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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 January 2025 (02.01.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/003358 A2

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

(21) International Application Number:
PCT/EP2024/068177

(22) International Filing Date:
27 June 2024 (27.06.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/510,918 29 June 2023 (29.06.2023) US

(71) Applicant: **UCB BIOPHARMA SRL** [BE/BE]; Allée de la Recherche, 60, 1070 Brussels (BE).

(72) Inventors: **BOWEN, Tyson David**; c/o IPD, UCB Biopharma SRL, Allée de la Recherche, 60, 1070 Brussels (BE). **RIEBER, Lila Herk**; c/o IPD, UCB Biopharma SRL, Allée de la Recherche, 60, 1070 Brussels (BE). **WANG, Meng**; c/o IPD, UCB Biopharma SRL, Allée de la Recherche, 60, 1070 Brussels (BE).

(74) Agent: **UCB INTELLECTUAL PROPERTY**; UCB, 208 Bath Road, Slough Berkshire SL1 3WE (GB).

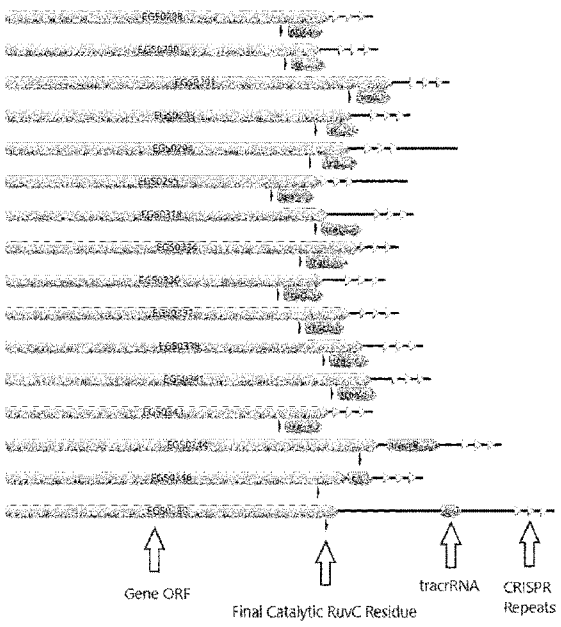
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ,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

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

(54) Title: NOVEL NUCLEIC ACID TARGETING SYSTEMS COMPRISING RNA-GUIDED NUCLEASES

(57) Abstract: The present invention provides novel nucleic acid targeting system comprising RNA-guided nuclease proteins for cleaving and/or modifying the target nucleotide of interest.

Figure 1



WO 2025/003358 A2

[Continued on next page]

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
— *of inventorship (Rule 4.17(iv))*

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*
— *with sequence listing part of description (Rule 5.2(a))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

NOVEL NUCLEIC ACID TARGETING SYSTEMS COMPRISING RNA-GUIDED NUCLEASES

[001] The present invention relates to novel RNA-guided nucleases (RGN) and nucleic acid targeting systems comprising such.

BACKGROUND

[002] Targeted genome editing or modification has been undergoing many changes in the past years since the discovery of novel technologies and systems. First systems relied on meganucleases, zinc finger fusion proteins or Transcription activator-like effector nucleases (TALENs), requiring the generation of chimeric nucleases with engineered, sequence- specific DNA-binding domains specific for each particular target sequence. RNA-guided nucleases (RGNs), such as the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated (Cas) proteins allow for the targeting of specific sequences by using a short RNA sequence that specifically hybridizes with a particular target sequence. Such CRISPR systems became popular and gained multiple uses in research, diagnostics and therapeutics due to the ease of production of target-specific short RNA sequences and use of such with the same RGN protein. Such RGNs can be used to edit genomes through the introduction of a sequence-specific, double -stranded break that is either repaired and introduces a mutation or repaired by introducing a stretch of heterologous DNA. Inactive versions RGNs has been also widely used to target specific DNA or RNA regions and in combination with other proteins allowed to study and modulate multiple cellular processes and provide a useful tool for gene function study and modulation of their activity.

[003] Type V-U4 CRISPR systems have been described in WO2018/035250 including some exemplary RGN sequences identified using bioinformatics methods. However, no guidance to how to make those RGNs functional for gene editing tools had been provided.

SUMMARY OF THE INVENTION

[004] The present invention provides type V-F5 (previously identified as Type V-U4) nucleic acid targeting systems comprising RGNs and RNA molecules, nucleic acid molecules encoding the same, and vectors and host cells comprising such nucleic acid molecules.

[005] Also provided are nucleic acid targeting systems for binding a target nucleic acid sequence of interest, wherein the system comprises a RGN polypeptide and one or more RNA sequences targeting the nucleic acid of interest.

[006] Thus, methods disclosed herein are drawn to binding a target sequence of interest, and in some embodiments, cleaving or modifying the target sequence of interest. The target sequence of interest can

be modified, for example, as a result of non-homologous end joining or homology-directed repair with an introduced donor sequence.

BRIEF DESCRIPTION OF THE DRAWINGS

[007] The present invention is described below by reference to the following figures.

[008] Figure 1 represents schematically the location of the RGNs, final catalytic RuvC residue, tracrRNA and CRISPR repeats within the locus.

[009] Figure 2 shows the activity of different truncations of EGS0293 RGN.

[010] Figure 3 shows the performance of several sgRNA truncation designs for EGS0293

[011] Figure 4 shows the performance of several sgRNA stabilization designs for EGS0293

[012] Figure 5 shows the performance of DNA affinity mutations of EGS0293

[013] Figure 6 shows the Recognition Sequence Diversity of Cas12 effector proteins**.** The first two principal components (PCs) in the PC decomposition of PAM position weight matrices of diverse Cas12-related effector proteins

DETAILED DESCRIPTION OF THE INVENTION

Definitions

[014] Table 1. Abbreviations used throughout the specification

Cas	CRISPR associated Sequence
Cas9	Cas protein 9
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
gRNA	long monomeric nucleic acid targeting RNAs
RGN	RNA-guided nuclease
ssDNA	Single stranded DNA
dsDNA	Double stranded DNA

[015] Table 2. Amino acids abbreviations

Abbreviation	**1 letter abbreviation**	**Amino acid name**
Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine

Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Pyl	O	Pyrrolysine
Ser	S	Serine
Sec	U	Selenocysteine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine

[016] Table 3. Nucleotide Code abbreviations

Abbreviation	Nucleotide
A	Adenine
G	Guanine
C	Cytosine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C or T)
N	Any nucleotide
W	Weak (A or T)
S	Strong (G or C)
M	Amino (A or C)
K	Keto (G or T)
B	Not A (G or C or T)
H	Not G (A or C or T)
D	Not C (A or G or T)
V	Not T (A or G or C)

[017] The following definitions are used throughout the description.

[018] The term "adeno-associated virus" or "AAV" as used interchangeably herein refers to a small virus belonging to the genus Dependovirus of the Parvoviridae family that infects humans and some other primate species. AAV is not currently known to cause disease and consequently the virus causes a very mild immune response.

[019] As used herein, a “biological sample” may contain whole cells and/or live cells and/or cell debris. The biological sample may contain (or be derived from) a “bodily fluid”. Bodily fluids may be obtained from a mammal organism, for example by puncture, or other collecting or sampling procedures.

[020] The term “Cas12f1” refers to type of an RGN that cleaves nucleic acid and is encoded by the CRISPR loci and is a part of the Type VF1 CRISPR system. The Cas12f1 protein commonly used is from an uncultured archaeon (Un1). The Cas12f1 protein may be mutated so that the nuclease activity is partly or completely inactivated. Cas12f1 RGNs are described in Harrington et al (2018). Science, 362(6416), 839–842 and Karvelis et al (2020) Nucleic acids research, 48(9), 5016–5023.

[021] The term “Cas12f5” or “c2c9” refers to type of an RGN that cleaves nucleic acid and is encoded by the CRISPR loci and is a part of a subtype of the Type V-F5 CRISPR system. The Cas12f5 protein consists of a Rec1 domain and tri-split RuvC domain and may be mutated so that the nuclease activity is partly or completely inactivated.

[022] The term “Cas9” refers to type of an RGN that cleaves nucleic acid and is encoded by the CRISPR loci and is a part of the Type II CRISPR system. The Cas9 protein commonly used is from bacterial species Streptococcus pyogenes. The Cas9 protein may be mutated so that the nuclease activity is partly or completely inactivated.

[023] The term "complement" or "complementary" as used herein means a nucleic acid can mean Watson-Crick or Hoogsteen base pairing between nucleotides or nucleotide analogs of nucleic acid molecules. The term "complementarity" refers to a property shared between two nucleic acid sequences, such that when they are aligned antiparallel to each other, the nucleotide bases at each position will be complementary.

[024] The term “CRISPR” (Clustered Regularly Interspaced Short Palindromic Repeats) refers to a family of DNA sequences found in the genomes of prokaryotic organisms such as bacteria and archaea. These sequences are derived from DNA fragments of bacteriophages that had previously infected the prokaryote. They are used to detect and destroy DNA from similar bacteriophages during subsequent infections.

[025] The term "CRISPR system" refers collectively to transcripts and other elements involved in the expression of or directing the activity of CRISPR-associated ("Cas") proteins, including sequences encoding a Cas protein, a tracr (trans -activating CRISPR) sequence (e.g. tracrRNA or an active partial tracrRNA), a tracr-mate sequence (containing a "direct repeat" and a tracrRNA-processed partial direct repeat in the context of an endogenous CRISPR system), a guide sequence (also referred herein to as a "spacer" in the context of an endogenous CRISPR system), or other sequences and transcripts from a CRISPR locus.

[026] The term “effective amount,” as used herein, refers to an amount of a biologically active agent that is sufficient to elicit a desired biological response. For example, in some embodiments, an effective

amount of a nuclease may refer to the amount of the nuclease that is sufficient to induce cleavage of a target site specifically bound and cleaved by the nuclease. In some embodiments, an effective amount of a recombinase may refer to the amount of the recombinase that is sufficient to induce recombination at a target site specifically bound and recombined by the recombinase. As will be appreciated by the skilled artisan, the effective amount of an agent, e.g., a nuclease, a recombinase, a hybrid protein, a fusion protein, a protein dimer, a complex of a protein (or protein dimer) and a polynucleotide, or a polynucleotide, may vary depending on various factors as, for example, on the desired biological response, the specific allele, genome, target site, cell, or tissue being targeted, and the agent being used.

[027] The term "enhancer" as used herein refers to non-coding DNA sequences containing multiple activator and repressor binding sites. Enhancers range from 200 bp to 1 kb in length and may be either proximal, 5' upstream to the promoter or within the first intron of the regulated gene, or distal, in introns of neighboring genes or intergenic regions far away from the locus. Through DNA looping, active enhancers contact the promoter dependently of the core DNA binding motif promoter specificity. 4 to 5 enhancers may interact with a promoter.

[028] As used herein, the term "fusion protein" refers to a chimeric protein created through the covalent or non-covalent joining of two or more genes, directly or indirectly, that originally coded for separate proteins. In some embodiments, the translation of the fusion gene results in a single polypeptide with functional properties derived from each of the original proteins.

[029] The term “gRNA”, also used interchangeably herein as a chimeric single guide RNA (“sgRNA”), refers to nucleic acid which is a fusion of two noncoding RNAs: a crRNA and a tracrRNA. “gRNA” is used interchangeably to refer to guide RNAs that exist as either single molecules or as a complex of two or more molecules. Typically, gRNAs that exist as single RNA species comprise two domains:(1) a domain that shares homology to a target nucleic acid (e.g., and directs binding of a Cas complex to the target); and (2) a domain that binds a Cas protein.

[030] An "isolated" or "purified" polypeptide, or biologically active portion thereof, is substantially or essentially free from components that normally accompany or interact with the polypeptide as found in its naturally occurring environment. Thus, an isolated or purified polypeptide is substantially free of other cellular material, or culture medium when produced by recombinant techniques, or substantially free of chemical precursors or other chemicals when chemically synthesized. A protein that is substantially free of cellular material includes preparations of protein having less than 30%, 20%, 10%, 5%, or 1% (by dry weight) of contaminating protein. When a protein or biologically active portion thereof is recombinantly produced, optimally culture medium represents less than 30%, 20%, 10%, 5%, or 1% (by dry weight) of chemical precursors or non-protein-of-interest chemicals.

[031] The term "leader sequence" refers to the final region of the CRISPR repeat before the reprogrammable spacer that does not base pair with the final portion of the tracrRNA known as the antirepeat. This leader sequence may be useful for interactions with other regions of the tracrRNA or with the Cas12f5 protein itself.

[032] The term "linker," as used herein, refers to a chemical group or a molecule linking two molecules or moieties, e.g., a binding domain and a cleavage domain of a nuclease. Typically, the linker is positioned between, or flanked by, two groups, molecules, or other moieties and connected to each one via a covalent bond, thus connecting the two. In some embodiments, the linker is an amino acid or a plurality of amino acids (e.g., a peptide or protein). In some embodiments, the linker is an organic molecule, group, polymer, or chemical moiety. In some embodiments, the linker is a polypeptide of 5-100 amino acids in length, for example, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 30-35, 35-40, 40-45, 45-50, 50-60, 60-70, 70-80, 80-90, 90-100, 100-150, or 150-200 amino acids in length. Longer or shorter linkers are also contemplated.

[033] The term "modification" in reference to a nucleic acid molecule refers to a change in the nucleotide sequence of the nucleic acid molecule, which can be a deletion, insertion, or substitution of one or more nucleotides, or a combination thereof.

[034] The term "mutation," as used herein, refers to a substitution of a residue within a sequence, e.g., a nucleic acid or amino acid sequence, with another residue, or a deletion or insertion of one or more residues within a sequence. Mutations are typically described herein by identifying the original residue followed by the position of the residue within the sequence and by the identity of the newly substituted residue. Various methods for making the amino acid substitutions (mutations) provided herein are well known in the art, and are provided by, for example, Green and Sambrook, Molecular Cloning: A Laboratory Manual (4th ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (2012)).

[035] As used herein, the terms "nucleic acid," "nucleic acid sequence," "nucleotide sequence," "oligonucleotide," and "polynucleotide" are interchangeable and refer to a polymeric form of nucleotides. The nucleotides may be deoxyribonucleotides (DNA), ribonucleotides (RNA), analogs thereof, or combinations thereof, and may be of any length. Polynucleotides may perform any function and may have any secondary and tertiary structures. The terms encompass known analogs of natural nucleotides and nucleotides that are modified in the base, sugar and/or phosphate moieties. Analogs of a particular nucleotide have the same base-pairing specificity (e.g., an analog of A base pairs with T). A polynucleotide may comprise one modified nucleotide or multiple modified nucleotides. Examples of modified nucleotides include fluorinated nucleotides, methylated nucleotides, and nucleotide analogs. Nucleotide structure may be modified before or after a polymer is assembled. Following polymerization,

polynucleotides may be additionally modified via, for example, conjugation with a labeling component or target binding component. A nucleotide sequence may incorporate non-nucleotide components. The terms also encompass nucleic acids comprising modified backbone residues or linkages, which are synthetic, naturally occurring, and non-naturally occurring, and have similar binding properties as a reference polynucleotide (e.g., DNA or RNA). Examples of such analogs include, but are not limited to, phosphorothioates, phosphoramidates, methyl phosphonates, chiral-methyl phosphonates, 2-O-methyl ribonucleotides, peptide-nucleic acids (PNAs), Locked Nucleic Acid (LNA™) (Exiqon, Inc., Woburn, MA) nucleosides, glycol nucleic acid, bridged nucleic acids, and morpholino structures. Polynucleotide sequences are displayed herein in the conventional 5' to 3' orientation unless otherwise indicated.

[036] The term “operably linked” as used herein means that expression of a gene is under the control of a promoter with which it is spatially connected. A promoter may be positioned 5′ (upstream) or 3′ (downstream) of a gene under its control. The distance between the promoter and a gene may be approximately the same as the distance between that promoter and the gene it controls in the gene from which the promoter is derived. As is known in the art, variation in this distance may be accommodated without loss of promoter function.

[037] The term “optional” or “optionally” means that the subsequent described event, circumstance or substituent may or may not occur, and that the description includes instances where the event or circumstance occurs and instances where it does not.

[038] As used herein, the terms "peptide," "polypeptide," and "protein" are interchangeable and refer to polymers of amino acids. A polypeptide may be of any length. It may be branched or linear, it may be interrupted by non-amino acids, and it may comprise modified amino acids. The terms may be used to refer to an amino acid polymer that has been modified through, for example, acetylation, disulfide bond formation, glycosylation, lipidation, phosphorylation, cross- linking, and/or conjugation (e.g., with a labeling component or ligand). Polypeptide sequences are displayed herein in the conventional N-terminal to C-terminal orientation. Polypeptides and polynucleotides can be made using routine techniques in the field of molecular biology (see, e.g., standard texts set forth above). Further, essentially any polypeptide or polynucleotide can be custom ordered from commercial sources.

[039] As used herein, "percentage of sequence identity" means the value determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (i. e. , gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched

positions, dividing the number of matched positions by the total number of positions in the window of comparison, and multiplying the result by 100 to yield the percentage of sequence identity.

[040] The term "promoter" as used herein means a synthetic or naturally-derived molecule which is capable of conferring, activating or enhancing expression of a nucleic acid in a cell. A promoter may comprise one or more specific transcriptional regulatory sequences to further enhance expression and/or to alter the spatial expression and/or temporal expression of same. A promoter may also comprise distal enhancer or repressor elements, which may be located as much as several thousand base pairs from the start site of transcription. A promoter may be derived from sources including viral, bacterial, fungal, plants, insects, and animals.

[041] The term “RNA-guided endonuclease” or “RGN” is used interchangeably herein and refer to a nuclease that forms a complex with (e.g., binds or associates with) one or more RNA that is not a target for cleavage.

[042] As used herein, "sequence identity" or "identity" in the context of two polynucleotides or polypeptide sequences makes reference to the residues in the two sequences that are the same when aligned for maximum correspondence over a specified comparison window. When percentage of sequence identity is used in reference to proteins it is recognized that residue positions which are not identical often differ by conservative amino acid substitutions, where amino acid residues are substituted for other amino acid residues with similar chemical properties (e.g., charge or hydrophobicity) and therefore do not change the functional properties of the molecule. When sequences differ in conservative substitutions, the percent sequence identity may be adjusted upwards to correct for the conservative nature of the substitution.

[043] Sequences that differ by such conservative substitutions are said to have "sequence similarity" or "similarity". Means for making this adjustment are well known to those of skill in the art. Typically this involves scoring a conservative substitution as a partial rather than a full mismatch, thereby increasing the percentage sequence identity. Thus, for example, where an identical amino acid is given a score of 1 and a non-conservative substitution is given a score of zero, a conservative substitution is given a score between zero and 1. The scoring of conservative substitutions is calculated, e.g. , as implemented in the program PC/GENE (Intelligenetics, Mountain View, California).

[044] As used herein the term “spacer sequence” or “spacer” refers to a part of gRNA nucleotide sequence that directly hybridizes with the target nucleotide sequence of interest.

[045] The term "subject" and "patient" as used herein interchangeably refers to any vertebrate, including, but not limited to, a mammal {e.g., cow, pig, camel, llama, horse, goat, rabbit, sheep, hamsters, guinea

pig, cat, dog, rat, and mouse, a non-human primate (for example, a monkey, such as a cynomolgous or rhesus monkey, chimpanzee, etc.) and a human). In some embodiments, the subject may be a human or a non-human. The subject or patient may be undergoing other forms of treatment.

[046] The term "target region", "target sequence" or "protospacer" as used interchangeably herein refers to the region of the target gene to which the CRISPR-based system targets.

[047] The terms "treatment," "treat," and "treating," refer to a clinical intervention aimed to reverse, alleviate, delay the onset of, or inhibit the progress of a disease or disorder, or one or more symptoms thereof, as described herein. As used herein, the terms "treatment," "treat," and "treating" refer to a clinical intervention aimed to reverse, alleviate, delay the onset of, or inhibit the progress of a disease or disorder, or one or more symptoms thereof, as described herein. In some embodiments, treatment may be administered after one or more symptoms have developed and/or after a disease has been diagnosed. In other embodiments, treatment may be administered in the absence of symptoms, e.g., to prevent or delay onset of a symptom or inhibit onset or progression of a disease. For example, treatment may be administered to a susceptible individual prior to the onset of symptoms (e.g., in light of a history of symptoms and/or in light of genetic or other susceptibility factors). Treatment may also be continued after symptoms have resolved, for example, to prevent or delay their recurrence

[048] The term "Type II CRISPR system" refers to effector system that carries out targeted DNA double-strand break in four sequential steps, using a single effector enzyme, Cas9, to cleave dsDNA. Compared to the Type I and Type III effector systems, which require multiple distinct effectors acting as a complex, the Type II effector system may function in alternative contexts such as eukaryotic cells. The Type II effector system consists of a long pre-crRNA, which is transcribed from the spacer-containing CRISPR locus, the Cas9 protein, and a tracrRNA, which is involved in pre-crRNA processing.

[049] The term "Type V-F5" refers to a novel type of CRISPR system provided in this disclosure comprising an effector protein, such as a RGN, located neara CRISPR repeat spacer array. No other common CRISPR proteins are found nearby. Additionally, the system comprises a trans-activating crRNA (tracrRNA) within 900 nt of the last catalytic RuvC domain of the potential Cas12f5, including regions within the ORF, which is capable of hybridizing with the CRISPR RNA (crRNA) expressed from the CRISPR array..

[050] The term "vector" as used herein means a nucleic acid sequence containing an origin of replication. A vector may be a viral vector, bacteriophage, bacterial artificial chromosome or yeast artificial chromosome. A vector may be a DNA or RNA vector. A vector may be a self- replicating extrachromosomal vector, or a DNA plasmid.

[051] Unless otherwise defined herein, scientific and technical terms used in connection with the present disclosure shall have the meanings that are commonly understood by those of ordinary skill in the art.

CRISPR systems

[052] The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) genomic locus is found in the genomes of many prokaryotes. CRISPR loci provide resistance to viruses and phages in prokaryotes. In this way, the CRISPR loci functions as a type of immune system to help defend prokaryotes against foreign invaders. In such system the response to such foreign invaders starts by cleaving the genome of invading viruses and plasmids and integrating segments (termed protospacers) of the genomic DNA into the CRISPR locus of the host organism. The segments that are integrated into the host genome are known as “spacers”, which mediate protection from subsequent attack by the same (or sufficiently related) virus or plasmid. Expression involves transcription of the CRISPR locus and subsequent enzymatic processing to produce short mature CRISPR RNAs (crRNA), each containing a single spacer sequence. Interference is induced after the CRISPR RNAs associate with Cas proteins to form effector complexes, which are then targeted to complementary protospacers in foreign genetic elements to induce nucleic acid degradation.

[053] Currently, two classes of CRISPR systems have been described, Class 1 and Class 2, based upon the genes encoding the effector component. Class 1 systems have a multi-subunit crRNA-effector complex, whereas Class 2 systems have a single effector protein. Typical examples of Class 2 effector proteins are Cas9 and Cpf1 (Cas12a).

[054] To date six types (Types I-VI) of CRISPR systems have been described (for an overview see Makarova *et al.*, Nature Reviews Microbiology (2015) 13:1-15). Class 1 systems comprise Type I, Type III and Type IV systems. Class 2 systems comprise Type II, Type V and Type VI systems.

[055] CRISPR loci include several short repeating sequences referred to as "repeats." The repeats can form hairpin structures and/or the repeats can be single-stranded sequences. The repeats occur in clusters. Repeats frequently diverge between species. Repeats are regularly interspaced with unique intervening sequences, referred to as "spacers," resulting in a repeat-spacer-repeat locus architecture. Spacers are sequences usually identical to or homologous to foreign invader sequences (such as viral sequences).

[056] In some cases, a spacer-repeat unit encodes a crisprRNA (crRNA). A crRNA refers to the mature form of the spacer-repeat unit. A crRNA contains a spacer sequence that is involved in targeting a target nucleic. crRNA has a region of complementarity to a potential DNA or RNA target sequence and in some cases, e.g., in currently characterized Type II systems, a second region that forms base-pair hydrogen bonds with a transactivating CRISPR RNA (tracrRNA) to form a secondary structure, typically to form at

least a stem structure. Complex formation between tracrRNA/crRNA and a Cas protein results in conformational change of the Cas protein that facilitates binding to DNA, nuclease activities of the Cas protein, and crRNA- guided site-specific DNA cleavage by the nuclease. For a Cas protein/tracrRNA/crRNA complex to cleave a DNA target sequence, the DNA target sequence is adjacent to a cognate protospacer adjacent motif (PAM).

[057] Usually, CRISPR locus comprises polynucleotide sequences encoding for CRISPR Associated Genes (cas) genes. Cas genes are involved in the biogenesis and/or the interference stages of crRNA function. Cas genes display extreme sequence diversity between different species and homologs. Some Cas proteins comprise a specific set of domain structures.

[058] Mature crRNAs are processed from a longer polycistronic CRISPR locus transcript, also referred to as pre-crRNA array. A pre-crRNA array comprises a plurality of crRNAs. The repeats in the pre-crRNA array are recognized by cas genes. Cas genes bind to the repeats and cleave the repeats. This action can liberate the plurality of crRNAs. crRNAs can be subjected to further events to produce the mature crRNA form such as trimming (e.g., with an exonuclease). A crRNA may comprise all, or some, of the CRISPR repeat sequences.

[059] Interference refers to the stage in the CRISPR system that is functionally responsible for combating infection by a foreign invader. CRISPR interference follows a similar mechanism to RNA interference, which results in target RNA degradation and/or destabilization. Currently characterized CRISPR systems perform interference of a target nucleic acid by coupling crRNAs and Cas genes, thereby forming CRISPR ribonucleoproteins (RNPs). crRNA of the RNP guides the RNP to foreign invader nucleic acid, (e.g. , by recognizing the foreign invader nucleic acid through hybridization). Hybridized target foreign invader nucleic acid- crRNA units are subjected to cleavage by Cas proteins. Target nucleic acid interference typically requires a protospacer adjacent motif (PAM) in a target nucleic acid.

[060] Currently CRISPR-Cas systems are divided into two main classes based on their effector molecules: class 1 and class 2. Class 1 is characterized by multi-unit effector molecules, while class 2 contains a single effector molecule. Class 1 systems comprise Type I, Type III, and Type IV systems. Class 2 systems comprise Type II, Type V, and Type VI systems.

[061] Type II system is commonly represented by cas9 genes. There are two strands of RNA in Type II systems: a crRNA and a tracrRNA. The duplex formed by the tracrRNA and crRNA is recognized by, and associates with Cas9, encoded by the cas9 gene, which combines the functions of the crRNA-effector complex with target DNA cleavage. Cas9 is directed to a target nucleic acid by a sequence of the crRNA that is complementary to, and hybridizes with, a sequence in the target nucleic acid.

[062] In Type V systems, nucleic acid target sequence binding involves a Cas12 protein and the crRNA, as does the nucleic acid target sequence cleavage. In Type V systems, the RuvC-like nuclease domain of Casl2 protein cleaves both strands of the nucleic acid target sequence in a sequential fashion (Swarts, *et al.* , Mol. Cell (2017) 66:221 -233), producing 5' overhangs, which differs from the fragments generated by Cas9 protein. There have been multiple subtypes of Type V systems identified so far (type V-A/B/C/D/E/F/G/H/I/K/L and CRISPR-Cas12j). All of them differ by the length of Cas protein, PAM sequence and whether they require tracrRNA for its functionality.

[063] Type V-A is represented by Cas12a protein. The Cas12a protein cleavage activity of Type V-A systems does not require hybridization of crRNA to tracrRNA to form a duplex; rather Type V-A systems use a single crRNA that has a stem-loop structure forming an internal duplex. Cas12a protein binds the crRNA in a sequence- and structure-specific manner by recognizing the stem loop and sequences adjacent to the stem loop, most notably the nucleotides 5' of the spacer sequence, which hybridizes to the nucleic acid target sequence. This stem-loop structure is typically in the range of 15 to 19 nucleotides in length. Substitutions that disrupt this stem-loop duplex abolish cleavage activity, whereas other substitutions that do not disrupt the stem-loop duplex do not abolish cleavage activity.

[064] In Type V-A systems, nucleic acid target sequence binding involves Cas12a and the crRNA, as does the nucleic acid target sequence cleavage. In Type V-A systems, the RuvC-like nuclease domain of Cas12a cleaves one strand of the double-stranded nucleic acid target sequence, and a putative nuclease domain cleaves the other strand of the double- stranded nucleic acid target sequence in a staggered configuration, producing 5' overhangs, which is different from the blunt ends generated by Cas9 cleavage. These 5' overhangs may facilitate insertion of DNA.

[065] The Cas12a cleavage activity of Type V systems also does not require hybridization of crRNA to tracrRNA to form a duplex, rather the crRNA of Type V systems uses a single crRNA that has a stem-loop structure forming an internal duplex. Cas12a binds the crRNA in a sequence and structure specific manner that recognizes the stem loop and sequences adjacent to the stem loop, most notably the nucleotide 5' of the spacer sequences that hybridizes to the nucleic acid target sequence. This stem-loop structure is typically in the range of 15 to 19 nucleotides in length. Substitutions that disrupt this stem-loop duplex abolish cleavage activity, whereas other substitutions that do not disrupt the stem-loop duplex do not abolish cleavage activity. In Type V systems, the crRNA forms a stem-loop structure at the 5 ' end, and the sequence at the 3' end is complementary to a sequence in a nucleic acid target sequence.

[066] Type V-F1 is represented by Cas12f1 protein. The Cas12f1 protein cleavage activity of Type V-F1 systems does require hybridization of crRNA to tracrRNA to form a duplex. Cas12f1 protein binds the tracrRNA/crRNA in a sequence- and structure-specific manner by recognizing the stem loops and

sequences adjacent to the stem loops, most notably the nucleotides 5' of the spacer sequence, which hybridizes to the nucleic acid target sequence. These stem-loop structure are typically in the range of 150 to 170 nucleotides in length for the tracrRNA and 28-34 nucleotides in length for the crRNA. Substitutions that disrupt these stem-loop duplex abolish cleavage activity, whereas other substitutions that do not disrupt the stem-loop duplex do not abolish cleavage activity.

[067] In Type V-F1 systems, nucleic acid target sequence binding involves Cas12f1 and the tracrRNA/crRNA, as does the nucleic acid target sequence cleavage. In Type V-F1 systems, the RuvC-like nuclease domain of Cas12f1 cleaves one strand of the double-stranded nucleic acid target sequence, and a putative nuclease domain cleaves the other strand of the double- stranded nucleic acid target sequence in a staggered configuration, producing 5' overhangs, which is different from the blunt ends generated by Cas9 cleavage. These 5' overhangs may facilitate insertion of DNA.

[068] The Cas12f1 cleavage activity of Type V systems also does require hybridization of crRNA to tracrRNA to form a duplex. Cas12f1 binds the tracrRNA/crRNA in a sequence and structure specific manner that recognizes the stem loops and sequences adjacent to the stem loop, most notably the nucleotide 5' of the spacer sequences that hybridizes to the nucleic acid target sequence. These stem-loop structure are typically in the range of 150 to 170 nucleotides in length for the tracrRNA and 28-34 nucleotides in length for the crRNA. Substitutions that disrupt this stem-loop duplex abolish cleavage activity, whereas other substitutions that do not disrupt the stem-loop duplex do not abolish cleavage activity. In Type V systems, the tracrRNA/crRNA forms stem-loop structures at the 5 ' end, and the sequence at the 3' end is complementary to a sequence in a nucleic acid target sequence.

[069] Other proteins associated with Type V crRNA and nucleic acid target sequence binding and cleavage include Cas12b, Cas12c, Cas12d, and Cas12e which are similar in length to Cas12a proteins, ranging from approximately 1000-1500 amino acids, but also require an additional RNA (either a tracRNA or a scoutRNA) (see for example Harrington et al, Molecular Cell,Volume 79, Issue 3, 2020,Pages 416-424). Still other proteins associated with Type V crRNA and nucleic acid target sequence binding and cleavage include Cas12f1, Cas12f2, Cas12f3, and Cas12g, which are smaller in length to Cas12a proteins, ranging from approximately 300-900 amino acids, but also require a tracrRNA.

[070] Type VI systems include the Cas13a protein (also known as Class 2 candidate 2 protein, or C2c2) which does not share sequence similarity with other CRISPR effector proteins (see Abudayyeh, *et al,* Science (2016) 353:aaf5573). Cas13a proteins have two HEPN domains and possess single-stranded RNA cleavage activity. Cas13a proteins are similar to Cas12a proteins in requiring a crRNA for nucleic acid target sequence binding and cleavage, but not requiring tracrRNA.

[071] While many of type V systems have been identified, the discovery and characterization of CRISPR systems is ongoing.

Method of identifying components of the RGN-based systems

[072] The present disclosure provides methods for identifying the RNA components of RGN-based nucleic acid targeting systems comprising the RGN polypeptides provided further herein. The methods comprise steps described below. Some of the steps could be replaced by alternative techniques that would be apparent to the skilled person.

[073] Genomic and /or metagenomic samples are searched for open reading frames (ORFs) and those that have predicted to be genes were selected. A hidden Markov model (HMM) was used to compare the putative genes to profiles of known Cas proteins. The identified Cas genes are subsequently grouped into operons, and the operon type is determined based on the presence of known signature genes. For each genome, the CRISPR arrays are identified based on the presence of regularly spaced repeats. The subtype of each CRISPR array is predicted using machine learning. Cas operons are considered linked to CRISPR arrays if they are less than 10 kilobases apart.

[074] Systems that fit the putative domain and CRISPR orientation for Cas12f5 were confirmed by predicting the structure computationally using neural network based models, similar to methods described, for example in Jumper et al (2021) Nature. V596m pp 583-589. These structural models are compared to each other, and to solved crystal structures to identify possible gRNA structures, and confirm the catalytic residue of the final RuvC domain of the proteins.

[075] The crRNA is held to be the last 14 bases of the CRISPR repeat followed by the reprogrammable spacer sequence. Regions in the identified CRIPSR operon are manually searched for potential tracrRNAs by searching for antirepeat sequences capable of hybridizing to approximately bases 1-10 of the putative crRNA sequence within 900 nt of the last catalytic RuvC domain of the RGN, including regions within the Open reading frame (ORF) of the corresponding RGN (Figure 1). Once identified, the putative tracrRNA, when joined together via a flexible linker, such as, for example, GAAA tetra loop, to the putative crRNA, the resulting sgRNA consists of 4 -6 stem loop sequences. The essential structure of the sgRNA consist of the antirepeat from the putative tracrRNA and approximately the 5-16 final 3' bases of the CRISPR repeat, of which the first 1-12 bases are complimentary to the antirepeat of the putative tracrRNA before 1-4 bases of unpaired "leader sequence" before the reprogrammable spacer sequence.

[076] In some embodiments, said genomic and metagenomic sequences are obtained from a sequence database such as Ensembl or NCBI genome databases.

Systems for binding to the target nucleotide sequence of interest

[077] The present disclosure provides a system (a nucleic acid targeting system) for binding a target sequence of interest, wherein the system comprises at least one RNA or a nucleotide sequence encoding the same, and at least one RGN or a nucleotide sequence encoding the same, as described further hereafter. The RNA hybridizes to the target sequence of interest and also binds to the RGN polypeptide, thereby directing the RGN polypeptide to the target sequence. In some of these embodiments, the RGN comprises an amino acid sequence set forth in Table 4 or an active variant or fragment thereof. In various embodiments, the RNA comprises 6 or more nucleotides of the CRISPR repeat sequence comprising the nucleotide sequence set forth in Table 5 or an active variant or fragment thereof. In particular the RNA comprises the 5-16 final 3' bases of the corresponding CRISPR repeat. In some embodiments, the gRNA comprises an RNA sequence comprising a nucleotide sequence set forth in Table 5, or an active variant or fragment thereof. In some embodiments, the gRNA comprises a CRPSR repeat sequence or partial CRISPR repeat sequence having sequence set forth in Table 5, or an active variant or fragment thereof. In particular embodiments, the system comprises a RGN and at least one gRNA, wherein the RGN and gRNA are not naturally complexed in nature. In some embodiments the system comprises an RNA (or gRNA) and an RGN as described above. The rules of identifying of RGN and gRNA scaffold sequences are provided above.

