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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 March 2023 (16.03.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/039378 A1

(51) International Patent Classification:
C12N 15/113 (2010.01) *C12N 9/12* (2006.01)

(21) International Application Number:
PCT/US2022/075992

(22) International Filing Date:
06 September 2022 (06.09.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/241,928 08 September 2021 (08.09.2021) US

(71) Applicant: **METAGENOMI, INC.** [US/US]; 1545 Park Avenue, Emeryville, California 94608 (US).

(72) Inventors: **THOMAS, Brian C.**; 1545 Park Avenue, Emeryville, California 94608 (US). **BROWN, Christopher**; 1545 Park Avenue, Emeryville, California 94608 (US). **CASTELLE, Cindy**; 1545 Park Avenue, Emeryville, California 94608 (US). **ALEXANDER, Lisa**; 1545 Park Avenue, Emeryville, California 94608 (US). **GONZALEZ-OSORIO, Liliana**; 1545 Park Avenue, Emeryville, California 94608 (US). **MATHEUS CARNEVALI, Paula**; 1545 Park Avenue, Emeryville, California 94608 (US). **CASTANZO, Dom**; 1545 Park Avenue, Emeryville, California 94608 (US).

(74) Agent: **TAVSHANJIAN, Brandon**; Wilson Sonsini Goodrich & Rosati, 650 Page Mill Road, Palo Alto, California 94304 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, 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, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, 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, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:
— *with international search report (Art. 21(3))*
— *with sequence listing part of description (Rule 5.2(a))*

(54) Title: CLASS II, TYPE V CRISPR SYSTEMS

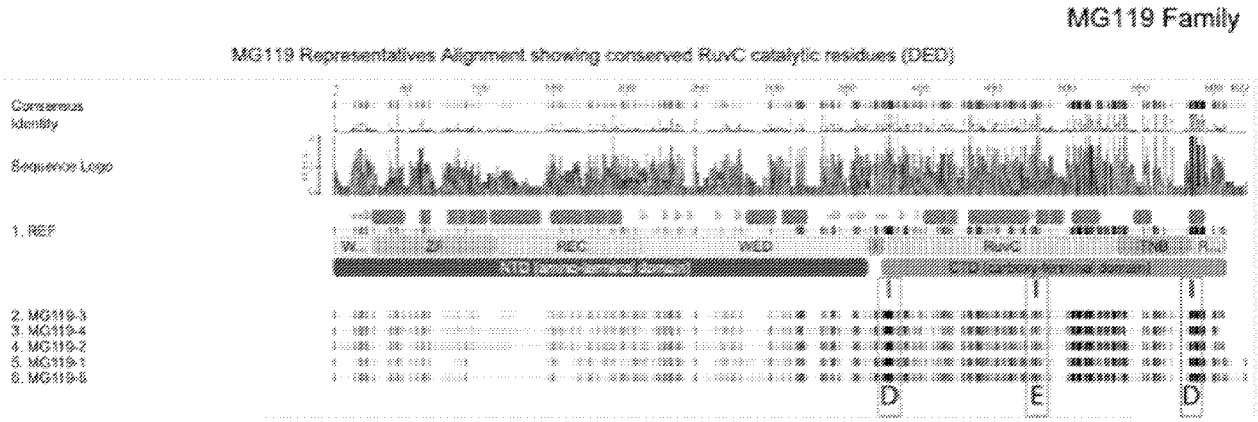


FIG. 2A

(57) Abstract: Described herein are methods, compositions, and systems derived from uncultivated microorganisms useful for gene editing involving novel Class II, Type V CRISPR-associated endonucleases.

WO 2023/039378 A1

CLASS II, TYPE V CRISPR SYSTEMS

RELATED APPLICATIONS

[0001] This application is related to PCT Patent Application No. PCT/US2021/021259 and to PCT Patent Application No. PCT/US2022/031849, each of which is incorporated herein by this reference in its entirety.

CROSS-REFERENCE

[0002] This application claims the benefit of U.S. Provisional Application No. 63/241,928, entitled "CLASS II, TYPE V CRISPR SYSTEMS", filed on September 8, 2021, which is incorporated herein by reference in its entirety.

