictyoptera*, *Notoptera*, *Grylloblattidae*, *Mantophasmatidae*, *Phasmatodea* , *Blattaria*, *Isoptera*, *Mantodea*, *Parapneuroptera*, *Psocoptera*, *Thysanoptera*, *Phthiraptera*, *Hemiptera*, *Endopterygota or Holometabola* , *Hymenoptera* , *Coleoptera*, *Strepsiptera*, *Raphidioptera*, *Megaloptera*, *Neuroptera* , *Mecoptera* , *Siphonaptera*, *Diptera*, *Trichoptera*, or *Lepidoptera.*

[00317] A cell is in some cases an insect cell. For example, in some cases, the cell is a cell of a mosquito, a grasshopper, a true bug, a fly, a flea, a bee, a wasp, an ant, a louse, a moth, or a beetle.

Introducing components into a target cell

[00318] A Cas9 guide RNA (or a nucleic acid comprising a nucleotide sequence encoding same), and/or a Cas9 fusion polypeptide (or a nucleic acid comprising a nucleotide sequence encoding same) and/or a donor polynucleotide can be introduced into a host cell by any of a variety of well-known methods.

[00319] Methods of introducing a nucleic acid into a cell are known in the art, and any convenient method can be used to introduce a nucleic acid (e.g., an expression construct) into a taret cell (e.g., eukaryotic cell, human cell, stem cell, progenitor cell, and the like). Suitable methods are described in more detail elsewhere herein and 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. Any or all of the components can be introduced into a cell as a composition (e.g., including any convenient combination of: a a CasY polypeptide, a CasY guide RNA, a donor polynucleotide, etc.) using known methods, e.g., such as nucleofection.

Donor Polynucleotide (donor template)

[00320] Guided by a CasY guide RNA, a CasY protein in some cases generates site-specific double strand breaks (DSBs) or single strand breaks (SSBs) (e.g., when the CasY protein is a nickase variant) within double-stranded DNA (dsDNA) target nucleic acids, which are repaired either by non-homologous end joining (NHEJ) or homology-directed recombination (HDR).

[00321] In some cases, contacting a target DNA (with a CasY protein and a CasY guide RNA) occurs under conditions that are permissive for nonhomologous end joining or homology-directed repair. Thus, in some cases, a subject method includes contacting the target DNA with a donor polynucleotide (e.g., by introducing the donor polynucleotide into a cell), wherein the donor polynucleotide, a portion of the donor polynucleotide, a copy of the donor polynucleotide, or a portion of a copy of the donor polynucleotide integrates into the target DNA. In some cases, the method does not comprise contacting a cell with a donor polynucleotide, and the target DNA is modified such that nucleotides within the target DNA are deleted.

[00322] In some cases, CasY guide RNA (or DNA encoding same) and a CasY protein (or a nucleic acid encoding same, such as an RNA or a DNA, e.g, one or more expression vectors) are coadministered (e.g., contacted with a target nucleic acid, administered to cells, etc.) with a donor polynucleotide sequence that includes at least a segment with homology to the target DNA sequence, the subject methods may be used to add, i.e. insert or replace, nucleic acid material to a target DNA sequence (e.g. to "knock in" a nucleic acid, e.g., one that encodes for a protein, an siRNA, an miRNA, etc.), to add a tag (e.g., 6xHis, a fluorescent protein (e.g., a green fluorescent protein; a yellow fluorescent protein, etc.), hemagglutinin (HA), FLAG, etc.), to add a regulatory sequence to a gene (e.g. promoter, polyadenylation signal, internal ribosome entry sequence (IRES), 2A peptide, start codon, stop codon, splice signal, localization signal, etc.), to modify a nucleic acid sequence (e.g., introduce a mutation, remove a disease causnig mutation by introducing a correct sequence), and the like. As such, a complex comprising a CasY guide RNA and CasY protein is useful in any *in vitro* or *in vivo* application in which it is desirable to modify DNA in a site-specific, i.e. "targeted", way, for example gene knock-out, gene knock-in, gene editing, gene tagging, etc., as used in, for example, gene therapy, e.g. to treat a disease or as an antiviral, antipathogenic, or anticancer therapeutic, the production of genetically modified organisms in agriculture, the large scale production of proteins by cells for therapeutic, diagnostic, or research purposes, the induction of iPS cells, biological research, the targeting of genes of pathogens for deletion or replacement, etc.

[00323] In applications in which it is desirable to insert a polynucleotide sequence into he genome where a target sequence is cleaved, a donor polynucleotide (a nucleic acid comprising a donor sequence) can also be provided to the cell. By a "donor sequence" or "donor polynucleotide" or "donor template" it

is meant a nucleic acid sequence to be inserted at the site cleaved by the CasY protein (e.g., after dsDNA cleavage, after nicking a target DNA, after dual nicking a target DNA, and the like). The donor polynucleotide can contain sufficient homology to a genomic sequence at the target site, e.g. 70%, 80%, 85%, 90%, 95%, or 100% homology with the nucleotide sequences flanking the target site, e.g. within about 50 bases or less of the target site, e.g. within about 30 bases, within about 15 bases, within about 10 bases, within about 5 bases, or immediately flanking the target site, to support homology-directed repair between it and the genomic sequence to which it bears homology. Approximately 25, 50, 100, or 200 nucleotides, or more than 200 nucleotides, of sequence homology between a donor and a genomic sequence (or any integral value between 10 and 200 nucleotides, or more) can support homology-directed repair. Donor polynucleotides can be of any length, e.g. 10 nucleotides or more, 50 nucleotides or more, 100 nucleotides or more, 250 nucleotides or more, 500 nucleotides or more, 1000 nucleotides or more, 5000 nucleotides or more, etc.

[00324] The donor sequence is typically not identical to the genomic sequence that it replaces. Rather, the donor sequence may contain at least one or more single base changes, insertions, deletions, inversions or rearrangements with respect to the genomic sequence, so long as sufficient homology is present to support homology-directed repair (e.g., for gene correction, e.g., to convert a disease-causing base pair ot a non disease-causing base pair). In some embodiments, the donor sequence comprises a non-homologous sequence flanked by two regions of homology, such that homology-directed repair between the target DNA region and the two flanking sequences results in insertion of the non-homologous sequence at the target region. Donor sequences may also comprise a vector backbone containing sequences that are not homologous to the DNA region of interest and that are not intended for insertion into the DNA region of interest. Generally, the homologous region(s) of a donor sequence will have at least 50% sequence identity to a genomic sequence with which recombination is desired. In certain embodiments, 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% sequence identity is present. Any value between 1% and 100% sequence identity can be present, depending upon the length of the donor polynucleotide.

[00325] The donor sequence may comprise certain sequence differences as compared to the genomic sequence, e.g. restriction sites, nucleotide polymorphisms, selectable markers (e.g., drug resistance genes, fluorescent proteins, enzymes etc.), etc., which may be used to assess for successful insertion of the donor sequence at the cleavage site or in some cases may be used for other purposes (e.g., to signify expression at the targeted genomic locus). In some cases, if located in a coding region, such nucleotide sequence differences will not change the amino acid sequence, or will make silent amino acid changes (i.e., changes which do not affect the structure or function of the protein). Alternatively, these

sequences differences may include flanking recombination sequences such as FLPs, loxP sequences, or the like, that can be activated at a later time for removal of the marker sequence.

[00326] In some cases, the donor sequence is provided to the cell as single-stranded DNA. In some cases, the donor sequence is provided to the cell as double-stranded DNA. It may be introduced into a cell in linear or circular form. If introduced in linear form, the ends of the donor sequence may be protected (e.g., from exonucleolytic degradation) by any convenient method and such methods are known to those of skill in the art. For example, one or more dideoxynucleotide residues can be added to the 3' terminus of a linear molecule and/or self-complementary oligonucleotides can be ligated to one or both ends. See, for example, Chang et al. (1987) Proc. Natl. Acad Sci USA 84:4959-4963; Nehls et al. (1996) Science 272:886-889. Additional methods for protecting exogenous polynucleotides from degradation include, but are not limited to, addition of terminal amino group(s) and the use of modified internucleotide linkages such as, for example, phosphorothioates, phosphoramidates, and O-methyl ribose or deoxyribose residues. As an alternative to protecting the termini of a linear donor sequence, additional lengths of sequence may be included outside of the regions of homology that can be degraded without impacting recombination. A donor sequence can be introduced into a cell as part of a vector molecule having additional sequences such as, for example, replication origins, promoters and genes encoding antibiotic resistance. Moreover, donor sequences can be introduced as naked nucleic acid, as nucleic acid complexed with an agent such as a liposome or poloxamer, or can be delivered by viruses (e.g., adenovirus, AAV), as described elsewhere herein for nucleic acids encoding a CasY guide RNA and/or a CasY fusion polypeptide and/or donor polynucleotide.

TRANSGENIC, NON-HUMAN ORGANISMS

[00327] As described above, in some cases, a nucleic acid (e.g., a recombinant expression vector) of the present disclosure (e.g., a nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure; a nucleic acid comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure; etc.), is used as a transgene to generate a transgenic non-human organism that produces a CasY polypeptide, or a CasY fusion polypeptide, of the present disclosure. The present disclosure provides a transgenic-non-human organism comprising a nucleotide sequence encoding a CasY polypeptide, or a CasY fusion polypeptide, of the present disclosure.