[078] The system for binding a target sequence of interest provided herein can be a ribonucleoprotein complex, which is at least one molecule of an RNA bound to at least one protein. The ribonucleoprotein complexes provided herein comprise at least one gRNA as the RNA component and an RGN as the protein component. Such ribonucleoprotein complexes can be purified from a cell or organism that naturally expresses an RGN polypeptide and has been engineered to express a particular gRNA that is specific for a target sequence of interest.

[079] Alternatively, the ribonucleoprotein complex can be purified from a cell or organism that has been transformed with polynucleotides that encode an RGN polypeptide and a gRNA and cultured under conditions to allow for the expression of the RGN polypeptide and guide RNA. Thus, methods are provided for making an RGN polypeptide or an RGN ribonucleoprotein complex. Such methods comprise culturing a cell comprising a nucleotide sequence encoding an RGN polypeptide under conditions in which the RGN polypeptide is expressed. In some embodiments the cell further comprises a nucleotide sequence encoding a gRNA. The RGN polypeptide or RGN ribonucleoprotein can then be purified from the cultured cells.

[080] Methods for purifying an RGN polypeptide or RGN ribonucleoprotein complex from a biological sample are known in the art (e.g., size exclusion and/or affinity chromatography, 2D-PAGE, HPLC, reversed-phase chromatography, immunoprecipitation). In particular, the RGN polypeptide can be

recombinantly produced and comprises a purification tag to aid in its purification, including but not limited to, glutathione-S-transferase (GST), chitin binding protein (CBP), maltose binding protein, thioredoxin (TRX), poly(NANP), tandem affinity purification (TAP) tag, myc, AcV5, AU1, AU5, E, ECS, E2, FLAG, HA, nus, Softag 1, Softag 3, Strep, SBP, Glu-Glu, HSV, KT3, S, Sl, T7, V5, VSV-G, 6xHis, lOxHis, biotin carboxyl carrier protein (BCCP), and calmodulin.

[081] Generally, the tagged RGN polypeptide or RGN ribonucleoprotein complex is purified using immobilized metal affinity chromatography. It will be appreciated that other similar methods known in the art may be used, including other forms of chromatography or for example immunoprecipitation, either alone or in combination.

[082] Some methods provided herein for binding and/or cleaving a target sequence of interest involve the use of an *in vitro* assembled RGN ribonucleoprotein complex. *In vitro* assembly of an RGN ribonucleoprotein complex can be performed using any method known in the art in which an RGN polypeptide is contacted with a guide RNA under conditions to allow for binding of the RGN polypeptide to the gRNA. The RGN polypeptide can be purified from a biological sample, cell lysate, or culture medium, produced via in vitro translation, or chemically synthesized. The gRNA can be purified from a biological sample, cell lysate, or culture medium, transcribed in vitro, or chemically synthesized. The RGN polypeptide and gRNA can be brought into contact in solution (e.g., buffered saline solution) to allow for in vitro assembly of the RGN ribonucleoprotein complex.

RNA-guided nucleases (RGNs)

[083] The present disclosure provides CRISPR-based nucleic acid targeting systems that comprise an RNA-guided nuclease (RGN) as defined in Table 4.

[084] Table 4. Novel RGNs

RGN code	SEQ ID NO	CRISPR Repeat length	Protein length
EGS0290	1	29	492
EGS0293	2	29	542
EGS0294	3	30	537
EGS0346	4	28	532
EGS0380	5	28	520
EGS0288	79	29	506
EGS0291	80	28	606
EGS0295	81	30	492
EGS0318	82	28	503
EGS0334	83	25	552
EGS0336	84	30	496

EGS0337	85	29	539
EGS0338	86	29	559
EGS0341	87	29	572
EGS0343	88	29	506
EGS0344	89	29	583

[085] An RGN provided herein binds to a target nucleotide sequence and hybridizes with the RNA molecule (crRNA) specific to the RNA-guided nuclease. The target sequence can then be subsequently cleaved by the RGN if the RGN polypeptide possesses nuclease activity. The presently disclosed RGNs can cleave nucleotides within a polynucleotide, functioning as an endonuclease. In some embodiments, the disclosed RGNs can cleave nucleotides of a target nucleotide sequence within any position of a polynucleotide and thus function as both an endonuclease and exonuclease.

[086] The presently disclosed RGNs can be wild-type sequences derived from bacterial or archaeal species. Alternatively, the RGNs can be variants or fragments of wild-type polypeptides. The wild-type RGN can be modified to alter nuclease activity or alter PAM specificity, for example. In some embodiments, the RGN is not naturally-occurring. Such RGN have a single functioning nuclease domain.

[087] In other embodiments, the RGNs lacks nuclease activity altogether or exhibits reduced nuclease activity and is referred to herein as nuclease-dead RGNs. Any method known in the art for introducing mutations into an amino acid sequence, such as PCR-mediated mutagenesis and site-directed mutagenesis, can be used for generating nuclease-dead RGNs. (e.g. US9,790,490).

[088] Alternatively, nuclease dead RGNs can be targeted to particular genomic locations to alter the expression of a desired sequence. In some embodiments, the binding of a nuclease-dead RNA-guided nuclease to a target sequence results in the repression of expression of the target sequence or a gene under transcriptional control by the target sequence by interfering with the binding of RNA polymerase or transcription factors within the targeted genomic region. In other embodiments, the RGN (e.g. , a nuclease- dead RGN) or its complexed gRNA further comprises an expression modulator that, upon binding to a target sequence, serves to either repress or activate the expression of the target sequence or a gene under transcriptional control by the target sequence. In some of these embodiments, the expression modulator modulates the expression of the target sequence or regulated gene through epigenetic mechanisms.

[089] In other embodiments, one or more of the nuclease-dead RGNs disclosed herein can be targeted to particular genomic locations to modify the sequence of a target polynucleotide through fusion to a base editing polypeptide, for example a deaminase polypeptide or active variant or fragment thereof that deaminates a nucleotide base, resulting in conversion from one nucleotide base to another. The base-

editing polypeptide can be fused to the RGN at its N-terminal or C-terminal end. Additionally, the base-editing polypeptide may be fused to the RGN via a peptide linker. A non-limiting example of a deaminase polypeptide that is useful for such compositions and methods include cytidine deaminase or the adenosine deaminase base editor described in Gaudelli *et al.* (2017) Nature 551:464-471, and WO2018/027078.

Structural elements of the RGN peptides

[090] The RGN proteins used in the present disclosure employ multiple domains distributed in a recognition lobe (REC) and a nuclease lobe (NUC) for substrate recognition and cleavage.

[091] In one embodiment, an RGN polypeptide of the disclosure comprises an amino-terminal domain (NTD) and a carboxy-terminal domain (CTD), which are connected by a linker loop. The NTD consists of two domains: the wedge (WED) and recognition (REC) domains. The CTD consists of the tri split RuvC domain, which is split by a second REC domain and a target nucleic acid-binding (TNB) domain and an unstructured tail that corresponds with the expression region of the tracrRNA, that is dispensable for activity. However, unlike Cas9, the RGN polypeptides of the present disclosure do not contain a HNH domain. The RGNs of the present disclosure may comprise one or more additional domains, e.g., one or more of a Rec domains.

[092] In certain embodiments, the RGN polypeptides provided herein are between 300 and 700 amino acids in size, between 400 and 600 amino acids in size, between 450 and 550 amino acids in size. Size variation may be dependent on the particular domain architecture of the RGN polypeptides provided herein.

RuvC domain

[093] The RuvC domain may comprise multiple subdomains: RuvC-I, RuvC-II and RuvC-III. The subdomains may be separated by other sequences on the amino acid sequence of the protein.

[094] Examples of RuvC domains include any polypeptides having a structural similarity and/or sequence similarity to a RuvC domain described in the art. For example, the RuvC domain may share a structural similarity and/or sequence similarity to a RuvC of Cas9. In some examples, the RuvC domain may have an amino acid sequence that share at least 80%, at least 85%, at least 90%, at least 95%, at least 99%, or 100% sequence identity with RuvC domains.

[095] In some examples, the RuvC domain comprise RuvC-I polypeptide, RuvC-II polypeptide, and RuvC-III polypeptide. Examples of the RuvC-I domain also include any polypeptides having a structural similarity and/or sequence similarity to a RuvC-I, II, and III domains described in the art, such as the corresponding domains of Cas9. The RuvC domain may have an amino acid sequence that share at least

80%, at least 85%, at least 90%, at least 95%, at least 99%, or 100% sequence identity with a RuvC domain of Cas9.

[096] The RuvC domain of Cas9 consists of a six-stranded mixed beta-sheet flanked by a-helices and two additional two-stranded antiparallel beta-sheets (see e.g., Nishimasu et al. Cell, 2014). The RuvC domain of Cas9 shares structural similarity with the retroviral integrase superfamily members characterized by an RNase H fold, such as Escherichia coli RuvC (PDB code 1HJR, 14% identity, root-mean-square deviation (rmsd) of 3.6 A for 126 equivalent Ca atoms) and Thermus thermophilus RuvC (PDB code 4LD0, 12% identity, rmsd of 3.4 A for 131 equivalent Ca atoms). E. coli RuvC is a 3-layer alpha-beta sandwich containing a 5-stranded beta-sheet sandwiched between 5 alpha-helices. RuvC nucleases have four catalytic residues (e.g., Asp7, Glu70, Hisl43 and Aspl46 in T. thermophilus RuvC), and cleave Holliday junctions (or structurally analogous cruciform junctions) through a two-metal mechanism. Asp 10 (Ala), Glu762, His983 and Asp986 of the Cas9 RuvC domain are located at positions similar to those of the catalytic residues of *T. thermophilus* RuvC.

REC domain

[097] The REC domain may comprise multiple subdomains: REC1 and REC2. The subdomains may be separated by other sequences on the amino acid sequence of the protein.

[098] Examples of REC domains include any polypeptides having a structural similarity and/or sequence similarity to a REC domain described in the art. For example, the REC domain may share a structural similarity and/or sequence similarity to a REC of Cas12a. In some examples, the REC domain may have an amino acid sequence that share at least 80%, at least 85%, at least 90%, at least 95%, at least 99%, or 100% sequence identity with REC domains.

[099] In some examples, the REC domain may have a Helix-turn-helix (HTH) DNA binding domain. HTH is the DNA-binding motif used in prokaryotic regulatory proteins such as Cro, CAP, and λ repressor and in many eukaryotic activators such as Myc, MyoD, E12, E47, and AP-4. In prokaryotic regulatory proteins, the HTH motif is a tightly packed amino acid structure consisting of an α-helical region is followed by a sharp β-turn and then another α-helical region. The HTH motif of the protein directly interacts with DNA, the second α helix (the "recognition helix") binding in the major groove of the DNA. In some examples, the REC domain may comprise a bridge helix (BH) domain. The bridge helix domain refers to a helix and arginine rich polypeptide. The bridge helix domain may be located next to anyone of the amino acid domains in the nucleic-acid guided nuclease. In one embodiment, the bridge helix domain is next to a RuvC domain, e.g., next to RuvC-I, RuvC-II, or RuvC-III subdomain. In one example, the bridge helix domain is between a RuvC-I and RuvC-II subdomains.

[100] The bridge helix domain may be from 10 to 100, from 20 to 60, from 30 to 50, e.g., 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46 or 47, 48, 49, or 50 amino acids in length. Examples of bridge helix includes the polypeptide of amino acids 60-93 of the sequence of *S. pyogenes* Cas9.

[101] In some examples, the REC domain comprises REC1 domain and REC2 domain. Examples of the REC1 domain also include any polypeptides having a structural similarity and/or sequence similarity to a REC1 and REC2 domains described in the art, such as the corresponding domains of Cas12a. The REC domain may have an amino acid sequence that share at least 80%, at least 85%, at least 90%, at least 95%, at least 99%, or 100% sequence identity with a REC domain of Cas12a. The REC domain of Cas12a consists of the REC1 and REC2 domains where REC1 comprises 13 alpha helices, and REC2 comprises ten alpha helices and two beta strands that form a small antiparallel sheet (see e.g., Yamano *et al.* (2016), Cell, 165, 4,Pages 949-962).

Target nucleic acid-binding domain

[102] Examples of Target nucleic acid-binding (TNB) domains include any polypeptides having a structural similarity and/or sequence similarity to a TNB domain described in the art. For example, the TNB domain may share a structural similarity and/or sequence similarity to a TNB of Cas12f1. In some examples, the TNB domain may have an amino acid sequence that share at least 80%, at least 85%, at least 90%, at least 95%, at least 99%, or 100% sequence identity with TNB domains.

[103] In some examples, the TNB domain may consist of a Zinc finger binding domain. A Zinc finger is a small protein structural motif that is characterized by the coordination of one or more zinc ions in order to stabilize the fold, typically consisting of a dual CXXC motif where X can be any amino acid and there are a variable number of amino acids between the two pairs of cytosines.

Modified RGN peptides

[104] The RGNs may comprise one or more modifications. The modified RGNs may be catalytically inactive (also referred as dead). A catalytically inactive or dead nuclease may have reduced or no nuclease activity compared to a wildtype counterpart nuclease. In some cases, a catalytically inactive or dead nuclease may have nickase activity. In some cases, a catalytically inactive or dead nuclease may not have nickase activity. Such a catalytically inactive or dead RGN may not make either double-strand or single-strand break on a target polynucleotide but may still bind or otherwise form complex with the target polynucleotide.

[105] In an embodiment, the RGN polypeptide comprises a mutation of the catalytic RuvC- residue corresponding to D289A, E388A or D486A (catalytic residues of RuvI, II, and III which are well known in the prior art) of SEQ ID NO:14 (mutated EGS0293) or equivalent residues of other RGN sequences

provided herein (see for example Kleinstiver, et al. (2019) Nat Biotechnol 37, 276–282). In one embodiment, the modifications of the RGN polypeptide may or may not cause an altered functionality. Some modifications will not result in an altered functionality include for instance codon optimization for expression into a particular host, or providing the nuclease with a particular marker. Modifications which may result in altered functionality may also include mutations, including point mutations, insertions, deletions, truncations (including split nucleases), etc., as well as chimeric RGNs (e.g., comprising domains from different orthologues or homologues) or fusion proteins.

[106] In an embodiment, the RGN polypeptide comprises a deletion of the unstructured tail at the carboxy terminus of the RGN polypeptide while maintaining the core catalytic activity of the enzyme (SEQ ID NO:6-13). This can be done to facilitate greater packaging into delivery vectors and to save on manufacturing costs. In some embodiments, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40 or more amino acids at the C-terminus might be deleted while retaining the catalytic activity.

[107] In some embodiments, the RGN polypeptide comprises mutations in the DNA binding pocket to increase affinity for DNA leading to enhanced binding activity. Such enhanced binding activity can lead to increased cleavage activity or can lead to increased activity of the fusion domain. In an embodiment, the RGN polypeptide comprises a mutation that increases the positive change of the of enzyme corresponding to T97R, and/or T101R, and/or N150R, and/or D153R, and/or N157R, and/or A190K or A190R, and/or E247R, and/or Q336R, and/or Q343K or Q343R, and/or N347K or N347R, Q373R, and/or D389K or D389R, and/orA424K or A424R, and/or V427R of SEQ ID NO: 2. In an example, the improved RGN is EGS0293v2_Q343K_N347K_A424R (SEQ ID NO: 98) (Figure 5).

[108] Fusion proteins may include, for example, fusions with heterologous domains or functional domains (e.g., localization signals, enzymes). In an embodiment, various different modifications may be combined (e.g., a mutated nuclease which is catalytically inactive and which further is fused to a functional domain, such as for instance to induce DNA methylation or another nucleic acid modification, such as, for example, a mutation, a deletion, an insertion, a replacement).

Localization signal sequences

[109] The RGNs can comprise at least one nuclear localization signal (NLS) to enhance transport of the RGN to the nucleus of a cell. Nuclear localization signals are known in the art and generally comprise a stretch of basic amino acids (see, e.g., Lange *et al.,* J. Biol. Chem. (2007) 282:5101-5105). In embodiments, the RGN comprises 2, 3, or more nuclear localization signals. The nuclear localization signal(s) can be a heterologous NLS. Non-limiting examples of nuclear localization signals useful for the presently disclosed RGNs are the nuclear localization signals of SV40 Large T-antigen, nucleopasmin,

and c-Myc (see. e.g., Ray *et al.* (2015) Bioconjug Chem 26(6): 1004-7). In particular embodiments, the RGN comprises the NLS sequence comprising the sequence of SEQ ID NO: 76 or 78. The RGN may comprise one or more NLS sequences at its N-terminus, C- terminus, or both the N-terminus and C-terminus. For example, the RGN may comprise two NLS sequences at the N-terminal region and four NLS sequences at the C-terminal region.

[110] Other localization signal sequences known in the art that localize polypeptides to particular subcellular location(s) can also be used to target the RGNs, including, but not limited to, plastid localization sequences, mitochondrial localization sequences, and dual-targeting signal sequences that target to both the plastid and mitochondria (see, e.g., Nassoury and Morse (2005) Biochim Biophys Acta 1743:5-19; Herrmann and Neupert (2003) IUBMB Life 55:219-225; Soil (2002) Curr Opin Plant Biol 5:529-535; Carrie and Small (2013) Biochim Biophys Acta 1833:253-259).

[111] In certain embodiments, the RGNs comprise at least one cell- penetrating domain that facilitates cellular uptake of the RGN. Cell-penetrating domains are known in the art and generally comprise stretches of positively charged amino acid residues (i.e., polycationic cell- penetrating domains), alternating polar amino acid residues and non-polar amino acid residues (i.e..amphipathic cell-penetrating domains), or hydrophobic amino acid residues (i.e.. hydrophobic cell- penetrating domains) (see, e.g., Milletti F. (2012) Drug Discov Today 17:850-860). A non-limiting example of a cell-penetrating domain is the trans-activating transcriptional activator (TAT) from the human immunodeficiency virus 1.

[112] The nuclear localization signal, plastid localization signal, mitochondrial localization signal, dual targeting localization signal, and/or cell-penetrating domain can be located at the amino-terminus (N-terminus), the carboxyl-terminus (C-terminus), or in an internal location of the RNA-guided nuclease.

Additional tags and labels

[113] The presently disclosed RGN polypeptides may comprise a detectable label or a purification tag. The detectable label or purification tag can be located at the N-terminus, the C-terminus, or an internal location of the RNA-guided nuclease, either directly or indirectly via a linker peptide. In some of these embodiments, the RGN component of the fusion protein is a nuclease-dead RGN. In other embodiments, the RGN component of the fusion protein is a RGN with nickase activity.

[114] RGNs that lack nuclease activity can be used to deliver a fused polypeptide, polynucleotide, or small molecule payload to a particular genomic location. In some of these embodiments, the RGN polypeptide or guide RNA can be fused to a detectable label to allow for detection of a particular sequence. As a non-limiting example, a nuclease-dead RGN can be fused to a detectable label (e.g.,

fluorescent protein) and targeted to a particular sequence associated with a disease to allow for detection of the disease-associated sequence.

[115] A detectable label is a molecule that can be visualized or otherwise observed. The detectable label may be fused to the RGN as a fusion protein (e.g., fluorescent protein) or may be a small molecule conjugated to the RGN polypeptide that can be detected visually or by other means. Detectable labels that can be fused to the presently disclosed RGNs as a fusion protein include any detectable protein domain, including but not limited to, a fluorescent protein or a protein domain that can be detected with a specific antibody. Non-limiting examples of fluorescent proteins include green fluorescent proteins (e.g., GFP, EGFP, ZsGreen) and yellow fluorescent proteins (e.g, YFP, EYFP, ZsYellow).

[116] RGN polypeptides can also comprise a purification tag, which is any molecule that can be utilized to isolate a protein or fused protein from a mixture (e.g., biological sample, culture medium). Non-limiting examples of purification tags include biotin, myc, maltose binding protein (MBP), and glutathione -S- transferase (GST).

Fusion proteins comprising the RGNs

[117] The presently disclosed RGNs can be fused to an effector domain (a fusion protein of an RGN and an effector domain), such as a cleavage domain, a deaminase domain, or an expression modulator domain, either directly or indirectly via a linker. Such effector domain can be located at the N-terminus, the C-terminus, or an internal location of the RNA-guided nuclease. In some embodiments, the RGN component of the fusion protein is a nuclease-dead RGN.

[118] RGNs that are fused to a polypeptide or domain can be separated or joined by a linker. In some embodiments, a linker joins a gRNA binding domain of an RNA guided nuclease and a base-editing polypeptide, such as a deaminase.

[119] In some embodiments, the RGN fusion protein comprises a cleavage domain, which is any domain that is capable of cleaving a polynucleotide (i.e.. RNA, DNA) and includes, but is not limited to, restriction endonucleases and homing endonucleases (see, e.g Linn *et al.* (eds.) Nucleases, Cold Spring Harbor Laboratory Press, 1993).

[120] In some embodiments, the RGN fusion protein comprises a deaminase domain that deaminates a nucleotide base, resulting in conversion from one nucleotide base to another, and includes, but is not limited to, a cytidine deaminase or an adenosine deaminase base editor.

[121] In some embodiments, the effector domain of the fusion protein can be an expression modulator domain, which is a domain that either serves to upregulate or downregulate transcription. The expression

modulator domain can be an epigenetic modification domain, a transcriptional repressor domain or a transcriptional activation domain.

[122] In some of these embodiments, the expression modulator of the RGN fusion protein comprises an epigenetic modification domain that covalently modifies DNA or histone proteins to alter histone structure and/or chromosomal structure without altering the DNA sequence, leading to changes in gene expression (i. e. , upregulation or downregulation). Non-limiting examples of epigenetic modifications include acetylation or methylation of lysine residues, arginine methylation, serine and threonine phosphorylation, and lysine ubiquitination and sumoylation of histone proteins, and methylation and hydroxymethylation of cytosine residues in DNA. Non-limiting examples of epigenetic modification domains include histone acetyltransferase domains, histone deacetylase domains, histone methyltransferase domains, histone demethylase domains, DNA methyltransferase domains, and DNA demethylase domains.

[123] In other embodiments, the expression modulator of the fusion protein comprises a transcriptional repressor domain, which interacts with transcriptional control elements and/or transcriptional regulatory proteins, such as RNA polymerases and transcription factors, to reduce or terminate transcription of at least one gene. Transcriptional repressor domains are known in the art and include, but are not limited to IKB, and Kruppel associated box (KRAB) domains.

[124] In yet other embodiments, the expression modulator of the fusion protein comprises a transcriptional activation domain, which interacts with transcriptional control elements and/or transcriptional regulatory proteins, such as RNA polymerases and transcription factors, to increase or activate transcription of at least one gene. Transcriptional activation domains are known in the art and include, but are not limited to, a VP16 activation domain and an NFAT activation domain.

[125] It is also envisaged that the nucleic acid-targeting effector protein-guide RNA complex as a whole may be associated with two or more functional domains. For example, there may be two or more functional domains associated with the nucleic acid-targeting effector protein, or there may be two or more functional domains associated with the guide RNA (via one or more adaptor proteins), or there may be one or more functional domains associated with the nucleic acid-targeting effector protein and one or more functional domains associated with the guide RNA (via one or more adaptor proteins).

[126] The fusion between the adaptor protein and the activator or repressor may include a linker. For example, Gly-Ser linkers GGGS can be used. They can be used in repeats of 3 or 6, 9 or even 12 or more, to provide suitable lengths, as required. Linkers can be used between the guide RNAs and the functional

domain (activator or repressor), or between the nucleic acid-targeting effector protein and the functional domain (activator or repressor).

GuideRNAs (gRNAs), tracrRNA, and crRNA

[127] The present disclosure provides RGNs that can bind to gRNAs. The term "gRNA" refers to a nucleotide sequence having sufficient complementarity with a target nucleotide sequence to hybridize with the target sequence and direct sequence-specific binding of an associated RNA- guided nuclease to the target nucleotide sequence. Thus, a RGN's respective gRNA is one or more RNA molecules (generally, one or two), that can bind to the RGN and guide the RGN to bind to a particular target nucleotide sequence, and in those instances wherein the RGN has nickase or nuclease activity, also cleave the target nucleotide sequence.

[128] In general, a gRNA comprises a CRISPR RNA (crRNA) and a trans-activating CRISPR RNA (tracrRNA). Native gRNAs that comprise both a crRNA and a tracrRNA generally comprise two separate RNA molecules that hybridize to each other through the repeat sequence of the crRNA and the anti-repeat sequence of the tracrRNA. Native direct repeat sequences within a CRISPR array generally range in length from 28 to 37 base pairs, although the length can vary between 23 bp to 55 bp. Spacer sequences within a CRISPR array generally range from 28 to 34 bp in length, although the length can be between 21 bp to 72 bp. Specifically, the spacer sequences of the present disclosure are normally between 35 and 46 nucleotides. Each CRISPR array generally comprises less than 60 units of the CRISPR repeat-spacer sequence. The CRISPRs are transcribed as part of a long transcript termed the primary CRISPR transcript, which comprises much of the CRISPR array. The primary CRISPR transcript is cleaved by Cas proteins to produce crRNAs or in some cases, to produce pre-crRNAs that are further processed by additional Cas proteins into mature crRNAs. Mature crRNAs comprise a spacer sequence and a CRISPR repeat sequence. In some embodiments in which pre-crRNAs are processed into mature (or processed) crRNAs, maturation involves the removal of one to six or more 5', 3', or 5' and 3' nucleotides. For the purposes of genome editing or targeting a particular target nucleotide sequence of interest, these nucleotides that are removed during maturation of the pre-crRNA molecule are not necessary for generating or designing a gRNA.

[129] More specifically, the length of the target DNA within the sequence of the gRNA with the complementary sequence is 17 to 23bp, 18 to 23bp, 19 to 23bp, more specifically 20 to 23bp, as more specifically, it may be a 21 to 23bp, but is not limited thereto.

[130] A CRISPR RNA (crRNA) comprises a spacer sequence and a CRISPR repeat sequence. The "spacer sequence" is the nucleotide sequence that directly hybridizes with the target nucleotide sequence of interest. The spacer sequence is engineered to be fully or partially complementary with the target

sequence of interest. In various embodiments, the spacer sequence can comprise from 8 nucleotides to 30 nucleotides, or more. For example, the spacer sequence can be 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, or more nucleotides in length. In some embodiments, the spacer sequence is 10 to 26 nucleotides in length, or 12 to 30 nucleotides in length. In particular embodiments, the spacer sequence is 20 nucleotides in length.

[131] A trans-activating CRISPR RNA (tracrRNA) molecule comprises a nucleotide sequence comprising a region that has sufficient complementarity to hybridize to a CRISPR repeat sequence of a crRNA, which is referred to herein as the anti-repeat region. In some embodiments, the tracrRNA molecule further comprises a region with secondary structure (e.g., stem-loop) or forms secondary structure upon hybridizing with its corresponding crRNA. In particular embodiments, the region of the tracrRNA that is fully or partially complementary to a CRISPR repeat sequence is at the 3' end of the molecule and the 5' end of the tracrRNA comprises secondary structure. This region of secondary structure generally comprises several hairpin structures, including the nexus hairpin, which is found adjacent to the anti - repeat sequence. There are often hairpins at the 5' end of the tracrRNA that can vary in structure and number.

[132] Table 5. RGNs and their corresponding consensus repeat sequence, tracrRNA, crRNA, sgRNA, and alternative sgRNA

RGN ID	CRISPR repeat SEQ ID	tracrRNA SEQ ID NO	crRNA SEQ ID NO	Partial CRISPR repeat SEQ ID NO	sgRNA SEQ ID NO	Alternative sgRNA SEQ ID NO
EGS0290	15	30	25	20	35	40-47
EGS0293	16	31	26	21	36	48-57
EGS0294	17	32	27	22	37	58-65
EGS0346	18	33	28	23	38	66-70
EGS0380	19	34	29	24	39	71-73
EGS0288	99	132	121	110	143	NA
EGS0291	100	133	122	111	144	154-159
EGS0295	101	134	123	112	145	NA
EGS0318	102	135	124	113	146	160
EGS0334	103	136	125	114	147	NA

EGS0336	104	137	126	115	148	NA
EGS0337	105	138	127	116	149	NA
EGS0338	106	139	128	117	150	NA
EGS0341	107	140	129	118	151	NA
EGS0343	108	141	130	119	152	NA
EGS0344	109	142	131	120	153	NA

[133] The gRNA can be a single gRNA or a dual-gRNA system. A single gRNA comprises the crRNA and tracrRNA on a single molecule of RNA, whereas a dual-gRNA system comprises a crRNA and a tracrRNA present on two distinct RNA molecules, hybridized to one another through at least a portion of the CRISPR repeat sequence of the crRNA and at least a portion of the tracrRNA, which may be fully or partially complementary to the CRISPR repeat sequence of the crRNA. In some of those embodiments wherein the gRNA is a single gRNA, the crRNA and tracrRNA are separated by a linker nucleotide sequence. In general, the linker nucleotide sequence is one that does not include complementary bases in order to avoid the formation of secondary structure within or comprising nucleotides of the linker nucleotide sequence. In some embodiments, the linker nucleotide sequence between the crRNA and tracrRNA is at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, or more nucleotides in length. In particular embodiments, the linker nucleotide sequence of a single gRNA is at least 4 nucleotides in length. In certain embodiments, the linker nucleotide sequence is the nucleotide sequence set forth in Table 5. In other embodiments, the linker nucleotide sequence is at least 6 nucleotides in length. In certain embodiments, the linker nucleotide sequence is RAAA.

[134] The single gRNA or dual-gRNA can be synthesized chemically or via in vitro transcription. Assays for determining sequence-specific binding between a RGN and a gRNA are known in the art and include, but are not limited to, in vitro binding assays between an expressed RGN and the gRNA, which can be tagged with a detectable label (e.g., biotin) and used in a pull-down detection assay in which the gRNA:RGN complex is captured via the detectable label (e.g., with streptavidin beads). A control gRNA with an unrelated sequence or structure to the gRNA can be used as a negative control for non-specific binding of the RGN to RNA.

[135] In certain embodiments, the gRNA can be introduced into a target cell, organelle, or embryo as an RNA molecule. The gRNA can be transcribed in vitro or chemically synthesized. In other embodiments, a nucleotide sequence encoding the gRNA is introduced into the cell, organelle, or embryo. In some of these embodiments, the nucleotide sequence encoding the gRNA is operably linked to a promoter (e.g., an

RNA polymerase III promoter). The promoter can be a native promoter or heterologous to the gRNA-encoding nucleotide sequence.

[136] In various embodiments, the gRNA can be introduced into a target cell, organelle, or embryo as a ribonucleoprotein complex, as described herein, wherein the gRNA is bound to an RNA-guided nuclease polypeptide. The gRNA directs an associated RNA-guided nuclease to a particular target nucleotide sequence of interest through hybridization of the gRNA to the target nucleotide sequence. A target nucleotide sequence can comprise DNA, RNA, or a combination of both and can be single-stranded or double -stranded. A target nucleotide sequence can be genomic DNA (i.e. , chromosomal DNA), plasmid DNA, or an RNA molecule (e.g. , messenger RNA, ribosomal RNA, transfer RNA, micro RNA, small interfering RNA). The target nucleotide sequence can be bound (and in some embodiments, cleaved) by an RNA-guided nuclease in vitro or in a cell. The chromosomal sequence targeted by the RGN can be a nuclear, plastid or mitochondrial chromosomal sequence. In some embodiments, the target nucleotide sequence is unique in the target genome.

Multiple gRNA molecules

[137] The present disclosure also provides methods for binding and/or modifying a target nucleotide sequence of interest. The methods include delivering a system comprising at least one gRNA or a polynucleotide encoding the same, and at least one fusion polypeptide comprises an RGN of the invention and a base-editing polypeptide, for example a cytidine deaminase or an adenosine deaminase, or a polynucleotide encoding the fusion polypeptide, to the target sequence or a cell, organelle, or embryo comprising the target sequence.

[138] One of ordinary skill in the art will appreciate that any of the presently disclosed methods can be used to target a single target sequence or multiple target sequences. Thus, methods comprise the use of a single RGN polypeptide in combination with multiple, distinct gRNAs, which can target multiple, distinct sequences within a single gene and/or multiple genes. Also encompassed herein are methods wherein multiple, distinct gRNAs are introduced in combination with multiple, distinct RGN polypeptides. These gRNAs and gRNA/RGN polypeptide systems can target multiple, distinct sequences within a single gene and/or multiple genes.

Protospacer adjacent motif (PAM) sequences

[139] The present disclosure also provides PAM (proto-spacer-adjacent Motif) sequences to the adjacent, target DNA sequence of the complementary chain (complementary strand) and base pair formation can be in sequence to include to, a gRNA, or a composition comprising a DNA coding for the gRNA.

[140] In the context of the RGNs disclosed herein, the target nucleotide sequence of the RGNs is adjacent to a sequence called protospacer adjacent motif (PAM). A protospacer adjacent motif is generally within 1 to 30 nucleotides from the target nucleotide sequence. A protospacer adjacent motif can be within 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides from the target nucleotide sequence. The PAM can be 5' or 3' of the target sequence depending on the RGN. In some embodiments, the PAM is 5' of the target sequence for the presently disclosed RGNs. Generally, the PAM is a consensus sequence of 3-4 nucleotides, but in particular embodiments, can be 2, 3, 4, 5, 6, 7, 8, 9, or more nucleotides in length. In particular, the PAM sequences of the presently disclosed RGN proteins are purine rich as opposed to pyrimidine rich PAM sequences of all other Type V systems.

[141] Table 6. PAM sequences for the RGNs

RGN ID	PAM sequence
EGS0290	AAG
EGS0293	RAAY
EGS0294	AAB
EGS0346	RAA
EGS0380	RAA
EGS0288	AAG
EGS0291	AA
EGS0295	AACG
EGS0318	RAAC
EGS0334	AAN
EGS0336	AAG
EGS0337	rAAy
EGS0338	VAC
EGS0341	RG
EGS0343	RAS
EGS0344	AAC

[142] In particular embodiments, the RGN having or an active variant or fragment thereof binds respectively a target nucleotide sequence adjacent to a PAM sequence set forth in Table 6. In some embodiments, the RGN binds to a guide sequence comprising a CRISPR repeat sequence set forth in Table 5, or an active variant or fragment thereof, and a tracrRNA sequence set forth in Table 5, or an active variant or fragment thereof.

[143] It is well-known in the art that PAM sequence specificity for a given nuclease enzyme is affected by enzyme concentration (see, e.g. , Karvelis et al. (2015) Genome Biol 16:253), which may be modified by altering the promoter used to express the RGN, or the amount of ribonucleoprotein complex delivered to the cell, organelle, or embryo.

[144] Upon recognizing its corresponding PAM sequence, the RGN can cleave the target nucleotide sequence at a specific cleavage site. As used herein, a cleavage site is made up of the two particular nucleotides within a target nucleotide sequence between which the nucleotide sequence is cleaved by an RGN. The cleavage site can comprise the Ist and 2nd , 2nd and 3rd , 3rd and 4th , 4th and 5th , 5th and 6th , 7th and 8th , or 8th and 9th nucleotides from the PAM in either the 5' or 3' direction. In some embodiments, the cleavage site may be over 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 nucleotides from the PAM in either the 5’ or 3’ direction. In some embodiments, the cleavage site is 4 nucleotides away from the PAM. In other embodiments, the cleavage site is at least 15 nucleotides away from the PAM. As RGNs can cleave a target nucleotide sequence resulting in staggered ends, in some embodiments, the cleavage site is defined based on the distance of the two nucleotides from the PAM on the positive (+) strand of the polynucleotide and the distance of the two nucleotides from the PAM on the negative (-) strand of the polynucleotide.

[145] The systems of the present disclosure exhibit novel PAM sequences compared to canonical Cas12 proteins, enabling a different targeting space, than reported Cas12 enzymes (Figure 6). The compact nature of the PAMs of the RGNs of this disclosure would enable their use as a tool for functional genomics screens. These screens involve systematically manipulating a large amount of the DNA and/or RNA of the host organism to perturb gene function and identify how genes function in space, in time, and in disease. This necessitates the ability to identify a target sequence in any region of interest to create a large and unbiased library. The alternative targeting space addressable by these RGNs compared to other known Type V systems shown in Figure 6 indicates the ability to target unique areas of the host genome and create unique gene perturbation screens to access regions unavailable to other Type V systems.