BACKGROUND

[0003] Cas enzymes along with their associated Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) guide ribonucleic acids (RNAs) appear to be a pervasive (~45% of bacteria, ~84% of archaea) component of prokaryotic immune systems, serving to protect such microorganisms against non-self nucleic acids, such as infectious viruses and plasmids by CRISPR-RNA guided nucleic acid cleavage. While the deoxyribonucleic acid (DNA) elements encoding CRISPR RNA elements may be relatively conserved in structure and length, their CRISPR-associated (Cas) proteins are highly diverse, containing a wide variety of nucleic acid-interacting domains. While CRISPR DNA elements have been observed as early as 1987, the programmable endonuclease cleavage ability of CRISPR/Cas complexes has only been recognized relatively recently, leading to the use of recombinant CRISPR/Cas systems in diverse DNA manipulation and gene editing applications.

SEQUENCE LISTING

[0004] The instant application contains a Sequence Listing which has been submitted electronically in XML format and is hereby incorporated by reference in its entirety. Said XML copy, created on September 6, 2022, is named 55921-732601_revised_2.xml and is 1,114,268 bytes in size.

SUMMARY

[0005] In some aspects, the present disclosure provides for an engineered nuclease system comprising: an endonuclease having at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and an engineered guide RNA, wherein said engineered guide RNA is configured to form a complex with said endonuclease and said

engineered guide RNA comprises a spacer sequence configured to hybridize to a target nucleic acid sequence.

[0006] In some embodiments, said guide RNA comprises a sequence with at least 80% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419, 432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474. In some embodiments, said endonuclease has at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629. In some embodiments, said guide RNA comprises a sequence with at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 414-419, 432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474.

[0007] In some aspects, the present disclosure provides for an engineered nuclease system comprising: an engineered guide RNA comprising a sequence with at least 80% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419, 432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474, and a class 2, type V Cas endonuclease configured to bind to said engineered guide RNA. In some embodiments, the engineered nuclease system further comprises a DNA repair template comprising a double-stranded DNA segment flanked by one or two single-stranded DNA segments. In some embodiments, said single-stranded DNA segments are conjugated to the 5′ ends of said double-stranded DNA segment. In some embodiments, said single stranded DNA segments are conjugated to the 3′ ends of said double-stranded DNA segment. In some embodiments, said single-stranded DNA segments have a length from 4 to 10 nucleotide bases.

[0008] In some embodiments, said single-stranded DNA segments have a nucleotide sequence complementary to a sequence within said spacer sequence. In some embodiments, said double-stranded DNA sequence comprises a barcode, an open reading frame, an enhancer, a promoter, a protein-coding sequence, a miRNA coding sequence, an RNA coding sequence, or a transgene. In some embodiments, said double-stranded DNA sequence is flanked by a nuclease cut site. In some embodiments, said nuclease cut site comprises a spacer and a PAM sequence. In some embodiments, said PAM comprises a sequence of any one of SEQ ID NOs: 433, 435, 437, 439,

441, 443, 445, 447, 449, 451, 453, 455, 457, 459, 461, 463, 465, 467, 469, 471, 473, and 475. In some embodiments, said system further comprises a source of Mg^{2+}. In some embodiments, said guide RNA comprises a hairpin comprising at least 8, at least 10, or at least 12 base-paired ribonucleotides. In some embodiments, said hairpin comprises 10 base-paired ribonucleotides. In some embodiments, said endonuclease comprises a sequence at least 75%, 80%, or 90% identical to any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof; and said guide RNA structure comprises a sequence at least 80%, or 90% identical to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419. In some embodiments, said endonuclease comprises a sequence at least about 75%, at least about 80%, at least about 85%, at least about at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629; and said guide RNA structure comprises a sequence at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 414-419, 432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474. In some embodiments, said sequence identity is determined by a BLASTP, CLUSTALW, MUSCLE, MAFFT algorithm, or a CLUSTALW algorithm with the Smith-Waterman homology search algorithm parameters. In some embodiments, said sequence identity is determined by said BLASTP homology search algorithm using parameters of a wordlength (W) of 3, an expectation (E) of 10, and a BLOSUM62 scoring matrix setting gap costs at existence of 11, extension of 1, and using a conditional compositional score matrix adjustment.