Transgenic, non-human animals

[00328] The present disclosure provides a transgenic non-human animal, which animal comprises a transgene comprising a nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide or a CasY fusion polypeptide. In some embodiments, the genome of the transgenic non-human animal comprises a nucleotide sequence encoding a CasY polypeptide,e or a CasY fusion polypeptide, of the present disclosure. In some cases, the transgenic non-human animal is homozygous for the genetic

modification. In some cases, the transgenic non-human animal is heterozygous for the genetic modification. In some embodiments, the transgenic non-human animal is a vertebrate, for example, a fish (e.g., salmon, trout, zebra fish, gold fish, puffer fish, cave fish, etc.), an amphibian (frog, newt, salamander, etc.), a bird (e.g., chicken, turkey, etc.), a reptile (e.g., snake, lizard, etc.), a non-human mammal (e.g., an ungulate, e.g., a pig, a cow, a goat, a sheep, etc.; a lagomorph (e.g., a rabbit); a rodent (e.g., a rat, a mouse); a non-human primate; etc.), etc. In some cases, the transgenic non-human animal is an invertebrate. In some cases, the transgenic non-human animal is an insect (e.g., a mosquito; an agricultural pest; etc.). In some cases, the transgenic non-human animal is an arachnid.

[00329] Nucleotide sequences encoding a a CasY polypeptide,e or a CasY fusion polypeptide, of the present disclosure can be under the control of (i.e., operably linked to) an unknown promoter (e.g., when the nucleic acid randomly integrates into a host cell genome) or can be under the control of (i.e., operably linked to) a known promoter. Suitable known promoters can be any known promoter and include constitutively active promoters (e.g., CMV promoter), inducible promoters (e.g., heat shock promoter, tetracycline-regulated promoter, steroid-regulated promoter, metal-regulated promoter, estrogen receptor-regulated promoter, etc.), spatially restricted and/or temporally restricted promoters (e.g., a tissue specific promoter, a cell type specific promoter, etc.), etc.

Transgenic plants

[00330] As described above, in some cases, a nucleic acid (e.g., a recombinant expression vector) of the present disclosure (e.g., a nucleic acid comprising a nucleotide sequence encoding a CasY polypeptide of the present disclosure; a nucleic acid comprising a nucleotide sequence encoding a CasY fusion polypeptide of the present disclosure; etc.), is used as a transgene to generate a transgenic plant that produces a CasY polypeptide, or a CasY fusion polypeptide, of the present disclosure. The present disclosure provides a transgenic plant comprising a nucleotide sequence encoding a CasY polypeptide, or a CasY fusion polypeptide, of the present disclosure.In some embodiments, the genome of the transgenic plant comprises a subject nucleic acid. In some embodiments, the transgenic plant is homozygous for the genetic modification. In some embodiments, the transgenic plant is heterozygous for the genetic modification.

[00331] Methods of introducing exogenous nucleic acids into plant cells are well known in the art. Such plant cells are considered "transformed," as defined above. Suitable methods include viral infection (such as double stranded DNA viruses), transfection, conjugation, protoplast fusion, electroporation, particle gun technology, calcium phosphate precipitation, direct microinjection, silicon carbide whiskers technology, Agrobacterium-mediated transformation and the like. The choice of method is generally dependent on the type of cell being transformed and the circumstances under which the transformation is taking place (i.e. in vitro, ex vivo, or in vivo).

[00332] Transformation methods based upon the soil bacterium *Agrobacterium tumefaciens* are particularly useful for introducing an exogenous nucleic acid molecule into a vascular plant. The wild type form of *Agrobacterium* contains a Ti (tumor-inducing) plasmid that directs production of tumorigenic crown gall growth on host plants. Transfer of the tumor-inducing T-DNA region of the Ti plasmid to a plant genome requires the Ti plasmid-encoded virulence genes as well as T-DNA borders, which are a set of direct DNA repeats that delineate the region to be transferred. An *Agrobacterium*-based vector is a modified form of a Ti plasmid, in which the tumor inducing functions are replaced by the nucleic acid sequence of interest to be introduced into the plant host.

[00333] *Agrobacterium*-mediated transformation generally employs cointegrate vectors or binary vector systems, in which the components of the Ti plasmid are divided between a helper vector, which resides permanently in the *Agrobacterium* host and carries the virulence genes, and a shuttle vector, which contains the gene of interest bounded by T-DNA sequences. A variety of binary vectors is well known in the art and are commercially available, for example, from Clontech (Palo Alto, Calif.). Methods of coculturing *Agrobacterium* with cultured plant cells or wounded tissue such as leaf tissue, root explants, hypocotyledons, stem pieces or tubers, for example, also are well known in the art. See, e.g., Glick and Thompson, (eds.), Methods in Plant Molecular Biology and Biotechnology, Boca Raton, Fla.: CRC Press (1993).

[00334] Microprojectile-mediated transformation also can be used to produce a subject transgenic plant. This method, first described by Klein et al. (Nature 327:70-73 (1987)), relies on microprojectiles such as gold or tungsten that are coated with the desired nucleic acid molecule by precipitation with calcium chloride, spermidine or polyethylene glycol. The microprojectile particles are accelerated at high speed into an angiosperm tissue using a device such as the BIOLISTIC PD-1000 (Biorad; Hercules Calif.).

[00335] A nucleic acid of the present disclosure (e.g., a nucleic acid (e.g., a recombinant expression vector) comprising a nucleotide sequence encoding a CasY polypeptide, or a CasY fusion polypeptide, of the present disclosure) may be introduced into a plant in a manner such that the nucleic acid is able to enter a plant cell(s), e.g., via an in vivo or ex vivo protocol. By "in vivo," it is meant in the nucleic acid is administered to a living body of a plant e.g. infiltration. By “ex vivo” it is meant that cells or explants are modified outside of the plant, and then such cells or organs are regenerated to a plant. A number of vectors suitable for stable transformation of plant cells or for the establishment of transgenic plants have been described, including those described in Weissbach and Weissbach, (1989) Methods for Plant Molecular Biology Academic Press, and Gelvin et al., (1990) Plant Molecular Biology Manual, Kluwer Academic Publishers. Specific examples include those derived from a Ti plasmid of *Agrobacterium tumefaciens*, as well as those disclosed by Herrera-Estrella et al. (1983) Nature 303: 209,

Bevan (1984) Nucl Acid Res. 12: 8711-8721, Klee (1985) Bio/Technolo 3: 637-642. Alternatively, non-Ti vectors can be used to transfer the DNA into plants and cells by using free DNA delivery techniques. By using these methods transgenic plants such as wheat, rice (Christou (1991) Bio/Technology 9:957-9 and 4462) and corn (Gordon-Kamm (1990) Plant Cell 2: 603-618) can be produced. An immature embryo can also be a good target tissue for monocots for direct DNA delivery techniques by using the particle gun (Weeks et al. (1993) Plant Physiol 102: 1077-1084; Vasil (1993) Bio/Technolo 10: 667-674; Wan and Lemeaux (1994) Plant Physiol 104: 37-48 and for Agrobacterium-mediated DNA transfer (Ishida et al. (1996) Nature Biotech 14: 745-750). Exemplary methods for introduction of DNA into chloroplasts are biolistic bombardment, polyethylene glycol transformation of protoplasts, and microinjection (Danieli et al Nat. Biotechnol 16:345-348, 1998; Staub et al Nat. Biotechnol 18: 333-338, 2000; O'Neill et al Plant J. 3:729-738, 1993; Knoblauch et al Nat. Biotechnol 17: 906-909; U.S. Pat. Nos. 5,451,513, 5,545,817, 5,545,818, and 5,576,198; in Intl. Application No. WO 95/16783; and in Boynton et al., Methods in Enzymology 217: 510-536 (1993), Svab et al., Proc. Natl. Acad. Sci. USA 90: 913-917 (1993), and McBride et al., Proc. Natl. Acad. Sci. USA 91: 7301-7305 (1994)). Any vector suitable for the methods of biolistic bombardment, polyethylene glycol transformation of protoplasts and microinjection will be suitable as a targeting vector for chloroplast transformation. Any double stranded DNA vector may be used as a transformation vector, especially when the method of introduction does not utilize Agrobacterium.

[00336] Plants which can be genetically modified include grains, forage crops, fruits, vegetables, oil seed crops, palms, forestry, and vines. Specific examples of plants which can be modified follow: maize, banana, peanut, field peas, sunflower, tomato, canola, tobacco, wheat, barley, oats, potato, soybeans, cotton, carnations, sorghum, lupin and rice.

[00337] The present disclosure provides transformed plant cells, tissues, plants and products that contain the transformed plant cells. A feature of the subject transformed cells, and tissues and products that include the same is the presence of a subject nucleic acid integrated into the genome, and production by plant cells of a CasY polypeptide, or a CasY fusion polypeptide, of the present disclosure. Recombinant plant cells of the present invention are useful as populations of recombinant cells, or as a tissue, seed, whole plant, stem, fruit, leaf, root, flower, stem, tuber, grain, animal feed, a field of plants, and the like.

[00338] Nucleotide sequences encoding a CasY polypeptide, or a CasY fusion polypeptide, of the present disclosure can be under the control of (i.e., operably linked to) an unknown promoter (e.g., when the nucleic acid randomly integrates into a host cell genome) or can be under the control of (i.e., operably linked to) a known promoter. Suitable known promoters can be any known promoter and include

constitutively active promoters, inducible promoters, spatially restricted and/or temporally restricted promoters, etc.

METHODS OF IDENTIFYING A CRISPR RNA-GUIDED ENDONUCLEASE

[00339] Provided are methods of identifying a CRISPR RNA-guided endonuclease. For example, in some embodiments, such a method includes a step of detetcing in a plurality of metagenomic nucleotide sequences, a nucleotide sequence encoding a Cas1 polypeptide. Cas1 proteins are known in the art and are present in the vicinity of CRISPR loci of Class 2 CRISPR systems, those CRISPR systems that include a single effector protein that functions as an endonuclease and does not require interaction with a complex of proteins in order to function properly. While the Cas 1 protein itself is involved in acquisition of new target sequences into the CRISPR locus, and thus is not the desired effector protein for identification by this method, the presence of a Cas1 protein in the vicinity of a CRISPR lcous is an indication that at least one of the other Cas proteins present near the locus may be an effector protein (an RNA-guided endonuclease).