Target nucleotide sequence

[146] The target polynucleotide of an RGN system can be any polynucleotide endogenous or exogenous to the eukaryotic cell. For example, the target polynucleotide can be a polynucleotide residing in the nucleus of the eukaryotic cell. The target polynucleotide can be a sequence coding a gene product (e.g., a protein) or a non-coding sequence (e.g., a regions or introns). The target sequence is generally associated with a PAM (protospacer adjacent motif. The precise sequence and length requirements for the PAM differ depending on the RGN used, but PAMs are typically 2-5 base pair sequences adjacent the protospacer (that is, the target sequence).

[147] A target nucleic acid can be single stranded DNA (ssDNA) or double stranded DNA (dsDNA). When the target DNA is single stranded, there is no preference or requirement for a PAM sequence in the target DNA. However, when the target DNA is dsDNA, a PAM is usually present adjacent to the target

sequence of the target DNA (e.g., see discussion of the PAM elsewhere herein). The source of the target DNA can be the same as the source of the sample, e.g., as described below.

[148] The source of the target DNA can be any source. In some cases, the target DNA is a viral DNA (e.g., a genomic DNA of a DNA virus). As such, subject method can be for detecting the presence of a viral DNA amongst a population of nucleic acids (e.g., in a sample). A subject method can also be used for the cleavage of non-target ssDNAs in the present of a target DNA. For example, if a method takes place in a cell, a subject method can be used to promiscuously cleave non-target ssDNAs in the cell (ssDNAs that do not hybridize with the guide sequence of the guide RNA) when a particular target DNA is present in the cell (e.g., when the cell is infected with a virus and viral target DNA is detected).

[149] The target polynucleotide of a RGN/RNA complex may be a disease-associated gene or polynucleotides or a gene/ polynucleotide associated with a biological pathway.

[150] Examples of possible target DNAs include, but are not limited to, viral DNAs such as: a papovavirus (e.g., human papillomavirus (HPV), polyoma virus); a hepadnavirus (e.g., Hepatitis B Virus (HBV)); a herpesvirus (e.g., herpes simplex virus (HSV), varicella zoster virus (VZV), Epstein-Barr virus (EBV), cytomegalovirus (CMV), herpes lymphotropic virus, Pityriasis Rosea, kaposi's sarcoma-associated herpesvirus); an adenovirus (e.g., atadenovirus, aviadenovirus, ichtadenovirus, mastadenovirus, siadeno virus); a poxvirus (e.g., smallpox, vaccinia virus, cowpox virus, monkeypox virus, orf virus, pseudocowpox, bovine papular stomatitis virus; tanapox virus, yaba monkey tumor virus; molluscum contagiosum virus (MCV)); a parvovirus (e.g., adeno-associated virus (AAV), Parvovirus B 19, human bocavirus, bufavirus, human parv4 G1); Gemini viridae; Nanoviridae; Phycodnaviridae; and the like. In some cases, the target DNA is parasite DNA. In some cases, the target DNA is bacterial DNA, e.g., DNA of a pathogenic bacterium.

RGNs complexes with RNA

[151] RGNs can be complexed to a gRNA (gRNA/RGN complex) in order to deliver Cas in proximity with a target nucleic acid sequence. The gRNA, is a polynucleotide that site-specifically guides a Cas nuclease, or a deactivated Cas nuclease, to a target nucleic acid region. The binding specificity is determined jointly by the complementary region on the cognate guide and a short DNA motif (protospacer adjacent motif or PAM) juxtaposed to the complementary region. The spacer present in the gRNA specifically hybridizes to a target nucleic acid sequence and determines the location of a Cas protein's site-specific binding and nucleolytic cleavage.

[152] RNA/Cas complexes can be produced using methods well known in the art. For example, the RNA of the complexes can be produced in vitro and RGN polypeptides can be recombinantly produced and then the RNA and RGN proteins can be complexed together using methods known in the art.

Additionally, cell lines constitutively expressing RGN proteins can be developed and can be transfected with the gRNA components, and complexes can be purified from the cells using standard purification techniques, such as but not limited to affinity, ion exchange and size exclusion chromatography. See, e.g. , Jinek M., et al, "A programmable dual-R A-guided DNA endonuclease in adaptive bacterial immunity," Science (2012) 337:816-821.

[153] Alternatively, the components, i.e., the gRNA and RGN polynucleotides may be provided separately to a cell, e.g., using separate constructs, or together, in a single construct, or in any combination, and complexes can be purified as above.

Variants of RGNs

[154] The present disclosure provides RGNs comprising at least 50, 100, 150, 200, 250, 300, 350, 400, 450 or more contiguous amino acid residues of the amino acid as provided above in Table 4. RNA-guided nucleases provided herein can comprise at least one nuclease domain (e.g ., DNase, RNase domain) and at least one RNA recognition and/or RNA binding domain to interact with gRNAs. Further domains that can be found in RNA-guided nucleases provided herein include, but are not limited to, DNA binding domains, helicase domains, protein-protein interaction domains, and dimerization domains. In specific embodiments, the RNA-guided nucleases provided herein can comprise at least 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% to one or more of a DNA binding domains, helicase domains, protein-protein interaction domains, and dimerization domains.

[155] While the activity of a variant or fragment may be altered compared to the polynucleotide or polypeptide of interest, the variant and fragment should retain the functionality of the polynucleotide or polypeptide of interest. For example, a variant or fragment may have increased activity, decreased activity, different spectrum of activity or any other alteration in activity when compared to the polynucleotide or polypeptide of interest.

[156] Fragments and variants of naturally-occurring RGN polypeptides, such as those disclosed herein, will retain sequence-specific, RNA-guided DNA-binding activity. In particular embodiments, fragments and variants of naturally-occurring RGN polypeptides, such as those disclosed herein, will retain nuclease activity (single -stranded or double-stranded).

[157] A biologically active variant of an RGN polypeptide of the invention may differ by as few as 1-15 amino acid residues, as few as 1-10, such as 6-10, as few as 5, as few as 4, as few as 3, as few as 2, or as few as 1 amino acid residue. In specific embodiments, the polypeptides can comprise an N- terminal or a C-terminal truncation, which can comprise at least a deletion of 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60,

65, 70, 75, 80, 85, 90, 95, 100, 150, 200, 250, 300, 350 amino acids or more from either the N or C terminus of the polypeptide.

[158] A biologically active variant of an RGN polypeptide of the invention may differ by as few as 1 or 2 amino acids.

Variants of tracrRNA and CRISPR RNA repeat sequence (crRNA)

[159] RGN proteins can have varying sensitivity to mismatches between a spacer sequence in a gRNA and its target sequence that affects the efficiency of cleavage. The CRISPR RNA repeat sequence comprises a nucleotide sequence that comprises a region with sufficient complementarity to hybridize to a tracrRNA. In various embodiments, the CRISPR RNA repeat sequence can comprise from 5 nucleotides to 30 nucleotides, or more. For example, the CRISPR repeat sequence can be 5,6,7,8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, or more nucleotides in length. In some embodiments, the CRISPR repeat sequence is 12 nucleotides in length In some embodiments, the degree of complementarity between a CRISPR repeat sequence and its corresponding tracrRNA sequence, when optimally aligned using a suitable alignment algorithm, is or more than 50%, 60%, 70%, 75%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or more. In particular embodiments, the CRISPR repeat sequence comprises the nucleotide sequence set forth in Table 5, or an active variant or fragment thereof that when comprised within a gRNA, is capable of directing the sequence-specific binding of an associated RNA-guided nuclease provided herein to a target sequence of interest. In certain embodiments, an active CRISPR repeat sequence variant of a wild-type sequence comprises a nucleotide sequence having at least 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or more sequence identity to the nucleotide sequence set forth in Table 5. In certain embodiments, an active CRISPR repeat sequence fragment of a wild-type sequence comprises at least 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 contiguous nucleotides from 3' end of the nucleotide sequence set forth in Table 5 (also listed in Table 10). . In certain embodiments, an active CRISPR repeat sequence fragment of a wild-type sequence comprises at least 10, 11, 12, 13, 15, 15, 16, 17, 18, 19, or 20 contiguous nucleotides from 3' end of the nucleotide sequence set forth in Table 10. In certain embodiments, an active CRISPR repeat sequence fragment of a wild-type sequence comprises a nucleotide sequence set forth in Table 11.

[160] Fragments and variants of naturally occurring CRISPR repeats, such as those disclosed herein, will retain the ability, when part of a gRNA (comprising a tracrRNA), to bind to and guide an RNA-guided nuclease (complexed with the gRNA) to a target nucleotide sequence in a sequence-specific manner.

[161] Fragments and variants of naturally occurring tracrRNAs, such as those disclosed herein, will retain the ability, when part of a gRNA (comprising a CRISPR RNA), to guide an RNA-guided nuclease (complexed with the gRNA) to a target nucleotide sequence in a sequence-specific manner.

[162] In various embodiments, the anti-repeat region of the tracrRNA that is fully or partially complementary to the CRISPR repeat sequence comprises from 3 nucleotides to 30 nucleotides, or more. For example, the region of base pairing between the tracrRNA anti-repeat sequence and the CRISPR repeat sequence can be 3,4,5,6,7,8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, or more nucleotides in length. In particular embodiments, the anti-repeat region of the tracrRNA that is fully or partially complementary to a CRISPR repeat sequence is 8 nucleotides in length. In some embodiments, the degree of complementarity between a CRISPR repeat sequence and its corresponding tracrRNA anti-repeat sequence, when optimally aligned using a suitable alignment algorithm, is or more than 50%, 60%, 70%, 75%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or more.

[163] In various embodiments, the entire tracrRNA can comprise from 60 nucleotides to more than 140 nucleotides. For example, the tracrRNA can be 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, or more nucleotides in length. In particular embodiments, the tracrRNA is 80 to 90 nucleotides in length, including 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, and 90 nucleotides in length. In certain embodiments, the tracrRNA is 107 nucleotides in length.

[164] In particular embodiments, the tracrRNA comprises the nucleotide sequence set forth in Table 5 or an active variant or fragment thereof that when comprised within a gRNA is capable of directing the sequence-specific binding of an associated RNA-guided nuclease provided herein to a target sequence of interest. In certain embodiments, an active tracrRNA sequence variant of a wild-type sequence comprises a nucleotide sequence having at least 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%,80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or more sequence identity to the nucleotide sequence set forth in Table 5. In certain embodiments, an active tracrRNA sequence fragment of a wild-type sequence comprises at least 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, or more contiguous nucleotides of the nucleotide sequence set forth in Table 5.

[165] In certain embodiments, the presently disclosed polynucleotides comprise or encode a CRISPR repeat comprising a nucleotide sequence having at least 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or greater identity to the nucleotide sequence set forth in Table 5.

[166] The presently disclosed polynucleotides can comprise or encode a tracrRNA comprising a nucleotide sequence having at least 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or greater identity to the nucleotide sequence set forth in Table 5.

[167] Biologically active variants of a CRISPR repeat or tracrRNA of the invention may differ by as few as 1-15 nucleotides, as few as 1-10, such as 6-10, as few as 5, as few as 4, as few as 3, as few as 2, or as few as 1 nucleotide. In specific embodiments, the polynucleotides can comprise a 5' or 3' truncation, which can comprise at least a deletion of 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80 nucleotides or more from either the 5' or 3' end of the polynucleotide.

Variants of spacer sequences

[168] In some embodiments, the degree of complementarity between a spacer sequence and its corresponding target sequence, when optimally aligned using a suitable alignment algorithm, is or more than 80%, 85%, 90, 95%, 96%, 97%, 98%, 99%, or more. In particular embodiments, the spacer sequence is free of secondary structure, which can be predicted using any suitable polynucleotide folding algorithm known in the art, including but not limited to mFold (see, e.g., Zuker and Stiegler (1981) Nucleic Acids Res . 9: 133-148) and RNAfold (see, e.g., Gruber et al. (2008) Cell l06(l):23-24).

Nucleotides encoding RNA-guided nucleases, crRNA, and/or tracrRNA

[169] The present disclosure provides polynucleotides comprising the presently disclosed crRNAs, tracrRNAs, and/or gRNAs and polynucleotides comprising a nucleotide sequence encoding the presently disclosed RGNs, crRNAs, tracrRNAs, and/or gRNAs. Presently disclosed polynucleotides include those comprising or encoding a CRISPR repeat sequence comprising the nucleotide sequence set forth in Table 5, or an active variant or fragment thereof that when comprised within a gRNA is capable of directing the sequence-specific binding of an associated RNA- guided nuclease to a target sequence of interest.

[170] The disclosure also provides polynucleotides comprising or encoding a tracrRNA comprising the nucleotide sequence set forth in Table 5, or an active variant or fragment thereof that when comprised within a gRNA is capable of directing the sequence- specific binding of an associated RNA-guided nuclease to a target sequence of interest. Polynucleotides are also provided that encode an RGN comprising the amino acid sequence set forth in Table 4, and active fragments or variants thereof that retain the ability to bind to a target nucleotide sequence in an RNA-guided sequence-specific manner.

[171] The expression cassette will include in the 5'-3' direction of transcription, a transcriptional (and, in some embodiments, translational) initiation region (i.e.. a promoter), an RGN-, a transcriptional initiation region (i.e.. a promoter), crRNA-, tracrRNA-and/or gRNA- encoding polynucleotide of the invention, and

a transcriptional (and in some embodiments, translational) termination region (i.e.. termination region) functional in the organism of interest. The promoters of the invention are capable of directing or driving expression of a coding sequence in a host cell. The regulatory regions (e.g., promoters, transcriptional regulatory regions, and translational termination regions) may be endogenous or heterologous to the host cell or to each other.

[172] Additional regulatory signals may include, but are not limited to, transcriptional initiation start sites, operators, activators, enhancers, other regulatory elements, ribosomal binding sites, an initiation codon, and termination signals.

[173] In preparing the expression cassette, the various DNA fragments may be manipulated, so as to provide for the DNA sequences in the proper orientation and, as appropriate, in the proper reading frame. Toward this end, adapters or linkers may be employed to join the DNA fragments or other manipulations may be involved to provide for convenient restriction sites, removal of superfluous DNA, removal of restriction sites, or the like. For this purpose, in vitro mutagenesis, primer repair, restriction, annealing, resubstitutions, e.g., transitions and transversions, may be involved.

[174] A number of promoters can be used in the practice of the invention. The promoters can be selected based on the desired outcome. The nucleic acids can be combined with constitutive, inducible, growth stage-specific, cell type-specific, tissue-preferred, tissue-specific, or other promoters for expression in the organism of interest.

[175] In some embodiments, the nucleotide comprises a tissue-preferred promoter. In some embodiments, the nucleic acid molecules encoding a RGN, crRNA-, tracrRNA-and/or gRNA comprise a cell type-specific promoter.

[176] The nucleic acid sequences encoding the RGNs, crRNA-, tracrRNA-and/or gRNA can be operably linked to a promoter sequence that is recognized by a phage RNA polymerase for example, for in vitro mRNA synthesis. For example, the promoter sequence can be a pol I, pol II, pol III, T7, T3, U6, CMV or SP6 promoter sequence or a variation of a T7, T3, U6, CMV or SP6 promoter sequence. In such embodiments, the expressed protein and/or RNAs can be purified for use in the methods of genome modification described herein. Any Pol II promoter or terminator could express the RGN. The choice of a promoter depends on how strongly RGN needs to be expressed and in what tissue type. In a preferred embodiment the RGN is expressed using is the CMV promoter. The gRNA can be expressed by Pol III promoters (e.g. U6 promoter).

[177] In certain embodiments, the polynucleotide encoding the RGN also can be linked to a polyadenylation signal (e.g., SV40 polyA signal, or sv40 polyA with rrnG terminator) and/or at least one

transcriptional termination sequence. Additionally, the sequence encoding the RGN also can be linked to sequence(s) encoding at least one nuclear localization signal, at least one cell- penetrating domain, and/or at least one signal peptide capable of trafficking proteins to particular subcellular locations.

[178] Additional regulatory signals include, but are not limited to, transcriptional initiation start sites, operators, activators, enhancers, other regulatory elements, ribosomal binding sites, an initiation codon, termination signals, and the like. See, for example, U.S. Pat. Nos. 5,039,523 and 4,853,331

[179] In preparing the expression cassette, the various DNA fragments may be manipulated, so as to provide for the DNA sequences in the proper orientation and, as appropriate, in the proper reading frame. Toward this end, adapters or linkers may be employed to join the DNA fragments or other manipulations may be involved to provide for convenient restriction sites, removal of superfluous DNA, removal of restriction sites, or the like. For this purpose, in vitro mutagenesis, primer repair, restriction, annealing, resubstitutions, e.g., transitions and transversions, may be involved.

Variants of polynucleotides

[180] For polynucleotides, conservative variants include those sequences that, because of the degeneracy of the genetic code, encode the native amino acid sequence of the gene of interest. Naturally occurring allelic variants such as these can be identified with the use of well-known molecular biology techniques, as, for example, with polymerase chain reaction (PCR) and hybridization techniques as outlined below. Variant polynucleotides also include synthetically derived polynucleotides, such as those generated, for example, by using site-directed mutagenesis but which still encode the polypeptide or the polynucleotide of interest. Generally, variants of a particular polynucleotide disclosed herein will have at least 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or more sequence identity to that particular polynucleotide as determined by sequence alignment programs and parameters described elsewhere herein.

[181] Variants of a particular polynucleotide disclosed herein (i.e. , the reference polynucleotide) can also be evaluated by comparison of the percent sequence identity between the polypeptide encoded by a variant polynucleotide and the polypeptide encoded by the reference polynucleotide. Percent sequence identity between any two polypeptides can be calculated using sequence alignment programs and parameters described elsewhere herein. Where any given pair of polynucleotides disclosed herein is evaluated by comparison of the percent sequence identity shared by the two polypeptides they encode, the percent sequence identity between the two encoded polypeptides is at least 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or more sequence identity.

[182] In particular embodiments, the presently disclosed polynucleotides encode an RGN polypeptide comprising an amino acid sequence having at least 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or greater identity to an amino acid sequence set forth in Table 4.

[183] Variant polynucleotides and proteins also encompass sequences and proteins derived from a mutagenic and recombinogenic procedure such as DNA shuffling. With such a procedure, one or more different RGN proteins disclosed herein is manipulated to create a new RGN protein possessing the desired properties. In this manner, libraries of recombinant polynucleotides are generated from a population of related sequence polynucleotides comprising sequence regions that have substantial sequence identity and can be homologously recombined in vitro or in vivo. For example, using this approach, sequence motifs encoding a domain of interest may be shuffled between the RGN sequences provided herein and other known RGN genes to obtain a new gene coding for a protein with an improved property of interest, such as an increased Km in the case of an enzyme. Strategies for such DNA shuffling are known in the art. See, for example, Stemmer (1994) Proc. Natl. Acad. Sci. USA

Codon-optimized sequences

[184] The nucleic acid molecules encoding RGNs and/or gRNA can be codon optimized for expression in a target cell or tissue of interest. Such polynucleotide coding sequence normally has its frequency of codon usage designed to mimic the frequency of preferred codon usage or transcription conditions of a particular host cell. Expression in the particular host cell or organism is enhanced as a result of the alteration of one or more codons at the nucleic acid level such that the translated amino acid sequence is not changed. Nucleic acid molecules can be codon optimized, either wholly or in part. Codon tables and other references providing preference information for a wide range of organisms are available in the art.

Vectors

[185] The polynucleotide encoding the RGN, and/or gRNA can be present in a vector or multiple vectors. Suitable vectors include plasmid vectors, phagemids, cosmids, artificial/mini-chromosomes, transposons, and viral vectors (e.g., lentiviral vectors, adeno-associated viral vectors, adenoviral vectors). The vector may comprise additional expression control sequences (e.g., enhancer sequences, Kozak sequences, polyadenylation sequences, transcriptional termination sequences), selectable marker sequences (e.g., antibiotic resistance genes), origins of replication, and the like. (see e.g. "Current Protocols in Molecular Biology" Ausubel et al, John Wiley & Sons, New York, 2003 or "Molecular Cloning: A Laboratory Manual" Sambrook & Russell, Cold Spring Harbor Press, Cold Spring Harbor, N.Y., 3rd edition, 2001).

[186] The vector may also comprise a selectable marker gene for the selection of transformed cells. Selectable marker genes are utilized for the selection of transformed cells or tissues. Marker genes include genes encoding antibiotic resistance, such as those encoding neomycin phosphotransferase II (NEO) and hygromycin phosphotransferase (HPT), as well as genes conferring resistance to herbicidal compounds, such as glufosinate ammonium, bromoxynil, imidazolinones, and 2,4-dichlorophenoxyacetate (2,4-D).

Delivery of the components to the target cells

[187] In some aspects, components of the present invention are delivered using nanoscale delivery systems, such as nanoparticles. Additionally, liposomes and other particulate delivery systems can be used. For example, vectors including the components of the present methods can be packaged in liposomes prior to delivery.

[188] As indicated, expression constructs comprising nucleotide sequences encoding the RGNs, and/or gRNA can be used to transform organisms of interest. Methods for transformation involve introducing a nucleotide construct into an organism of interest.

[189] The methods of the invention do not require a particular method for introducing a nucleotide construct to a host organism, only that the nucleotide construct gains access to the interior of a target cell. The host cell can be a eukaryotic or prokaryotic cell. In a particular embodiment, the eukaryotic host cell is a plant cell, a mammalian cell, or an insect cell. Methods for introducing nucleotide constructs into host cells are known in the art including, but not limited to, stable transformation methods, transient transformation methods, and virus-mediated methods.

[190] It is recognized that other exogenous or endogenous nucleic acid sequences or DNA fragments may also be incorporated into the host cell. Transformation of a host cell may be performed by infection, transfection, microinjection, electroporation, microprojection, biolistics or particle bombardment, electroporation, silica/carbon fibers, ultrasound mediated, PEG mediated, calcium phosphate co-precipitation, polycation DMSO technique, DEAE dextran procedure, and viral mediated, liposome mediated and other similar methods. Viral -mediated introduction of a polynucleotide encoding an RGN, and/or gRNA includes retroviral, lentiviral, adenoviral, and adeno-associated viral mediated introduction and expression.

[191] Transformation may result in stable or transient incorporation of the nucleic acid into the cell.

[192] The cells that have been transformed may be grown into a transgenic organism using well-known methods. Alternatively, cells that have been transformed may be introduced into an organism. These cells could have originated from the organism, wherein the cells are transformed in an ex vivo approach.

[193] The polynucleotides encoding the RGNs, and/or gRNAs can also be used to transform any prokaryotic cells, including but not limited to, archaea and bacteria.

[194] The polynucleotides encoding the RGNs, and/or gRNAs can be used to transform any eukaryotic cells, including but not limited to animal (e.g., mammals, insects, fish, birds, and reptiles), fungi, amoeba, algae, and yeast cells.

[195] Conventional viral and non-viral based gene transfer methods can be used to introduce nucleic acids in mammalian cells or target tissues. Such methods can be used to administer nucleic acids encoding components of a CRISPR system to cells in culture, or in a host organism. Non-viral vector delivery systems include DNA plasmids, RNA (e.g. a transcript of a nucleic acid described herein), naked nucleic acid, and nucleic acid complexed with a delivery vehicle, such as a liposome. Viral vector delivery systems include DNA and RNA viruses, which have either episomal or integrated genomes after delivery to the cell.

[196] Methods of non-viral delivery of nucleic acids include lipofection, nucleofection, microinjection, biolistics, virosomes, liposomes, immunoliposomes, polycation or lipid: nucleic acid conjugates, naked DNA, artificial virions, and agent-enhanced uptake of DNA. Lipofection is described in, e.g., US 5,049,386 and lipofection reagents are wildly available commercially. Delivery can be to cells (e.g. in vitro or ex vivo administration) or target tissues (e.g. in vivo administration).

Viral delivery for therapeutic applications

[197] The use of viral based systems for the delivery of nucleic acids allows targeting a virus to specific cells and trafficking the viral payload to the nucleus. Viral vectors can be administered directly to patients (*in vivo*) or they can be used to treat cells in vitro, and the modified cells may optionally be administered to patients (*ex vivo*). Conventional viral based systems could include retroviral, lentivirus, adenoviral, adeno-associated and herpes simplex virus vectors for gene transfer. Integration in the host genome is possible with the retrovirus, lentivirus, and adeno-associated virus gene transfer methods, often resulting in long term expression of the inserted transgene. Additionally, high transduction efficiencies have been observed in many different cell types and target tissues.

[198] The tropism of a retrovirus can be altered by incorporating foreign envelope proteins, expanding the potential target population of target cells. Lentiviral vectors are retroviral vectors that are able to transduce or infect non-dividing cells and typically produce high viral titers. Selection of a retroviral gene transfer system would therefore depend on the target tissue. Retroviral vectors are comprised of cis-acting long terminal repeats with packaging capacity for up to 6-10 kb of foreign sequence. The minimum cis-acting LTRs are sufficient for replication and packaging of the vectors, which are then used to integrate

the therapeutic gene into the target cell to provide permanent transgene expression. Widely used retroviral vectors include those based upon murine leukemia virus (MuLV), gibbon ape leukemia virus (GaLV), Simian Immuno deficiency virus (SIV), human immuno deficiency virus (HIV), and combinations thereof (see, e.g., Buchscher et al., J. Viral. 66:2731-2739 (1992); Johann et al., J. Viral. 66: 1635-1640 (1992); Sommnerfelt et al., Viral. 176:58-59 (1990); Wilson et al., J. Viral. 63:2374-2378 (1989); Miller et al., 7. Viral. 65:2220-2224

Transient expression and gene therapy applications

[199] In applications where transient expression is preferred, adenoviral based systems may be used. Adenoviral based vectors are capable of very high transduction efficiency in many cell types and do not require cell division. With such vectors, high titer and levels of expression have been obtained. This vector can be produced in large quantities in a relatively simple system. Adeno-associated virus ("AAV") vectors may also be used to transduce cells with target nucleic acids. Construction of recombinant AAV vectors are described in a number of publications, including U.S. 5,173,414. Packaging cells are typically used to form virus particles that are capable of infecting a host cell.

[200] Viral vectors used in gene therapy are usually generated by producing a cell line that packages a nucleic acid vector into a viral particle. The vectors typically contain the minimal viral sequences required for packaging and subsequent integration into a host, other viral sequences being replaced by an expression cassette for the polynucleotide(s) to be expressed. The missing viral functions are typically supplied in trans by the packaging cell line. For example, AAV vectors used in gene therapy typically only possess ITR sequences from the AAV genome which are required for packaging and integration into the host genome. Viral DNA is packaged in a cell line, which contains a helper plasmid encoding the other AAV genes, namely rep and cap, but lacking ITR sequences.

[201] The cell line may also be infected with adenovirus as a helper. The helper virus promotes replication of the AAV vector and expression of AAV genes from the helper plasmid. The helper plasmid is not packaged in significant amounts due to a lack of ITR sequences. Contamination with adenovirus can be reduced by, e.g., heat treatment to which adenovirus is more sensitive than AAV. Additional methods for the delivery of nucleic acids to cells are known to those skilled in the art. See, for example, US20030087817

Methods of modifying target nucleotide sequence

[202] In one aspect, the disclosure provides methods of modifying a target polynucleotide in a eukaryotic cell, which may be performed in vivo, ex vivo or in vitro. In some embodiments, the method comprises sampling a cell or population of cells from a human or non-human animal or plant (including microalgae)

and modifying the cell or cells. Culturing may occur at any stage ex vivo. The cell or cells may even be re-introduced into the non-human animal or plant (including micro-algae).

[203] The present disclosure provides methods for binding, cleaving, and/or modifying a target nucleotide sequence of interest. The methods include delivering a system comprising at least one gRNA or a polynucleotide encoding the same, and at least one RGN polypeptide or a polynucleotide encoding the same to the target sequence or a cell, organelle, or embryo comprising the target sequence. In some of these embodiments, the RGN comprises the amino acid sequence as disclosed above, or an active variant or fragment thereof. In various embodiments, the gRNA comprises a CRISPR repeat sequence comprising the nucleotide sequence as provided above, or an active variant or fragment thereof. In a particular embodiment, the gRNA comprising the nucleotide sequence as provided above, or an active variant or fragment thereof. The RGN of the system may be nuclease dead RGN, or may be a fusion polypeptide. In some embodiments, the fusion polypeptide comprises a base-editing polypeptide, for example a cytidine deaminase or an adenosine deaminase. In particular embodiments, the RGN and/or gRNA is heterologous to the cell, organelle, or embryo to which the RGN and/or gRNA (or polynucleotide(s) encoding at least one of the RGN and gRNA) are introduced.

[204] In those embodiments wherein the method comprises delivering a polynucleotide encoding a gRNA and/or an RGN polypeptide, the cell or embryo can then be cultured under conditions in which the gRNA and/or RGN polypeptide are expressed. In various embodiments, the method comprises contacting a target sequence with an RGN ribonucleoprotein complex. The RGN ribonucleoprotein complex may comprise an RGN that is nuclease dead or has nickase activity. In some embodiments, the RGN of the ribonucleoprotein complex is a fusion polypeptide comprising a base-editing polypeptide.

[205] In certain embodiments, the method comprises introducing into a cell, organelle, or embryo comprising a target sequence an RGN ribonucleoprotein complex. The RGN ribonucleoprotein complex can be one that has been purified from a biological sample, recombinantly produced and subsequently purified, or in vitro- assembled as described herein. In those embodiments wherein the RGN ribonucleoprotein complex that is contacted with the target sequence or a cell organelle, or embryo has been assembled in vitro, the method can further comprise the in vitro assembly of the complex prior to contact with the target sequence, cell, organelle, or embryo.

[206] A purified or in vitro assembled RGN ribonucleoprotein complex can be introduced into a cell, organelle, or embryo using any method known in the art, including, but not limited to electroporation. Alternatively, an RGN polypeptide and/or polynucleotide encoding or comprising the gRNA can be introduced into a cell, organelle, or embryo using any method known in the art.

[207] Upon delivery to or contact with the target sequence or cell, organelle, or embryo comprising the target sequence, the gRNA directs the RGN to bind to the target sequence in a sequence-specific manner. In those embodiments wherein the RGN has nuclease activity, the RGN polypeptide cleaves the target sequence of interest upon binding. The target sequence can subsequently be modified via endogenous repair mechanisms, such as non-homologous end joining, or homology-directed repair with a provided donor polynucleotide.

[208] Methods to measure binding of an RGN polypeptide to a target sequence are known in the art and include chromatin immunoprecipitation assays, gel mobility shift assays, DNA pull-down assays, reporter assays, microplate capture and detection assays. Likewise, methods to measure cleavage or modification of a target sequence are known in the art and include in vitro or in vivo cleavage assays wherein cleavage is confirmed using PCR, sequencing, or gel electrophoresis, with or without the attachment of an appropriate label (e.g., radioisotope, fluorescent substance) to the target sequence to facilitate detection of degradation products. Alternatively, the nicking triggered exponential amplification reaction (NTEXPAR) assay can be used (see, e.g., Zhang et al. (2016) Chem. Sci. 7:4951-4957). In vivo cleavage can be evaluated using the Surveyor assay (Guschin et al. (2010) Methods Mol Biol 649:247-256).

[209] In some embodiments, the methods involve the use of a single type of RGN complexed with more than one gRNA. The more than one guide RNA can target different regions of a single gene or can target multiple genes.

[210] In those embodiments wherein a donor polynucleotide is not provided, a double -stranded break introduced by an RGN polypeptide can be repaired by a non-homologous end-joining (NHEJ) repair process. Due to the error-prone nature of NHEJ, repair of the double -stranded break can result in a modification to the target sequence. Modification of the target sequence can result in the expression of an altered protein product or inactivation of a coding sequence.

[211] In those embodiments wherein a donor polynucleotide is present, the donor sequence in the donor polynucleotide can be integrated into or exchanged with the target nucleotide sequence during the course of repair of the introduced double-stranded break, resulting in the introduction of the exogenous donor sequence. A donor polynucleotide thus comprises a donor sequence that is desired to be introduced into a target sequence of interest. In some embodiments, the donor sequence alters the original target nucleotide sequence such that the newly integrated donor sequence will not be recognized and cleaved by the RGN. Integration of the donor sequence can be enhanced by the inclusion within the donor polynucleotide of flanking sequences that have substantial sequence identity with the sequences flanking the target nucleotide sequence, allowing for a homology-directed repair process. In those embodiments wherein the RGN polypeptide introduces double -stranded staggered breaks, the donor polynucleotide can comprise a

donor sequence flanked by compatible overhangs, allowing for direct ligation of the donor sequence to the cleaved target nucleotide sequence comprising overhangs by a non-homologous repair process during repair of the double -stranded break.

[212] In various embodiments, a method is provided for binding a target nucleotide sequence and detecting the target sequence, wherein the method comprises introducing into a cell, organelle, or embryo at least one guide RNA or a polynucleotide encoding the same, and at least one RGN polypeptide or a polynucleotide encoding the same, expressing the guide RNA and/or RGN polypeptide (if coding sequences are introduced), wherein the RGN polypeptide is a nuclease-dead RGN and further comprises a detectable label, and the method further comprises detecting the detectable label. The detectable label may be fused to the RGN as a fusion protein (e.g., fluorescent protein) or may be a small molecule conjugated to or incorporated within the RGN polypeptide that can be detected visually or by other means.

Methods of modulating gene expression

[213] Also provided herein are methods for modulating the expression of a target sequence or a gene of interest under the regulation of a target sequence. The methods comprise introducing into a cell, organelle, or embryo at least one gRNA or a polynucleotide encoding the same, and at least one RGN polypeptide or a polynucleotide encoding the same, expressing the gRNA and/or RGN polypeptide (if coding sequences are introduced), wherein the RGN polypeptide is a nuclease-dead RGN. In some of these embodiments, the nuclease-dead RGN is a fusion protein comprising an expression modulator domain (i.e., epigenetic modification domain, transcriptional activation domain or a transcriptional repressor domain) as described herein.

Methods for detecting ssDNA

[214] An RGN polypeptide of the present disclosure, once activated by detection of a target DNA (double or single stranded), can cleave non-targeted single stranded DNA (ssDNA). Once an RGN polypeptide is activated by a gRNA, after hybridization of gRNA with a target sequence of a target DNA, the protein becomes a nuclease that promiscuously cleaves ssDNAs. Thus, when the target DNA is present in the sample, the result is cleavage of ssDNAs in the sample, which can be detected using any common detection method (such as using a labeled single stranded DNA).

[215] Hence, the present disclosure provides systems and methods for detecting a target DNA (double stranded or single stranded) in a sample. In some cases, a detector DNA is used that is single stranded (ssDNA) and does not hybridize with the gRNA (i.e., the detector ssDNA is a non-target ssDNA). Such methods comprise steps of: (a) contacting the sample with: (i) an RGN polypeptide; (ii) a gRNA

comprising: a region that binds to the RGN polypeptide, and a spacer sequence that hybridizes with the target DNA; and (iii) a detector DNA that is single stranded and does not hybridize with the spacer sequence; and (b) measuring a detectable signal produced by cleavage of the single stranded detector DNA by the RGN polypeptide, thereby detecting the target DNA.

[216] The contacting step of a subject method can be carried out in a composition comprising divalent metal ions. The contacting step can be carried out outside of a cell. The contacting step can be carried out inside a cell. The contacting step can be carried out in a cell in vitro. The contacting step can be also carried out in a cell *ex vivo*. The contacting step can be carried out in a cell *in vivo*.

[217] In some embodiments the sample is contacted for 2 hours or less (e.g., 1.5 hours or less, 1 hour or less, 40 minutes or less, 30 minutes or less, 20 minutes or less, 10 minutes or less, or 5 minutes or less, or 1 minute or less), under conditions that provide for trans cleavage of the detector DNA. Conditions that provide for trans cleavage of the detector DNA include temperature conditions such as from 17°C to 39°C (e.g., 37°C).

[218] Methods for detecting ssDNA have been described for example in Chen, *et al* (2018). Science 360(6387), 436–439 or Kaminski, *et al.* Nat Biomed Eng 5, 643–656 (2021).

Kits

[219] In one aspect, the invention provides kits containing any one or more of the elements disclosed in the above methods and compositions. In some embodiments, the kit comprises a vector system and instructions for using the kit. In some embodiments, the vector system comprises (a) a first regulatory element operably linked to a gRNA sequence and one or more insertion sites for inserting a guide sequence downstream of the gRNA sequence, wherein when expressed, the gRNA directs sequence-specific binding of a CRISPR complex to a target sequence in a eukaryotic cell, wherein the CRISPR complex comprises a CRISPR enzyme complexed with (1) the gRNA sequence that is hybridized to the target sequence, and (2) a second regulatory element operably linked to an enzyme coding sequence encoding said CRISPR enzyme comprising a nuclear localization sequence. Elements may be provided individually or in combinations, and may be provided in any suitable container, such as a vial, a bottle, or a tube.