[0009] In some aspects, present disclosure provides for an engineered guide ribonucleic acid (RNA) polynucleotide comprising: a DNA-targeting segment comprising a nucleotide sequence that is complementary to a target sequence in a target DNA molecule; and a protein-binding segment comprising two complementary stretches of nucleotides that hybridize to form a double-stranded RNA (dsRNA) duplex, wherein said two complementary stretches of nucleotides are covalently linked to one another with intervening nucleotides, and wherein said engineered guide ribonucleic acid polynucleotide is capable of forming a complex with a type 2, class V Cas endonuclease. In some embodiments, said type 2, class V Cas endonuclease is derived from an uncultivated organism. In some embodiments, said Cas endonuclease has at least 75% sequence

identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629, and targeting said complex to said target sequence of said target DNA molecule. In some embodiments, said DNA-targeting segment is positioned 3′ of both of said two complementary stretches of nucleotides. In some embodiments, said protein binding segment comprises a sequence having at least 70%, at least 80%, or at least 90% identity to the non-degenerate nucleotides of SEQ ID NO: 410-419. In some embodiments, said double-stranded RNA (dsRNA) duplex comprises at least 5, at least 8, at least 10, or at least 12 ribonucleotides.

[0010] In some aspects, the present disclosure provides for a deoxyribonucleic acid polynucleotide encoding any of the engineered guide RNAs disclosed herein.

[0011] In some aspects, the present disclosure provides for a nucleic acid comprising an engineered nucleic acid sequence optimized for expression in an organism, wherein said nucleic acid encodes a class 2, type V Cas endonuclease, and wherein said endonuclease is derived from an uncultivated microorganism, wherein the organism is not said uncultivated organism. In some embodiments, said endonuclease comprises a variant having at least 70% or at least 80% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629. In some embodiments, said endonuclease comprises a sequence encoding one or more nuclear localization sequences (NLSs) proximal to an N- or C-terminus of said endonuclease. In some embodiments, said NLS comprises a sequence selected from SEQ ID NOs: 630-645. In some embodiments, said NLS comprises SEQ ID NO: 631. In some embodiments, said NLS is proximal to said N-terminus of said endonuclease. In some embodiments, said NLS comprises SEQ ID NO: 630. In some embodiments, said NLS is proximal to said C-terminus of said endonuclease. In some embodiments, said organism is prokaryotic, bacterial, eukaryotic, fungal, plant, mammalian, rodent, or human.

[0012] In some aspects, the present disclosure provides for an engineered vector comprising a nucleic acid sequence encoding a class 2, type V Cas endonuclease, wherein said endonuclease is derived from an uncultivated microorganism.

[0013] In some aspects, the present disclosure provides for an engineered vector comprising any nucleic acid disclosed herein. In some embodiments, the vector is a plasmid, a minicircle, a CELiD, an adeno-associated virus (AAV) derived virion, a lentivirus, or an adenovirus.

[0014] In some aspects the present disclosure provides for a cell comprising any engineered vector disclosed herein.

[0015] In some aspects, the present disclosure provides for a method of manufacturing an endonuclease, comprising cultivating any cell disclosed herein.

[0016] In some aspects, the present disclosure provides for a method for binding, cleaving, marking, or modifying a double-stranded deoxyribonucleic acid polynucleotide, comprising:

contacting said double-stranded deoxyribonucleic acid polynucleotide with a class 2, type V Cas endonuclease in complex with an engineered guide RNA configured to bind to said endonuclease and said double-stranded deoxyribonucleic acid polynucleotide; wherein said double-stranded deoxyribonucleic acid polynucleotide comprises a protospacer adjacent motif (PAM); and wherein said guide RNA structure comprises a sequence at least 80%, or 90% identical to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419. In some embodiments, said double-stranded deoxyribonucleic acid polynucleotide comprises a first strand comprising a sequence complementary to a sequence of said engineered guide RNA and a second strand comprising said PAM. In some embodiments, said PAM is directly adjacent to the 5′ end of said sequence complementary to said sequence of said engineered guide RNA. In some embodiments, said PAM comprises a sequence of any one of SEQ ID NOs: 433, 435, 437, 439, 441, 443, 445, 447, 449, 451, 453, 455, 457, 459, 461, 463, 465, 467, 469, 471, 473, and 475. In some embodiments, said class 2, type V Cas endonuclease is derived from an uncultivated microorganism. In some embodiments, said class 2, type V Cas endonuclease further comprises a PAM interacting domain. In some embodiments, said double-stranded deoxyribonucleic acid polynucleotide is a eukaryotic, plant, fungal, mammalian, rodent, or human double-stranded deoxyribonucleic acid polynucleotide.