[00340] As used herein, the term "metagenomics" means the parallel analysis of nucleic acids recovered from multiple microorganisms (e.g., bacteria, archaea, etc.) in a sample, e.g., an environmental sample such as a sample that contains an unknown quantity of prokaryotes (bacteria/archaea) and may contain prokaryotes that have never before discovered and/or characterized. Nucleic acids can be recovered from such a sample by any convenient method, and generally the nucleic acids are recovered together from the entire sample such that prior to analysis it is unknown from which microorganism any given nucleic acid molecule originated. In some embodiments, the sample contains an unknown mixture and/or quantity of microorganisms. The nucleic acids can then be sequenced to generate a plurality of metagenomic sequences. In some cases, a subject method of identifying a CRISPR RNA-guided endonuclease includes a step of isolating a sample (e.g., an environmental sample). In some cases, a subject method of identifying a CRISPR RNA-guided endonuclease includes a step of isolating nucleic acids from the sample and/or assaying the sample to generated a plurality of metagenomic nucleotide sequences from the sample.

[00341] Once a Cas1 protein is identified, a subject method of identifying a CRISPR RNA-guided endonuclease can include a step of detecting a CRISPR array (a repeat-spacer-repeat array) in the vicinity of the Cas1-encoding nucleotide sequence. The method can then include a step of cloning (e.g., from a nucleic acid sample from which the plurality of metagenomic nucleotide sequences was derived) a CRISPR locus comprising the detected CRISPR array, into an expression vector to generate a recombinant CRISPR locus expression vector. The CRISPR locus can then be tested for function by assaying the recombinant CRISPR locus expression vector for the ability to cleave a target nucleic acid. Any convenient assay can be used. In some embodiments the assaying step includes introducing the

recombinant CRISPR locus expression vector and a target nucleic into a cell, e.g., a heterologous host cell such as an *E. coli* cell. For example, refer to the PAM depletion assays of the working examples below (Figure 5). In some cases, the step of assaying includes introducing into a population of host cells (e.g., E. coli cells) a plasmid library wherein each plasmid of the libary has 4 to 10 (e.g., 5 to 10, 5 to 8, 6 to 10, 6 to 8, 5, 6, 7, 8,) nucleotides randomized 5' and/or 3' of a target sequence. The host cells can already contain the recombinant CRISPR locus expression vector to be tested, or the recombinant CRISPR locus expression vector can be introduced after the library. Only test CRISPR loci that are functional, and therefore include a functional CRISPR RNA-guided endonuclease, will result in the ability to cleave plasmids that have the target sequence. The reason for included the randomized sequences 5' and 3' of the target sequence is that one might not know the PAM sequence required for the desired endonuclease at the outset of the experiment.

[00342] If the expression vector can cleave a target nucleic acid (e.g., one with an appropriate target sequence and PAM, such as a target sequence that matches as least one spacer of the CRISPR array), then the CRISPR locus comprises a nucleotide sequence encoding a candidate CRISPR RNA-guided endonuclease. Thus, one can then identify an open reading frame from the CRISPR locus that encodes a CRISPR RNA-guided endonuclease . In some cases, it is desirable to identify a previously unknown CRISPR RNA-guided endonuclease, and thus in some cases, the identified polypeptide that has less than 20% amino acid sequence identity (e.g., less than 15%, less than 10%, less than 5% amino acid sequence identity) to the amino acid sequence of a known CRISPR RNA-guided endonuclease polypeptide.

EXAMPLES OF NON-LIMITING ASPECTS OF THE DISCLOSURE

[00343] Aspects, including embodiments, of the present subject matter described above may be beneficial alone or in combination, with one or more other aspects or embodiments. Without limiting the foregoing description, certain non-limiting aspects of the disclosure, numbered 1-123 are provided below. As will be apparent to those of skill in the art upon reading this disclosure, each of the individually numbered aspects may be used or combined with any of the preceding or following individually numbered aspects. This is intended to provide support for all such combinations of aspects and is not limited to combinations of aspects explicitly provided below:

Aspects

1. A composition comprising:

a) a CasY polypeptide, or a nucleic acid molecule encoding the CasY polypeptide; and

b) a CasY guide RNA, or one or more DNA molecules encoding the CasY guide RNA.

2. The composition of 1, wherein the CasY polypeptide comprises an amino acid sequence having 50% or more identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

3. The composition of 1 or 2, wherein the CasY guide RNA comprises a nucleotide sequence having 80% or more identity with the crRNA sequence set forth in any one of SEQ ID NOs: 11-15.

4. The composition of 1 or 2, wherein the CasY polypeptide is fused to an NLS sequence.

5. The composition of any one of 1-4, wherein the composition comprises a lipid.

6. The composition of any one of 1-4, wherein a) and b) are within a liposome.

7. The composition of any one of 1-4, wherein a) and b) are within a particle.

8. The composition of any one of 1-7, comprising one or more of: a buffer, a nuclease inhibitor, and a protease inhibitor.

9. The composition of any one of 1-8, wherein the CasY polypeptide comprises an amino acid sequence having 85% or more identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

10. The composition of any one of 1-9, wherein the CasY polypeptide is a nickase that can cleave only one strand of a double-stranded target nucleic acid molecule.

11. The composition of any one of 1-9, wherein the CasY polypeptide is a catalytically inactive CasY Polypeptide (dCasY).

12. The composition of 10 or 11, wherein the CasY polypeptide comprises one or more mutations at a position corresponding to those selected from: D672, E769, and D935 of SEQ ID NO: 1.

13. The composition of any one of 1-12, further comprising a DNA donor template.

14. A CasY fusion polypeptide comprising: a CasY polypeptide fused to a heterologous polypeptide.

15. The CasY fusion polypeptide of 14, wherein the CasY polypeptide comprises an amino acid sequence having 50% or more identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

16. The CasY fusion polypeptide of 14, wherein the CasY polypeptide comprises an amino acid sequence having 85% or more identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

17. The CasY fusion polypeptide of any one of 14-16, wherein the CasY polypeptide is a nickase that can cleave only one strand of a double-stranded target nucleic acid molecule.

18. The CasY fusion polypeptide of any one of 14-17, wherein the CasY polypeptide is a catalytically inactive CasY Polypeptide (dCasY).

19. The CasY fusion polypeptide of 17 or 18, wherein the CasY polypeptide comprises one or more mutations at a position corresponding to those selected from: D672, E769, and D935 of SEQ ID NO: 1.

20. The CasY fusion polypeptide of any one of 14-19, wherein the heterologous polypeptide is fused to the N-terminus and/or the C-terminus of the CasY polypeptide.

21. The CasY fusion polypeptide of any one of 14-20, comprising an NLS.

22. The CasY fusion polypeptide of any one of 14-21, wherein the heterologous polypeptide is a targeting polypeptide that provides for binding to a cell surface moiety on a target cell or target cell type.

23. The CasY fusion polypeptide of any one of 14-21, wherein the heterologous polypeptide exhibits an enzymatic activity that modifies target DNA.

24. The CasY fusion polypeptide of 23, wherein the heterologous polypeptide exhibits an one or more enzymatic activities selected from: nuclease activity, methyltransferase activity, demethylase activity, DNA repair activity, DNA damage activity, deamination activity, dismutase activity, alkylation activity, depurination activity, oxidation activity, pyrimidine dimer forming activity, integrase activity, transposase activity, recombinase activity, polymerase activity, ligase activity, helicase activity, photolyase activity and glycosylase activity.

25. The CasY fusion polypeptide of 24, wherein the heterologous polypeptide exhibits an one or more enzymatic activities selected from: nuclease activity, methyltransferase activity, demethylase activity, deamination activity, depurination activity, integrase activity, transposase activity, and recombinase activity.

26. The CasY fusion polypeptide of any one of 14-21, wherein the heterologous polypeptide exhibits an enzymatic activity that modifies a target polypeptide associated with a target nucleic acid.

27. The CasY fusion polypeptide of 26, wherein the heterologous polypeptide exhibits histone modification acitivity.

28. The CasY fusion polypeptide of 26 or 27, wherein the heterologous polypeptide exhibits an one or more enzymatic activities selected from: methyltransferase activity, demethylase activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity, demyristoylation activity, glycosylation activity (e.g., from O-GlcNAc transferase) and deglycosylation activity.

29. The CasY fusion polypeptide of 28, wherein the heterologous polypeptide exhibits an one or more enzymatic activities selected from: methyltransferase activity, demethylase activity, acetyltransferase activity, and deacetylase activity.

30. The CasY fusion polypeptide of any one of 14-21, wherein the heterologous polypeptide is an endosomal escape polypeptide.

31. The CasY fusion polypeptide of 30, wherein the endosomal escape polypeptide comprises an amino acid sequence selected from: GLFXALLXLLXSLWXLLLXA (SEQ ID NO:94), and GLFHALLHLLHSLWHLLLHA (SEQ ID NO:95), wherein each X is independently selected from lysine, histidine, and arginine.

32. The CasY fusion polypeptide of any one of 14-21, wherein the heterologous polypeptide is a chloroplast transit peptide.