[220] In some embodiments, the kit includes instructions in one or more languages. In some embodiments, a kit comprises one or more reagents for use in a process utilizing one or more of the elements described herein. Reagents may be provided in any suitable container. For example, a kit may provide one or more reaction or storage buffers. Reagents may be provided in a form that is usable in a particular assay, or in a form that requires addition of one or more other components before use (e.g. in

concentrate or lyophilized form). A buffer can be any buffer, including but not limited to a sodium carbonate buffer, a sodium bicarbonate buffer, a borate buffer, a Tris buffer, a MOPS buffer, a HEPES buffer, and combinations thereof. In some embodiments, the buffer is alkaline. In some embodiments, the buffer has a pH from 6 to 10.

[221] In some embodiments, the kit comprises one or more oligonucleotides corresponding to a guide sequence for insertion into a vector so as to operably link the guide sequence and a regulatory element. In some embodiments, the kit comprises a homologous recombination template polynucleotide. In one aspect, the invention provides methods for using one or more elements of a CRISPR system. The CRISPR complex of the invention provides an effective means for modifying a target polynucleotide. The CRISPR complex of the disclosure has a wide variety of utility including modifying (e.g., deleting, inserting, translocating, inactivating, activating) a target polynucleotide in a multiplicity of cell types. As such the CRISPR complex of the invention has a broad spectrum of applications in, e.g., gene therapy, drug screening, disease diagnosis, and prognosis. An exemplary CRISPR complex comprises a CRISPR enzyme complexed with a guide sequence hybridized to a target sequence within the target polynucleotide.

Cells comprising the RGN-based systems

[222] Provided herein are cells and organisms comprising a target sequence of interest that has been modified using a process or the system based an RGN, and/or gRNA as described herein. Also are provided cells and organisms comprising the system for binding a target sequence of interest comprising an RGN, and/or gRNA as described herein.

[223] In some of these embodiments, the RGN comprises the amino acid sequence as disclosed above, or an active variant or fragment thereof. In various embodiments, the gRNA comprises a CRISPR repeat sequence comprising the nucleotide sequence as disclosed above, or an active variant or fragment thereof. In particular embodiments, the gRNA comprises the nucleotide sequence as disclosed above, or an active variant or fragment thereof. The modified cells can be eukaryotic (e.g., mammalian, plant, insect cell) or prokaryotic. Also provided are organelles and embryos comprising at least one nucleotide sequence that has been modified by a process utilizing an RGN and/or gRNA as described herein. The genetically modified cells, organisms, organelles, and embryos can be heterozygous or homozygous for the modified nucleotide sequence.

[224] The chromosomal modification of the cell, organism, organelle, or embryo can result in altered expression (up-regulation or down-regulation), inactivation, or the expression of an altered protein product or an integrated sequence. In those instances wherein the chromosomal modification results in either the inactivation of a gene or the expression of a non-functional protein product, the genetically

modified cell, organism, organelle, or embryo is referred to as a “knock-out”. The knock out phenotype can be the result of a deletion mutation (i.e.. deletion of at least one nucleotide), an insertion mutation (i.e.. insertion of at least one nucleotide), or a nonsense mutation (/. e. , substitution of at least one nucleotide such that a stop codon is introduced).

[225] Alternatively, the chromosomal modification of a cell, organism, organelle, or embryo can produce a “knock-in”, which results from the chromosomal integration of a nucleotide sequence that encodes a protein. In some of these embodiments, the coding sequence is integrated into the chromosome such that the chromosomal sequence encoding the wild-type protein is inactivated, but the exogenously introduced protein is expressed.

[226] In other embodiments, the chromosomal modification results in the production of a variant protein product. The expressed variant protein product can have at least one amino acid substitution and/or the addition or deletion of at least one amino acid. The variant protein product encoded by the altered chromosomal sequence can exhibit modified characteristics or activities when compared to the wild-type protein, including but not limited to altered enzymatic activity or substrate specificity.

[227] In yet other embodiments, the chromosomal modification can result in an altered expression pattern of a protein. As a non-limiting example, chromosomal alterations in the regulatory regions controlling the expression of a protein product can result in the overexpression or downregulation of the protein product or an altered tissue or temporal expression pattern.

Pharmaceutical compositions

[228] The polypeptides, nucleic acids and vectors of the present disclosure may be in a form of a pharmaceutical composition. The pharmaceutical composition may comprise 1 ng to 10 mg of DNA encoding the RGN/gRNA- based system or RGN/gRNA-based system protein component, i.e., the fusion protein. The pharmaceutical composition may comprise 1 ng to 10 mg of the DNA of the modified lentiviral vector. The pharmaceutical composition may comprise 1 ng to 10 mg of the DNA of the modified AAV vector and a nucleotide sequence encoding the site-specific nuclease. The pharmaceutical compositions according to the present invention can be formulated according to the mode of administration to be used. In cases where pharmaceutical compositions are injectable pharmaceutical compositions, they are sterile, pyrogen free and particulate free. An isotonic formulation is preferably used. Generally, additives for isotonicity may include sodium chloride, dextrose, mannitol, sorbitol and lactose. In some cases, isotonic solutions such as phosphate buffered saline are preferred. Stabilizers include gelatin and albumin. In some embodiments, a vasoconstriction agent is added to the formulation.

[229] The composition may further comprise a pharmaceutically acceptable excipient. The pharmaceutically acceptable excipient may be functional molecules as vehicles, adjuvants, carriers, or diluents. The pharmaceutically acceptable excipient may be a transfection facilitating agent, which may include surface active agents, such as immune-stimulating complexes (ISCOMS), Freunds incomplete adjuvant, LPS analog including monophosphoryl lipid A, muramyl peptides, quinone analogs, vesicles such as squalene and squalene, hyaluronic acid, lipids, liposomes, calcium ions, viral proteins, polyanions, polycations, or nanoparticles, or other known transfection facilitating agents.

[230] The transfection facilitating agent can be a polyanion, polycation, including poly-L- glutamate (LGS), or lipid. The transfection facilitating agent is poly-L-glutamate, and more preferably, the poly-L-glutamate is present in the composition for genome editing in skeletal muscle or cardiac muscle at a concentration less than 6 mg/ml. The transfection facilitating agent may also include surface active agents such as immune-stimulating complexes (ISCOMS), Freunds incomplete adjuvant, LPS analog including monophosphoryl lipid A, muramyl peptides, quinone analogs and vesicles such as squalene and squalene, and hyaluronic acid may also be used administered in conjunction with the genetic construct. In some embodiments, the DNA vector encoding the composition may also include a transfection facilitating agent such as lipids, liposomes, including lecithin liposomes or other liposomes known in the art, as a DNA-liposome mixture (see for example W09324640), calcium ions, viral proteins, polyanions, polycations, or nanoparticles, or other known transfection facilitating agents. Preferably, the transfection facilitating agent is a polyanion, polycation, including poly-L-glutamate (LGS), or lipid.

[231] The sequences included in the present invention are shown in Tables 8-15:

[232] Table 8. RGN sequences

Name	Sequence	SEQ ID NO
EGS0290	MTQVETLRAYKFALDPTAAQLADLARHAGAARWAFNHALGSKVAAHRQ WRAAVDALVAGGMPEPQARKAVKVPVPGNQAIKKTLNGLKGDSRSDPA LPDGFHGPPRPCPWWHEVSTYAFQSAFIDADRAWGNWLDSLTGRRAGG PVGYPRFKRKGRARDSFRLHHDVKKPTIRLDGYRRLQLPRLGSIRVHD SGKRLARLIAKGHAVIQSVTVFRGGNRWYASVLAKVQQDIPDKPTRAQ TGRGTVGVDWGIKHLATLSQPLDPADPATLHVANPRHLDAWNRQLATA QRALSRTERGSKRRAKAARRVGAIQHRIAQRRATTVHLLTKRLATGFA TVAVEDLNVRGMSASARGTVEQPGSRVRQKAGLNRGILDAAPGELRRQ LTYKTSWYGSTLAVLDRWHPSSKTCSSCGTAKPKLTLSERVFTCTTCG LTIDRDHNAAVNIARHAVVPLVEGDANARRSPRPADAGRDGHAERQKR EGPPPGGSPRRE	1
EGS0293	MATTDKKDKDKEKLRAYKFRLDPNQAQTIALYQAAGAARYTYNMLTAY NLEVNRLRDDYWKKRHDEGISDADIKKELNALTKEDKRYKQLKYGAFG TQYLTPEKKRHEQAEHRIENGEDPSVVWNQETERSANPWLHTANQRVL VSGLQNASDAWDNFWASRTGKRAGRLVGAPRFKKKGISRDSFTVPAAE TMGAYGTAYLRGEAAYQQGKRTITDYRHVRLSYLGVIRTYDSTKPLVK	2

	AVAAGAEIRSYTVSRNADRWYVSFLVKFSEPIRRSATKRARAAGAVGV DLGVKYLASLSDSEAPQRFPNLKFAEGLPSLENPRWSEASSRRLHKLQ RALARSQKGSNRRSRLVKQIARLHHMTALRRESNLHQLTKKLSTGYTL VGLEDLNVSGMTASAKGTVENPGKNVAQKSGLNRVVLDAAFGVFRYQL EYKTAWYGSTLEKIDRYFASSQTCSECGRKAKTKLTLRDRVFDCAYCG NMMDRDFNAAVNICREAQRLFNKKLASENGESLNGRGSRGALQGAETV EASRPPTSHRRGSP	
EGS0294	MTSTTLAPEEPLMVFRGARFRLDPTGEQQGILSQQAGAARVAYNMMCT LNKDILEARSQLYSTLIKDGKTKDKAKKELKAAAKEDPSLAIVWARDF DKNYITPERNRHKHAAQRIAAGENPVDVWNPDEERFNEPWLHTANRRV LRSGQKQYEQALDNFFKSQNGSRAGQKMGKPRFKTKIRSTDSFTIDAV DVSSSTTLIRDIGPKDHARYKTGEASTGIIADYRHVRLSHLGTFRVFG STKALVRQLDRGGRIKSCTVSRSADRWYVSFLVELPIEIARSTPTKKQ YKAGAVGIDLGVKSLAALSTGEIIPNPRFLRTADKKIKKLQRKIARCQ KGSKNSIRLKRRLARCHHELALQRAGYLNELTSMLASSFSAIALEDLN VAGMTSSARGTVENPGKNVKQKAGLNRSILDISPGRIRTLLEYKCTDR GVELQVIDRFFPSSQLCSSCGSKTTIPLAQRIYHCDVCGGVIDRDVNA AINIVYEAKRLVEQKCSEHSAPEGAEDKRPWSVHLPSTQYVDGLYTRK RQGPSGHSS	**3**
EGS0346	MSYLCARGRFNLSGFSTGSMGVLPVRASLFALLFTTLHGARISAHWYG MSKENDHTVTCIKICLEPNKAQRAQFASFAGSARWAYNFALAIKIGYQ KRWFEARKQFIESGLDEKAAGKKASEQVGRMPNYMSIATNEWTQLRDE VCPWYPEVPRRVFVGGFQRADAAFKNWFDSKSGRRSGAAMGWPKFKSK SKSRESFVIANDVQPAFVANLNRYIKTGELADMDYRHIKVPKCGEVRL TPGSAGQLRQLGRTMLAEAKTGELITRITSGTISRLGDRWYVSLVISG PFVPDAISTKRQRRNGVVGVDLGSGRFYATTSEGLSIINPKFVSKYEQ ELARANRALAKTAKGSAARKKALARLRRVHARSALARDGFSHQVSAWL TSQFAGVAVEKFDLASMLASAKGTVEKPGKNVDVKARFNAHLADVGIA STIDKLLYKGKRDGCRVQVVNTLDNSSTTCAKCGHTCVCGPEQKTFTC PDCGYNAPRQLNSAQYIRQLATVGFDELGLDMTASLTPDTGKRPIAFM TSAH	**4**
EGS0380	MSITTKTSDTSPDTVERHVAFKFCLNPTKAQLRALARDGGASRATYNM LVGYNADVMRNCNEYWGKRREEGASDDEIKAELKTLAKEDPRYKILGY MAYGKVLTAEGARHRACAKVIEEGAPVEEVWGADERSKEPWLHTINRR VLVSGTQSADKALKNFFDSRTGKRAGGKMGAPKFKSRAENSDSFTIPA PDTMGGYGTAYLRGEPAYKHAKETMKRRGEKGNPVISDYRHVRLAHLG TMRIHGNTRRMMRTIRNGGVIKSFTVSQVADRWYVSFLIETTVPAPKP TRKQKNAGAVGVDLGVKYMAALSDTNAPKRFSPDSGVNFTHETSPTVE NPRWLKRTEKRITKLQQKIARQVKGSNRRKVTVRKLAKTHHLVALRRE TGLHQLTKNLTTRYALIGIEDLNVTGMTASASGTIENPGKNVAQKAGL NRAVLDVSFGAFRTQLEYKASWYGSAVQVIDRYYPSSQTCSNCGKRPD AKLTLSDRVYKCEHCHTEIDRDLNAAINIRREAERLHAEA	**5**
EGS0288	MASTKTVVLRAFKFTLAPTATQDQQLLRWCGNARLAFNYALASKRAAH TEWRAQVDALVTSGVKEPVARKRVTGPKTPTKPAVYKAFIAERGDTRE GLDGVCPWAHEINTHVFQSAFIDADRAWKNWLDSFKGTRKGRRVGYPR FKKRGRARDAFRLHHTVTKPTIRFSTHRRLRLPTFGEVRLHDSARTLV RQIDRGTAVVQSVTVSRAGHRWYASVLCKVEMDLPSGPTRSQQAAGTV GIDFGVKALAALSKPLVPDRPESTLLPNPRHLAKAAHRLKRAQQTLSR RQKGSARREKARRRVARLHHEVAVRRQSALHQITKRLTTRFATIAVED LHVSGMTRSARGTMDKPGRKVRQKAGLNRAILDASMAEARRQITYKTS WYGSRLAVLDRWWPSSKTCSACGWQNPSLTLADRVFECAQCGLTLDRD LNAARNIEQHAVQVASGTGETQNARGEPVRLPRPRAEKQGSTKREDTG PPGPVPPRRSDPPTPPNPRQGQAKLF	**79**
EGS0291	MAQAEAPRRLRAYKFALDPTEAQLREFEQHAGSARWAYNHANAILSRY SDTLRNRWNAWIAQHHGLSREQLYALPDRERTAIQAAARAAVKAENAQ LAAELRIIDDHRKRVTHKGKPSVEPGEQPAEDAPERAYQLWRERVELA	**80**

	RLHAEDPQAYRAERKRILDEIRPLVNATKRKLIEQGAYRPTAMDISTL WREIRDLPPDEGGSPWWPEVSIYAFTSGFAHAETAWKNYLESLAGRRA GRPVGKPRFKKKRRSRRSFTLYGSVKLVTYRRIQVPSIGSVRLHGSAK RLHRALERRGGIIKSITISQGGHRWYASVLVDELDITPGRETQRGPSR RQRDRGAVGVDLGVHHLVALSDPNEKTLDNPRHLRKARKRLLKAQRAM SRRRGPDKRTGQEPSRRWVKARNRVARLHHELAVRRAGHLHEITKRLA TSYELVAIEDLNVAGMTRSARGTIDQPGRGVRAKAGLNRSILDTSPAE FRRQLQYKASWYGATVAVIDRWAPTSRTCSSCGAVKAKLSLAERTFFC EHCGMELDRDINAARNILAFAQSAYPGEGKALNACGGSVSPGSQSVVQ AGADEAGRPARKPRRSSRGSDPPATPTTRA	
EGS0295	MTDVELSAYRFALDTTPAQLTMLRQHAGAARWAYNHALGVKFAALDER KTVIAGLVEQGLDPKTSAAQAPKIPTKPAIQKALNTTKGDDRISAAGD CPWWHTVSTYAFQSAFADADTAWKNWLASLTGKRSGPIGAPRFKSKHR SRDSFRIHHDVNNPTIRPDDGYRRIIVPRLGSLRVHDSTKRLKRAIDR GAVIQSVTISRGGHRWYASILVKAPAAHAAPTRRQRQAGTVGVDLGVH HLAALSTGDIIDNPRHLAAGQKRLTKAQRALSRTEKGSNRRRRAAARV GRRHHEITERRATTLHTLTKHLATNWATVAIEDLNVAGMTRSARGTID NPGTNVRAKAGLSRAILDTSPGELRRQLTYKTGWYGSTLAICDRWFPS SQQCCECKVRTKLRLSQRVFTCPACGYGPIDRDVHAARNIAAYAAVAS DTGETLTARRDTAEAPTRVGRRRGAVDAGRPHRETGAATPAEQPAGHP KHADQRTLPLVS	**81**
EGS0318	MNRAYKFRLDPNQAQKAELMRCVGAARYTYNLLNAYNLQILRNEQEYR NTRNAEGADYETINGEIRKLRKKDPAYKFLGHAEYEKRYLTPEKQRHE AIAQAITDGADPAVVWSETERFAEPWLHTIARRVLVSGIKNADKAWDN YNKSRMKQRAGARMGIPRFKRKGVSRDSFTVPHETTGAYGAYYHKKDP EYARRKVQLKRRGISAKPTITDYRHVRLASLGVIRTHNTTKPLVKAVR AGAEIKSFTVSRAADHWYVSILVELTRPSTAPTRAQRSAGAVGVDLGV RYLAALSDEQAPQRFAQYPSLEFTSDGAPTLANPRWARAAEKRLVRLQ RALSRAQKGSKRRARIVQQIARHHHLVALRRESGLHQVSKRLATGYTL IGLEDLAVAGMTASAAGTIEAPGKNVRQKAGLNRSILDAAFSTLRRQL EYKASWYGSQVQIIDRFFASSQTCSACGARAKTKLDLRVRVFECAACG VRIDRDVNAARNIRAEAVRMYEA	**82**
EGS0334	MSVATDAPTQRLRAFKHRLDPDASQLELLGQYAGAARVAFNMLTAHNR AALQAGWDRRRQLAEAGVPAEELAGRMKAERAADPALKVAGYQQFATE HLTPMVRAHREAAEAIAAGADPGQVWADERYAQPWMHTVPRRVLVSGL QNAAKATENWMASACGARAGRRVGLPRFKKKGRSRDSFTIPAPEVIGA AGTAYKRGEARGGVITDHRHLRLASLGTIRTMDKTTRLVRSCRRGAVV RSVTISQAGGHWYASVLVAEPVTLRRGPSRRQRANGTVGVDLGVKHLA ALSTGELIPNARAGQAQARRLARLQRALARTERGSRRRERVRRQISAL HHQVALRRTGMLHEVSTRLARDYAIVALEDLNVAGMTRSAAGTIEHPG KNVAAKAGLNRAILDAGLGTLRAQLHYKTSWAGSQVKMIDRFAPSSKT CSRCGAVKATLSLRERVYECEVCALVIDRDVNAAINIRAWAVQEGTGH RVELAQGNGESQNGRRAAGRTPPSGGPTGWQRRSVKPAPRGAGRSSRA TGWSSHTPHIEREHAEADVFGPVR	**83**
EGS0336	MAQVLRAFRYALDPSPPQLETLQRCAGNARLAFNFALAMKKDAHQRWR DEVDILITTGLSEKDARTRLKGTSKIPNKPDVYKAFQHLRGDAAQGID GIAPWHAEIPTYVFQSAFQDADRAWKNWLDSYTGKRAGRRVGYPRFKK KFRSRDSFRLHHDAKKPKPALRLEGYRRLRLGGALKTVRLHGSAKPLH RLVSSGRAVVQSVTVSRGGTRWYASVLCKVETDVPDPTHRQKANGRIG LDWGLNHLAALSTPLDGHHLVDNPRHLRHASKRLTKAQRALSRTQKGS ARRRRAAARVGKLHHLVAEQRATFLHTLTKRLTTTFACVAIEDLNVAG MTRSARGTVQLPGKNVSSKAGLNRSILDAAPAELRRQLEYKTSWYGSH IAILDRWFPSSKTCSGCGWRNPSLPLSEREFVCAECGLRLDRDLNAAR NIAAHAEVPASGTGAPGRGESVNARGGCVSPRLLRESGQHPSKREDAA PSGPAPPRRSNPPTFP	**84**

EGS0337	MSQDKKDETVDHRAFKFRLDPTDAQLSLLAQSAGAARVGYNMLLGHNV AAYQARQELHASLLDSGHNEHDAKSAVTDRKTHDPSLQTLSYQSFSTH YLTPEITRHRIASDAIKAGADPSKVWNDPRYETPWMHTIPRRVFISGL QHANTAYTNWFASMRGDRAGARVGRPRFKKKGRCRDSFTIPAPEAMGA KGAPYKRGEPRSGVIEDYRHVRLAFLGVLRTHNSTKRLVRACQRGGKI KSFTVSRNADRWYVSFLVASPARAAVTTKRQRANGAVGVDVGVHSLAA LSNGEVFSNPRHGQLAQKKLKRAQRKLARTQKGSKGRARAAQRVGRLQ HLVTLQRETTAHHLTKYLATHYCAVAIEDLNVSGMTRSAKGTMEHPGK NVKAKSGLNRSILDASFGRIHQQLRYKMGWSGADLQIIGRFVPSSKMC SECGTVKSKLPLSERVYECEHCGLKIDRDVNAAINILHAGAGSITPSN SPLTRGSLNGRGDQTCLSPSGERITRGDSRSVKADPIWCRSPQTSNRC SSLTVRGTALK	85
EGS0338	MAKKTSTDAETMNRAYKFRLDPTQAQKAELMRCAGAARCTYNLLNEYN LQILRNEQEYRRTRAAEGIDHKTITSELKKLGKKDPEYKFLGQAEYEK RYLTPEKQRHEAIAQAIADGADPAIVWPETERFTEPWLHTIARRVLVS GIKNADKAWDNYNKSRMKQRAGARMGVPRRKRKGVSRDSFTIPAPEAI GAYGTYYHKKDPEYARRKAQLKRRGINAAPTIEDYRHVRLSHLGVIRT HNTTKPLVKAVRAGAQIRSYTVSRTADRWYVSILVELTRPSATPTRAQ RAAGAVGVDLGVRYLAALSDEWAPQRFARYPSLEFTGDGAPILANPCW ARAAERRLVRLQRALARAQKGSRRRARLVQQIARHHHLVALRRESGLH QVSKRLSTGYTLIGLEDLAVTGMTASAAGTVEAPGKNVRQKAGLNRSI LDAAFSTLRRQLEYKSGWYGSQVQIIDRFFASSQTCSACGARAKTKLD LRVRVFECAACGVRIDRDVNAARNIRAEAVRMYEAQLAPGTGESLNGR GAAGSDAAGSVVLGDVALDASRPAATGGGSP	86
EGS0341	MAEKTSTDDGVLYRAYKFRLDPNQAQKTELMRCVGAARHTYNLLTTYN LQILHNKQEYRRTRAAEGADHETINGELKKLREKDPAYQILYKAGYKT GNEYVPGYRQLYLTPEIKRHRAASKAIAAGADPAVVWPETERYSEPWL HSINSRVLDSGLLNAGKAWKNYKESLMHLRAGARMGVPRRKRKGVSRD SFTVDRDPNNDEVGAYGTPYKKGTPEFARRTAQLKRRGIKTAPTIEDY RHVRLSHLGVIRTHNTTKPLVKAVRAGANVKSYTVSRAANRWYVSILV KLSRTPATPTRAQRSAGAVGVDLGVRYLAALSDEQAPQRFAQYPSLEF TDNGAPTLANPRWARTAEKRLVRLQRALARAQKGSNRRARLVRQIARH HHLVALRRESGLHQVSKRLATGYTLIGLEDLAVAGMTASAAGTIKAPG KNVRQKAGLNRSILDAAFGTLRRQLEYKASWYGSQVQIIDRFFASSQT CSACGARAKTKLGLHVRVFECAACGARIDRDVNAARNIRAEAVRMYEA QLAPGTGESLNGRGATDSDAAGSVVLGDVALDVSRPAASGGGSP	87
EGS0343	MVQTEILKAFRFALDPTSVQVAALSRHAGAARWAFNHALAAKVGAHER WRAEVAKLVGDGVPEEQARRQVRVPVPMKPAIQKALNAVKGDSRKGLD GACPWWHEVNTYAFQSAFIDADQAWKNWLDSLSGKRVGRRVGYPRFKK KGRARDSFRLHHDVKKPSIRLAGYRRLRLPRIGEVRLHDSGKRLARLI DRGDAVVQSVTVSRGGHRWYASVLCKVTVQVPDRPSRRQRERGAVGVD LGVKVLAALSKPLVVDDPSSALVRNPQHLRQAERRLLKAQRALARTQK GSARREKAKRRVGRAHHEVAVRRHAALHQITKRLTTGFAVVALEDLNV AGMTRSARGTVAAPGKNVRQKAGLNRVILDSAPAELRRQVNYKATWYG SELAVADRWFPSSKTCSGCGWQNPHLKLSDRVFRCTDCGLVMDRDMNA ARNIERHAVLIDRNVACDRRETLNARGASIRPTTSRGGRHDASKREGS GARPESPQQSDLLTLPTLNNETATPP	88
EGS0344	MFERTTTKAAFLSPLDLRPTQATDLERFAGTTRWAFNWANALLEAHHQ AYEGRRQQAARHLFGLGPEQLDELRVLANGTRDENGKKAKGDPVKRRE YESIQKATKKAVSEENKALGAEMKLWDEHRSLVVHKGRPLLTPGDEPA LDAPPLAHRLYARRVELAGIQKTDPDYYAEQRKKEREAITPNVVAMKR DLMAKGAYFPSEYDLQYIWRTVRDLPKEEGGSPWWPECPTILFYDGIN RARTAWKNWMDSASGARKGPPVGMPRFKSKYKAKDTFTITNPNRSVIK FETYRRIAITGIGSMRLHRGAKLLARRIAAGQAEITSATISRSGTAWY VSVLCTVHTTARTAPSKAQRSRGAVGVDWGVRALATTSKPIALTPGKP ASRTVPAEKYGAAMSQKIARAQRQLARMPKGSSRRRKAARHVADLQHL	89

	VAQRRASSVHQLSKALAQSFEIVAIEGLNVRGMTKSAKGTVENPGKNI RQKAGLNRAILDATPGELKRQLEYKTKKYGSRLVELDTWYPSSKTCSR CGWVHPKLKLSMRTFRCQQCGLVEDRDFNAAVNIERQGITHIVKENEG TDDREEG	

[233] Table 9. Truncated Protein Sequences and improved variants

ID	Sequence	SEQ ID NO
EGS0290v2	MTQVETLRAYKFALDPTAAQLADLARHAGAARWAFNHALGSKVAAHRQ WRAAVDALVAGGMPEPQARKAVKVPVPGNQAIKKTLNGLKGDSRSDPA LPDGFHGPPRPCPWWHEVSTYAFQSAFIDADRAWGNWLDSLTGRRAGG PVGYPRFKRKGRARDSFRLHHDVKKPTIRLDGYRRLQLPRLGSIRVHD SGKRLARLIAKGHAVIQSVTVFRGGNRWYASVLAKVQQDIPDKPTRAQ TGRGTVGVDWGIKHLATLSQPLDPADPATLHVANPRHLDAWNRQLATA QRALSRTERGSKRRAKAARRVGAIQHRIAQRRATTVHLLTKRLATGFA TVAVEDLNVRGMSASARGTVEQPGSRVRQKAGLNRGILDAAPGELRRQ LTYKTSWYGSTLAVLDRWHPSSKTCSSCGTAKPKLTLSERVFTCTTCG LTIDRDHNAAVNIARH	6
EGS0293v2	MATTDKKDKDKEKLRAYKFRLDPNQAQTIALYQAAGAARYTYNMLTAY NLEVNRLRDDYWKKRHDEGISDADIKKELNALTKEDKRYKQLKYGAFG TQYLTPEKKRHEQAEHRIENGEDPSVVWNQETERSANPWLHTANQRVL VSGLQNASDAWDNFWASRTGKRAGRLVGAPRFKKKGISRDSFTVPAAE TMGAYGTAYLRGEAAYQQGKRTITDYRHVRLSYLGVIRTYDSTKPLVK AVAAGAEIRSYTVSRNADRWYVSFLVKFSEPIRRSATKRARAAGAVGV DLGVKYLASLSDSEAPQRFPNLKFAEGLPSLENPRWSEASSRRLHKLQ RALARSQKGSNRRSRLVKQIARLHHMTALRRESNLHQLTKKLSTGYTL VGLEDLNVSGMTASAKGTVENPGKNVAQKSGLNRVVLDAAFGVFRYQL EYKTAWYGSTLEKIDRYFASSQTCSECGRKAKTKLTLRDRVFDCAYCG NMMDRDFNAAVNICREAQRLFNK	7
EGS0293v3	MRAYKFRLDPNQAQTIALYQAAGAARYTYNMLTAYNLEVNRLRDDYWK KRHDEGISDADIKKELNALTKEDKRYKQLKYGAFGTQYLTPEKKRHEQ AEHRIENGEDPSVVWNQETERSANPWLHTANQRVLVSGLQNASDAWDN FWASRTGKRAGRLVGAPRFKKKGISRDSFTVPAAETMGAYGTAYLRGE AAYQQGKRTITDYRHVRLSYLGVIRTYDSTKPLVKAVAAGAEIRSYTV SRNADRWYVSFLVKFSEPIRRSATKRARAAGAVGVDLGVKYLASLSDS EAPQRFPNLKFAEGLPSLENPRWSEASSRRLHKLQRALARSQKGSNRR SRLVKQIARLHHMTALRRESNLHQLTKKLSTGYTLVGLEDLNVSGMTA SAKGTVENPGKNVAQKSGLNRVVLDAAFGVFRYQLEYKTAWYGSTLEK IDRYFASSQTCSECGRKAKTKLTLRDRVFDCAYCGNMMDRDFNAAVNI CREAQRLFNK	8
EGS0293v4	MATTDKKDKDKEKLRAYKFRLDPNQAQTIALYQAAGAARYTYNMLTAY NLEVNRLRDDYWKKRHDEGISDADIKKELNALTKEDKRYKQLKYGAFG TQYLTPEKKRHEQAEHRIENGEDPSVVWNQETERSANPWLHTANQRVL VSGLQNASDAWDNFWASRTGKRAGRLVGAPRFKKKGISRDSFTVPAAE TMGAYGTAYLRGEAAYQQGKRTITDYRHVRLSYLGVIRTYDSTKPLVK AVAAGAEIRSYTVSRNADRWYVSFLVKFSEPIRRSATKRARAAGAVGV DLGVKYLASLSDSEAPQRFPNLKFAEGLPSLENPRWSEASSRRLHKLQ RALARSQKGSNRRSRLVKQIARLHHMTALRRESNLHQLTKKLSTGYTL VGLEDLNVSGMTASAKGTVENPGKNVAQKSGLNRVVLDAAFGVFRYQL EYKTAWYGSTLEKIDRYFASSQTCSECGRKAKTKLTLRDRVFDCAYCG NMMDRDFNAAVNICREAQRLFNKKLASENGESLN	9
EGS0293v5	MATTDKKDKDKEKLRAYKFRLDPNQAQTIALYQAAGAARYTYNMLTAY NLEVNRLRDDYWKKRHDEGISDADIKKELNALTKEDKRYKQLKYGAFG TQYLTPEKKRHEQAEHRIENGEDPSVVWNQETERSANPWLHTANQRVL VSGLQNASDAWDNFWASRTGKRAGRLVGAPRFKKKGISRDSFTVPAAE	10

	TMGAYGTAYLRGEAAYQQGKRTITDYRHVRLSYLGVIRTYDSTKPLVK AVAAGAEIRSYTVSRNADRWYVSFLVKFSEPIRRSATKRARAAGAVGV DLGVKYLASLSDSEAPQRFPNLKFAEGLPSLENPRWSEASSRRLHKLQ RALARSQKGSNRRSRLVKQIARLHHMTALRRESNLHQLTKKLSTGYTL VGLEDLNVSGMTASAKGTVENPGKNVAQKSGLNRVVLDAAFGVFRYQL EYKTAWYGSTLEKIDRYFASSQTCSECGRKAKTKLTLRDRVFDCAYCG NMMDRDFNAAVNICREAQRLFNKKLASENGESLNGRGSRGALQGAETV EAS	
EGS0294v2	MTSTTLAPEEPLMVFRGARFRLDPTGEQQGILSQQAGAARVAYNMMCT LNKDILEARSQLYSTLIKDGKTKDKAKKELKAAAKEDPSLAIVWARDF DKNYITPERNRHKHAAQRIAAGENPVDVWNPDEERFNEPWLHTANRRV LRSGQKQYEQALDNFFKSQNGSRAGQKMGKPRFKTKIRSTDSFTIDAV DVSSSTTLIRDIGPKDHARYKTGEASTGIIADYRHVRLSHLGTFRVFG STKALVRQLDRGGRIKSCTVSRSADRWYVSFLVELPIEIARSTPTKKQ YKAGAVGIDLGVKSLAALSTGEIIPNPRFLRTADKKIKKLQRKIARCQ KGSKNSIRLKRRLARCHHELALQRAGYLNELTSMLASSFSAIALEDLN VAGMTSSARGTVENPGKNVKQKAGLNRSILDISPGRIRTLLEYKCTDR GVELQVIDRFFPSSQLCSSCGSKTTIPLAQRIYHCDVCGGVIDRDVNA AINIVYEAKRLVEQKC	11
EGS0294v3	MVFRGARFRLDPTGEQQGILSQQAGAARVAYNMMCTLNKDILEARSQL YSTLIKDGKTKDKAKKELKAAAKEDPSLAIVWARDFDKNYITPERNRH KHAAQRIAAGENPVDVWNPDEERFNEPWLHTANRRVLRSGQKQYEQAL DNFFKSQNGSRAGQKMGKPRFKTKIRSTDSFTIDAVDVSSSTTLIRDI GPKDHARYKTGEASTGIIADYRHVRLSHLGTFRVFGSTKALVRQLDRG GRIKSCTVSRSADRWYVSFLVELPIEIARSTPTKKQYKAGAVGIDLGV KSLAALSTGEIIPNPRFLRTADKKIKKLQRKIARCQKGSKNSIRLKRR LARCHHELALQRAGYLNELTSMLASSFSAIALEDLNVAGMTSSARGTV ENPGKNVKQKAGLNRSILDISPGRIRTLLEYKCTDRGVELQVIDRFFP SSQLCSSCGSKTTIPLAQRIYHCDVCGGVIDRDVNAAINIVYEAKRLV EQKC	12
EGS0346v2	MSYLCARGRFNLSGFSTGSMGVLPVRASLFALLFTTLHGARISAHWYG MSKENDHTVTCIKICLEPNKAQRAQFASFAGSARWAYNFALAIKIGYQ KRWFEARKQFIESGLDEKAAGKKASEQVGRMPNYMSIATNEWTQLRDE VCPWYPEVPRRVFVGGFQRADAAFKNWFDSKSGRRSGAAMGWPKFKSK SKSRESFVIANDVQPAFVANLNRYIKTGELADMDYRHIKVPKCGEVRL TPGSAGQLRQLGRTMLAEAKTGELITRITSGTISRLGDRWYVSLVISG PFVPDAISTKRQRRNGVVGVDLGSGRFYATTSEGLSIINPKFVSKYEQ ELARANRALAKTAKGSAARKKALARLRRVHARSALARDGFSHQVSAWL TSQFAGVAVEKFDLASMLASAKGTVEKPGKNVDVKARFNAHLADVGIA STIDKLLYKGKRDGCRVQVVNTLDNSSTTCAKCGHTCVCGPEQKTFTC PDCGYNAPRQLNSAQYIRQLATVGFDELG	13
dEGS0293v2	MATTDKKDKDKEKLRAYKFRLDPNQAQTIALYQAAGAARYTYNMLTAY NLEVNRLRDDYWKKRHDEGISDADIKKELNALTKEDKRYKQLKYGAFG TQYLTPEKKRHEQAEHRIENGEDPSVVWNQETERSANPWLHTANQRVL VSGLQNASDAWDNFWASRTGKRAGRLVGAPRFKKKGISRDSFTVPAAE TMGAYGTAYLRGEAAYQQGKRTITDYRHVRLSYLGVIRTYDSTKPLVK AVAAGAEIRSYTVSRNADRWYVSFLVKFSEPIRRSATKRARAAGAVGV ALGVKYLASLSDSEAPQRFPNLKFAEGLPSLENPRWSEASSRRLHKLQ RALARSQKGSNRRSRLVKQIARLHHMTALRRESNLHQLTKKLSTGYTL VGLEDLNVSGMTASAKGTVENPGKNVAQKSGLNRVVLDAAFGVFRYQL EYKTAWYGSTLEKIDRYFASSQTCSECGRKAKTKLTLRDRVFDCAYCG NMMDRAFNAAVNICREAQRLFNK	14
EGS0288v2	MASTKTVVLRAFKFTLAPTATQDQQLLRWCGNARLAFNYALASKRAAH TEWRAQVDALVTSGVKEPVARKRVTGPKTPTKPAVYKAFIAERGDTRE GLDGVCPWAHEINTHVFQSAFIDADRAWKNWLDSFKGTRKGRRVGYPR FKKRGRARDAFRLHHTVTKPTIRFSTHRRLRLPTFGEVRLHDSARTLV	90

	RQIDRGTAVVQSVTVSRAGHRWYASVLCKVEMDLPSGPTRSQQAAGTV GIDFGVKALAALSKPLVPDRPESTLLPNPRHLAKAAHRLKRAQQTLSR RQKGSARREKARRRVARLHHEVAVRRQSALHQITKRLTTRFATIAVED LHVSGMTRSARGTMDKPGRKVRQKAGLNRAILDASMAEARRQITYKTS WYGSRLAVLDRWWPSSKTCSACGWQNPSLTLADRVFECAQCGLTLDRD LNAARNIEQHAVQV	
EGS0291v2	MAQAEAPRRLRAYKFALDPTEAQLREFEQHAGSARWAYNHANAILSRY SDTLRNRWNAWIAQHHGLSREQLYALPDRERTAIQAAARAAVKAENAQ LAAELRIIDDHRKRVTHKGKPSVEPGEQPAEDAPERAYQLWRERVELA RLHAEDPQAYRAERKRILDEIRPLVNATKRKLIEQGAYRPTAMDISTL WREIRDLPPDEGGSPWWPEVSIYAFTSGFAHAETAWKNYLESLAGRRA GRPVGKPRFKKKRRSRRSFTLYGSVKLVTYRRIQVPSIGSVRLHGSAK RLHRALERRGGIIKSITISQGGHRWYASVLVDELDITPGRETQRGPSR RQRDRGAVGVDLGVHHLVALSDPNEKTLDNPRHLRKARKRLLKAQRAM SRRRGPDKRTGQEPSRRWVKARNRVARLHHELAVRRAGHLHEITKRLA TSYELVAIEDLNVAGMTRSARGTIDQPGRGVRAKAGLNRSILDTSPAE FRRQLQYKASWYGATVAVIDRWAPTSRTCSSCGAVKAKLSLAERTFFC EHCGMELDRDINAARNILAFAQSA	91
EGS0295v2	MTDVELSAYRFALDTTPAQLTMLRQHAGAARWAYNHALGVKFAALDER KTVIAGLVEQGLDPKTSAAQAPKIPTKPAIQKALNTTKGDDRISAAGD CPWWHTVSTYAFQSAFADADTAWKNWLASLTGKRSGPIGAPRFKSKHR SRDSFRIHHDVNNPTIRPDDGYRRIIVPRLGSLRVHDSTKRLKRAIDR GAVIQSVTISRGGHRWYASILVKAPAAHAAPTRRQRQAGTVGVDLGVH HLAALSTGDIIDNPRHLAAGQKRLTKAQRALSRTEKGSNRRRRAAARV GRRHHEITERRATTLHTLTKHLATNWATVAIEDLNVAGMTRSARGTID NPGTNVRAKAGLSRAILDTSPGELRRQLTYKTGWYGSTLAICDRWFPS SQQCCECKVRTKLRLSQRVFTCPACGYGPIDRDVHAARNIAAYAAVA	92
EGS0334v2	MSVATDAPTQRLRAFKHRLDPDASQLELLGQYAGAARVAFNMLTAHNR AALQAGWDRRRQLAEAGVPAEELAGRMKAERAADPALKVAGYQQFATE HLTPMVRAHREAAEAIAAGADPGQVWADERYAQPWMHTVPRRVLVSGL QNAAKATENWMASACGARAGRRVGLPRFKKKGRSRDSFTIPAPEVIGA AGTAYKRGEARGGVITDHRHLRLASLGTIRTMDKTTRLVRSCRRGAVV RSVTISQAGGHWYASVLVAEPVTLRRGPSRRQRANGTVGVDLGVKHLA ALSTGELIPNARAGQAQARRLARLQRALARTERGSRRRERVRRQISAL HHQVALRRTGMLHEVSTRLARDYAIVALEDLNVAGMTRSAAGTIEHPG KNVAAKAGLNRAILDAGLGTLRAQLHYKTSWAGSQVKMIDRFAPSSKT CSRCGAVKATLSLRERVYECEVCALVIDRDVNAAINIRAWAVQEGTGH RV	93
EGS0337v2	MSQDKKDETVDHRAFKFRLDPTDAQLSLLAQSAGAARVGYNMLLGHNV AAYQARQELHASLLDSGHNEHDAKSAVTDRKTHDPSLQTLSYQSFSTH YLTPEITRHRIASDAIKAGADPSKVWNDPRYETPWMHTIPRRVFISGL QHANTAYTNWFASMRGDRAGARVGRPRFKKKGRCRDSFTIPAPEAMGA KGAPYKRGEPRSGVIEDYRHVRLAFLGVLRTHNSTKRLVRACQRGGKI KSFTVSRNADRWYVSFLVASPARAAVTTKRQRANGAVGVDVGVHSLAA LSNGEVFSNPRHGQLAQKKLKRAQRKLARTQKGSKGRARAAQRVGRLQ HLVTLQRETTAHHLTKYLATHYCAVAIEDLNVSGMTRSAKGTMEHPGK NVKAKSGLNRSILDASFGRIHQQLRYKMGWSGADLQIIGRFVPSSKMC SECGTVKSKLPLSERVYECEHCGLKIDRDVNAAINILHAGAGSITPSN S	94
EGS0338v2	MAKKTSTDAETMNRAYKFRLDPTQAQKAELMRCAGAARCTYNLLNEYN LQILRNEQEYRRTRAAEGIDHKTITSELKKLGKKDPEYKFLGQAEYEK RYLTPEKQRHEAIAQAIADGADPAIVWPETERFTEPWLHTIARRVLVS GIKNADKAWDNYNKSRMKQRAGARMGVPRRKRKGVSRDSFTIPAPEAI GAYGTYYHKKDPEYARRKAQLKRRGINAAPTIEDYRHVRLSHLGVIRT HNTTKPLVKAVRAGAQIRSYTVSRTADRWYVSILVELTRPSATPTRAQ RAAGAVGVDLGVRYLAALSDEWAPQRFARYPSLEFTGDGAPILANPCW	95

	ARAAERRLVRLQRALARAQKGSRRRARLVQQIARHHHLVALRRESGLH QVSKRLSTGYTLIGLEDLAVTGMTASAAGTVEAPGKNVRQKAGLNRSI LDAAFSTLRRQLEYKSGWYGSQVQIIDRFFASSQTCSACGARAKTKLD LRVRVFECAACGVRIDRDVNAARNIRAEAVRMYEA	
EGS0341v2	MAEKTSTDDGVLYRAYKFRLDPNQAQKTELMRCVGAARHTYNLLTTYN LQILHNKQEYRRTRAAEGADHETINGELKKLREKDPAYQILYKAGYKT GNEYVPGYRQLYLTPEIKRHRAASKAIAAGADPAVVWPETERYSEPWL HSINSRVLDSGLLNAGKAWKNYKESLMHLRAGARMGVPRRKRKGVSRD SFTVDRDPNNDEVGAYGTPYKKGTPEFARRTAQLKRRGIKTAPTIEDY RHVRLSHLGVIRTHNTTKPLVKAVRAGANVKSYTVSRAANRWYVSILV KLSRTPATPTRAQRSAGAVGVDLGVRYLAALSDEQAPQRFAQYPSLEF TDNGAPTLANPRWARTAEKRLVRLQRALARAQKGSNRRARLVRQIARH HHLVALRRESGLHQVSKRLATGYTLIGLEDLAVAGMTASAAGTIKAPG KNVRQKAGLNRSILDAAFGTLRRQLEYKASWYGSQVQIIDRFFASSQT CSACGARAKTKLGLHVRVFECAACGARIDRDVNAARNIRAEAVRMYEA	96
EGS0343v2	MVQTEILKAFRFALDPTSVQVAALSRHAGAARWAFNHALAAKVGAHER WRAEVAKLVGDGVPEEQARRQVRVPVPMKPAIQKALNAVKGDSRKGLD GACPWWHEVNTYAFQSAFIDADQAWKNWLDSLSGKRVGRRVGYPRFKK KGRARDSFRLHHDVKKPSIRLAGYRRLRLPRIGEVRLHDSGKRLARLI DRGDAVVQSVTVSRGGHRWYASVLCKVTVQVPDRPSRRQRERGAVGVD LGVKVLAALSKPLVVDDPSSALVRNPQHLRQAERRLLKAQRALARTQK GSARREKAKRRVGRAHHEVAVRRHAALHQITKRLTTGFAVVALEDLNV AGMTRSARGTVAAPGKNVRQKAGLNRVILDSAPAELRRQVNYKATWYG SELAVADRWFPSSKTCSGCGWQNPHLKLSDRVFRCTDCGLVMDRDMNA ARNIERHAVLIDRNV	97
EGS0293v2_Q343K_N347K_A424R	MATTDKKDKDKEKLRAYKFRLDPNQAQTIALYQAAGAARYTYNMLTAY NLEVNRLRDDYWKKRHDEGISDADIKKELNALTKEDKRYKQLKYGAFG TQYLTPEKKRHEQAEHRIENGEDPSVVWNQETERSANPWLHTANQRVL VSGLQNASDAWDNFWASRTGKRAGRLVGAPRFKKKGISRDSFTVPAAE TMGAYGTAYLRGEAAYQQGKRTITDYRHVRLSYLGVIRTYDSTKPLVK AVAAGAEIRSYTVSRNADRWYVSFLVKFSEPIRRSATKRARAAGAVGV DLGVKYLASLSDSEAPQRFPNLKFAEGLPSLENPRWSEASSRRLHKLQ RALARSKKGSKRRSRLVKQIARLHHMTALRRESNLHQLTKKLSTGYTL VGLEDLNVSGMTASAKGTVENPGKNVAQKSGLNRVVLDARFGVFRYQL EYKTAWYGSTLEKIDRYFASSQTCSECGRKAKTKLTLRDRVFDCAYCG NMMDRDFNAAVNICREAQRLFNK	98

[234] Table 10. CRISPR repeat sequences of the RGN polypeptides

Name	**Sequence**	**SEQ ID NO**
EGS0290 repeat	GTCCGGCCCGGGCGCGCAGGGGCTGACCG	15
EGS0293 repeat	GTCTGCCCCGCGTGTGCATGGATGGTTCC	16
EGS0294 repeat	TGTTCTTCCCACGCACACGAGGAAGATCCC	17
EGS0346 repeat	GTCTTGCCCGTGCGCGTGTGGGGTGATC	18
EGS0380 repeat	GTTTTCCCCGTACACATGGGGATGGCTC	19
EGS0288 repeat	GTGCGGCCCGGGCGCGCAGGGGGTGACCG	99

EGS0291 repeat	GTCTGTTCCGTGCGCACAGGGGTAGCCC	100
EGS0295 repeat	GCGCACCCCACGCGCGCACAGGTGAGTCCG	101
EGS0318 repeat	GTGTTCCCCGTATGCGCGGGGGTGAGCT	102
EGS0334 repeat	GCCCCGCGCACGCAGGGATGAGCCC	103
EGS0336 repeat	CGTCGGGTCCACGCGCACAGGGAACGACCG	104
EGS0337 repeat	GAGCGGCCCTCGCGCACAGGGATGAGTCC	105
EGS0338 repeat	GCGTTCCCCGCGTGTGCGGGGGTGAGCTC	106
EGS0341 repeat	GTATTCCCCGTATGCACGGGGATGAGCCG	107
EGS0343 repeat	GTCGGCCCCACGCGTGTAGGGAGCGATCG	108
EGS0344 repeat	GTGCTCCCCACGCGCGTACGGGTGGTCCC	109

[235] Table 11. Part of the CRISPR repeat sequence of the RNA

Name	**Sequence**	**SEQ ID NO**
EGS0290 partial repeat	GGGGCTGACCG	20
EGS0293 partial repeat	ATGGATGGTTCC	21
EGS0294 partial repeat	ACACGAGGAAGATCCC	22
EGS0346 partial repeat	TGTGGGGTGATC	23
EGS0380 partial repeat	TGGGGATGGCTC	24
EGS0288 partial repeat	GGGGGTGACCG	110
EGS0291 partial repeat	GGGGTAGCCC	111
EGS0295 partial repeat	GCGCGCACAGGTGAGTCCG	112
EGS0318 partial repeat	GGGGGTGAGCTG	113
EGS0334 partial repeat	GGGATGAGCCC	114
EGS0336 partial repeat	GGGAACGACCG	115

EGS0337 partial repeat	AGGGATGAGTCC	116
EGS0338 partial repeat	GGGGGTGAGCTC	117
EGS0341 partial repeat	GGGATGAGCCG	118
EGS0343 partial repeat	GTAGGGAGCGATCG	119
EGS0344 partial repeat	CGGGTGGTCCC	120

[236] Table 12. crRNA sequences of the RGN polypeptides

Name	**Sequence**	**SEQ ID NO**
EGS0290 crRNA	GGGGCTGACCGnnnnnnnnnnnnnnnnnnn	**25**
EGS0293 crRNA	ATGGATGGTTCCnnnnnnnnnnnnnnnnnnn	**26**
EGS0294 crRNA	ACACGAGGAAGATCCCnnnnnnnnnnnnnnnnnnn	**27**
EGS0346 crRNA	TGTGGGGTGATCnnnnnnnnnnnnnnnn	**28**
EGS0380 crRNA	TGGGGATGGCTCnnnnnnnnnnnnnnnn	**29**
EGS0288 crRNA	GGGGGUGACCGNNNNNNNNNNNNNNNNNNN	**121**
EGS0291 crRNA	GGGGUAGCCCNNNNNNNNNNNNNNNNNNN	**122**
EGS0295 crRNA	GCGCGCACAGGUGAGUCCGNNNNNNNNNNNNNNNNNNN	**123**
EGS0318 crRNA	GGGGGUGAGCUGNNNNNNNNNNNNNNNNNNN	**124**
EGS0334 crRNA	GGGATGAGCCCNNNNNNNNNNNNNNNNNNN	**125**
EGS0336 crRNA	GGGAACGACCGNNNNNNNNNNNNNNNNNNN	**126**
EGS0337 crRNA	AGGGATGAGTCCNNNNNNNNNNNNNNNNNNN	**127**
EGS0338 crRNA	GGGGGTGAGCTCNNNNNNNNNNNNNNNNNNN	**128**
EGS0341 crRNA	GGGAUGAGCCGNNNNNNNNNNNNNNNNNNN	**129**
EGS0343 crRNA	GUAGGGAGCGAUCGNNNNNNNNNNNNNNNNNNN	**130**
EGS0344 crRNA	CGGGUGGUCCCNNNNNNNNNNNNNNNNNNN	**131**

[237] Table 13. tracrRNA sequences of the RGN polypeptides

Name	Sequence	SEQ ID NO
EGS0290 tracrRNA	CGTCGTCCCCCTGGTAGAGGGGGACGCTAACGCCCGCAGAAGTCCACG CCCGGCCGATGCAGGCCGGGACGGACACGCCGAAAGGCAGAAGCGGGA AGGCCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCCCCGACA GCCCC	30
EGS0293 tracrRNA	ACTCGCCTCAGAAAATGGGGAGAGTCTAAACGGACGTGGAAGTCGAGG CGCTCTTCAGGGTGCTGAGACTGTGGAAGCGTCAAGACCACCTACGAG TCATCGTAGAGGGTCACCGTAGATGAGTAATCATCTGCCCATCTAT	31
EGS0294 tracrRNA	GTGAACACTCGGCTCCGGAGGGGGCAGAGGATAAACGGCCGTGGAGTG TACATCTCCCATCAACGCAGTATGTTGATGGACTGTATACGAGGAAGC GGCAAGGCCCCAGCGGTCATAGCAGTTGAGTAATTGGCTGCTCTTTAT CGTGT	32
EGS0346 tracrRNA	GCCTTCGATAAGAGGGTTTAACGCCCACATTTTAAGGGGCCAGGCTGT AAAGGCTTGGTTTCGGTGGGAAGGCCACATTTTGTGGTCACCGTAGTC GAGTAATCGGCTACCACTTCACA	33
EGS0380 tracrRNA	ATAGGCTTGGGAAAGCCCGAGCCACCCTGTGTATCTCACTAAGATGCA CAGCAATGCCCCTTTAGCTCAGTTTGGTCAGAGCTACGGACTTTTA	34
EGS0288 tracrRNA	GUACAGGUCGCCUCCGGUACGGGGGAGACCCAAAACGCCCGUGGAGAA CCCGUCAGACUCCCCCGCCCCCGGGCGGAGAAGCAGGGCUCGACGAAG CGGGAAGACACCGGCCCUCCCGGGCCGGUGCCACCUCGGCGGAGCGAU CCGCCGACACCCCC	132
EGS0291 tracrRNA	UCCGCCUACCCCGGGGAGGGGAAGGCGUUAAAUGCCUGCGGAGGCUCU GUAAGUCCUGGAUCACAGUCUGUGGUUCAGGCUGGAGCUGAUGAAGCU GGCAGACCUGCGCGGAAACCGCGCAGGUCAUCCCGGGGGAGUGAUCCC CCGGCUACCCC	133
EGS0295 tracrRNA	UAUGCGGCCGUCGCCUCCGACACCGGGGAGACGUUAACCGCCCGCCGA GAUACCGCGGAAGCCCCCACACGUGUGGGUCGCCGCCGCGGUGCCGUU GACGCGGGAAGACCACACCGGGAAACCGGUGCGGCCACCCCAGCAGAG CAACCUGCUGGUCACCCAAAGCACGC	134
EGS0318 tracrRNA	GAGGCUGUGCGGAUGUAUGAGGCGCAACUCGCCCCCGGCAUGGGGGAG AGUCUAAACGGACGCGGAGUUACUGACUCUGAUGCUGCUGUUUCGGUG GUGUUGGGGGAUGCGGCGUUGGAUGCGUCAAGACCAGCCGCCAUGGGC GGCGGGUCACCGUAGGCAAGUAAUUGUCUACUCAUCCCC	135
EGS0334 tracrRNA	TGGGCCGTGCAGGAGGGCACGGGTCACCGGGTCGAGCTCGCCCAGGGC AATGGGGAGAGCCAAAACGGGCGCCGAGCTGCCGGCCGGACCCCACCA TCGGGTGGGCCCACCGGGTGGCAGCGCCGAAGCGTCAAGCCCGCACCG CGCGGTGCGGGCAGATCCAGCCGGGCAACTGGCTGGTCATCCC	136
EGS0336 tracrRNA	CACGCGGAAGTGCCAGCTTCTGGCACCGGTGCCCCTGGTAGAGGGGAA TCCGTGAACGCCCGTGGAGGGTGTGTAAGTCCCCGGCTCCTTCGGGAG TCTGGGCAGCACCCATCGAAACGGGAAGACGCTGCGCCCTCTGGGCCA GCGCCACCTCGGCGGAGCAACCCGCCGACGTTCCC	137
EGS0337 tracrRNA	GCCGGTGCTGGAAGTATCACCCCAAGCAACTCACCTCTGACAAGGGGG AGTCTAAATGGGCGTGGAGATCAAACCTGCCTTTCACCTTCGGGCGAG AGGATCACGAGGGGCGATTCACGAAGCGTCAAGGCCGACCCAATTTGG TGCCGGTCACCGCAAACGAGTAATCGTTGCTCATCTCT	138
EGS0338 tracrRNA	GAGGCTGTGCGGATGTATGAGGCGCAACTCGCCCCCGGCACGGGGGAG AGTCTAAACGGACGCGGAGCTGCGGGCTCTGATGCTGCCGGTTCGGTG GTATTGGGGGATGTGGCGTTGGATGCGTCAAGGCCAGCCGCCACTGGC GGCGGGTCACCGTAGGCGAGTGATCGTCTACTCATCCCC	139
EGS0341 tracrRNA	GAGGCUGUACGGAUGUAUGAGGCGCAACUCGCCCCCGGCACGGGGGAG AGUCUAAACGGACGCGGAGCUACUGACUCUGAUGCUGCCGGUUCGGUG	140

	GUAUUGGGGGAUGUGGCGUUGGAUGUGUCAAGGCCCGCCGCUUCUGGC GGCGGGUCACCGUAGGCGAGUAAUCGUCUACUCAUCCC	
EGS0343 tracrRNA	CCACGCAGUGCUGAUCGAUCGAAACGUCGCCUGCGAUAGGCGGGAGAC GCUAAACGCCCGUGGAGCAUCCAUAAGACCAACCACCUCUCGGGGCGG UAGGCACGACGCAUCGAAGCGGGAAGGCUCCGGCGCUCGGCCUGAGUC ACCUCAGCAGAGUGAUCUGCUGACGCUCCCAAC	**141**
EGS0344 tracrRNA	GCGCAAAAACCACGCGCGACCCCCUGCUGGACACUCACAACACGUGAC GCUCCCUGGAACAGAUGGGGUGCCAUCCGGCGCCCUGGAGCAGCCUCA UAAAGCGGAGCACGCCCGGCUCGCCGGACGCCGCCGAGGCACUGACCC GCCCCCGGAUGUGAAUCCAGGGGCGAGUUGGAGACCUGACAAGCCCCU CGCGGGCCGGCCAGGCCAUCCCAGUCGAGUAAUCGACUGGCCACCCG	**142**

[238] Table 14. gRNA sequences of the RGN polypeptides

Name	**Sequence**	**SEQ ID NO**
EGS0290 sgRNA	CGTCGTCCCCCTGGTAGAGGGGGACGCTAACGCCCGCAGAAGTCCACG CCCGGCCGATGCAGGCCGGGACGGACACGCCGAAAGGCAGAAGCGGGA AGGCCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCCCCGACA GCCCCGAAAGGGGCTGACCGNNNNNNNNNNNNNNNNNNNN	**35**
EGS0293 sgRNA	ACTCGCCTCAGAAAATGGGGAGAGTCTAAACGGACGTGGAAGTCGAGG CGCTCTTCAGGGTGCTGAGACTGTGGAAGCGTCAAGACCACCTACGAG TCATCGTAGAGGGTCACCGTAGATGAGTAATCATCTGCCCATCTATGA AAATGGATGGTTCCnnnnnnnnnnnnnnnnnnnn	**36**
EGS0294 sgRNA	GTGAACACTCGGCTCCGGAGGGGGCAGAGGATAAACGGCCGTGGAGTG TACATCTCCCATCAACGCAGTATGTTGATGGACTGTATACGAGGAAGC GGCAAGGCCCCAGCGGTCATAGCAGTTGAGTAATTGGCTGCTCTTTAT CGTGTTGTGTTGAAGGAAAATCGTGGTAAGTGGTGATGGATTCGCTGG TCAGGAAGTGTTCTTCCCACGCACACGAGGAAGATCCCNNNNNNNNNN NNNNNNNNNNNNN	**37**
EGS0346 sgRNA	GCCTTCGATAAGAGGGTTTAACGCCCACATTTTAAGGGGCCAGGCTGT AAAGGCTTGGTTTCGGTGGGAAGGCCACATTTTGTGGTCACCGTAGTC GAGTAATCGGCTACCACTTCACAgaaaTGTGGGGTGATCnnnnnnnnn nnnnnnn	**38**
EGS0380 sgRNA	AUAGGCUUGGGAAAGCCCGAGCCACCCUGUGUAUCUCACUAAGAUGCA CAGCAAUGCCCCUUUAGCUCAGUUUGGUCAGAGCUACGGACUUUUAGA AAUGGGGAUGGCUCNNNNNNNNNNNNNNNNN	**39**
EGS0288 sgRNA	GUACAGGUCGCCUCCGGUACGGGGGAGACCCAAAACGCCCGUGGAGAA CCCGUCAGACUCCCCCGCCCCCGGGCGGAGAAGCAGGGCUCGACGAAG CGGGAAGACACCGGCCCUCCCGGGCCGGUGCCACCUCGGCGGAGCGAU CCGCCGACACCCCCGAAAGGGGGUGACCGNNNNNNNNNNNNNNNNNN N	**143**
EGS0291 sgRNA	UCCGCCUACCCCGGGGAGGGGAAGGCGUUAAAUGCCUGCGGAGGCUCU GUAAGUCCUGGAUCACAGUCUGUGGUUCAGGCUGGAGCUGAUGAAGCU GGCAGACCUGCGCGGAAACCGCGCAGGUCAUCCCGGGGGAGUGAUCCC CCGGCUACCCCGAAAGGGGUAGCCCNNNNNNNNNNNNNNNNNNNN	**144**
EGS0295 sgRNA	UAUGCGGCCGUCGCCUCCGACACCGGGGAGACGUUAACCGCCCGCCGA GAUACCGCGGAAGCCCCCACACGUGUGGGUCGCCGCCGCGGUGCCGUU GACGCGGGAAGACCACACCGGGAAACCGGUGCGGCCACCCCAGCAGAG CAACCUGCUGGUCACCCAAAGCACGCGAAAGCGCGCACAGGUGAGUCC GNNNNNNNNNNNNNNNNNNN	**145**
EGS0318 sgRNA	GAGGCUGUGCGGAUGUAUGAGGCGCAACUCGCCCCCGGCAUGGGGGAG AGUCUAAACGGACGCGGAGUUACUGACUCUGAUGCUGCUGUUUCGGUG GUGUUGGGGGAUGCGGCGUUGGAUGCGUCAAGACCAGCCGCCAUGGGC	**146**

	GGCGGGUCACCGUAGGCAAGUAAUUGUCUACUCAUCCCCGAAAGGGGG UGAGCUGNNNNNNNNNNNNNNNNNN	
EGS0334 sgRNA	TGGGCCGTGCAGGAGGGCACGGGTCACCGGGTCGAGCTCGCCCAGGGC AATGGGGAGAGCCAAAACGGGCGCCGAGCTGCCGGCCGGACCCCACCA TCGGGTGGGCCCACCGGGTGGCAGCGCCGAAGCGTCAAGCCCGCACCG CGCGGTGCGGGCAGATCCAGCCGGGCAACTGGCTGGTCATCCCGAAAG GGATGAGCCCNNNNNNNNNNNNNNNNNN	**147**
EGS0336 sgRNA	CACGCGGAAGTGCCAGCTTCTGGCACCGGTGCCCCTGGTAGAGGGGAA TCCGTGAACGCCCGTGGAGGGTGTGTAAGTCCCCGGCTCCTTCGGGAG TCTGGGCAGCACCCATCGAAACGGGAAGACGCTGCGCCCTCTGGGCCA GCGCCACCTCGGCGGAGCAACCCGCCGACGTTCCCGAAAGGGAACGAC CGNNNNNNNNNNNNNNNNNN	**148**
EGS0337 sgRNA	GCCGGTGCTGGAAGTATCACCCCAAGCAACTCACCTCTGACAAGGGGG AGTCTAAATGGGCGTGGAGATCAAACCTGCCTTTCACCTTCGGGCGAG AGGATCACGAGGGGCGATTCACGAAGCGTCAAGGCCGACCCAATTTGG TGCCGGTCACCGCAAACGAGTAATCGTTGCTCATCTCTGAAAAGGGAT GAGTCCNNNNNNNNNNNNNNNNNN	**149**
EGS0338 sgRNA	GAGGCTGTGCGGATGTATGAGGCGCAACTCGCCCCCGGCACGGGGGAG AGTCTAAACGGACGCGGAGCTGCGGGCTCTGATGCTGCCGGTTCGGTG GTATTGGGGGATGTGGCGTTGGATGCGTCAAGGCCAGCCGCCACTGGC GGCGGGTCACCGTAGGCGAGTGATCGTCTACTCATCCCCGAAAGGGGG TGAGCTCNNNNNNNNNNNNNNNNNN	**150**
EGS0341 sgRNA	GAGGCUGUACGGAUGUAUGAGGCGCAACUCGCCCCCGGCACGGGGGAG AGUCUAAACGGACGCGGAGCUACUGACUCUGAUGCUGCCGGUUCGGUG GUAUUGGGGGAUGUGGCGUUGGAUGUGUCAAGGCCCGCCGCUUCUGGC GGCGGGUCACCGUAGGCGAGUAAUCGUCUACUCAUCCCGAAAGGGAUG AGCCGNNNNNNNNNNNNNNNNNN	**151**
EGS0343 sgRNA	CCACGCAGUGCUGAUCGAUCGAAACGUCGCCUGCGAUAGGCGGGAGAC GCUAAACGCCCGUGGAGCAUCCAUAAGACCAACCACCUCUCGGGGCGG UAGGCACGACGCAUCGAAGCGGGAAGGCUCCGGCGCUCGGCCUGAGUC ACCUCAGCAGAGUGAUCUGCUGACGCUCCCAACGAAAGUAGGGAGCGA UCGNNNNNNNNNNNNNNNNNN	**152**
EGS0344 sgRNA	GCGCAAAAACCACGCGCGACCCCCUGCUGGACACUCACAACACGUGAC GCUCCCUGGAACAGAUGGGGUGCCAUCCGGCGCCCUGGAGCAGCCUCA UAAAGCGGAGCACGCCCGGCUCGCCGGACGCCGCCGAGGCACUGACCC GCCCCCGGAUGUGAAUCCAGGGGCGAGUUGGAGACCUGACAAGCCCCU CGCGGGCCGGCCAGGCCAUCCCAGUCGAGUAAUCGACUGGCCACCCGG AAACGGGUGGUCCCNNNNNNNNNNNNNNNNNN	**153**

[239] Table 15. Improved sgRNA designs

Name	Sequence	SEQ ID NO
EGS0290 sgRNA v2	CGTCGTCCCCCTGGTAGAGGGGGACGCTAACGCCCGCGAAAGCGGGAA GGCCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCCCCGACAG CCCCGAAAGGGGCTGACCG	**40**
EGS0290 sgRNA v2.1	CGGCGTCCCCCTGGTAGAGGGGGACGCTACCGCCCGCGAAAGCGGGAA GGCCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCCCCGACAG CCCCGAAAGGGGCTGACCGNNNNNNNNNNNNNNNNNN	**41**
EGS0290 sgRNA v2.2	CGTCGTCCCCCTGGTAGAGGGGGACGCGAACGCCCGCGAAAGCGGGAA GGCCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCCCCGACAG CCCCGAAAGGGGCTGACCGNNNNNNNNNNNNNNNNNN	**42**
EGS0290 sgRNA v2.3	CGGCGTCCCCCTGGTAGAGGGGGACGCGACCGCCCGCGAAAGCGGGAA GGCCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCCCCGACAG CCCCGAAAGGGGCTGACCGNNNNNNNNNNNNNNNNNN	**43**

EGS0290 sgRNA v2.4	GCGTCCCCCTGGTAGAGGGGGACGCGACCGCCCGCGAAAGCGGGAAGG CCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCCCCGACAGCC CCGAAAGGGGCTGACCGNNNNNNNNNNNNNNNNNN	**44**
EGS0290 sgRNA v2.5	CGCCGTCCCCCTGGTAGAGGGGGACGCTAGCGCCCGCGAAAGCGGGAA GGCCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCCCCGACAG CCCCGAAAGGGGCTGACCGNNNNNNNNNNNNNNNNNN	**45**
EGS0290 sgRNA v2.6	CGTCGTCCGCCTGGTAGAGGCGGACGCTAACGCCCGCGAAAGCGGGAA GGCCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCCCCGACAG CCCCGAAAGGGGCTGACCGNNNNNNNNNNNNNNNNNN	**46**
EGS0290 sgRNA v2.7	CGTCGTCCGCCTGGTAGAGGCGGACGCTAACGCCCGCGAAAGCGGGAA GGCCCACCACCCGGTGGGTCACCTCGGCGGGAGTAATCCCGCCGACAG CCCCGAAAGGGGCTGACCGNNNNNNNNNNNNNNNNNN	**47**
EGS0293 sgRNA v2	ACTCGCCTCAGAAAATGGGGAGAGTCTAAACGGACGTGGAAGCGTCAA GACCACCTACGAGTCATCGTAGAGGGTCACCGTAGATGAGTAATCATC TGCCCATCTATGAAAATGGATGGTTCCNNNNNNNNNNNNNNNNNN	**48**
EGS0293 sgRNA v2.1	CCTCGCCTCAGAAAATGGGGCGAGGCTAAACGGACGTGGAAGCGTCAA GACCACCTACGAGTCATCGTAGAGGGTCACCGTAGATGAGTAATCATC TGCCCATCTATGAAAATGGATGGTTCCNNNNNNNNNNNNNNNNNN	**49**
EGS0293 sgRNA v2.2	ACTCGCCTCAGAAAATGGGGAGAGTCTAAACGGACGTGGAAGCGTCAA GACCCCCTACGAGTCATCGTAGGGGGTCACCGTAGATGAGTAATCATC TGCCCATCTATGAAAATGGATGGTTCCNNNNNNNNNNNNNNNNNN	**50**
EGS0293 sgRNA v2.3	CCTCGCCTCAGAAAATGGGGCGAGGCTAAACGGACGTGGAAGCGTCAA GACCCCCTACGAGTCATCGTAGGGGGTCACCGTAGATGAGTAATCATC TGCCCATCTATGAAAATGGATGGTTCCNNNNNNNNNNNNNNNNNN	**51**
EGS0293 sgRNA v2.4	CCTCGCCTCAGAAAATGGGGCGAGGGACGTGGAAGCGTCAAGACCCCC TACGAGTCATCGTAGGGGGTCACCGTAGATGAGTAATCATCTGCCCAT CTATGAAAATGGATGGTTCCNNNNNNNNNNNNNNNNNN	**52**
EGS0293 sgRNA v2.5	ACTCGCCTCAGAAAATGGGGCGAGTCTAAACGGACGTGGAAGCGTCAA GACCACCTACGAGTCATCGTAGAGGGTCACCGTAGATGAGTAATCATC TGCCCATCTATGAAAATGGATGGTTCCNNNNNNNNNNNNNNNNNN	**53**
EGS0293 sgRNA v3	CGGACGTGGAAGCGTCAAGACCACCTACGAGTCATCGTAGAGGGTCAC CGTAGATGAGTAATCATCTGCCCATCTATGAAAATGGATGGTTCCNNN NNNNNNNNNNNNNNN	**54**
EGS0293 sgRNA v4	AGACCACCTACGAGTCATCGTAGAGGGTCACCGTAGATGAGTAATCAT CTGCCCATCTATGAAAATGGATGGTTCCNNNNNNNNNNNNNNNNNN	**55**
EGS0293 sgRNA v5	ACTCGCCTCAGAAAATGGGGAGAGTCTAAACGGACGTGGAAGCGTCAA GACCACCGAGGGTCACCGTAGATGAGTAATCATCTGCCCATCTATGAA AATGGATGGTTCCNNNNNNNNNNNNNNNNNN	**56**
EGS0293 sgRNA v6	CGGACGTGGAAGCGTCAAGACCACCGAGGGTCACCGTAGATGAGTAAT CATCTGCCCATCTATGAAAATGGATGGTTCCNNNNNNNNNNNNNNNNN NN	**57**
EGS0294 sgRNA v2	GTGAACACTCGGCTCCGGAGGGGGCAGAGGATAAACGGCCGTGGAGTG TACATCTCCCATCAACGCAGTATGTTGATGGACTGTATACGAGGAAGC GGCAAGGCCCCAGCGGTCATAGCAGTTGAGTAATTGGCTGCTCTTTAT CGTGTTGTGTGAAAACGCACACGAGGAAGATCCCNNNNNNNNNNNNN NNNNN	**58**
EGS0294 sgRNA v3	GTGAACACTCGGCTCCGGAGGGGGCAGAGGATAAACGGCCGTGGAGTG TACATCTCCGACTGTATACGAGGAAGCGGCAAGGCCCCAGCGGTCATA GCAGTTGAGTAATTGGCTGCTCTTTATCGTGTTGTGTGAAAACGCACA CGAGGAAGATCCCNNNNNNNNNNNNNNNNNN	**59**
EGS0294 sgRNA v3.1	GTGAACACTCGGCTCCGGAGGGGGCAGAGGATAAACGGCCGCGGAGTG TACATCTCCGACTGTATACGAGGAAGCGGCAAGGCCCCAGCGGTCATA GCAGTTGAGTAATTGGCTGCTCTTTATCGTGTTGTGTGAAAACGCACA CGAGGAAGATCCCNNNNNNNNNNNNNNNNNN	**60**
EGS0294 sgRNA v3.2	GTGAACACTCGGCTCCGGAGGGGGCAGAGGATAAACGGCCGTGGAGTG TACATCTCCGACTGTATACGAGGAAGCGGCAAGGCCCCAGCGGCCATA	**61**