33. The CasY fusion polypeptide of 32, wherein the chloroplast transit peptide comprises an amino acid sequence selected from
MASMISSSAVTTVSRASRGQSAAMAPFGGLKSMTGFPVRKVNTDITSITSNGGRVKCMQVWPPI GKKKFETLSYLPPLTRDSRA (SEQ ID NO:83);
MASMISSSAVTTVSRASRGQSAAMAPFGGLKSMTGFPVRKVNTDITSITSNGGRVKS (SEQ ID NO:84);
MASSMLSSATMVASPAQATMVAPFNGLKSSAAFPATRKANNDITSITSNGGRVNCMQVWPPIEK KKFETLSYLPDLTDSGGRVNC (SEQ ID NO:85);
MAQVSRICNGVQNPSLISNLSKSSQRKSPLSVSLKTQQHPRAYPISSSWGLKKSGMTLIGSELRPL KVMSSVSTAC (SEQ ID NO:86);
MAQVSRICNGVWNPSLISNLSKSSQRKSPLSVSLKTQQHPRAYPISSSWGLKKSGMTLIGSELRPL KVMSSVSTAC (SEQ ID NO:87);
MAQINNMAQGIQTLNPNSNFHKPQVPKSSSFLVFGSKKLKNSANSMLVLKKDSIFMQLFCSFRIS ASVATAC (SEQ ID NO:88);
MAALVTSQLATSGTVLSVTDRFRRPGFQGLRPRNPADAALGMRTVGASAAPKQSRKPHRFDRR CLSMVV (SEQ ID NO:89);
MAALTTSQLATSATGFGIADRSAPSSLLRHGFQGLKPRSPAGGDATSLSVTTSARATPKQQRSVQ RGSRRFPSVVVC (SEQ ID NO:90);
MASSVLSSAAVATRSNVAQANMVAPFTGLKSAASFPVSRKQNLDITSIASNGGRVQC (SEQ ID NO:91);
MESLAATSVFAPSRVAVPAARALVRAGTVVPTRRTSSTSGTSGVKCSAAVTPQASPVISRSAAAA (SEQ ID NO:92); and
MGAAATSMQSLKFSNRLVPPSRRLSPVPNNVTCNNLPKSAAPVRTVKCCASSWNSTINGAAATT NGASAASS (SEQ ID NO:93).

34. The CasY fusion polypeptide of any one of 14-21, wherein the heterologous polypeptide is protein that increases or decreases transcription.

35. The CasY fusion polypeptide of 34, wherein the heterologous polypeptide is a transcriptional repressor domain.

36. The CasY fusion polypeptide of 34, wherein the heterologous polypeptide is a transcriptional activation domain.

37. The CasY fusion polypeptide of any one of 14-21, wherein the heterologous polypeptide is a protein biding domain.

38. A nucleic acid molecule encoding the CasY fusion polypeptide of any one of 14-37.

39. The nucleic acid molecule of 38, wherein the nucleotide sequence encoding the CasY fusion polypeptide is operably linked to a promoter.

40. The nucleic acid molecule of 39, wherein the promoter is functional in a eukaryotic cell.

41. The nucleic acid molecule of 40, wherein the promoter is functional in one or more of: a plant cell, a fungal cell, an animal cell, cell of an invertebrate, a fly cell, a cell of a vertebrate, a mammalian cell, a primate cell, a non-human primate cell, and a human cell.

42. The nucleic acid molecule of any one of 39-41, wherein the promoter is one or more of: a constitutive promoter, an inducible promoter, a cell type-specific promoter, and a tissue-specific promoter.

43. The nucleic acid molecule of any one of 38-42, wherein the DNA molecule is a recombinant expression vector.

44. The nucleic acid molecule of 43, wherein the recombinant expression vector is a recombinant adenoassociated viral vector, a recombinant retroviral vector, or a recombinant lentiviral vector.

45. The nucleic acid molecule of 39, wherein the promoter is functional in a prokaryotic cell.

46. The nucleic acid molecule of 38, wherein the nucleic acid molecule is an mRNA.

47. One or more nucleic molecules encoding:

(a) a CasY guide RNA; and

(b) a CasY polypeptide.

48. The one or more nucleic acid molecules of 47, wherein the CasY polypeptide comprises an amino acid sequence having 50% or more identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

49. The one or more nucleic acid molecules of 47, wherein the CasY polypeptide comprises an amino acid sequence having 85% or more identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

50. The one or more nucleic acid molecules of any one of 47-49, wherein the CasY guide RNA comprises a nucleotide sequence having 80% or more identity with the crRNA sequence set forth in any one of SEQ ID NOs: 11-15.

51. The one or more nucleic acid molecules of any one of 47-50, wherein the CasY polypeptide is fused to an NLS sequence.

52. The one or more nucleic acid molecules of any one of 47-51, wherein said one or more nucleic acid molecules comprises a nucleotide sequence encoding the CasY guide RNA that is operably linked to a promoter.

53. The one or more nucleic acid molecules of any one of 47-52, wherein said one or more nucleic acid molecules comprises a nucleotide sequence encoding the CasY polypeptide that is operably linked to a promoter.

54. The one or more nucleic acid molecules of 52 or 53, wherein the promoter operably linked to the nucleotide sequence encoding thea CasY guide RNA, and/or the promoter operably linked to the nucleotide sequence encoding the CasY polypeptide, is functional in a eukaryotic cell.

55. The one or more nucleic acid molecules of 54, wherein the promoter is functional in one or more of: a plant cell, a fungal cell, an animal cell, cell of an invertebrate, a fly cell, a cell of a vertebrate, a mammalian cell, a primate cell, a non-human primate cell, and a human cell.

56. The one or more nucleic acid molecules of any one of 53-55, wherein the promoter is one or more of: a constitutive promoter, an inducible promoter, a cell type-specific promoter, and a tissue-specific promoter.

57. The one or more nucleic acid molecules of any one of 47-56, wherein the one or more nucleic acid molecules is one or more recombinant expression vectors.

58. The one or more nucleic acid molecules of 57, wherein the one or more recombinant expression vectors are selected from: one or more adenoassociated viral vectors, one or more recombinant retroviral vectors, or one or more recombinant lentiviral vectors.

59. The one or more nucleic acid molecules of 53, wherein the promoter is functional in a prokaryotic cell.

60. A eukaryotic cell comprising one or more of:

a) a CasY polypeptide, or a nucleic acid molecule encoding the CasY polypeptide,

b) a CasY fusion polypeptide, or a nucleic acid molecule encoding the CasY fusion polypeptide, and

c) a CasY guide RNA, or a nucleic acid molecule encoding the CasY guide RNA.

61. The eukaryotic cell of 60, comprising the nucleic acid molecule encoding the CasY polypeptide, wherein said nucleic acid molecule is integrated into the genomic DNA of the cell.

62. The eukaryotic cell of 60 or 61, wherein the eukaryotic cell is a plant cell, a mammalian cell, an insect cell, an arachnid cell, a fungal cell, a bird cell, a reptile cell, an amphibian cell, an invertebrate cell, a mouse cell, a rat cell, a primate cell, a non-human primate cell, or a human cell.

63. A cell comprising a CasY fusion polypeptide, or a nucleic acid molecule encoding the CasY fusion polypeptide.

64. The cell of 63, wherein the cell is a prokaryotic cell.

65. The cell of 63 or 64, comprising the nucleic acid molecule encoding the CasY fusion polypeptide, wherein said nucleic acid molecule is integrated into the genomic DNA of the cell.

66. A method of modifying a target nucleic acid, the method comprising contacting the target nucleic acid with:

a) a CasY polypeptide; and

b) a CasY guide RNA comprising a guide sequence that hybridizes to a target sequence of the target nucleic acid,

wherein said contacting results in modification of the target nucleic acid by the CasY polypeptide.

67. The method of 66, wherein said modification is cleavage of the target nucleic acid.

68. The method of 66 or 67, wherein the target nucleic acid is selected from: double stranded DNA, single stranded DNA, RNA, genomic DNA, and extrachromosomal DNA.

69. The method of any of 66-68, wherein said contacting takes place *in vitro* outside of a cell.

70. The method of any of 66-68, wherein said contacting takes place inside of a cell in culture.

71. The method of any of 66-68, wherein said contacting takes place inside of a cell *in vivo.*

72. The method of 70 or 71, wherein the cell is a eukaryotic cell.

73. The method of 72, wherein the cell is selected from: a plant cell, a fungal cell, a mammalian cell, a reptile cell, an insect cell, an avian cell, a fish cell, a parasite cell, an arthropod cell, a cell of an invertebrate, a cell of a vertebrate, a rodent cell, a mouse cell, a rat cell, a primate cell, a non-human primate cell, and a human cell.

74. The method of 70 or 71, wherein the cell is a prokaryotic cell.

75. The method of any one of 66-74, wherein said contacting results in genome editing.

76. The method of any one of 66-75, wherein said contacting comprises: introducing into a cell: (a) the CasY polypeptide, or a nucleic acid molecule encoding the the CasY polypeptide, and (b) the CasY guide RNA, or a nucleic acid molecule encoding the the CasY guide RNA.

77. The method of 76, wherein said contacting further comprises: introducing a DNA donor template into the cell.

78. The method of any one of 66-77, wherein the CasY guide RNA comprises a nucleotide sequence having 80% or more identity with the crRNA sequence set forth in any one of SEQ ID NOs: 11-15.

79. The method of any one of 66-78, wherein the CasY polypeptide is fused to an NLS sequence.

80. A method of modulating transcription from a target DNA, modifying a target nucleic acid, or modifying a protein associated with a target nucleic acid, the method comprising contacting the target nucleic acid with:

a) a CasY fusion polypeptide comprising a CasY polypeptide fused to a heterologous polypeptide; and

b) a CasY guide RNA comprising a guide sequence that hybridizes to a target sequence of the target nucleic acid.

81. The method of 80, wherein the CasY guide RNA comprises a nucleotide sequence having 80% or more identity with the crRNA sequence set forth in any one of SEQ ID NOs: 11-15.