	GCAGTTGAGTAATTGGCTGCTCTTTATCGTGTTGTGTGAAAACGCACA CGAGGAAGATCCCNNNNNNNNNNNNNNNNNN	
EGS0294 sgRNA v3.3	GTGAACACGCGGCTCCGGAGGGGGCCGCGGATAAACGGCCGTGGAGTG TACATCTCCGACTGTATACGAGGAAGCGGCAAGGCCCCAGCGGTCATA GCAGTTGAGTAATTGGCTGCTCTTTATCGTGTTGTGTGAAAACGCACA CGAGGAAGATCCCNNNNNNNNNNNNNNNNNN	**62**
EGS0294 sgRNA v3.4	GTGAACACGCGGCTCCGGAGGGGGCCGCGGATAAACGGCCGCGGAGTG TACATCTCCGACTGTATACGAGGAAGCGGCAAGGCCCCAGCGGTCATA GCAGTTGAGTAATTGGCTGCTCTTTATCGTGTTGTGTGAAAACGCACA CGAGGAAGATCCCNNNNNNNNNNNNNNNNNN	**63**
EGS0294 sgRNA v3.5	GCGGCTCCGGAGGGGGCCGCGGATAAACGGCCGCGGAGTGTACATCTC CGACTGTATACGAGGAAGCGGCAAGGCCCCAGCGGTCATAGCAGTTGA GTAATTGGCTGCTCTTTATCGTGTTGTGTGAAAACGCACACGAGGAAG ATCCCNNNNNNNNNNNNNNNNNN	**64**
EGS0294 sgRNA v4	GTGAACACTCGGCTCCGGAGGGGGCAGAGGATAAACGGCCGTGGAGTG TACATCTCCCATCAACGCAGTATGTTGATGGACTGTATACGAGGAAGC GGCAAGGCCCCAGCGGTCATAGCAGTTGAGTAATTGGCTGCTCTTTAT CGTGTGAAAACACGAGGAAGATCCCNNNNNNNNNNNNNNNNNNNNN N	**65**
EGS0346 sgRNA v2	GCCUUCGAUAAGAGGGUUUAACGCCCACAUUUUAAGGGGAAACUCGGU GGGAAGGCCACAUUUUGUGGUCACCGUAGUCGAGUAAUCGGCUACCAC UUCACAGAAAUGUGGGGUGAUCNNNNNNNNNNNNNNNNNNN	**66**
EGS0346 sgRNA v2.1	GCCUUCGAUAAGAGGGUUUAACGCCCACAUUCUAAGGGGAAACUCGGU GGGAAGGCCACAUCUUGUGGUCACCGUAGUCGAGUAAUCGGCUACCAC UUCACAGAAAUGUGGGGUGAUCNNNNNNNNNNNNNNNNNNN	**67**
EGS0346 sgRNA v2.2	GCCCUCGAUAAGAGGGUUUAACGCCCACAUUCUAAGGGGAAACCCGGU GGGAAGGCCACAUCUUGUGGUCACCGUAGUCGAGUAAUCGGCUACCAC CUCGCAGAAACGCGGGGUGAUCNNNNNNNNNNNNNNNNNNN	**68**
EGS0346 sgRNA v3	GCCUUCGAUAAGAGGGUUUAACGCCCACGAAAGUGGGAAGGCCACAUC UUGUGGUCACCGUAGUCGAGUAAUCGGCUACCACUUCACAGAAAUGUG GGGUGAUCNNNNNNNNNNNNNNNNNNN	**69**
EGS0346 sgRNA v3.1	GCCCUCGAUAAGAGGGUUUAACGCCCACGAAAGUGGGAAGGCCACAUC UUGUGGUCACCGUAGUCGAGUAAUCGGCUACCACCUCACAGAAAUGUG GGGUGAUCNNNNNNNNNNNNNNNNNNN	**70**
EGS0380 sgRNA v1.1	AUAGGCUUGGGAAAGCCCGAGCCACCCUGUGUAUCUCACUAAGAUGCA CAGCAAUGCCCCUUUAGCUCAGUUUGGUCAGAGCUACGGACUUCUAGA AAUGGGGAUGGCUCNNNNNNNNNNNNNNNNN	**71**
EGS0380 sgRNA v1.2	AUAGGCUUGGGAAAGCCCGAGCCACCCUGUGUAUCUCACUAAGAUGCA CAGCAAUGCCCCUUUAGCUCAGUUUGGUCAGAGCUACGGACCUCUAGA AAUGGGGAUGGCUCNNNNNNNNNNNNNNNNN	**72**
EGS0380 sgRNA v1.3	AUAGGCUUGGGAAAGCCCGAGCCACCCUGUGUAUCUCACUAAGAUGCA CAGCAAUGCCCCUUUAGCUCAGUUUGGUCAGAGCUACGGAUCUCUAGA AAUGGGGAUGGCUCNNNNNNNNNNNNNNNNN	**73**
EGS0291 sgRNA v4	UCCGCCUACCCCGGGGAGGGGAAGGCGUUAAAUGCCUGCGGAGGCUCU GUAAGGAAACUGGAGCUGAUGAAGCUGGCAGACCUGCGCGGAAACCGC GCAGGUCAUCCCGGGGGAGUGAUCCCCCGGCUACCCCGAAAGGGGUAG CCCNNNNNNNNNNNNNNNNNNN	**154**
EGS0291 sgRNA v4.1	UCCGCCUACCCCGGGGAGGGGAAGGCGUUAAAUGCCUGCUUAGGCUCU GUAAGGAAACUGGAGCUUGAAGCUGGCAGACCUGCGCGGAAACCGCGC AGGUCAUCCCGGGGGAGUGAUCCCCCGGCUACCCCGAAAGGGGUAGCC CNNNNNNNNNNNNNNNNNNN	**155**
EGS0291 sgRNA v4.2	UCCGCCUUCCCCGGGGAGGGGAAGGCGUUAAAUGCCUGCGGAGGCUCU GUAAGGAAACUGGAGCUGAUGAAGCUGGCAGACCUGCGCGGAAACCGC GCAGGUCAUCCCGGGGGAGUGAUCCCCCGGCUACCCCGAAAGGGGUAG CCCNNNNNNNNNNNNNNNNNNN	**156**

EGS0291 sgRNA v4.4	UCCGCCUACCCCGGGGAGGGGAAGGCGUUAAAUGCCUGCGGCUCUGUA AGGAAACUGGAGCUGCUGGCAGACCUGCGCGGAAACCGCGCAGGUCAU CCCGGGGGAGUGAUCCCCCGGCUACCCCGAAAGGGGUAGCCCNNNNNN NNNNNNNNNNNNN	**157**
EGS0291 sgRNA v4.7	UCCGCCUACCCCGGGGAGGGGAAGGCGUUAAAUGCCUGCGGAGGCUCU GUAAGGAAACUGGAGCUGAUGAAGCUGGCAGACCUGCGCGGAAACCGC GCAGGUCAUCCCGGGGGAGUGAUCCCCCGGCUACCGAAAGGUAGCCCN NNNNNNNNNNNNNNNNNN	**158**
EGS0291 sgRNA v4.8	UCCGCCUACCCCGGGGAGGGGAAGGCGUUAAAUGCCUGCGGCUCUGUA AGGAAACUGGAGCUGCUGGCAGACCUGCGCGGAAACCGCGCAGGUCAU CCCGGGGGAGUGAUCCCCCGGCUACCGAAAGGUAGCCCNNNNNNNNNN NNNNNNNNNN	**159**
EGS0318 sgRNA v2.1	GAGGCTGTGCGGATGTATGAGGCGCAACTCGCCCCCGGCACGGGGGAG AGTCTAAACGGACGCGGAGTCGCAGGCTCTGATGCTGCCGGTTTGGTG GTGTTGGGGGATGTGGCGTTGGATGCGTCAAGGCCAGCCGCCACTGGC GGCGGGTCACCGTAGGCGAGTAATCGTCTACTCATCCCCGAAAGGGGA TGAGCCCNNNNNNNNNNNNNNNNNNN	**160**

EXAMPLES

Example 1: Identification of novel RNA-Guided Nucleases

[240] The identification of RGNs was performed based on the methods described for example in Russel *et al.* (2020) The CRISPR Journal. V.3, no.6, pp. 462-469. Metagenomic samples were searched for open reading frames (ORFs) and those that were predicted to be genes were selected. A hidden Markov model (HMM) was used to compare the putative genes to profiles of known Cas proteins. The identified Cas genes were grouped into operons, and the operon type was determined based on the presence of known signature genes. For each genome, the CRISPR arrays were identified based on the presence of regularly spaced repeats. The subtype of each CRISPR array was predicted using machine learning. Cas operons were linked to CRISPR arrays if they were less than 10 kilobases apart.

Example 2: gRNA identification

[241] Systems that fit the putative domain and CRISPR orientation for Cas12f5 were confirmed by predicting the structure computationally using neural network based models, similar to methods described, for example in Jumper et al (2021) Nature. V596m pp 583-589. These structural models were compared to each other, and to solved crystal structures to identify possible gRNA structures, and confirm the catalytic residue of the final RuvC domain of the proteins. The crRNA was held to be the last 14 bases of the CRISPR repeat followed by the reprogrammable spacer sequence. Regions in the identified CRIPSR operon were manually searched for potential tracrRNAs by searching for antirepeat sequences capable of hybridizing to approximately bases 1-10 of the putative crRNA sequence within 400 nt of the last catalytic RuvC domain of the potential Cas12f5, including regions within the ORF. Once identified, the putative tracrRNA, when joined together via a flexible linker, such as a GAAA tetra loop, to the putative crRNA, where there was sufficient hybridization between the antirepeat and the repeat regions,

the complex adopted a sgRNA form consisting of four to six stem loop sequences. The second stem loop of the structure was typically the longest and most poorly paired, and was able to be truncated down to various lengths while maintaining catalytic activity, as long as a single transcript was produced that maintained the core structure of the sgRNA. The final stem loop of the sgRNA consisted of the antirepeat from the putative tracrRNA and approximately bases 1-10 of the CRISPR repeat before consisting of 1-4 bases of unpaired "leader sequence" before the reprogrammable spacer sequence.

Example 3: Determination of PAM requirements for each RGN through Bacterial PAM Depletion

[242] PAM requirements for each RGN were determined using a bacterial PAM depletion assay essentially adapted from Kleinstiver et al. (2015) Nature 523:481-485 and Zetsche et al. (2015) Cell 163:759-771. Briefly, two plasmid libraries (C2 and T2) were generated in a pUCl8 backbone (ampR), with each containing a distinct 23bp protospacer (target) sequence flanked by 8 random nucleotides (i.e., the PAM region). The target sequence and flanking PAM region of library T2 and library C2 for each RGN are set forth in Table 16.

[243] Table 16. Target sequences

Library	Sequence	SEQ ID NO
C2	NNNNNNNNTCTCCGACGGATGTCTCCCTTCTTGTGAGCAAGG	74
T2	NNNNNNNNATCCTGTCCCTAGTGGCCCCTCTTGTGAGCAAGG	75

[244] The libraries were separately electroporated into T7 Express *E. coli* (NEB) cells harboring pET28b expression vectors containing an RGN of the invention (codon optimized for E. coli) along with a potential cognate sgRNA containing a spacer sequence corresponding to the protospacer in C2 or T2. Sufficient library plasmid was used in the transformation reaction to obtain >10^8 cfu. Both the RGN and sgRNA in the pET28b backbone were under the control of separate T7 promoters. The transformation reaction was allowed to recover for 1 hr after which it was diluted into LB media containing carbenicillin and kanamycin and grown overnight. The following day the mixture was diluted into self-inducing Overnight Express™ Instant TB Medium (Millipore Sigma) to allow expression of the RGN and sgRNA, and grown for an additional 4h at 37C and then shifted to 30C for an additional 16h after which the cells were spun down and plasmid DNA was isolated with a Mini-prep kit (Qiagen, Germantown, MD). In the presence of the appropriate sgRNA, plasmids containing a PAM that is recognizable by the RGN will be cleaved resulting in their removal from the population. Plasmids containing PAMs that are not recognizable by the RGN, or that are transformed into bacteria not containing an appropriate sgRNA, will survive and replicate. The PAM and protospacer regions of uncleaved plasmids were PCR-amplified and prepared for sequencing following published protocols (l6s-metagenomic library prep guide 15044223B,

Illumina, San Diego, CA). Deep sequencing (55bp paired end reads) was performed on a NextSeq (Illumina). Typically, 1-4M reads were obtained per amplicon. PAM regions were extracted, counted, and normalized to total reads for each sample. PAMs that lead to plasmid cleavage were identified by being underrepresented when compared to controls (i.e., when the library is transformed into *E. coli* containing the RGN but lacking an appropriate sgRNA). To identify the PAM requirements for a novel RGN, an enrichment value was computed for each kmer as the difference between the library size-normalized read counts in the control sample and in the targeting sample. This value was rounded to the nearest integer for positive numbers and set to zero for negative numbers. Enrichment values were then summed across all kmers to yield a position frequency matrix, which was represented visually as a sequence logo using the command line utility weblogo. Those RGNs with consistency among the most enriched kmers—sequence logo information content > 0.2 when including the top 100 enriched kmers—and with qualitatively consistent PAMs across plasmid libraries (T2 and C2) were deemed to have bonafide PAMs. The final PAM for these RGNs were obtained by summing counts across both plasmid libraries, normalizing counts, computing kmer enrichment values, summing across kmers to yield a position frequency matrix, then visually representing the PAM as a sequence logo using the command line utility weblogo.

Example 4: Active Truncations of the Protein

[245] Truncations to the proteins were created to identify just a conserved active domain. Activity was impaired with small truncations at the amino terminus (Figure 2). The carboxyl terminus of the protein was able to be truncated down to having at least 9 amino acids past the final RuvC Catalytic residue without impacting catalytic activity. Truncated sequences are listed in Table 9.

Example 5: Improved gRNA designs

[246] Improvements to the editing ability of novel RGNs can be accomplished by changing the sgRNA scaffolds by altering the tracrRNA and crRNA linkage, removing hairpin mismatches, strengthening hairpins by swapping A:U base pairs for G:C base pairs, altering the starting position of the tracrRNA, removing non-protein contacting regions of the sgRNA to minimize the design. To this end novel sgRNA designs were developed and tested for genomic editing according to Table 15. Bacterial Editing was confirmed with all v1 sgRNA designs to identify the PAM sequence. Improved eukaryotic editing was observed with EGS0293 when truncating the second stem loop (sgRNA v2), but activity was lost or impaired when removing the first stem loop (sgRNA v3) or truncating the third stem loop (sgRNA v5) (Figures 2 and 3).

Example 6: Trans Activated DNA Cleavage by Cas12f5 proteins

[247] gRNAs were synthesized by in vitro transcription of the gRNA cassettes with a GeneArt™ Precision gRNA Synthesis Kit (ThermoFisher). Activity was confirmed by combing the purified RGN along with the gRNA in 20 mM HEPES, pH 6.8 at 37°C, 500 mM NaCl, 1 mM DTT and 5 mM MgCl2 (Reaction Buffer) for 30 min at 37C. The ribonucleo-protein (RNP) complex was then added in excess to linear dsDNA or ssDNA or no target DNA that matched the target sequence of the gRNA along with M13 ssDNA in the reaction buffer and incubated at 37C for a time course. The reaction was then with EDTA, Proteinase K, and RNase A before being run on an agarose gel. Trans activated cleavage of the M13 ssDNA was visually confirmed. Degradation of m13 ssDNA was apparent only in the presence of IVT RNA and the presence of a ssDNA or dsDNA Activator target (Figure 7).

Example 7: Demonstration of gene editing activity on endogenous targets in mammalian cells

[248] The RGN was codon optimized for human expression and cloned into expression cassettes with a Nterm SV40 NLS, and a Cterm FLAGtag and c-myc NLS under control of a CMV promoter for mammalian expression. The sequences are set forth in Table 17.

[249] Table 17. NLS and FLAG tag sequences

Tag	Sequence	SEQ ID NO
NTerm SV40 NLS	CCCAAGAAGAAGAGGAAGGTG	76
CTerm FLAGtag	GACTATAAGGACCACGACGGAGACTACAAGGATCATGATATTGATTACAAAGACGATGACGATAAG	77
CTerm c-myc NLS	CCTGCTGCCAAACGCGTTAAACTAGAC	78

[250] . gRNA expression constructs encoding a single gRNA each under the control of a human RNA polymerase III U6 promoter were produced and introduced into an expression vector containing GFP under control of a CMV promoter. Guides were design to targeted regions of selected genes with the appropriate PAM for the system. The constructs described were introduced into mammalian cells. One day prior to transfection, HEK293T cells (Sigma) were plated in 24-well dishes in Dulbecco's modified Eagle medium (DMEM) plus 10% (vol/vol) fetal bovine serum (Gibco) and 1% Penicillin-Streptomycin (Gibco). The next day when the cells were at 50-60% confluency, 500 ng of a RGN expression plasmid plus 500 ng of a single gRNA expression plasmid were co-transfected using 1.5 uL of Lipofectamine 3000 (Thermo Scientific) per well, following the manufacturer's instructions. After 48 hours of growth, total genomic DNA was harvested using a genomic DNA isolation kit (Machery-Nagel) according to the manufacturer's instructions.

[251] The total genomic DNA was then analyzed to determine the rate of editing in the targeted gene. Oligonucleotides were produced to be used for PCR amplification and subsequent analysis of the amplified genomic target site. All PCR reactions were performed using 10 uL of 2X Master Mix Platinum SuperFi DNA polymerase (Thermo Scientific) in a 20 uL reaction including 0.5 uM of each primer specific for each guide using a program of: 98°C, 1 min; 35 cycles of [98°C, 10 sec; 65°C, 15 sec; 72°C, 30 sec]; 72°C, 5 min; 12°C, forever. Primers for PCR#2 include Nextera Read 1 and Read 2 Transposase Adapter overhang sequences for Illumina sequencing.

[252] Following the PCR amplification, DNA was cleaned using a PCR cleanup kit (Zymo) according to the manufacturer's instructions and eluted in water. Products containing the Illumina overhang sequences underwent library preparation following the Illumina 16S Metagenomic Sequencing Library protocol. Deep sequencing was performed on an Illumina NextSeq platform. Typically, 200,000 of 150 bp paired-end reads (2 x 100,000 reads) are generated per amplicon. The reads were analyzed using CRISPResso (Pinello, et al. 2016 Nature Biotech, 34:695-697) to calculate the rates of editing. Output alignments were hand-curated to confirm insertion and deletion sites as well as identify microhomology sites at the recombination sites. The overall rates of editing for actively edited samples are shown in Table 18.

[253] Table 18. Active Eukaryotic Editing

Entity	RGN	Target Gene	% of reads with INDEL in window	Average INDEL size	gRNA
INDEL5284	EGS0293	CRYGS	1	2	sgRNA v1
INDEL5285	EGS0293	CRYGS	0.6	4.9	sgRNA v2
INDEL5349	EGS0293	RNU6_1	1.9	2.9	sgRNA v2
INDEL6053	EGS0293v3	CCN6	7.4	5.8	sgRNA v2
INDEL6049	EGS0293	CCN6	6.2	9	sgRNA v2
INDEL6051	EGS0293v2	CCN6	5.9	7.5	sgRNA v2
INDEL6048	EGS0293	CCN6	4.3	7.7	sgRNA v1
INDEL6050	EGS0293v2	CCN6	4.1	7.6	sgRNA v1
INDEL6033	EGS0293	OR10D3	3.3	6.6	sgRNA v2
INDEL6052	EGS0293v3	CCN6	3.2	6.4	sgRNA v1
INDEL6081	EGS0293	OR10D3	2	6.5	sgRNA v2
INDEL6035	EGS0293v2	OR10D3	1.8	6.4	sgRNA v2
INDEL6041	EGS0293	RNU6_1	1.6	3.5	sgRNA v2
INDEL6083	EGS0293v2	OR10D3	1.6	3.6	sgRNA v2
INDEL6034	EGS0293v2	OR10D3	0.9	9.3	sgRNA v1

INDEL6025	EGS0293	CRYGS	0.8	3.5	sgRNA v2
INDEL6043	EGS0293v2	RNU6_1	0.8	12.8	sgRNA v2
INDEL6036	EGS0293v3	OR10D3	0.7	2.3	sgRNA v1
INDEL6045	EGS0293v3	RNU6_1	0.7	4.9	sgRNA v2
INDEL6113	EGS0293	ZNF605	0.7	2.8	sgRNA v2
INDEL6024	EGS0293	CRYGS	0.6	1	sgRNA v1
INDEL6044	EGS0293v3	RNU6_1	0.6	2.9	sgRNA v1
INDEL6080	EGS0293	OR10D3	0.6	3.6	sgRNA v1
INDEL6037	EGS0293v3	OR10D3	0.5	4.4	sgRNA v2
INDEL6064	EGS0293	CNTNAP2	0.5	1.5	sgRNA v1
INDEL6085	EGS0293v3	OR10D3	0.5	3.7	sgRNA v2
INDEL6104	EGS0293	PRPF4B	0.5	1.9	sgRNA v1
INDEL6042	EGS0293v2	RNU6_1	0.4	3.8	sgRNA v1
INDEL6066	EGS0293v2	CNTNAP2	0.4	1.9	sgRNA v1
INDEL6084	EGS0293v3	OR10D3	0.4	1.9	sgRNA v1
INDEL6089	EGS0293	OR10D3	0.4	5.2	sgRNA v2
INDEL6067	EGS0293v2	CNTNAP2	0.3	1.9	sgRNA v2
INDEL6093	EGS0293v3	OR10D3	0.3	5	sgRNA v2
INDEL6065	EGS0293	CNTNAP2	0.2	3.7	sgRNA v2
INDEL6069	EGS0293v3	CNTNAP2	0.2	1.8	sgRNA v2
INDEL6109	EGS0293v3	PRPF4B	0.2	1.9	sgRNA v2
INDEL6117	EGS0293v3	ZNF605	0.2	1.1	sgRNA v2
INDEL6227	EGS0293	OR10D3	0.4	3.7	sgRNA v1
INDEL6243	EGS0293	CCN6	3	4.2	sgRNA v1
INDEL6259	EGS0293	CNTNAP2	0.3	1.4	sgRNA v1
INDEL6228	EGS0293	OR10D3	4.5	7.9	sgRNA v2
INDEL6236	EGS0293	RNU6_1	0.7	4.8	sgRNA v2
INDEL6244	EGS0293	CCN6	6.5	9.7	sgRNA v2
INDEL6276	EGS0293	OR10D3	0.7	1	sgRNA v2
INDEL6308	EGS0293	ZNF605	0.4	1.9	sgRNA v2
INDEL6222	EGS0293v2	CRYGS	1	9.8	sgRNA v1
INDEL6246	EGS0293v2	CCN6	3.2	3	sgRNA v1
INDEL6262	EGS0293v2	CNTNAP2	0.7	10.7	sgRNA v1
INDEL6312	EGS0293v2	CRYGS	0.5	6.6	sgRNA v2
INDEL6320	EGS0293v2	OR10D3	1.7	4.5	sgRNA v2
INDEL6336	EGS0293v2	CCN6	6.7	8.4	sgRNA v2
INDEL6368	EGS0293v2	OR10D3	1	7.7	sgRNA v2
INDEL6376	EGS0293v2	OR10D3	0.7	2.8	sgRNA v2
INDEL6322	EGS0293v4	OR10D3	0.4	1	sgRNA v1

INDEL6330	EGS0293v4	RNU6_1	0.7	2	sgRNA v1
INDEL6338	EGS0293v4	CCN6	0.9	2.9	sgRNA v1
INDEL6323	EGS0293v4	OR10D3	1.8	12.1	sgRNA v2
INDEL6339	EGS0293v4	CCN6	6.6	8.8	sgRNA v2
INDEL6371	EGS0293v4	OR10D3	1.5	8.6	sgRNA v2
INDEL6317	EGS0293v5	CRYGS	1.3	5.6	sgRNA v1
INDEL6325	EGS0293v5	OR10D3	0.6	1	sgRNA v1
INDEL6373	EGS0293v5	OR10D3	0.3	4.7	sgRNA v1
INDEL6318	EGS0293v5	CRYGS	0.6	4.8	sgRNA v2
INDEL6326	EGS0293v5	OR10D3	3.1	9.2	sgRNA v2
INDEL6334	EGS0293v5	RNU6_1	1.4	5.6	sgRNA v2
INDEL6342	EGS0293v5	CCN6	9.3	6.6	sgRNA v2
INDEL6350	EGS0293v5	CCN6	0.4	5.7	sgRNA v2
INDEL6358	EGS0293v5	CNTNAP2	0.3	1.9	sgRNA v2
INDEL6374	EGS0293v5	OR10D3	2.2	6	sgRNA v2
INDEL6530	EGS0293	CRYGS	1.3	4.9	sgRNA v2
INDEL6532	EGS0293v2	CRYGS	0.9	2	sgRNA v2
INDEL6534	EGS0293v4	CRYGS	2.6	3.4	sgRNA v2
INDEL6536	EGS0293v5	CRYGS	2.4	3.3	sgRNA v2
INDEL6538	EGS0293	OR10D3	2.1	3.9	sgRNA v2
INDEL6539	EGS0293	OR10D3	0.4	3.5	sgRNA v5
INDEL6540	EGS0293v2	OR10D3	3.2	5.2	sgRNA v2
INDEL6542	EGS0293v4	OR10D3	2	4.1	sgRNA v2
INDEL6544	EGS0293v5	OR10D3	0.4	3.5	sgRNA v2
INDEL6545	EGS0293v5	OR10D3	0.9	13.3	sgRNA v5
INDEL6546	EGS0293	RNU6_1	1.9	7.5	sgRNA v2
INDEL6547	EGS0293	RNU6_1	0.7	2	sgRNA v5
INDEL6548	EGS0293v2	RNU6_1	1.9	5.7	sgRNA v2
INDEL6554	EGS0293	CCN6	6.9	8.1	sgRNA v2
INDEL6555	EGS0293	CCN6	5.9	4.4	sgRNA v5
INDEL6556	EGS0293v2	CCN6	6.9	9.8	sgRNA v2
INDEL6557	EGS0293v2	CCN6	3	6.2	sgRNA v5
INDEL6558	EGS0293v4	CCN6	10.2	8.2	sgRNA v2
INDEL6560	EGS0293v5	CCN6	11.4	8	sgRNA v2
INDEL6561	EGS0293v5	CCN6	5.6	5	sgRNA v5
INDEL6568	EGS0293v5	CCN6	0.6	3.9	sgRNA v2
INDEL6574	EGS0293v4	CNTNAP2	0.3	2.8	sgRNA v2
INDEL6587	EGS0293	OR10D3	0.5	1	sgRNA v5
INDEL6588	EGS0293v2	OR10D3	3.7	4.8	sgRNA v2

INDEL6590	EGS0293v4	OR10D3	5.3	3.4	sgRNA v2
INDEL6591	EGS0293v4	OR10D3	0.4	3.6	sgRNA v5
INDEL6592	EGS0293v5	OR10D3	4.4	4.3	sgRNA v2
INDEL6593	EGS0293v5	OR10D3	1	3.8	sgRNA v5
INDEL6594	EGS0293	OR10D3	1.8	2.9	sgRNA v2
INDEL6596	EGS0293v2	OR10D3	1.1	3.1	sgRNA v2
INDEL6610	EGS0293	PRPF4B	1.1	3.2	sgRNA v2
INDEL7940	EGS0293v2	RNF2	5.5	7.4	sgRNA v2
INDEL7866	EGS0293v2	CCN6	2.1	5.8	sgRNA v2
INDEL7941	EGS0293v2	RNU6_1	2.1	4.1	sgRNA v2
INDEL7938	EGS0293v2	LCA5	1.4	2.8	sgRNA v2
INDEL7863	EGS0293v2	CXCR4	0.3	3.7	sgRNA v2
INDEL7870	EGS0293v2	CNTNAP2	0.3	6.2	sgRNA v2
INDEL7847	EGS0293v2	CRYGS	0.2	4.4	sgRNA v2
INDEL7948	EGS0293v2	RNF2	0.2	3.8	sgRNA v2
INDEL7956	EGS0293v2	RNF2	0.2	2.4	sgRNA v2
INDEL7686	EGS0293v2	CRYGS	0.7	4.2	sgRNA v2
INDEL7689	EGS0293v2	CRYGS	0.2	3.9	sgRNA v2.5
INDEL7694	EGS0293v2	OR10D3	2.1	3.3	sgRNA v2
INDEL7710	EGS0293v2	CCN6	8.2	8.2	sgRNA v2
INDEL7713	EGS0293v2	CCN6	0.3	2.8	sgRNA v2.5
INDEL7726	EGS0293v2	CNTNAP2	0.5	5.3	sgRNA v2
INDEL7736	EGS0293v2	CRYGS	0.3	3.9	sgRNA v2.4
INDEL7742	EGS0293v2	OR10D3	1.3	3	sgRNA v2
INDEL7743	EGS0293v2	OR10D3	0.1	10.5	sgRNA v2.2
INDEL7744	EGS0293v2	OR10D3	0.3	5.4	sgRNA v2.4
INDEL7750	EGS0293v2	OR10D3	0.5	3.4	sgRNA v2
INDEL7752	EGS0293v2	RNU6_1	0.1	5.3	sgRNA v2.4
INDEL7759	EGS0293v2	CCN6	0.1	2.6	sgRNA v2.2
INDEL7760	EGS0293v2	CCN6	5.7	4.7	sgRNA v2.4
INDEL7766	EGS0293v2	PRPF4B	0.1	2.6	sgRNA v2
INDEL7768	EGS0293v2	OR10D3	0.2	13	sgRNA v2.4
INDEL8426	EGS0346	OR10D3	0.1	1	sgRNA v1
INDEL8427	EGS0346	OR10D3	0.1	1	sgRNA v2
INDEL8428	EGS0346	OR10D3	0.1	5.2	sgRNA v2.1
INDEL8522	EGS0346v2	OR10D3	0.4	10.5	sgRNA v1
INDEL8526	EGS0346v2	OR10D3	0.1	1.2	sgRNA v3
INDEL9105	EGS0380	OR10D3	0.1	1	sgRNA v1.1
INDEL9106	EGS0380	OR10D3	0.1	6.8	sgRNA v1.2

INDEL11816	EGS0288	AASDHPPT	2.84	3.7	sgRNA v3
INDEL11817	EGS0288v2	AASDHPPT	4.88	10.9	sgRNA v3
INDEL11824	EGS0288	CRYGS	4.33	5.4	sgRNA v3
INDEL11825	EGS0288v2	CRYGS	5.46	4.3	sgRNA v3
INDEL11832	EGS0288	MTA3	0.8	4.4	sgRNA v3
INDEL11833	EGS0288v2	MTA3	0.4	6.2	sgRNA v3
INDEL11835	EGS0291	CNTNAP2	0.72	3.2	sgRNA v4
INDEL11837	EGS0291v2	CNTNAP2	2.22	6	sgRNA v4
INDEL11845	EGS0291v2	CRYGS	0.22	4.7	sgRNA v4
INDEL11848	EGS0288	RNU6_1	1.36	1.5	sgRNA v3
INDEL11849	EGS0288v2	RNU6_1	3.04	4.6	sgRNA v3
INDEL11851	EGS0291	EMX1	2.01	4.8	sgRNA v4
INDEL11853	EGS0291v2	EMX1	0.83	2.6	sgRNA v4
INDEL11856	EGS0288	ZNF605	0.26	8.6	sgRNA v3
INDEL11857	EGS0288v2	ZNF605	0.68	7.5	sgRNA v3
INDEL11864	EGS0288	LCA5	5.13	5.5	sgRNA v3
INDEL11865	EGS0288v2	LCA5	5.45	5.2	sgRNA v3
INDEL11866	EGS0291	LCA5	1.48	2	sgRNA v3
INDEL11867	EGS0291	LCA5	7.58	4.7	sgRNA v4
INDEL11868	EGS0291v2	LCA5	1.15	3.4	sgRNA v3
INDEL11869	EGS0291v2	LCA5	6.33	5	sgRNA v4
INDEL11875	EGS0291	MTA3	0.78	7.8	sgRNA v4
INDEL11877	EGS0291v2	MTA3	0.61	8	sgRNA v4
INDEL11881	EGS0288v2	NPY	2.03	8.1	sgRNA v3
INDEL11889	EGS0288v2	OR10D3	2.04	9.4	sgRNA v3
INDEL11890	EGS0291	OR10D3	3.08	6.4	sgRNA v3
INDEL11891	EGS0291	OR10D3	8.16	7.2	sgRNA v4
INDEL11892	EGS0291v2	OR10D3	1.99	3.8	sgRNA v3
INDEL11893	EGS0291v2	OR10D3	8.11	6	sgRNA v4
INDEL11897	EGS0288v2	PRPF4B	5.27	5.2	sgRNA v3
INDEL11906	EGS0291	ZNF605	0.68	5.1	sgRNA v3
INDEL11909	EGS0291v2	ZNF605	2.14	7	sgRNA v4
INDEL11913	EGS0318	CRYGS	8.01	6.2	sgRNA v2
INDEL11914	EGS0318	CRYGS	9.08	3.3	sgRNA v2.1
INDEL11921	EGS0318	OR10D3	8.09	6.9	sgRNA v2
INDEL11922	EGS0318	OR10D3	12.9	6.5	sgRNA v2.1
INDEL11929	EGS0318	RNU6_1	0.6	7.1	sgRNA v2
INDEL11938	EGS0318	CNTNAP2	1.52	2.5	sgRNA v2.1
INDEL11945	EGS0318	OR10D3	11.28	3.2	sgRNA v2
INDEL11946	EGS0318	OR10D3	12.8	5.3	sgRNA v2.1
INDEL11953	EGS0318	CNTNAP2	4.46	8	sgRNA v2

INDEL11960	EGS0295	PRPF4B	0.54	1	sgRNA v3
INDEL11961	EGS0318	CXCR4	6.05	4.6	sgRNA v2
INDEL11962	EGS0318	CXCR4	5.62	4.4	sgRNA v2.1
INDEL11969	EGS0318	CXCR4	8.26	5.8	sgRNA v2
INDEL11970	EGS0318	CXCR4	5.41	3	sgRNA v2.1
INDEL11976	EGS0295	PTPN3	0.4	3.9	sgRNA v3
INDEL11977	EGS0318	EMX1	1.01	2.3	sgRNA v2
INDEL11978	EGS0318	EMX1	3.53	4.8	sgRNA v2.1
INDEL11986	EGS0318	GTF2A1	2.14	3.3	sgRNA v2.1
INDEL11993	EGS0318	LCA5	3.45	5.3	sgRNA v2
INDEL11994	EGS0318	LCA5	6.13	6.6	sgRNA v2.1
INDEL12001	EGS0318	RNF2	8.68	8	sgRNA v2
INDEL12002	EGS0318	RNF2	6.15	7.2	sgRNA v2.1
INDEL12585	EGS0291	AASDHPPT	4.12	12.8	sgRNA v4
INDEL12589	EGS0291v2	AASDHPPT	5.55	6.1	sgRNA v4
INDEL12591	EGS0291v2	NPY	1.93	3.9	sgRNA v4
INDEL12598	EGS0291v2	FAM160B1	6.12	5.2	sgRNA v4
INDEL12600	EGS0291	CNTNAP2	1.77	4.8	sgRNA v4
INDEL12601	EGS0291	B2M	1.13	7.1	sgRNA v4
INDEL12602	EGS0291	GAPDH	0.85	2.5	sgRNA v4
INDEL12603	EGS0291	OR10D3	0.14	1.9	sgRNA v4
INDEL12604	EGS0291v2	CNTNAP2	1.53	4.8	sgRNA v4
INDEL12605	EGS0291v2	B2M	0.79	10.6	sgRNA v4
INDEL12606	EGS0291v2	GAPDH	0.6	2	sgRNA v4
INDEL12608	EGS0291	CRYGS	0.25	1	sgRNA v4
INDEL12610	EGS0291	GTF2A1	2.04	2.4	sgRNA v4
INDEL12612	EGS0291v2	CRYGS	0.92	4.7	sgRNA v4
INDEL12614	EGS0291v2	GTF2A1	0.94	2.6	sgRNA v4
INDEL12617	EGS0291	CCN6	1.4	1.6	sgRNA v4
INDEL12619	EGS0291	PRPF4B	1.85	6.7	sgRNA v4
INDEL12620	EGS0291v2	EMX1	0.78	6.7	sgRNA v4
INDEL12621	EGS0291v2	CCN6	3.28	5.9	sgRNA v4
INDEL12622	EGS0291v2	HEATR6	5.54	8	sgRNA v4
INDEL12625	EGS0291	CLPB	3.77	12.2	sgRNA v4
INDEL12629	EGS0291v2	CLPB	2.68	8	sgRNA v4
INDEL12631	EGS0291v2	PRPF4B	3.27	6	sgRNA v4
INDEL12636	EGS0291v2	LCA5	5.79	5.3	sgRNA v4
INDEL12637	EGS0291v2	CNTNAP2	0.36	3.7	sgRNA v4
INDEL12638	EGS0291v2	LCA5	0.46	4.7	sgRNA v4
INDEL12639	EGS0291v2	PTPN3	1.49	4.9	sgRNA v4
INDEL12640	EGS0291	MTA3	2.89	10.4	sgRNA v4