82. The method of 80 or 81, wherein the CasY fusion polypeptide comprises an NLS sequence.

83. The method of any of 80-82, wherein said modification is not cleavage of the target nucleic acid.

84. The method of any of 80-83, wherein the target nucleic acid is selected from: double stranded DNA, single stranded DNA, RNA, genomic DNA, and extrachromosomal DNA.

85. The method of any of 80-84, wherein said contacting takes place *in vitro* outside of a cell.

86. The method of any of 80-84, wherein said contacting takes place inside of a cell in culture.

87. The method of any of 80-84, wherein said contacting takes place inside of a cell *in vivo.*

88. The method of 86 or 87, wherein the cell is a eukaryotic cell.

89. The method of 88, wherein the cell is selected from: a plant cell, a fungal cell, a mammalian cell, a reptile cell, an insect cell, an avian cell, a fish cell, a parasite cell, an arthropod cell, a cell of an invertebrate, a cell of a vertebrate, a rodent cell, a mouse cell, a rat cell, a primate cell, a non-human primate cell, and a human cell.

90. The method of 86 or 87, wherein the cell is a prokaryotic cell.

91. The method of any one of 80-90, wherein said contacting comprises: introducing into a cell: (a) the CasY fusion polypeptide, or a nucleic acid molecule encoding the the CasY fusion polypeptide, and (b) the CasY guide RNA, or a nucleic acid molecule encoding the the CasY guide RNA.

92. The method of any one of 80-91, wherein the CasY polypeptide is a catalytically inactive CasY Polypeptide (dCasY).

93. The method of any one of 80-92, wherein the CasY polypeptide comprises one or more mutations at a position corresponding to those selected from: D672, E769, and D935 of SEQ ID NO: 1.

94. The method of any one of 80-93, wherein the heterologous polypeptide exhibits an enzymatic activity that modifies target DNA.

95. The method of 94, wherein the heterologous polypeptide exhibits an one or more enzymatic activities selected from: nuclease activity, methyltransferase activity, demethylase activity, DNA repair activity, DNA damage activity, deamination activity, dismutase activity, alkylation activity, depurination activity, oxidation activity, pyrimidine dimer forming activity, integrase activity, transposase activity, recombinase activity, polymerase activity, ligase activity, helicase activity, photolyase activity and glycosylase activity.

96. The method of 95, wherein the heterologous polypeptide exhibits an one or more enzymatic activities selected from: nuclease activity, methyltransferase activity, demethylase activity, deamination activity, depurination activity, integrase activity, transposase activity, and recombinase activity.

97. The method of any one of 80-93, wherein the heterologous polypeptide exhibits an enzymatic activity that modifies a target polypeptide associated with a target nucleic acid.

98. The method of 97, wherein the heterologous polypeptide exhibits histone modification acitivity.

99. The method of 97 or 98, wherein the heterologous polypeptide exhibits an one or more enzymatic activities selected from: methyltransferase activity, demethylase activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity, demyristoylation activity, glycosylation activity (e.g., from O-GlcNAc transferase) and deglycosylation activity.

100. The method of 99, wherein the heterologous polypeptide exhibits an one or more enzymatic activities selected from: methyltransferase activity, demethylase activity, acetyltransferase activity, and deacetylase activity.

101. The method of any one of 80-93, wherein the heterologous polypeptide is protein that increases or decreases transcription.

102. The method of 101, wherein the heterologous polypeptide is a transcriptional repressor domain.

103. The method of 101, wherein the heterologous polypeptide is a transcriptional activation domain.

104. The method of any one of 80-93, wherein the heterologous polypeptide is a protein biding domain.

105. A transgenic, multicellular, non-human organism whose genome comprises a transgene comprising a nucleotide sequence encoding one or more of:

a) a CasY polypeptide,

b) a CasY fusion polypeptide, and

c) a CasY guide RNA.

106. The transgenic, multicellular, non-human organism of 105, wherein the CasY polypeptide comprises an amino acid sequence having 50% or more amino acid sequence identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

107. The transgenic, multicellular, non-human organism of 105, wherein the CasY polypeptide comprises an amino acid sequence having 85% or more amino acid sequence identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

108. The transgenic, multicellular, non-human organism of any one of 105-107, wherein the organism is a plant, a monocotyledon plant, a dicotyledon plant, an invertebrate animal, an insect, an arthropod, an arachnid, a parasite, a worm, a cnidarian, a vertebrate animal, a fish, a reptile, an amphibian, an ungulate, a bird, a pig, a horse, a sheep, a rodent, a mouse, a rat, or a non-human primate.

109. A system comprising:

a) a CasY polypeptide and a CasY guide RNA;

b) a CasY polypeptide, a CasY guide RNA, and a DNA donor template;

c) a CasY fusion polypeptide and a CasY guide RNA;

d) a CasY fusion polypeptide, a CasY guide RNA, and a DNA donor template;

e) an mRNA encoding a CasY polypeptide, and a CasY guide RNA;

f) an mRNA encoding a CasY polypeptide; a CasY guide RNA, and a DNA donor template;

g) an mRNA encoding a CasY fusion polypeptide, and a CasY guide RNA;

h) an mRNA encoding a CasY fusion polypeptide, a CasY guide RNA, and a DNA donor template;

i) one or more recombinant expression vectors comprising: i) a nucleotide sequence encoding a CasY polypeptide; and ii) a nucleotide sequence encoding a CasY guide RNA;

j) one or more recombinant expression vectors comprising: i) a nucleotide sequence encoding a CasY polypeptide; ii) a nucleotide sequence encoding a CasY guide RNA; and iii) a DNA donor template;

k) one or more recombinant expression vectors comprising: i) a nucleotide sequence encoding a CasY fusion polypeptide; and ii) a nucleotide sequence encoding a CasY guide RNA; and

l) one or more recombinant expression vectors comprising: i) a nucleotide sequence encoding a CasY fusion polypeptide; ii) a nucleotide sequence encoding a CasY guide RNA; and a DNA donor template.

110. The CasY system of 109, wherein the CasY polypeptide comprises an amino acid sequence having 50% or more amino acid sequence identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

111. The CasY system of 109, wherein the CasY polypeptide comprises an amino acid sequence having 85% or more amino acid sequence identity to the amino acid sequence set forth in SEQ ID NO:1 or SEQ ID NO:2 (or the amino acid sequence set forth in any one of SEQ ID NOs:1-8).

112. The CasY system of any of 109-111, wherein the donor template nucleic acid has a length of from 8 nucleotides to 1000 nucleotides.

113. The CasY system of any of 109-111, wherein the donor template nucleic acid has a length of from 25 nucleotides to 500 nucleotides.

114. A kit comprising the CasY system of any one of 109-113.

115. The kit of 114, wherein the components of the kit are in the same container.

116. The kit of 114, wherein the components of the kit are in separate containers.

117. A sterile container comprising the CasY system of any one of 109-116.

118. The sterile container of 117, wherein the container is a syringe.

119. An implantable device comprising the CasY system of any one of 109-116.

120. The implantable device of 119, wherein the CasY system is within a matrix.

121. The implantable device of 119, wherein the CasY system is in a reservoir.

122. A method of identifying a CRISPR RNA-guided endonuclease, the method comprising:

detecting, in a plurality of metagenomic nucleotide sequences, a nucleotide sequence encoding a Cas1 polypeptide;

detecting a CRISPR array in the vicinity of the Cas1-encoding nucleotide sequence;

cloning, from a nucleic acid sample from which the plurality of metagenomic nucleotide sequences was derived, a CRISPR locus comprising the detected CRISPR array, into an expression vector to generate a recombinant CRISPR locus expression vector;

assaying the recombinant CRISPR locus expression vector for the ability to cleave a target nucleic acid, wherein a CRISPR locus that has the ability to cleave a target nucleic acid comprises a nucleotide sequence encoding a CRISPR RNA-guided endonuclease.

identifying, in the CRISPR locus, an open reading frame encoding a polypeptide that has less than 20% amino acid sequence identity to the amino acid sequence of a known CRISPR RNA-guided endonuclease polypeptide.

123. The method of 122, wherein said assaying comprises introducing the recombinant CRISPR locus expression vector and a target nucleic into a cell.

EXAMPLES

[00344] 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

[00345] The work described herein includes an analysis of metagenomic samples of microbial communities from groundwater, sediments, and acid mine drainage. New Class 2 CRISPR-Cas systems were identified that are not represented among cultured organisms.

[00346] **Figure 3. CasY domains and similarity searches.** (**panel a**) Schematic domain representation for CasY inferred from distant homolog alignments with AcCpf1, using HHpred. Conserved catalytic residues are marked by red bars above the proteins. CasY contains a RuvC split domain in the C-terminal region (RuvC-I, RuvC-II, and RuvC-III), and a large novel N-terminal domain. Below the schematic are displayed top hits based on the following searches: (1) BLAST search against all the proteins in NCBI (NR database, including model and environmental proteins). (2) Profile hidden markov model (HMM) search based on models built using all the Cas proteins described in Makarova et al. Nat Rev Microbiol. 2015 Nov;13(11):722-36, and Shmakov et al. Mol Cell. 2015 Nov 5;60(3):385-97). (3) Distant homolog search based on HHpred. Hits are color-coded based on their significance, and the hit range and E-value is provided. Notably, CasY had only local hits. The 812 N-terminal amino acid of CasY had only one very minor partial hit. Combined, these finding indicate CasY is a new Cas protein. (**panel b**) Different CasY-contaning CRISPR loci scaffolds were constructed from sequence data.

Example 2

[00347] **Figure 4. Schematic diagram of Diagram of CasY and C2c3 loci.** Interference proteins are shown in green, acquisition proteins in red. Repeats folded using RNA structure are shown to the right revealing a strong hairpin at the 5' end, suggesting self processing of the CRISPR array by CasY..

[00348] **Figure 5 (panels a-d) PAM dependent plasmid interference by CasY.** (**panel a**) PAM depletion assays were conducted with CasY. *E. coli* containing the CasY CRISPR locus were transformed with a plasmid library with 7 nucleotides randomized 5' or 3' of the target sequence. The target plasmid was selected for and transformants were pooled. The randomized region was amplified and prepared for deep sequencing. Depleted sequences were identified and used to generate a PAM logo. (**panel b**) the generated PAM logo for CasY.1 showed a strong preference for sequences containing a 5'-TA-3' flanking sequence 5' of the target. A 3' PAM was not detected. (**panel c**) Four differnet PAMs were assayed directly to verify the PAM determined from the PAM depletion assay. (**panel d**) the generated PAM logos for CasY.2 showed a preference for sequences containing 5'-YR-3' and/or 5'-TR-3' (e.g., 5'-DTR-3') (lower threshold and higher threshold, respectively) flanking sequence 5' of the target (where Y is a T or C; R is an A or G; and D is an A, G, or T). A 3' PAM was not detected.

[00349] **Figure 6.** (**panel a**) 'repeat' sequences from naturaly occurring CasY guide RNAs (For CasY loci Y1-Y6). (**panel b**) **Diagram of CasY RNA guided DNA cleavage.** CasY protein binds to a

crRNA (the CasY guide RNA) in the repeat region (black, repeat; red, spacer). Base pairing of the guide sequence of the guide RNA to the target sequence (blue) containing the correct protospacer adjacent motif (PAM) results in double stranded cleavage of the target DNA.

Example 3: New CRISPR-Cas systems from uncultivated microbes

[00350] CRISPR-Cas adaptive immune systems have revolutionized genome engineering by providing programmable enzymes capable of site-specific DNA cleavage. However, current CRISPR-Cas technologies are based solely on systems from cultured bacteria, leaving untapped the vast majority of enzymes from organisms that have not been isolated. The data provied herein show, using cultivation-independent genome-resolved metagenomics, identification of new CRISPR-Cas systems, including the first reported Cas9 in the archaeal domain of life. This divergent Cas9 enzyme was found in little-studied nanoarchaea as part of an active CRISPR-Cas system. In bacteria, two previously unknown systems were discovered, CRISPR-CasX and CRISPR-CasY, which are among the most streamlined systems yet identified. Notably, all required functional components were identified by metagenomics, which allowed validation of robust RNA-guided DNA interference activity in *E. coli*. The data herein show that interrogation of environmental microbial communities combined with experiments in living cells allows access to an unprecedented diversity of genomes whose content will expand the repertoire of microbe-based biotechnologies.

RESULTS

[00351] Terabase-scale metagenomic datasets from groundwater, sediment, and acid mine drainage microbial communities were analyzed, seeking class 2 CRISPR-Cas systems that are not represented among cultured organisms. The first Cas9 proteins in domain Archaea were identified and two new CRISPR-Cas systems were discovered, CRISPR-CasX and CRISPR-CasY, in uncultivated bacteria (Figure 7). Notably, both the archaeal Cas9 and CasY were encoded exclusively in the genomes of organisms from lineages with no known isolated representatives.

First identification of archaeal Cas9

[00352] One of the hallmarks of CRISPR-Cas9 was its presumed presence only in the bacterial domain. It was therefore surprising to discover Cas9 proteins encoded in genomes of the nanoarchaea ARMAN-1 (*Candidatus* Micrarchaeum acidiphilum ARMAN-1) and ARMAN-4 (*Candidatus* Parvarchaeum acidiphilum ARMAN-4) in acid-mine drainage (AMD) metagenomic datasets. These findings expand the occurrence of Cas9-containing CRISPR systems to another domain of life.

[00353] The ARMAN-4 *cas9* gene was found in 16 different samples in the same genomic context, but with no other adjacent *cas* genes (despite being centrally located in several DNA sequence contigs > 25 kbp), and with only one adjacent CRISPR repeat-spacer unit (Figure 13). The lack of a typical CRISPR array and *cas*1, which encodes the universal CRISPR integrase, points to a system with

no capacity to acquire new spacers. No target could be identified for the spacer sequence, but given the conservation of the locus in samples collected over several years, its function in a "single-target" CRISPR-Cas system cannot be ruled out at this time.

[00354] Conversely, the CRISPR-Cas locus in ARMAN-1, recovered from 15 different samples, includes large CRISPR arrays adjacent to *cas1*, *cas2*, *cas4* and *cas9* genes. Numerous alternative ARMAN-1 CRISPR arrays with a largely conserved end (likely comprised of the oldest spacers) and a variable region into which many distinct spacers have been incorporated were reconstructed (Figure 8a and Figure 14). Based on this hypervariability in spacer content, these data show that the ARMAN-1 CRISPR-Cas9 system is active in the sampled populations.

[00355] Remarkably, 56 of the putative spacer targets (protospacers) of the ARMAN-1 CRISPR-Cas9 system were located on a single 10 kbp genome fragment that is likely an ARMAN-1 virus, given that it encodes a high density of short hypothetical proteins (Figure 8b). Indeed, cryo-electron tomographic reconstructions often identified viral particles attached to ARMAN cells. ARMAN-1 protospacers also derived from a putative transposon within the genome of ARMAN-2 (another nanoarchaeon) and a putative mobile element in the genomes of *Thermoplasmatales* archaea, including that of I-plasma from the same ecosystem (Figure 15). Direct cytoplasmic "bridges" were observed between ARMAN and *Thermoplasmatales* cells, implying a close relationship between them. The ARMAN-1 CRISPR-Cas9 may thus defend against transposon propagation between these organisms, a role that is reminiscent of piRNA-mediated defense against transposition in the eukaryotic germ line.

[00356] Active DNA-targeting CRISPR-Cas systems use 2 to 4 bp protospacer-adjacent motifs (PAMs) located next to target sequences for self versus non-self discrimination. Examining sequences adjacent to the genomic target sequences indeed revealed a strong 'NGG' PAM preference in ARMAN-1 (Figure 8c). Cas9 also employs two separate transcripts, CRISPR RNA (crRNA) and trans-activating CRISPR RNA (tracrRNA), for RNA-guided DNA cleavage. A putative tracrRNA was identified in the vicinity of both ARMAN-1 and ARMAN-4 CRISPR-Cas9 systems (Figure 16). Previously, it was suggested that type II CRISPR systems were absent in archaea due to a lack of the host factor, RNase III, responsible for crRNA-tracrRNA guide complex maturation. Notably, no RNase III homologs have been identified in the ARMAN-1 genome (estimated to be 95% complete) and no internal promoters are predicted for the CRISPR array, suggesting an as-yet undetermined mechanism of guide RNA production. Biochemical experiments to test cleavage activity of ARMAN-1 and ARMAN-4 Cas9 proteins purified from both *E. coli* and yeast and *in vivo E. coli* targeting assays did not reveal any detectable activity (see Figure 21 and Figure 17).

CRISPR-CasX is a new dual-RNA-guided CRISPR system

[00357] In addition to Cas9, only three families of class 2 Cas effector proteins have been discovered and experimentally validated: Cpf1, C2c1, and C2c2. Another gene, *c2c3*, which was identified only on small DNA fragments, has been suggested to also encode such a protein family. A new type of class 2 CRISPR-Cas system was found in the genomes of two bacteria recovered repeatedly from groundwater and sediment samples. The high conservation of this system in two organisms belonging to different phyla, Deltaproteobacteria and Planctomycetes, suggests a recent cross-phyla transfer. This newly described system includes Cas1, Cas2, Cas4 and an uncharacterized ~980 aa protein, referred to herein as CasX. The CRISPR arrays associated with each CasX had highly similar repeats of 37 base pairs, spacers of 33–34 base pairs, and a putative tracrRNA between the Cas operon and the CRISPR array (Figure 7b). BLAST searches revealed only weak similarity (e-value > 1×10^{-4}) to transposases, with similarity restricted to specific regions of the CasX C-terminus. Distant homology detection and protein modeling identified a RuvC domain near the CasX C-terminal end, with organization reminiscent of that found in type V CRISPR-Cas systems (Figure 18). The rest of the CasX protein (630 N-terminal amino acids) showed no detectable similarity to any known protein, suggesting this is a novel class 2 effector. The combination of tracrRNA and separate Cas1, Cas2 and Cas4 proteins is unique among type V systems. Further, CasX is considerably smaller than any known type V proteins: 980 aa compared to a typical size of larger than 1,200 aa for Cpf1, C2c1 and C2c3.

[00358] It was next wondered whether, despite its small size and non-canonical locus content, CasX would be capable of RNA-guided DNA targeting analogous to Cas9 and Cpf1 enzymes. To test this possibility, a plasmid encoding a minimal CRISPR-CasX locus including *casX*, a short repeat-spacer array and intervening noncoding regions was synthesized. When expressed in *E. coli*, this minimal locus blocked transformation by a plasmid bearing a target sequence identified by metagenomic analysis (Figure 9a–c, Figure 19). Furthermore, interference with transformation occurred only when the spacer sequence in the mini-locus matched the protospacer sequence in the plasmid target. To identify a PAM sequence for CasX, the transformation assay was repeated in *E. coli* using a plasmid containing either a 5′ or 3′ randomized sequence adjacent to the target site. This analysis revealed a stringent preference for the sequence 'TTCN' located immediately 5′ of the protospacer sequence (Figure 9d). No 3′ PAM preference was observed (Figure 19). Consistent with this finding, 'TTCA' was the sequence found upstream of the putative Deltaproteobacteria CRISPR-CasX protospacer that was identified in the environmental samples. Notably, both CRISPR-CasX loci share the same PAM sequence, in line with their high degree of CasX protein homology.