INDEL12643	EGS0291	RNF2	0.79	7.7	sgRNA v4
INDEL12644	EGS0291v2	MTA3	0.78	6.7	sgRNA v4
INDEL12645	EGS0291v2	CNTNAP2	0.23	6.1	sgRNA v4
INDEL12646	EGS0291v2	LCA5	0.93	3.2	sgRNA v4
INDEL12647	EGS0291v2	RNF2	2.28	5.7	sgRNA v4
INDEL12654	EGS0291v2	LINC01589	1.81	1	sgRNA v4
INDEL12656	EGS0291	OR10D3	4.44	8.1	sgRNA v4
INDEL12666	EGS0291	MTA3	0.7	2.1	sgRNA v4
INDEL12668	EGS0291v2	PRPF4B	0.25	1.9	sgRNA v4
INDEL12669	EGS0291v2	EMX1	1.62	14.3	sgRNA v4
INDEL12670	EGS0291v2	MTA3	1	2.3	sgRNA v4
INDEL12672	EGS0291	ZNF605	0.9	4.1	sgRNA v4
INDEL12673	EGS0291	EMX1	0.36	2.1	sgRNA v4
INDEL12674	EGS0291	MTA3	2.77	7.1	sgRNA v4
INDEL12675	EGS0291	ZNF605	0.31	5.7	sgRNA v4
INDEL12676	EGS0291v2	ZNF605	0.88	3	sgRNA v4
INDEL12678	EGS0291v2	MTA3	3.22	7.5	sgRNA v4
INDEL12679	EGS0291v2	ZNF605	0.38	7.2	sgRNA v4
INDEL12690	EGS0336	CRYGS	0.62	4.9	sgRNA v3
INDEL12699	EGS0337	CRYGS	0.16	1.5	sgRNA v2
INDEL12721	EGS0334v2	EMX1	0.24	11.3	sgRNA v3
INDEL12748	EGS0337v2	OR10D3	0.24	1	sgRNA v2
INDEL12749	EGS0338	EMX1	0.43	6.4	sgRNA v2
INDEL12750	EGS0338v2	EMX1	0.67	4.3	sgRNA v2
INDEL12757	EGS0338	EMX1	5.85	8	sgRNA v2
INDEL12765	EGS0338	GAPDH	0.69	8.9	sgRNA v2
INDEL12766	EGS0338v2	GAPDH	0.5	12.7	sgRNA v2
INDEL14696	EGS0341	AASDHPPT	1.38	6.3	sgRNA v3
INDEL14697	EGS0341v2	AASDHPPT	1.99	5.5	sgRNA v3
INDEL14699	EGS0343v2	AASDHPPT	2.99	8.1	sgRNA v3
INDEL14704	EGS0341	CRYGS	1.52	4.1	sgRNA v3
INDEL14705	EGS0341v2	CRYGS	1.96	3.9	sgRNA v3
INDEL14706	EGS0343	CRYGS	3.04	5.2	sgRNA v3
INDEL14707	EGS0343v2	CRYGS	2.51	4.4	sgRNA v3
INDEL14712	EGS0341	LCA5	20.08	8.2	sgRNA v3
INDEL14713	EGS0341v2	LCA5	18.41	8.5	sgRNA v3
INDEL14720	EGS0341	MTA3	0.42	7.4	sgRNA v3
INDEL14721	EGS0341v2	MTA3	0.5	7.9	sgRNA v3
INDEL14723	EGS0343v2	LCA5	7.96	7	sgRNA v3
INDEL14728	EGS0341	MTA3	1.69	5.3	sgRNA v3
INDEL14729	EGS0341v2	MTA3	1.22	4.4	sgRNA v3

INDEL14730	EGS0343	MTA3	1.73	8.6	sgRNA v3
INDEL14731	EGS0343v2	MTA3	1.48	8.8	sgRNA v3
INDEL14738	EGS0343	MTA3	0.12	1.8	sgRNA v3
INDEL14744	EGS0341	NPY	9.23	9.1	sgRNA v3
INDEL14745	EGS0341v2	NPY	7.22	9	sgRNA v3
INDEL14752	EGS0341	OR10D3	16.28	9.2	sgRNA v3
INDEL14753	EGS0341v2	OR10D3	11.58	9.5	sgRNA v3
INDEL14754	EGS0343	NPY	6.46	9.9	sgRNA v3
INDEL14755	EGS0343v2	NPY	5.35	9.2	sgRNA v3
INDEL14760	EGS0341	PRPF4B	10.81	6.4	sgRNA v3
INDEL14761	EGS0341v2	PRPF4B	11	6.6	sgRNA v3
INDEL14762	EGS0343	OR10D3	9.78	9.1	sgRNA v3
INDEL14763	EGS0343v2	OR10D3	6.81	8.8	sgRNA v3
INDEL14768	EGS0341	RNU6_1	1.8	6	sgRNA v3
INDEL14769	EGS0341v2	RNU6_1	2.05	4.7	sgRNA v3
INDEL14776	EGS0341	ZNF605	2.63	8.2	sgRNA v3
INDEL14777	EGS0341v2	ZNF605	1.66	7.4	sgRNA v3
INDEL14778	EGS0343	RNU6_1	2.96	6.8	sgRNA v3
INDEL14779	EGS0343v2	RNU6_1	1.92	7.5	sgRNA v3
INDEL14780	EGS0344	GAPDH	0.76	5.3	sgRNA v3
INDEL14784	EGS0341	ZNF605	13.42	8.3	sgRNA v3
INDEL14785	EGS0341v2	ZNF605	15.9	7.6	sgRNA v3
INDEL14786	EGS0343	ZNF605	0.91	6	sgRNA v3
INDEL14787	EGS0343v2	ZNF605	0.22	4.7	sgRNA v3
INDEL14792	EGS0291v2	AASDHPPT	3.39	7.4	sgRNA v4
INDEL14793	EGS0291v2	AASDHPPT	2.32	7.5	sgRNA v4.1
INDEL14794	EGS0291v2	AASDHPPT	2.83	7.4	sgRNA v4.2
INDEL14795	EGS0291v2	AASDHPPT	1.26	6	sgRNA v4.4
INDEL14796	EGS0291v2	AASDHPPT	4.94	8.8	sgRNA v4.7
INDEL14797	EGS0291v2	AASDHPPT	0.77	4.8	sgRNA v4.8
INDEL14800	EGS0291v2	CCN6	2.35	4.8	sgRNA v4
INDEL14801	EGS0291v2	CCN6	2.57	5.2	sgRNA v4.1
INDEL14802	EGS0291v2	CCN6	3.37	5	sgRNA v4.2
INDEL14803	EGS0291v2	CCN6	3.15	4.8	sgRNA v4.4
INDEL14805	EGS0291v2	CCN6	1.03	4.8	sgRNA v4.8
INDEL14808	EGS0291v2	CLPB	2.21	7.6	sgRNA v4
INDEL14809	EGS0291v2	CLPB	1.29	7.6	sgRNA v4.1
INDEL14810	EGS0291v2	CLPB	2.98	7.7	sgRNA v4.2
INDEL14811	EGS0291v2	CLPB	1.38	6.2	sgRNA v4.4
INDEL14812	EGS0291v2	CLPB	4.37	7.3	sgRNA v4.7
INDEL14813	EGS0291v2	CLPB	0.46	7.1	sgRNA v4.8

INDEL14816	EGS0291v2	FAM160B1	3.55	6.6	sgRNA v4
INDEL14817	EGS0291v2	FAM160B1	4.54	7.6	sgRNA v4.1
INDEL14818	EGS0291v2	FAM160B1	3.77	7.4	sgRNA v4.2
INDEL14819	EGS0291v2	FAM160B1	2.71	6.2	sgRNA v4.4
INDEL14820	EGS0291v2	FAM160B1	5.87	6.6	sgRNA v4.7
INDEL14821	EGS0291v2	FAM160B1	1.42	8.4	sgRNA v4.8
INDEL14824	EGS0291v2	HEATR6	4.79	7	sgRNA v4
INDEL14825	EGS0291v2	HEATR6	2.02	6.9	sgRNA v4.1
INDEL14826	EGS0291v2	HEATR6	4.36	6.7	sgRNA v4.2
INDEL14827	EGS0291v2	HEATR6	2.2	7.8	sgRNA v4.4
INDEL14828	EGS0291v2	HEATR6	6.83	6.9	sgRNA v4.7
INDEL14829	EGS0291v2	HEATR6	1.81	8.4	sgRNA v4.8
INDEL14832	EGS0291v2	LCA5	4.2	4.3	sgRNA v4
INDEL14833	EGS0291v2	LCA5	3.55	5	sgRNA v4.1
INDEL14834	EGS0291v2	LCA5	3	5.1	sgRNA v4.2
INDEL14835	EGS0291v2	LCA5	2.4	5.3	sgRNA v4.4
INDEL14836	EGS0291v2	LCA5	3.38	4.2	sgRNA v4.7
INDEL14837	EGS0291v2	LCA5	0.1	6	sgRNA v4.8
INDEL14840	EGS0291v2	LINC01589	0.2	4.1	sgRNA v4
INDEL14841	EGS0291v2	LINC01589	0.88	5	sgRNA v4.1
INDEL14842	EGS0291v2	LINC01589	0.64	3.8	sgRNA v4.2
INDEL14843	EGS0291v2	LINC01589	0.43	4.7	sgRNA v4.4
INDEL14844	EGS0291v2	LINC01589	0.49	3.7	sgRNA v4.7
INDEL14845	EGS0291v2	LINC01589	0.22	3.7	sgRNA v4.8
INDEL14848	EGS0291v2	MTA3	3.76	7.6	sgRNA v4
INDEL14849	EGS0291v2	MTA3	1.62	6.7	sgRNA v4.1
INDEL14850	EGS0291v2	MTA3	3.59	7.1	sgRNA v4.2
INDEL14851	EGS0291v2	MTA3	1.13	5.4	sgRNA v4.4
INDEL14852	EGS0291v2	MTA3	4.72	6.8	sgRNA v4.7
INDEL14853	EGS0291v2	MTA3	0.28	6.6	sgRNA v4.8
INDEL14856	EGS0291v2	NPY	0.92	3.3	sgRNA v4
INDEL14857	EGS0291v2	NPY	1.24	6.6	sgRNA v4.1
INDEL14858	EGS0291v2	NPY	0.93	5	sgRNA v4.2
INDEL14859	EGS0291v2	NPY	1.71	5.9	sgRNA v4.4
INDEL14860	EGS0291v2	NPY	1.01	4.7	sgRNA v4.7
INDEL14861	EGS0291v2	NPY	0.42	5.1	sgRNA v4.8
INDEL14864	EGS0291v2	OR10D3	8.06	6.6	sgRNA v4
INDEL14865	EGS0291v2	OR10D3	3.69	6.5	sgRNA v4.1
INDEL14866	EGS0291v2	OR10D3	6.43	6.8	sgRNA v4.2
INDEL14867	EGS0291v2	OR10D3	5.69	7	sgRNA v4.4
INDEL14868	EGS0291v2	OR10D3	8.84	7	sgRNA v4.7

INDEL14869	EGS0291v2	OR10D3	3.17	5.5	sgRNA v4.8
INDEL14872	EGS0291v2	PRPF4B	5.22	8.1	sgRNA v4
INDEL14873	EGS0291v2	PRPF4B	1.88	6	sgRNA v4.1
INDEL14874	EGS0291v2	PRPF4B	4.46	7.6	sgRNA v4.2
INDEL14875	EGS0291v2	PRPF4B	1.8	7.7	sgRNA v4.4
INDEL14876	EGS0291v2	PRPF4B	7.09	8	sgRNA v4.7
INDEL14877	EGS0291v2	PRPF4B	0.31	5.5	sgRNA v4.8
INDEL14880	EGS0291v2	RNF2	2.56	6.3	sgRNA v4
INDEL14881	EGS0291v2	RNF2	3.3	5	sgRNA v4.1
INDEL14882	EGS0291v2	RNF2	4.35	6.8	sgRNA v4.2
INDEL14883	EGS0291v2	RNF2	1.22	5.3	sgRNA v4.4

Example 8: Demonstration of base editing activity on endogenous targets in mammalian cells

[254] The coding sequence of the identified RGN is codon-optimized for expression in mammalian cells and introduced into the expression cassette, which produces a fusion protein that includes a NLS tag at its N-terminal end operably linked to a codon optimized known eukaryotic cytosine deaminase sequence (APOBEC3A) at its C-terminal end. The deaminase is operably linked to a flexible amino acid linker at their C-terminal end, and the amino acid linker is operably linked to the RNA guided nuclease at its C-terminal end, that has been mutated to have an inactive RuvC domain (dEGS0293v2_D289A_D486A) (That is, it has been mutated into RGN that is catalytically dead). The RNA-guided DNA binding polypeptide is operably linked to a flexible amino acid linker at their C-terminal end, and the amino acid linker is operably linked to an uracil protecting peptide (developed in house). The uracil protecting peptide is operably linked to a flexible amino acid linker at their C-terminal end, and the amino acid linker is operably linked to a second NLS at its C-terminal end. Each of these expression cassettes is introduced into a vector capable of driving the expression of the fusion protein in mammalian cells. A vector capable of expressing gRNA to target the deaminase-RGN-UPP fusion protein to the determined genomic location was also produced. These guide RNAs can guide the deaminase-RGN-UPP fusion protein to the target genome sequence for base editing.

[255] Using liposome transfection, vectors capable of expressing the deaminase-RGN-UPP fusion protein and guide RNAs were transfected into HEK293T cells. For liposome transfection, the day before transfection, the cells were distributed in a 24-well plate of growth medium (DMEM + 10% fetal bovine serum + 1% penicillin/streptomycin) at 1.3×10^{5} cells/well. According to the manufacturer's instructions, use Lipofectamine® 3000 reagent (Thermo Fisher Scientific) to transfect 500 ng deaminase-RGN fusion expression vector and 500 ng guide RNA expression vector. 48-72 hours after liposome transfection, genomic DNA is harvested from the transfected cells, and the DNA is sequenced and analyzed for the

presence of targeted cytosine base editing mutations using CRISPResso2 (Clement K, et al Nat Biotechnol. 2019 Mar; 37(3):224-226. doi: 10.1038/s41587-019-0032-3. PubMed PMID: 30809026).

[256] Tables 19, 20 and 21 show the editing rate of cytidine bases for each deaminase-RGN-UPP fusion protein and the rate for targeted cytosine deamination for the deaminase-RGN-UPP targeted to the same region as the catalytically dead RGN-UPP. Active cytosine base editing was defined as a greater than 5x increase of C>D SNP base editing along the targeted window of the deaminase-RGN-UPP under investigation, and >60% of specific C>T SNP base editing at highly mutated cytosines. With a catalytically dead nuclease domain, the RGN will not generate a detectable INDEL formation by itself.. When fused with an active cytosine deaminase that acts on the opposite strand a cytosine will be turned into a uracil. The uracil is rapidly removed from the DNA leaving an abasic site, and eventually a gap, on the strand opposite the strand bound by the gRNA. This can result in a double stranded break which is repaired through non-homologous end joining (NHEJ) and detectable INDEL formation, however, with the presence of an active UPP, the converted uracil is protected from removal and the abasic site is never removed and NHEJ does not occur. This also leads to predominantly C>T conversions because the uracil created by the deaminase is protected and not removed before the strand is replicated, it will be read as a thymine and an adenosine will be inserted across from it. Then when the uracil is eventually removed, it will be replaced by thymine during the excision repair process fixing the mutation at C>T.

[257] Table 19. INDEL and Max Cytosine Base editing results for constructs.

Effector	**Max % INDEL Target 1**	**Max % Base Editing Target 1**	**Max % INDEL Target 2**	**Max % Base Editing Target 2**
A3A.dEGS0293v2.UPP12	0	2.7	0	9.1
dEGS0293v2	0	0	0	0

[258] Table 20. Target 1 Specific cytosine editing results

		A3A.dEGS0293v2.UPP12				**dEGS0293v2**			
	Base (down stream from PAM)	C>T (total %)	C>T (% of SNPs)	C>G (total %)	C>G (% of SNPs)	C>T (total %)	C>T (% of SNPs)	C>G (total %)	C>G (% of SNPs)
C	3	0.291698	96.47355	0.003046	1.007557	0.005041	26.92308	0.006001	32.05128
C	4	1.776085	99.19183	0.005331	0.297746	0.006001	25	0.004081	17
C	9	2.713633	99.33092	0.006855	0.250906	0.007201	33.33333	0.007441	34.44444
C	11	2.53313	99.46172	0.006855	0.269139	0.007201	29.12621	0.004081	16.50485
C	12	1.92003	99.21291	0.011424	0.590319	0.005521	24.73118	0.012002	53.76344
C	17	0.833206	68.28964	0.222391	18.22722	0.044887	10.26908	0.223716	51.18067

[259] Table 21. Target 2 Specific cytosine editing results

		A3A.dEGS0293v2.UPP12				dEGS0293v2			
	Base (downstream from PAM)	C>T (total %)	C>T (% of SNPs)	C>G (total %)	C>G (% of SNPs)	C>T (total %)	C>T (% of SNPs)	C>G (total %)	C>G (% of SNPs)
C	8	9.072732	99.59171	0.016347	0.17944	0.007903	14.16667	0.032078	57.5
C	11	8.431651	98.67746	0.0552	0.64602	0.013017	9.210526	0.087866	62.17105
C	14	6.545605	87.06983	0.951433	12.65599	0.007903	7.589286	0.075314	72.32143
C	15	7.495143	98.22715	0.100687	1.319548	0.010228	5.378973	0.166434	87.53056
C	17	3.15494	93.82133	0.147358	4.382133	0.013947	11.58301	0.033008	27.41313
C	18	3.29448	98.6381	0.026297	0.787346	0.008368	6.271777	0.076709	57.49129
C	20	0.47856	94.12861	0.015162	2.982293	0.010228	17.1875	0.027429	46.09375

Example 9: Engineered variants with increased activity

To test if the performance could be further improved, the predicted protein structure of EGS0293v2 was manually searched for amino acids with potential to interact with either the sgRNA or the target DNA. The identified residues then were mutated to a positively charged arginine or lysine. Multiple targets for each mutation were then tested in HEK293T cells and NHEJ activity for each target was compared to the wildtype enzyme. Those that showed neutral or improved performance were combined with additional mutations identified with predicted protein structure and assayed again (Figure 5 and Table 22).

[260] Table 22. Active Eukaryotic Editing of Mutated EGS0293v2

Entity	RGN	Target Gene	% of reads with INDEL in window	Average INDEL size	gRNA
INDEL11304	EGS0293v2	EMX1	0.05	1.1	sgRNA v2
INDEL11328	EGS0293v2	RNF2	4.27	9	sgRNA v2
INDEL11312	EGS0293v2	GTF2A1	0.02	1.1	sgRNA v2
INDEL11256	EGS0293v2	RNU6_1	0.21	4.8	sgRNA v2
INDEL11296	EGS0293v2	CXCR4	0.24	2.7	sgRNA v2
INDEL11240	EGS0293v2	CRYGS	0.72	3.6	sgRNA v2
INDEL11280	EGS0293v2	CNTNAP2	0.06	1	sgRNA v2
INDEL11264	EGS0293v2	CNTNAP2	0.35	4	sgRNA v2
INDEL11288	EGS0293v2	CXCR4	0.82	2	sgRNA v2
INDEL11320	EGS0293v2	LCA5	6.01	6.4	sgRNA v2
INDEL9896	EGS0293v2	CRYGS	0		sgRNA v2
INDEL9904	EGS0293v2	OR10D3	2.7	5.8	sgRNA v2

INDEL9912	EGS0293v2	RNU6_1	0.6	4.6	sgRNA v2
INDEL9920	EGS0293v2	CNTNAP2	1.1	4.8	sgRNA v2
INDEL9936	EGS0293v2	CNTNAP2	2.6	7.8	sgRNA v2
INDEL9944	EGS0293v2	CXCR4	1.1	2	sgRNA v2
INDEL9952	EGS0293v2	CXCR4	0		sgRNA v2
INDEL9960	EGS0293v2	EMX1	0.1	1.2	sgRNA v2
INDEL9968	EGS0293v2	GTF2A1	0.3	4	sgRNA v2
INDEL9976	EGS0293v2	LCA5	8.2	5.5	sgRNA v2
INDEL9984	EGS0293v2	RNF2	3.9	7.4	sgRNA v2
INDEL11409	EGS0293v2_A190R_Q343K_N347K	GTF2A1	0.86	2.7	sgRNA v2
INDEL11401	EGS0293v2_A190R_Q343K_N347K	EMX1	0.08	1.1	sgRNA v2
INDEL11361	EGS0293v2_A190R_Q343K_N347K	CNTNAP2	0.84	3.9	sgRNA v2
INDEL11385	EGS0293v2_A190R_Q343K_N347K	CXCR4	2.89	2.4	sgRNA v2
INDEL11345	EGS0293v2_A190R_Q343K_N347K	OR10D3	4.95	7.5	sgRNA v2
INDEL11377	EGS0293v2_A190R_Q343K_N347K	CNTNAP2	2.69	7.6	sgRNA v2
INDEL11417	EGS0293v2_A190R_Q343K_N347K	LCA5	10.12	6.7	sgRNA v2
INDEL11353	EGS0293v2_A190R_Q343K_N347K	RNU6_1	0.62	4.2	sgRNA v2
INDEL11369	EGS0293v2_A190R_Q343K_N347K	OR10D3	6.83	7.4	sgRNA v2
INDEL11337	EGS0293v2_A190R_Q343K_N347K	CRYGS	0.72	2.6	sgRNA v2
INDEL11425	EGS0293v2_A190R_Q343K_N347K	RNF2	3.45	7.3	sgRNA v2
INDEL9996	EGS0293v2_D389K	CRYGS	0		sgRNA v2
INDEL10004	EGS0293v2_D389K	OR10D3	1.5	3.2	sgRNA v2
INDEL10020	EGS0293v2_D389K	CNTNAP2	0.4	3.5	sgRNA v2
INDEL10028	EGS0293v2_D389K	OR10D3	2.4	9.7	sgRNA v2
INDEL10036	EGS0293v2_D389K	CNTNAP2	1.8	6.5	sgRNA v2
INDEL10044	EGS0293v2_D389K	CXCR4	1.1	1.8	sgRNA v2
INDEL10052	EGS0293v2_D389K	CXCR4	0.3	3.6	sgRNA v2
INDEL10060	EGS0293v2_D389K	EMX1	0		sgRNA v2
INDEL10076	EGS0293v2_D389K	LCA5	6.7	6.7	sgRNA v2
INDEL9997	EGS0293v2_D389R	CRYGS	0		sgRNA v2
INDEL10005	EGS0293v2_D389R	OR10D3	2	7.5	sgRNA v2
INDEL10013	EGS0293v2_D389R	RNU6_1	0.3	1.8	sgRNA v2
INDEL10029	EGS0293v2_D389R	OR10D3	4.4	6.3	sgRNA v2
INDEL10037	EGS0293v2_D389R	CNTNAP2	4.2	7.6	sgRNA v2
INDEL10045	EGS0293v2_D389R	CXCR4	0.4	1.9	sgRNA v2
INDEL10053	EGS0293v2_D389R	CXCR4	0		sgRNA v2
INDEL10061	EGS0293v2_D389R	EMX1	0.2	1.1	sgRNA v2
INDEL10069	EGS0293v2_D389R	GTF2A1	1.5	4.8	sgRNA v2
INDEL10077	EGS0293v2_D389R	LCA5	10.6	7.8	sgRNA v2
INDEL11370	EGS0293v2_E247R_Q343K_N347K	OR10D3	6.07	6.5	sgRNA v2
INDEL11418	EGS0293v2_E247R_Q343K_N347K	LCA5	10.66	7.9	sgRNA v2

INDEL11346	EGS0293v2_E247R_Q343K_N347K	OR10D3	5.38	8.1	sgRNA v2
INDEL11426	EGS0293v2_E247R_Q343K_N347K	RNF2	5.32	8.1	sgRNA v2
INDEL11378	EGS0293v2_E247R_Q343K_N347K	CNTNAP2	2.87	9.4	sgRNA v2
INDEL11410	EGS0293v2_E247R_Q343K_N347K	GTF2A1	1.14	2	sgRNA v2
INDEL11354	EGS0293v2_E247R_Q343K_N347K	RNU6_1	0.34	2	sgRNA v2
INDEL11362	EGS0293v2_E247R_Q343K_N347K	CNTNAP2	0.39	3.9	sgRNA v2
INDEL11394	EGS0293v2_E247R_Q343K_N347K	CXCR4	0.03	1	sgRNA v2
INDEL11402	EGS0293v2_E247R_Q343K_N347K	EMX1	0.06	1.3	sgRNA v2
INDEL11338	EGS0293v2_E247R_Q343K_N347K	CRYGS	0.8	2	sgRNA v2
INDEL11287	EGS0293v2_N157R_Q343K_N347K	CNTNAP2	10.17	9.1	sgRNA v2
INDEL11303	EGS0293v2_N157R_Q343K_N347K	CXCR4	1.54	3.8	sgRNA v2
INDEL11271	EGS0293v2_N157R_Q343K_N347K	CNTNAP2	1.96	3.5	sgRNA v2
INDEL11327	EGS0293v2_N157R_Q343K_N347K	LCA5	12.66	6.5	sgRNA v2
INDEL11295	EGS0293v2_N157R_Q343K_N347K	CXCR4	1.84	2	sgRNA v2
INDEL11255	EGS0293v2_N157R_Q343K_N347K	OR10D3	11.29	11.3	sgRNA v2
INDEL11319	EGS0293v2_N157R_Q343K_N347K	GTF2A1	6.03	8.2	sgRNA v2
INDEL11335	EGS0293v2_N157R_Q343K_N347K	RNF2	10.49	8.7	sgRNA v2
INDEL11263	EGS0293v2_N157R_Q343K_N347K	RNU6_1	0.41	6.9	sgRNA v2
INDEL11247	EGS0293v2_N157R_Q343K_N347K	CRYGS	2.56	5.6	sgRNA v2
INDEL9994	EGS0293v2_N347K	CRYGS	1.8	3.4	sgRNA v2
INDEL10002	EGS0293v2_N347K	OR10D3	3.9	7.5	sgRNA v2
INDEL10010	EGS0293v2_N347K	RNU6_1	0.6	4.9	sgRNA v2
INDEL10018	EGS0293v2_N347K	CNTNAP2	0.2	2.9	sgRNA v2
INDEL10026	EGS0293v2_N347K	OR10D3	8.3	6.7	sgRNA v2
INDEL10034	EGS0293v2_N347K	CNTNAP2	1.7	8	sgRNA v2
INDEL10050	EGS0293v2_N347K	CXCR4	0.2	4.5	sgRNA v2
INDEL10058	EGS0293v2_N347K	EMX1	0		sgRNA v2
INDEL10066	EGS0293v2_N347K	GTF2A1	1.4	6	sgRNA v2
INDEL10074	EGS0293v2_N347K	LCA5	12.3	8.2	sgRNA v2
INDEL10082	EGS0293v2_N347K	RNF2	6.9	9.7	sgRNA v2
INDEL9995	EGS0293v2_N347R	CRYGS	2.3	3.7	sgRNA v2
INDEL10003	EGS0293v2_N347R	OR10D3	1.8	5.5	sgRNA v2
INDEL10011	EGS0293v2_N347R	RNU6_1	0.9	8.2	sgRNA v2
INDEL10019	EGS0293v2_N347R	CNTNAP2	0.8	4.7	sgRNA v2
INDEL10027	EGS0293v2_N347R	OR10D3	8.1	6.3	sgRNA v2
INDEL10035	EGS0293v2_N347R	CNTNAP2	3.1	6.7	sgRNA v2
INDEL10043	EGS0293v2_N347R	CXCR4	1.3	1.6	sgRNA v2
INDEL10051	EGS0293v2_N347R	CXCR4	0.7	2.7	sgRNA v2
INDEL10059	EGS0293v2_N347R	EMX1	0.1	1.1	sgRNA v2
INDEL10075	EGS0293v2_N347R	LCA5	12.3	8	sgRNA v2
INDEL10083	EGS0293v2_N347R	RNF2	6.8	9.5	sgRNA v2

INDEL11273	EGS0293v2_Q343K	OR10D3	4.45	6.6	sgRNA v2
INDEL11249	EGS0293v2_Q343K	OR10D3	3.24	5.1	sgRNA v2
INDEL11241	EGS0293v2_Q343K	CRYGS	1.17	3	sgRNA v2
INDEL11281	EGS0293v2_Q343K	CNTNAP2	0.42	4	sgRNA v2
INDEL11305	EGS0293v2_Q343K	EMX1	0.05	1.2	sgRNA v2
INDEL11321	EGS0293v2_Q343K	LCA5	7.86	7.2	sgRNA v2
INDEL11289	EGS0293v2_Q343K	CXCR4	0.66	2	sgRNA v2
INDEL11265	EGS0293v2_Q343K	CNTNAP2	0.3	2.9	sgRNA v2
INDEL11297	EGS0293v2_Q343K	CXCR4	0.03	1.1	sgRNA v2
INDEL11257	EGS0293v2_Q343K	RNU6_1	0.39	4.9	sgRNA v2
INDEL11329	EGS0293v2_Q343K	RNF2	3.62	7.2	sgRNA v2
INDEL9902	EGS0293v2_Q343K	CRYGS	2.6	4.5	sgRNA v2
INDEL9910	EGS0293v2_Q343K	OR10D3	2.7	10.7	sgRNA v2
INDEL9918	EGS0293v2_Q343K	RNU6_1	0.9	3.8	sgRNA v2
INDEL9934	EGS0293v2_Q343K	OR10D3	3.4	6.1	sgRNA v2
INDEL9950	EGS0293v2_Q343K	CXCR4	1.4	1.9	sgRNA v2
INDEL9958	EGS0293v2_Q343K	CXCR4	0.4	2.5	sgRNA v2
INDEL9966	EGS0293v2_Q343K	EMX1	0.4	1	sgRNA v2
INDEL9974	EGS0293v2_Q343K	GTF2A1	0.8	2.2	sgRNA v2
INDEL9982	EGS0293v2_Q343K	LCA5	7.2	6.2	sgRNA v2
INDEL9990	EGS0293v2_Q343K	RNF2	5.5	9.3	sgRNA v2
INDEL11282	EGS0293v2_Q343K_N347K	CNTNAP2	1.93	11.6	sgRNA v2
INDEL11322	EGS0293v2_Q343K_N347K	LCA5	6.56	4.7	sgRNA v2
INDEL11298	EGS0293v2_Q343K_N347K	CXCR4	0.26	2.7	sgRNA v2
INDEL11290	EGS0293v2_Q343K_N347K	CXCR4	3.71	3.2	sgRNA v2
INDEL11314	EGS0293v2_Q343K_N347K	GTF2A1	1.13	4.4	sgRNA v2
INDEL11266	EGS0293v2_Q343K_N347K	CNTNAP2	1.95	3.6	sgRNA v2
INDEL11306	EGS0293v2_Q343K_N347K	EMX1	0.07	1.2	sgRNA v2
INDEL11274	EGS0293v2_Q343K_N347K	OR10D3	6.41	5.4	sgRNA v2
INDEL11242	EGS0293v2_Q343K_N347K	CRYGS	2.52	5.1	sgRNA v2
INDEL11330	EGS0293v2_Q343K_N347K	RNF2	5.74	9.9	sgRNA v2
INDEL11250	EGS0293v2_Q343K_N347K	OR10D3	2.25	7.2	sgRNA v2
INDEL11258	EGS0293v2_Q343K_N347K	RNU6_1	0.48	3.2	sgRNA v2
INDEL11366	EGS0293v2_Q343K_N347K_A424R	CNTNAP2	6.05	8.2	sgRNA v2
INDEL11382	EGS0293v2_Q343K_N347K_A424R	CNTNAP2	10.75	10.8	sgRNA v2
INDEL11342	EGS0293v2_Q343K_N347K_A424R	CRYGS	3.64	4.2	sgRNA v2
INDEL11406	EGS0293v2_Q343K_N347K_A424R	EMX1	0.05	1.2	sgRNA v2
INDEL11374	EGS0293v2_Q343K_N347K_A424R	OR10D3	8.22	6.7	sgRNA v2
INDEL11414	EGS0293v2_Q343K_N347K_A424R	GTF2A1	3.95	5.5	sgRNA v2
INDEL11422	EGS0293v2_Q343K_N347K_A424R	LCA5	7.19	8.5	sgRNA v2
INDEL11358	EGS0293v2_Q343K_N347K_A424R	RNU6_1	3.72	5.2	sgRNA v2

INDEL11390	EGS0293v2_Q343K_N347K_A424R	CXCR4	12.84	6.2	sgRNA v2
INDEL11398	EGS0293v2_Q343K_N347K_A424R	CXCR4	5.54	6	sgRNA v2
INDEL11350	EGS0293v2_Q343K_N347K_A424R	OR10D3	8.03	11.5	sgRNA v2
INDEL9903	EGS0293v2_Q343R	CRYGS	3.6	3.5	sgRNA v2
INDEL9911	EGS0293v2_Q343R	OR10D3	3.2	9.4	sgRNA v2
INDEL9919	EGS0293v2_Q343R	RNU6_1	0.2	1	sgRNA v2
INDEL9927	EGS0293v2_Q343R	CNTNAP2	0.6	3.7	sgRNA v2
INDEL9943	EGS0293v2_Q343R	CNTNAP2	1.9	9	sgRNA v2
INDEL9959	EGS0293v2_Q343R	CXCR4	0		sgRNA v2
INDEL9975	EGS0293v2_Q343R	GTF2A1	0.3	1.9	sgRNA v2
INDEL9983	EGS0293v2_Q343R	LCA5	4.1	6.9	sgRNA v2
INDEL9991	EGS0293v2_Q343R	RNF2	3.7	8.1	sgRNA v2

Prophetic Example 10: CRISPR Activation and inhibition with Cas12f2 variants

[261] For CRISPR activation (CRISPRa) the catalytically dead dEGS0293v2_D289A_D486A is codon optimized for human expression and cloned into expression cassettes with a Nterm FLAGtag and SV40 NLS and a Cterm Nucleoplasmin NLS and a VPR activation domain under control of a hUbC promoter for mammalian expression. With a mutated catalytic domain, the construct will not be able to cleave the dsDNA but will still bind to the dsDNA in the presence of a targeting sgRNA. While bound to the DNA, the fused activator domains will recruit expression proteins to the same location and increase expression of the targeted gene of interest. For CRISPR inhibition (CRISPRi) the catalytically dead dEGS0293v2_D289A_D486A is codon optimized for human expression and cloned into expression cassettes with a Nterm FLAGtag and SV40 NLS and a Cterm Nucleoplasmin NLS and a KRAB repression domain under control of a hUbC promoter for mammalian expression. While bound to the DNA, the fused repressor domain will inhibit the recruitment of expression proteins to the same location and repress expression of the targeted gene of interest. Tag sequences are set forth in Table 17, activator domain sequences in Table 23, and repressor domain sequences in Table 24.