[00359] Examples of both single-RNA and dual-RNA guided systems exist among type V CRISPR loci. Environmental meta-transcriptomic data was used to determine whether CasX requires a

tracrRNA for DNA targeting activity. This analysis revealed a non-coding RNA transcript with a sequence complementary to the CRISPR repeat encoded between the Cas2 open reading frame and the CRISPR array (Figure 10a). Transcriptomic mapping further suggests that the CRISPR RNA (crRNA) is processed to include 22 nts of the repeat and 20 nts of the adjacent spacer, similar to the crRNA processing that occurs in CRISPR-Cas9 systems (Figure 10a). Furthermore, a 2-nt 3′ overhang was identified, consistent with RNase III-mediated processing of the crRNA-tracrRNA duplex (Figure 10b). To determine the dependence of CasX activity on the putative tracrRNA, this region was deleted from the minimal CRISPR-CasX locus described above, and repeated the plasmid interference assays. Deletion of the putative tracrRNA-encoding sequence from the CasX plasmid abolished the robust transformation interference observed in its presence (Figure 10c). Together, these results establish CasX as a new functional DNA-targeting, dual-RNA guided CRISPR enzyme.

CRISPR-CasY, a system found exclusively in bacterial lineages lacking isolates

[00360] Another new class 2 Cas protein encoded in the genomes of certain candidate phyla radiation (CPR) bacteria was identified. These bacteria typically have small cell sizes (based on cryo-TEM data and enrichment via filtration), very small genomes and a limited biosynthetic capacity, indicating they are most likely symbionts. The new ~1,200 aa Cas protein, referred to herein as CasY, appears to be part of a minimal CRISPR-Cas system that includes, at most, Cas1 and a CRISPR array (Figure 11a). Most of the CRISPR arrays have unusually short spacers of 17–19 nts, but one system, which lacks Cas1 (CasY.5), has longer spacers (27–29 nts). The six examples of CasY proteins identified had no significant sequence similarity to any protein in public databases. A sensitive search using profile models (HMMs) built from published Cas proteins[3,4] indicated that four of the six CasY proteins had local similarities (e-values 4×10^{-11} – 3×10^{-18}) to C2c3 in the C-terminal region overlapping the RuvC domains and a small region (~45 aa) of the N-terminus (see Figure 18). C2c3 are putative type V Cas effectors that were identified on short contigs with no taxonomic affiliation, and have not been validated experimentally. Like CasY, the C2c3 were found next to arrays with short spacers and Cas1, but with no other Cas proteins. Notably, two of the CasY proteins identified in the current study had no significant similarity to C2c3, despite sharing significant sequence similarity (best Blast hits: e-values 6×10^{-85}, 7×10^{-75}) with the other CasY proteins.

[00361] Given the low homology of CRISPR-CasY to any experimentally validated CRISPR loci, it was next wondered whether this system confers RNA-guided DNA interference, but due to the short spacer length reliable information did not exist about a possible PAM motif that might be required for such activity. To work around this, the entire CRISPR-CasY.1 locus was synthesized with a shortened CRISPR array and introduced into *E. coli* on a plasmid vector. These cells were then challenged in a transformation assay using a target plasmid with a sequence matching a spacer sequence in the array and

containing an adjacent randomized 5′ or 3′ region to identify a possible PAM. Analysis of transformants revealed depletion of sequences containing a 5′ TA directly adjacent to the targeted sequence (Figure 11b). Using this identified PAM sequence, the CasY.1 locus was tested against plasmids containing a single PAM. Plasmid interference was demonstrated only in the presence of a target containing the identified 5′ TA PAM sequence (Figure 11c). Thus, these data show that CRISPR-CasY has DNA interference activity.

Discussion

[00362] New class 2 CRISPR-Cas adaptive immune systems in genomes from uncultivated bacteria and archaea were identified and characterized. Evolutionary analysis of Cas1 (Figure 12a), which is universal to active CRISPR loci, suggested that the archaeal Cas9 system described here does not clearly fall into any existing type II subtype. The Cas1 phylogeny (as well as the existence of *cas4*) clustered it together with type II-B systems, yet the sequence of Cas9 was more similar to type II-C proteins (Figure 20). Thus, the archaeal type II system may have arisen as a fusion of type II-C and II-B systems (Figure 12b). Likewise, Cas1 phylogenetic analyses indicated that the Cas1 from the CRISPR-CasX system is distant from any other known type V system. Type V systems have been suggested to be the result of the fusion of a transposon with the adaptation module (Cas1–Cas2) from an ancestral type I system. It is therefore hypothesized that the CRISPR-CasX system emerged following a fusion event different from those that gave rise to the previously described type V systems. Strikingly, both CRISPR-CasY and the putative C2c3 systems seem to lack Cas2, a protein thought to be essential for integrating DNA into the CRISPR locus. Given that all CRISPR-Cas systems are thought be descendants of an ancestral type I system that contained both Cas1 and Cas2, CRISPR-CasY and C2c3 systems may either have different ancestry than the rest of the CRISPR-Cas systems, or alternatively, Cas2 might have been lost during their evolutionary history.

[00363] The discovery described herein of Cas9 in archaea and two previously unknown CRISPR-Cas systems in bacteria used extensive DNA and RNA sequence datasets obtained from complex natural microbial communities. In the case of CasX and CasY, genome context was critical to prediction of functions that would not have been evident from unassembled sequence information. Further, the identification of a putative tracrRNA as well as targeted viral sequences uncovered through analysis of the metagenomic data guided functional testing. Interestingly, some of the most compact CRISPR-Cas loci identified to date were discovered in organisms with very small genomes. A consequence of small genome size is that these organisms likely depend on other community members for basic metabolic requirements, and thus they have remained largely outside the scope of traditional cultivation-based methods. The limited number of proteins that are required for interference make these minimal systems especially valuable for the development of new genome editing tools. Importantly, it is

shown herein that metagenomic discoveries related to CRISPR-Cas systems are not restricted to *in silico* observations, but can be introduced into an experimental setting where their function can be tested. Given that virtually all environments where life exists can now be probed by genome-resolved metagenomic methods, it is anticipated that the combined computational-experimental approach described herein will greatly expand the diversity of known CRISPR-Cas systems, providing new technologies for biological research and clinical applications.

METHODS

Metagenomics and metatranscriptomics

[00364] Metagenomic samples from three different sites were analyzed: (1) Acid mine drainage (AMD) samples collected between 2006 and 2010 from the Richmond Mine, Iron Mountain, California (2) Groundwater and sediment samples collected between 2007 and 2013 from the Rifle Integrated Field Research (IFRC) site, adjacent to the Colorado River near Rifle, Colorado. (3) Groundwater collected in 2009 and 2014 from Crystal Geyser, a cold, CO_2-driven geyser on the Colorado Plateau in Utah.

[00365] For the AMD data, DNA extraction methods and short read sequencing were reported by Denef and Banfield (2012) and Miller et al. (2011). For the Rifle data, DNA and RNA extraction, as well as sequencing, assembly, and genomic reconstructed were described by Anantharaman et al. (2016) and Brown et al. (2015). For samples from Crystal Geyser, methods follow those described by Probst et al (2016) and Emerson et al. (2015). Briefly, DNA was extracted from samples using the PowerSoil DNA Isolation Kit (MoBio Laboratories Inc., Carlsbad, CA, USA). RNA was extracted from 0.2 μm filters collected from six 2011 Rifle groundwater samples, as described by Brown et al. (2015). DNA was sequenced on Illumina HiSeq2000 platform, and Metatranscriptomic cDNA on 5500XL SOLiD platform. For the newly reported Crystal Geyser data and reanalysis of the AMD data, sequences were assembled using IDBA-UD. DNA and RNA (cDNA) read-mapping used to determine sequencing coverage and gene expression, respectively, was performed using Bowtie2. Open reading frames (ORFs) were predicted on assembled scaffolds using Prodigal. Scaffolds from the Crystal Geyser dataset were binned on the basis of differential coverage abundance patterns using a combination of ABAWACA, ABAWACA2 (https://github.com/CK7) Maxbin2, and tetranucleotide frequency using Emergent Self-Organizing Maps (ESOM). Genomes were manually curated using % GC content, taxonomic affiliation, and genome completeness. Scaffolding errors were corrected using ra2.py (https://github.com/christophertbrown).

CRISPR-Cas computation analysis

[00366] The assembled contigs from the various samples were scanned for known Cas proteins using Hidden Markov Model (HMMs) profiles, which were built using the HMMer suite, based on alignments from Makarova *et al.* and Shmakov *et al.*. CRISPR arrays were identified using a local version of the CrisprFinder software. Loci that contained both Cas1 and a CRISPR array were further analyzed if

one of the ten ORFs adjacent to the *cas1* gene encoded for an uncharacterized protein larger than 800 aa, and no known *cas* interference genes were identified on the same contig. These large proteins were further analyzed as potential class 2 Cas effectors. The potential effectors were clustered to protein families based on sequence similarities using MCL. These protein families were expanded by building HMMs representing each of these families, and using them to search the metagenomic datasets for similar Cas proteins. To make sure that the protein families are indeed new, known homologs were searched using BLAST against NCBI's non-redundant (nr) and metagenomic (env_nr) protein databases, as well as HMM searches against the UniProt KnowledegeBase. Only proteins with no full-length hits (> 25% of the protein's length) were considered novel proteins. Distant homology searches of the putative Cas proteins were performed using HHpred from the HH-suite. High scoring HHpred hits were used to infer domain architecture based on comparison to resolved crystal structures, and secondary structure that was predicted by JPred4. The HMM database, including the newly discovered Cas proteins are available in Supplementary Data 1.

[00367] Spacer sequences were determined from the assembled data using CrisprFinder. CRASS was used to locate additional spacers in short DNA reads of the relevant samples. Spacer targets (protospacers) were then identified by BLAST searches (using "-task blastn-short") against the relevant metagenomic assemblies for hits with ≤ 1 mismatch to spacers. Hits belonging to contigs that contained an associated repeat were filtered out (to avoid identifying CRISPR arrays as protospacers). Protospacer adjacent motifs (PAMs) were identified by aligning regions flanking the protospacers and visualized using WebLogo. RNA structures were predicted using mFold. CRISPR array diversity was analyzed by manually aligning spacers, repeats and flanking sequences from the assembled data. Manual alignments and contig visualizations were performed with Geneious 9.1.

[00368] For the phylogenetic analyses of Cas1 and Cas9 proteins of the newly identified systems were used along with the proteins from Makarova *et al.* and Shmakov *et al.*. A non-redundant set was compiled by clustering together proteins with ≥ 90% identity using CD-HIT. Alignments were produced with MAFFT, and maximum-likelihood phylogenies were constructed using RAxML with PROTGAMMALG as the substitution model and 100 bootstrap samplings. Cas1 tree were rooted using the branch leading to casposons. Trees were visualized using FigTree 1.4.1 (http://tree.bio.ed.ac.uk/software/figtree/) and iTOL v3.

Generation of heterologous plasmids

[00369] Metagenomic contigs were made into minimal CRISPR interference plasmids by removing proteins associated with acquisition for CasX and reducing the size of the CRISPR array for both CasX and CasY. The minimal locus was synthesized as Gblocks (Integrated DNA Technology) and assembled using Gibson Assembly.

PAM depletion assay

[00370] PAM depletion assays were conducted as previously described with modification. Plasmid libraries containing randomized PAM sequences were assembled by annealing a DNA oligonucleotide containing a target with a 7 nt randomized PAM region with a primer and extended with Klenow Fragment (NEB). The double stranded DNA was digested with EcoRI and NcoI and ligated into a pUC19 backbone. The ligated library was transformed into DH5α and $>10^8$ cells were harvested and the plasmids extracted and purified. 200 ng of the pooled library was transformed into electrocompetent *E. coli* harboring a CRISPR locus or a control plasmid with no locus. The transformed cells were plated on selective media containing carbenicillin (100 mg L^{-1}) and chloramphenicol (30 mg L^{-1}) for 30 hours at 25°C. Plasmid DNA was extracted and the PAM sequence was amplified with adapters for Illumina sequencing. The 7 nt PAM region was extracted and PAM frequencies calculated for each 7 nt sequence. PAM sequences depleted above the specified threshold were used to generate a WebLogo.

Plasmid Interference

[00371] Putative targets identified from metagenomic sequence analysis or PAM depletion assays were cloned into a pUC19 plasmid. 10 ng of target plasmid were transformed into electrocompetent *E. coli* (NEB Stable) containing the CRISPR loci plasmid. Cells were recovered for 2 hrs at 25°C and an appropriate dilution was plated on selective media. Plates were incubated at 25°C and colony forming units were counted. All plasmid interference experiments were performed in triplicate and electrocompetent cells were prepared independently for each replicate.

ARMAN-Cas9 protein expression and purification

[00372] Expression constructs for Cas9 from ARMAN-1 (AR1) and ARMAN-4 (AR4) were assembled from gBlocks (Integrated DNA Technologies) that were codon-optimized for *E. coli*. The assembled genes were cloned into a pET-based expression vector as an N-terminal His_6-MBP or His_6 fusion protein. Expression vectors were transformed into BL21(DE3) *E. coli* cells and grown in LB broth at 37°C. For protein expression, cells were induced during mid-log phase with 0.4 mM IPTG (isopropyl β-D-1-thiogalactopyranoside) and incubated overnight at 16°C. All subsequent steps were conducted at 4°C. Cell pellets were resuspended in lysis buffer (50mM Tris-HCl pH 8, 500 mM NaCl, 1 mM TCEP, 10 mM Imidazole) 0.5% Triton X-100 and supplemented with Complete protease inhibitor mixture (Roche) before lysis by sonication. Lysate was clarified by centrifugation at 15000g for 40 min and applied to Superflow Ni-NTA agarose (Qiagen) in batch. The resin was washed with extensively with Wash Buffer A (50 mM Tris-HCl pH 8, 500 mM NaCl, 1mM TCEP, 10 mM Imidazole) followed by 5 column volumes of Wash Buffer B (50 mM Tris-HCl pH 8, 1M NaCl, 1 mM TCEP, 10 mM Imidazole). Protein was eluted off of Ni-NTA resin with Elution Buffer (50 mM Tris-HCl pH 8, 500 mM NaCl, 1 mM TCEP, 300 mM Imidazole). The His_6-MBP tag was removed by TEV protease during overnight dialysis against

Wash Buffer A. Cleaved Cas9 was removed from the affinity tag through a second Ni-NTA agarose column. The protein was dialyzed into IEX Buffer A (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 1 mM TCEP, 5% glycerol) before application to a 5 mL Heparin HiTrap column (GE Life Sciences). Cas9 was eluted over a linear NaCl (0.3-1.5 M) gradient. Fractions were pooled and concentrated with a 30 kDa spin concentrator (Thermo Fisher). When applicable, Cas9 was further purified via size-exclusion chromatography on an Superdex 200 pg column (GE Life Sciences) and stored in IEX Buffer A for subsequent cleavage assays. For yeast expression, AR1-Cas9 was cloned into a Gal1/10 His6-MBP TEV Ura S. cerevisiae expression vector (Addgene plasmid # 48305). The vector was transformed into a BY4741 URA3 strain and cultures were grown in MEDIA at 30˚C. At an OD600 of ~0.6, protein expression was induced with 2% w/v galactose and incubated overnight at 16˚C. Protein purification was performed as above.

RNA in vitro transcription and oligonucleotide purification

[00373] *In vitro* transcription reactions were performed as previously described[65] using synthetic DNA templates containing a T7 promoter sequence. All *in vitro* transcribed guide RNAs and target RNAs or DNAs were purified via denaturing PAGE. Double-stranded target RNAs and DNAs were hybridized in 20 mM Tris HCl pH 7.5 and 100 mM NaCl by incubation at 95°C for 1 min, followed by slow-cooling to room temperature. Hybrids were purified by native PAGE.

In vitro cleavage assays

[00374] Purified DNA and RNA oligonucleotides were radiolabeled using T4 polynucleotide kinase (NEB) and [γ-32P] ATP (Perkin-Elmer) in 1x PNK buffer for 30 min at 37˚C. PNK was heat inactivated at 65°C for 20 min and free ATP was removed from the labeling reactions using illustra Microspin G-25 columns (GE Life Sciences). CrRNA and tracrRNAs were mixed in equimolar quantities in 1x refolding buffer (50 mM Tris HCl pH 7.5, 300 mM NaCl, 1 mM TCEP, 5% glycerol) and incubated at 70˚C for 5 min and then slow-cooled to room temperature. The reactions were supplemented to 1 mM final metal concentration and subsequently heated at 50˚C for 5 min. After slow-cooling to room temperature, refolded guides were placed on ice. Unless noted for buffer, salt concentration, Cas9 was reconstituted with an equimolar amount of guide in 1x cleavage buffer (50 mM Tris HCl pH 7.5, 300 mM NaCl, 1 mM TCEP, 5% glycerol, 5 mM divalent metal) at 37˚C for 10 min. Cleavage reactions were conducted in 1x cleavage buffer with a 10x excess of Cas9-guide complex over radiolabeled target at 37˚C or the indicated temperature. Reactions were quenched in an equal volume of gel loading buffer supplemented with 50mM EDTA. Cleavage products were resolved on 10% denaturing PAGE and visualized by phosphorimaging.

In vivo *E. coli* interference assays

[00375] *E. coli* transformation assays for AR1- and AR4-Cas9 were conducted as previously published[66]. Briefly, E. coli transformed with guide RNAs were made electrocompetent. Cells were then transformed with 9 fmol of plasmid encoding wild-type or catalytically inactive Cas9 (dCas9). A dilution series of recovered cells was plated on LB plates with selective antibiotics. Colonies were counted after 16 hr at 37˚C.

[00376] **Table 1.** Details regarding the organisms and genomic location in which the CRISPR-Cas system were identified, as well as information on the number and average length of reconstructed spacers, and repeats length (NA, not available). ARMAN-1 spacers were reconstructed from 16 samples.

Taxonomic group	Cas effector	NCBI Accession	Coordinates	Repeat length	# spacers	Spacers avg. length
ARMAN-1	Cas9	MOEG01000017	1827..7130	36	271	34.5
ARMAN-4	Cas9	KY040241	11779..14900	36	1	36
Deltaproteobacteria	CasX	MGPG01000094	4319..9866	37	5	33.6
Planctomycetes	CasX	MHYZ01000150	1..5586	37	7	32.3
Candidatus Katanobacteria	CasY.1	MOEH01000029	459..5716	26	14	17.1
Candidatus Vogelbacteria	CasY.2	MOEJ01000028	7322..13087	26	18	17.3
Candidatus Vogelbacteria	CasY.3	MOEK01000006	1..4657	26	12	17.3
Candidatus Parcubacteria	CasY.4	KY040242	1..5193	25	13	18.4
Candidatus Komeilibacteria	CasY.5	MOEI01000022	2802..7242	36	8	26
Candidatus Kerfeldbacteria	CasY.6	MHKD01000036	11503..15366	NA	NA	NA

[00377] 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.