[262] A vector capable of expressing a single sgRNA under the control of a human RNA polymerase III U6 promoter is produced and introduced into an expression vector containing GFP under control of a CMV promoter. Guides are designed to targeted regions upstream of selected genes with the appropriate PAM for the system, thereby directing the fusion protein to the targeted gene. The constructs described are introduced into mammalian cells. On day of transfection, K562 cells (Sigma) are combined with SF Nucleofector solution (Lonza) and 600 ng of a RGN expression plasmid plus 500 ng of a single crRNA expression plasmid and nucleofected at FF120 per manufactures protocol (Lonza) and plated in duplicate 96-well dishes in Dulbecco's modified Eagle medium (RPMI) plus 10% (vol/vol) fetal bovine serum (Gibco) and 1% Penicillin-Streptomycin (Gibco). After 72 hours of growth, cells are washed and

harvested in PBS+1% BSA and stained with APC-labeled target specific antibody (BioLegend). Samples are run on Aria (BD).

[263] Percentage increase of the expressed targeted protein for the gene of interest is measured. Successful gene activation is defined as at least a 3-fold increase in the detectable amount of targeted protein.

[264] Table 23. DNA sequence of Activator domain

Activator Domain	Sequence	Seq ID No
VP64	GACGCATTGGACGATTTTGATCTGGATATGCTGGGAAGTGACGCCCTCGATG ATTTTGACCTTGACATGCTTGGTTCGGATGCCCTTGATGACTTTGACCTCGA CATGCTCGGCAGTGACGCCCTTGATGATTTCGACCTGGACATGCTG	161
RelA (p65) AD	CCTACACAGGCCGGCGAGGGCACACTGTCTGAAGCTCTGCTGCAGCTGCAGT TCGACGACGAGGATCTGGGAGCCCTGCTGGGAAACAGCACCGATCCTGCCGT GTTCACCGACCTGGCCAGCGTGGACAACAGCGAGTTCCAGCAGCTGCTGAAC CAGGGCATCCCTGTGGCCCCTCACACCACCGAGCCCATGCTGATGGAATACC CCGAGGCCATCACCCGGCTCGTGACAGGCGCTCAGAGGCCTCCTGATCCAGC TCCTGCCCCTCTGGGAGCACCAGGCCTGCCTAATGGACTGCTGTCTGGCGAC GAGGACTTCAGCTCTATCGCCGATATGGATTTCTCAGCCTTGCTG	162
Rta AD	CGGGATTCCAGGGAAGGGATGTTTTTGCCGAAGCCTGAGGCCGGCTCCGCTA TTAGTGACGTGTTTGAGGGCCGCGAGGTGTGCCAGCCAAAACGAATCCGGCC ATTTCATCCTCCAGGAAGTCCATGGGCCAACCGCCCACTCCCCGCCAGCCTC GCACCAACACCAACCGGTCCAGTACATGAGCCAGTCGGGTCACTGACCCCGG CACCAGTCCCTCAGCCACTGGATCCAGCGCCCGCAGTGACTCCCGAGGCCAG TCACCTGTTGGAGGATCCCGATGAAGAGACGAGCCAGGCTGTCAAAGCCCTT CGGGAGATGGCCGATACTGTGATTCCCCAGAAGGAAGAGGCTGCAATCTGTG GCCAAATGGACCTTTCCCATCCGCCCCCAAGGGGCCATCTGGATGAGCTGAC AACCACACTTGAGTCCATGACCGAGGATCTGAACCTGGACTCACCCCTGACC CCGGAATTGAACGAGATTCTGGATACCTTCCTGAACGACGAGTGCCTCTTGC ATGCCATGCATATCAGCACAGGACTGTCCATCTTCGACACATCTCTGTTT	163
VPR Domain	GGAAGCGAGGCCAGCGGTTCCGGACGGGCTGACGCATTGGACGATTTTGATC TGGATATGCTGGGAAGTGACGCCCTCGATGATTTTGACCTTGACATGCTTGG TTCGGATGCCCTTGATGACTTTGACCTCGACATGCTCGGCAGTGACGCCCTT GATGATTTCGACCTGGACATGCTGATTAACTCTAGAAGTTCCGGATCTCCGA AAAAGAAACGCAAAGTTGGTAGCCAGTACCTGCCCGACACCGACGACCGGCA CCGGATCGAGGAAAAGCGGAAGCGGACCTACGAGACATTCAAGAGCATCATG AAGAAGTCCCCCTTCAGCGGCCCCACCGACCCTAGACCTCCACCTAGAAGAA TCGCCGTGCCCAGCAGATCCAGCGCCAGCGTGCCAAAACCTGCCCCCCAGCC TTACCCCTTCACCAGCAGCCTGAGCACCATCAACTACGACGAGTTCCCTACC ATGGTGTTCCCCAGCGGCCAGATCTCTCAGGCCTCTGCTCTGGCTCCAGCCC CTCCTCAGGTGCTGCCTCAGGCTCCTGCTCCTGCACCAGCTCCAGCCATGGT GTCTGCACTGGCTCAGGCACCAGCACCCGTGCCTGTGCTGGCTCCTGGACCT CCACAGGCTGTGGCTCCACCAGCCCCTAAACCTACACAGGCCGGCGAGGGCA CACTGTCTGAAGCTCTGCTGCAGCTGCAGTTCGACGACGAGGATCTGGGAGC	164

	CCTGCTGGGAAACAGCACCGATCCTGCCGTGTTCACCGACCTGGCCAGCGTG GACAACAGCGAGTTCCAGCAGCTGCTGAACCAGGGCATCCCTGTGGCCCCTC ACACCACCGAGCCCATGCTGATGGAATACCCCGAGGCCATCACCCGGCTCGT GACAGGCGCTCAGAGGCCTCCTGATCCAGCTCCTGCCCCTCTGGGAGCACCA GGCCTGCCTAATGGACTGCTGTCTGGCGACGAGGACTTCAGCTCTATCGCCG ATATGGATTTCTCAGCCTTGCTGGGCTCTGGCAGCGGCAGCCGGGATTCCAG GGAAGGGATGTTTTTGCCGAAGCCTGAGGCCGGCTCCGCTATTAGTGACGTG TTTGAGGGCCGCGAGGTGTGCCAGCCAAAACGAATCCGGCCATTTCATCCTC CAGGAAGTCCATGGGCCAACCGCCCACTCCCCGCCAGCCTCGCACCAACACC AACCGGTCCAGTACATGAGCCAGTCGGGTCACTGACCCCGGCACCAGTCCCT CAGCCACTGGATCCAGCGCCCGCAGTGACTCCCGAGGCCAGTCACCTGTTGG AGGATCCCGATGAAGAGACGAGCCAGGCTGTCAAAGCCCTTCGGGAGATGGC CGATACTGTGATTCCCCAGAAGGAAGAGGCTGCAATCTGTGGCCAAATGGAC CTTTCCCATCCGCCCCCAAGGGGCCATCTGGATGAGCTGACAACCACACTTG AGTCCATGACCGAGGATCTGAACCTGGACTCACCCCTGACCCCGGAATTGAA CGAGATTCTGGATACCTTCCTGAACGACGAGTGCCTCTTGCATGCCATGCAT ATCAGCACAGGACTGTCCATCTTCGACACATCTCTGTTT	

[265] Table 24. DNA sequence of repressor domains

Repressor Domain	Sequence	Seq ID No
KRAB	GATGCTAAGTCACTGACTGCCTGGTCCCGGACACTGGTGACCTTCAAGGATG TGTTTGTGGACTTCACCAGGGAGGAGTGGAAGCTGCTGGACACTGCTCAGCA GATCCTGTACAGAAATGTGATGCTGGAGAACTATAAGAACCTGGTTTCCTTG GGTTATCAGCTTACTAAGCCAGATGTGATCCTCCGGTTGGAGAAGGGAGAAG AGCCCTGGCTGGTGGAGAGAGAAATTCACCAAGAGACCCATCCTGATTCAGA GACTGCATTTGAAATCAAATCATCAGTT	165

Prophetic Example 11: Functional Genomics Screens with c2c9

[266] Genome Wide protein coding knock out libraries are designed by first identifying all possible PAM matches on both DNA strands inferred to cut within a protein-coding exon. crRNAs corresponding to all PAM matches are inferred. The specificity of all gRNAs is computed using GuideScan2 (Schmidt et al. 2022, bioRxiv, https://doi.org/10.1101/2022.05.02.490368), based on all potential off-targets up to 5 mismatches (with a maximum of 1,000 off-targets considered for each crRNA) and where each off-target site is weighted for probability of cutting based on a cutting frequency determination matrix. The cutting frequency determination matrix is based on either an approximation, given knowledge of related RGN(s), or an empirical estimation computed from systematic variation and testing of known, high efficiency crRNAs. Efficiency of all crRNAs is computed using either an in-house machine learning model of efficiency trained on empirical in-house data or using established models from related RGN(s). Accessibility of each crRNA target site is computed as the proportion of unique cellular contexts in

ENCODE for which the crRNA target site falls wholly within a DNase-seq narrow- or broad peak as called using the standardized ENCODE pipeline. crRNAs that target exons shared across transcripts are preferrable to those that target rarely utilized alternative exons. All crRNA target sites are scored for 'inclusion-by-expression', which is an estimate of the inclusion of an exon across transcripts and is estimated by exon-level expression across diverse RNA-seq data sources and normalized within-gene and across the length of the gene to control for the 3'-bias of RNA-seq. crRNAs that target exons earlier in genes are preferrable because indels at these locations are more likely to induce nonsense-mediated decay than those in later exons. All crRNA target sites are scored for exon-level 'exon priority', which is computed as a nonlinear function of order of the targeted exon and the total number of exons in the gene model. All crRNAs per gene are ranked by a linear combination of specificity, efficiency, accessibility, inclusion-by-expression, and exon priority. The top N crRNAs are iteratively selected per gene such that the midpoints of no two selected crRNA target sites are less than 30 bp from one another.

[267] Genome Wide protein-coding inhibition (CRISPRi) and activation (CRISPRa) libraries are designed by first identifying all possible PAM matches on both DNA strands inferred to cut within 300 bp of each annotated protein-coding transcription start site (TSS). crRNAs corresponding to all PAM matches are inferred. The specificity of all crRNAs is computed using GuideScan2 (Schmidt et al. 2022), based on all potential off-targets up to 5 mismatches (with a maximum of 1,000 off-targets considered for each crRNA) and where each off-target site is weighted for probability of cutting based on a cutting frequency determination matrix. The cutting frequency determination matrix is based on either an approximation, given knowledge of related RGN(s), or an empirical estimation computed from systematic variation and testing of known, high efficiency crRNAs. Efficiency of all crRNAs is computed using an in-house machine learning model of efficiency trained on empirical in-house data or using established models from related RGN(s). Accessibility of each crRNA target site is computed as the proportion of unique cellular contexts in ENCODE for which the crRNA target site falls wholly within a DNase-seq narrow- or broad peak as called using the standardized ENCODE pipeline. All crRNA target sites are assigned an activation/inhibition-specific distance weight based on the primary genomic sequence distance of the cut site to the TSS as input in a nonlinear function that peaks, for CRISPRa, 50 bp upstream from the TSS and decays in both directions and that peaks, for CRISPRi, 50 bp downstream from the TSS and decays in both directions. All crRNAs per gene are ranked by a linear combination of specificity, efficiency, accessibility, and distance weight. The top N crRNAs are iteratively selected per gene such that the midpoints of no two selected crRNA target sites will be less than 30 bp from one another.

[268] Separate libraries of sgRNA are created for gene knockout, gene activation, and gene repression. These sequences are synthesized in parallel on arrays, cloned as a pool into a lentiviral transfer plasmid, and packaged into lentivirus for pooled delivery to the cell type of interest. The relevant protein construct for the appropriate screen (such as catalytically active for gene knock out, dEGS0293-VPR for CRISPRa, and dEGS0293-KRAB for CRISPRi) is codon optimized for human expression and cloned into expression cassettes with a Nterm FLAGtag and SV40 NLS and a Cterm Nucleoplasmin NLS with a P2A cleavable linker and a selectable marker under control of a hUbC promoter for mammalian expression in a plasmid capable of being packaged into a lentivirus. The constructs are then packaged into a lentivirus for delivery to the cell type of interest. Cell types are then transduced with a lentivirus to stably express the effector protein. The pooled sgRNA libraries are then transduced to stably express the the targeted sgRNA. For all libraries, cells with differential target gene expression are isolated by fluorescence-activated cell sorting and the responsible sgRNAs recovered from the lentiviral vector in the genomic DNA by PCR. These recovered libraries are then subjected to high-throughput next generation DNA sequencing to identify the genomic sequences associated with altering the expression of each gene.

WHAT IS CLAIMED IS:

1. A nucleic acid targeting system, comprising

a) a polypeptide comprising an RNA-guided nuclease (RGN) protein comprising an amino acid sequence that is at least 80% identical to the amino acid sequence of SEQ ID NO: 1, 2, 3, 4, 5, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, or 89; and

b) an RNA molecule binding to RGN protein (gRNA) and targeting a nucleic acid sequence of interest, said gRNA comprising in 5' to 3' orientation:

i. a first region that binds to the RGN comprising the sequence of SEQ ID NO: 20, 21, 22, 23, 24, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, or 120 respectively or a variant thereof;

ii. a second region comprising a spacer sequence complimentary to the target nucleic acid sequence; and

2. The nucleic acid targeting system of claim 1, wherein the target nucleic acid is a DNA.

3. The nucleic acid targeting system of claim 1, wherein the first region that binds to the RGN additionally comprises the sequence of SEQ ID NO: 30, 31, 32, 33, 34, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, or 142 respectively or a variant thereof.

4. The nucleic acid targeting system of claim 1, wherein the target nucleic acid sequence is adjacent to a PAM sequence.

5. The nucleic acid targeting system of claim 4, wherein the PAM sequence is on the non-target strand of the target nucleic acid.

6. The nucleic acid targeting system of claim 1, wherein said RGN polypeptide is a nickase or nuclease dead.

7. The nucleic acid targeting system of claim 1, wherein said RGN polypeptide is fused to a nuclear localization signal (NLS).

8. The nucleic acid targeting system of claim 1, wherein said RGN polypeptide is fused to a heterologous polypeptide.

9. The nucleic acid targeting system of claim 1, wherein the target nucleic acid sequence is within a eukaryotic cell.

10. The nucleic acid targeting system of claim 1, wherein said RGN polypeptide is nuclease dead, and wherein the RGN polypeptide is operably linked to a base-editing polypeptide.

11. One or more isolated polynucleotides encoding the nucleic acid targeting system of any one of claims 1-10.

12. The isolated polynucleotides of claim 11, wherein the polynucleotide sequences encoding of the nucleic acid targeting system have been codon optimized for optimal expression in a target cell or organism.

13. One or more vectors comprising one or more isolated polynucleotides of claim 11.

14. The one or more vectors of claim 13, wherein said vector is a lentiviral or an AAV vector.

15. A vector comprising polynucleotides encoding said nucleic acid targeting system of any one of claims 1-10.

16. The vector of claim 15, wherein said vector is a lentiviral or an AAV vector.

17. A cell comprising the nucleic acid targeting system of any one of claims 1-10, a polynucleotide of claim 11, or a vector of claim 15.

18. A cell according to claim 17, wherein said cell is an eucaryotic cell.

19. A composition comprising the DNA targeting system of any one of claims 1-10, one or more polynucleotides of claim 9, or one or more vectors of claim 10 or 12.

20. A method for binding to a target DNA sequence comprising contacting the DNA targeting system according to any one of claims 1-10, with said target DNA sequence or a cell comprising the target DNA sequence.

21. A method for cleaving and/or modifying a target nucleic acid sequence, comprising

1) contacting the target nucleic acid sequence with a nucleic acid targeting system of any one of claims 1-10; and
2) incubating said nucleic acid targeting system with the target nucleic acid for the time and under conditions sufficient for the cleaving and/or modification to occur.

22. The method of claim 21, wherein said target nucleic acid sequence is a DNA.

23. The method of claim 21, wherein said modified target DNA sequence comprises insertion of heterologous nucleic acid sequence into the target DNA sequence.

24. The method of claim 21, wherein said modified target DNA sequence comprises deletion of at least one nucleotide from the target DNA sequence

25. The method of claim 21, wherein said modified target DNA sequence comprises mutation of at least one nucleotide in the target DNA sequence

26. The method of claim 21, wherein the target nucleic acid sequence is within a cell.

27. The method of claim 26, wherein the cell is a eukaryotic cell.

28. The method of claim 26, further comprising

culturing the cell under conditions sufficient for expression of the RGN polypeptide and selecting a cell comprising said modified target nucleic acid sequence.

29. A pharmaceutical composition comprising the nucleic acid targeting system of any one of claims 1-10, one or more polynucleotides of claim 11, or one or more vectors of claim 13 or 15.

30. The nucleic acid targeting system of claim 1, wherein said RGN polypeptide comprises a mutation corresponding to T97R, and/or T101R, and/or N150R, and/or D153R, and/or N157R, and/or A190K or A190R, and/or E247R, and/or Q336R, and/or Q343K or Q343R, and/or N347K or N347R, Q373R, and/or D389K or D389R, and/orA424K or A424R, and/or V427R of SEQ ID NO: 2.

31. A nucleic acid targeting system, comprising

a) a polypeptide comprising an RNA-guided nuclease (RGN) protein comprising an amino acid sequence that is at least 80% identical to the amino acid sequence of SEQ ID NO: 98

b) an RNA molecule binding to RGN protein (gRNA) and targeting a nucleic acid sequence of interest, said gRNA comprising in 5' to 3' orientation:

i. a first region that binds to the RGN comprising the sequence of SEQ ID NO: 21, or a variant thereof;

ii. a second region comprising a spacer sequence complimentary to the target nucleic acid sequence; and

Figure 1

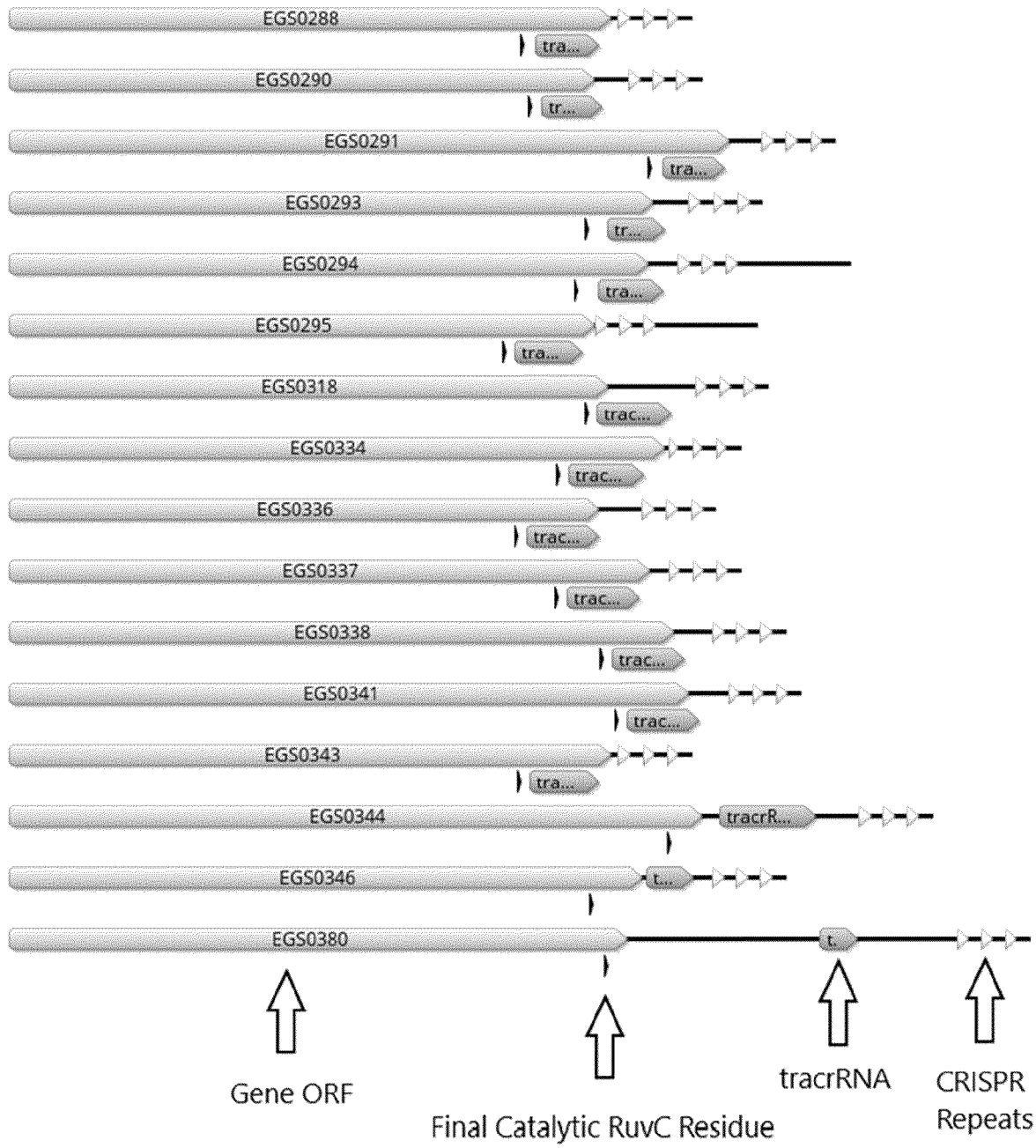

EGS0288
tra...
EGS0290
tr...
EGS0291
tra...
EGS0293
tr...
EGS0294
tra...
EGS0295
tra...
EGS0318
trac...
EGS0334
trac...
EGS0336
trac...
EGS0337
trac...
EGS0338
trac...
EGS0341
trac...
EGS0343
tra...
EGS0344
tracrR...
EGS0346
t...
EGS0380
t...
Gene ORF
Final Catalytic RuvC Residue
tracrRNA
CRISPR Repeats

Figure 2

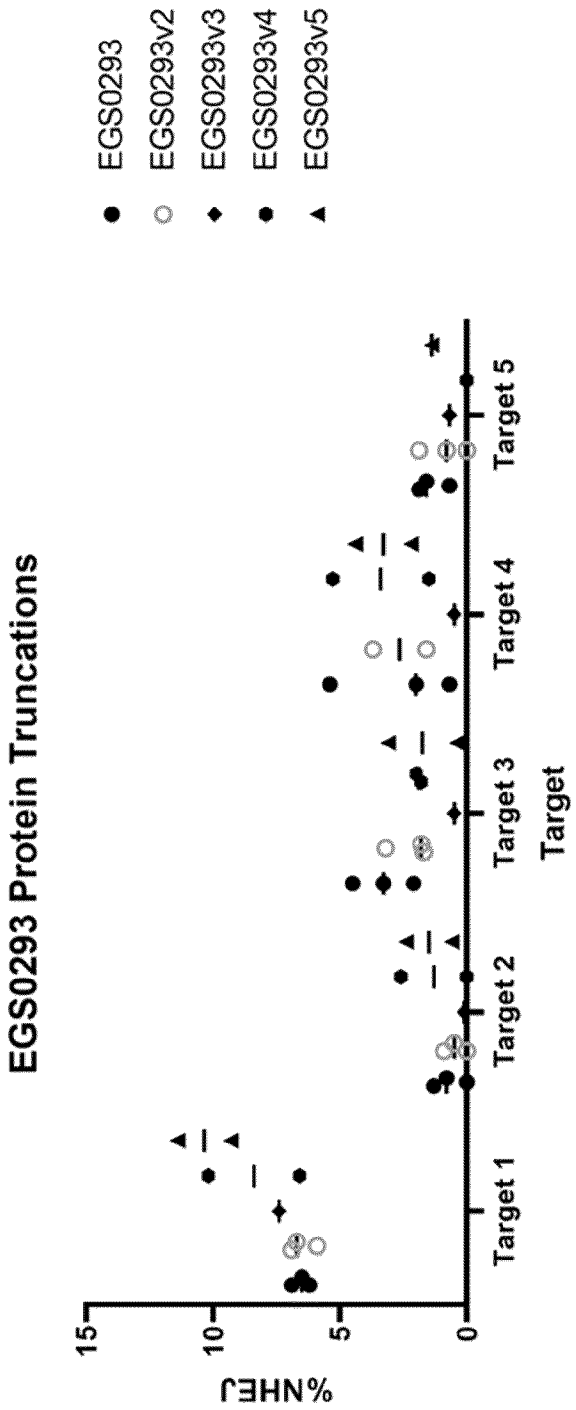

EGS0293
EGS0293v2
EGS0293v3
EGS0293v4
EGS0293v5
EGS0293 Protein Truncations
%NHEJ
15
10
5
0
Target 1
Target 2
Target 3
Target 4
Target 5
Target

Figure 3

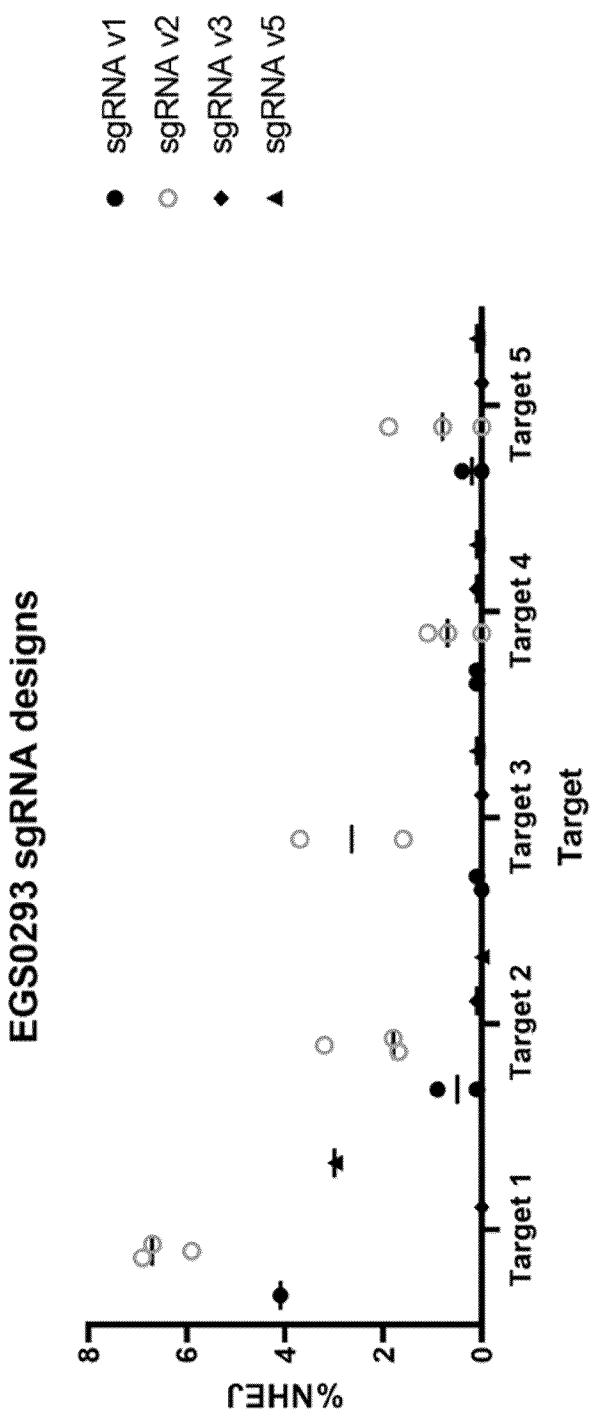

sgRNA v1
sgRNA v2
sgRNA v3
sgRNA v5
EGS0293 sgRNA designs
%NHEJ
0
2
4
6
8
Target 1
Target 2
Target 3
Target 4
Target 5
Target

Figure 4

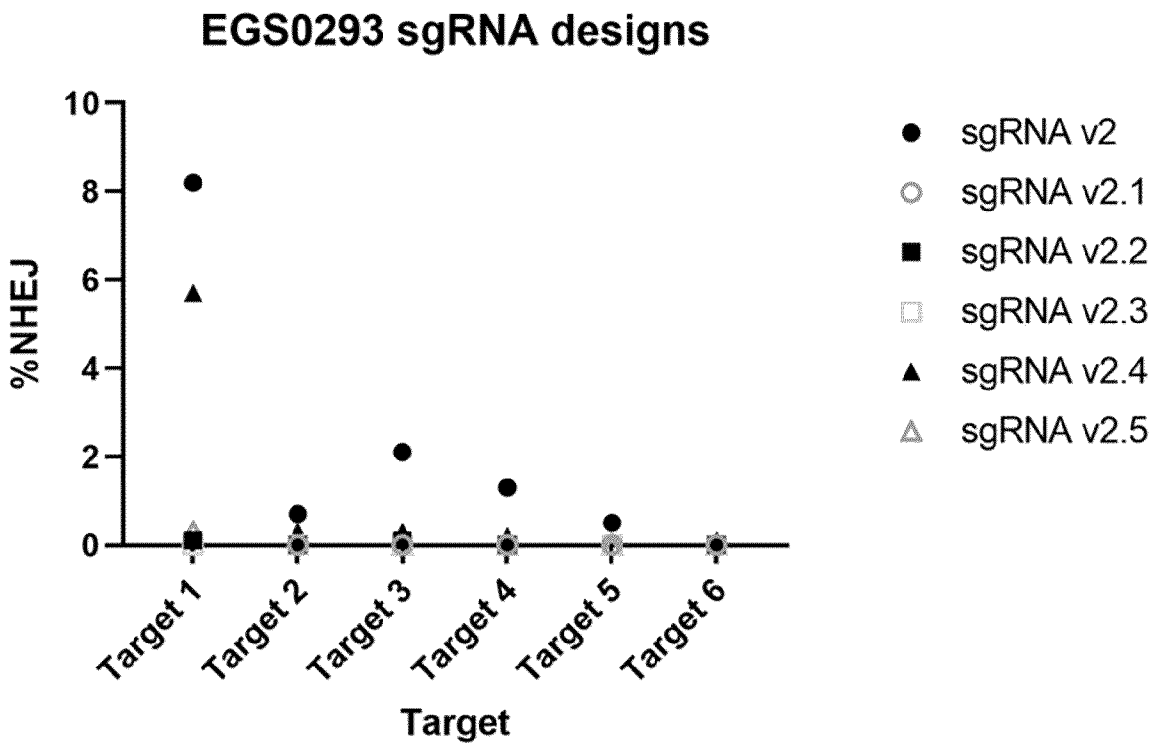

EGS0293 sgRNA designs
%NHEJ
10
8
6
4
2
0
Target 1
Target 2
Target 3
Target 4
Target 5
Target 6
Target
sgRNA v2
sgRNA v2.1
sgRNA v2.2
sgRNA v2.3
sgRNA v2.4
sgRNA v2.5

Figure 5

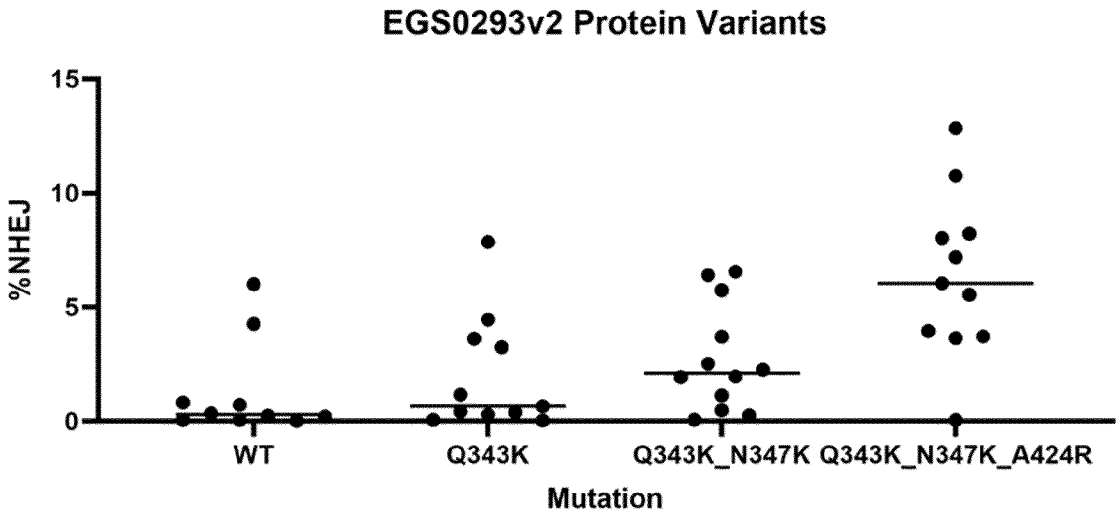
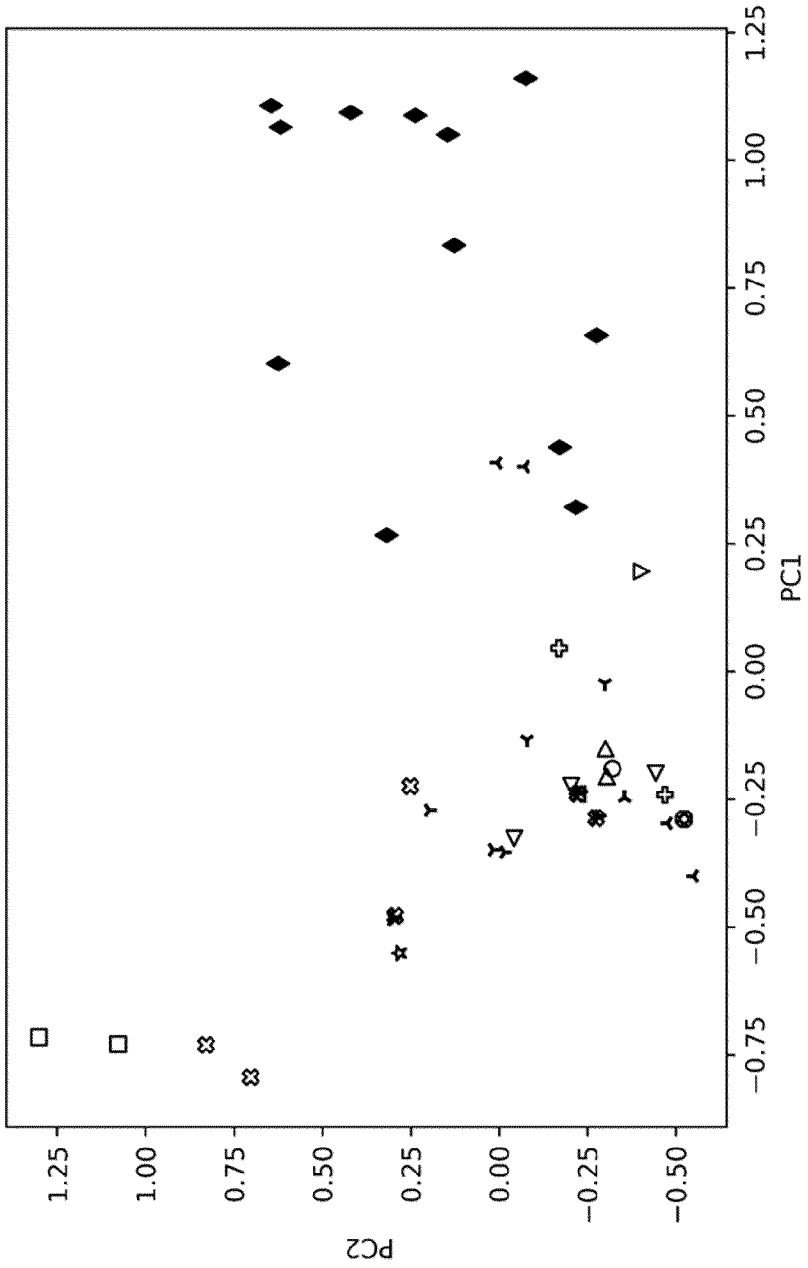

EGS0293v2 Protein Variants
%NHEJ
15
10
5
0
WT
Q343K
Q343K_N347K
Q343K_N347K_A424R
Mutation

Figure 6


Cas12a
Cas12b
Cas12c
Cas12d
Cas12e
Cas12h
Cas12i
Cas12j
Cas12k
Cas12l
Cas12m
Cas12f1
Cas12f2
c2c9
PC1
PC2
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25