[0017] In some aspects, the present disclosure provides a method of modifying a target nucleic acid locus, said method comprising delivering to said target nucleic acid locus said engineered nuclease system of any one of claims 1-29, wherein said endonuclease is configured to form a complex with said engineered guide ribonucleic acid structure, and wherein said complex is configured such that upon binding of said complex to said target nucleic acid locus, said complex modifies said target nucleic acid locus. In some embodiments, modifying said target nucleic acid locus comprises binding, nicking, cleaving, or marking said target nucleic acid locus. In some embodiments, said target nucleic acid locus comprises deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). In some embodiments, said target nucleic acid comprises genomic DNA, viral DNA, viral RNA, or bacterial DNA. In some embodiments, said target nucleic acid locus is in vitro. In some embodiments, said target nucleic acid locus is within a cell. In some embodiments, said cell is a prokaryotic cell, a bacterial cell, a eukaryotic cell, a fungal cell, a plant cell, an animal cell, a mammalian cell, a rodent cell, a primate cell, a human cell, or a primary cell. In some embodiments, said cell is a primary cell. In some embodiments, said primary cell is a T cell. In some embodiments, said primary cell is a hematopoietic stem cell (HSC). In some embodiments, delivering said engineered nuclease system to said target nucleic acid locus comprises delivering any nucleic acid as disclosed herein or any vector as disclosed herein. In some embodiments, delivering said engineered nuclease system to said target nucleic

acid locus comprises delivering a nucleic acid comprising an open reading frame encoding said endonuclease. In some embodiments, said nucleic acid comprises a promoter to which said open reading frame encoding said endonuclease is operably linked. In some embodiments, delivering said engineered nuclease system to said target nucleic acid locus comprises delivering a capped mRNA containing said open reading frame encoding said endonuclease. In some embodiments, delivering said engineered nuclease system to said target nucleic acid locus comprises delivering a translated polypeptide. In some embodiments, delivering said engineered nuclease system to said target nucleic acid locus comprises delivering a deoxyribonucleic acid (DNA) encoding said engineered guide RNA operably linked to a ribonucleic acid (RNA) pol III promoter. In some embodiments, said endonuclease induces a single-stranded break or a double-stranded break at or proximal to said target locus. In some embodiments, said endonuclease induces a staggered single stranded break within or 3′ to said target locus.

[0018] In some aspects, the present disclosure provides a host cell comprising an open reading frame encoding a heterologous endonuclease having at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof. In some embodiments, said endonuclease has at least 75% sequence identity to any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof. In some embodiments, said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629. In some embodiments, said host cell is an E. coli cell. In some embodiments, said E. coli cell is a λDE3 lysogen or said E. coli cell is a BL21(DE3) strain. In some embodiments, said E. coli cell has an ompT lon genotype. In some embodiments, said open reading frame is operably linked to a T7 promoter sequence, a T7-lac promoter sequence, a lac promoter sequence, a tac promoter sequence, a trc promoter sequence, a ParaBAD promoter sequence, a PrhaBAD promoter sequence, a T5 promoter sequence, a cspA promoter sequence, an araPBAD promoter, a strong leftward promoter from phage lambda (pL promoter), or any combination thereof. In some embodiments, said open reading frame comprises a sequence encoding an affinity tag linked in-frame to a sequence encoding said endonuclease. In some embodiments, said affinity tag is an immobilized metal affinity chromatography (IMAC) tag. In some embodiments, said IMAC tag is a polyhistidine tag. In some embodiments, said affinity tag is a myc tag, a human influenza hemagglutinin (HA) tag, a maltose binding protein (MBP) tag, a glutathione S-transferase (GST) tag, a streptavidin tag, a FLAG tag, or any combination thereof. In some embodiments, said

affinity tag is linked in-frame to said sequence encoding said endonuclease via a linker sequence encoding a protease cleavage site. In some embodiments, said protease cleavage site is a tobacco etch virus (TEV) protease cleavage site, a PreScission® protease (PSP) cleavage site, a Thrombin cleavage site, a Factor Xa cleavage site, an enterokinase cleavage site, or any combination thereof. In some embodiments, said open reading frame is codon-optimized for expression in said host cell. In some embodiments, said open reading frame is provided on a vector. In some embodiments, said open reading frame is integrated into a genome of said host cell.

[0019] In some aspects, the present disclosure provides a culture comprising any host cell disclosed herein in compatible liquid medium.

[0020] In some aspects, the present disclosure provides a method of producing an endonuclease, comprising cultivating any host cell disclosed herein in compatible growth medium. In some embodiments, the method further comprises inducing expression of said endonuclease by addition of an additional chemical agent or an increased amount of a nutrient. In some embodiments, the method further comprises isolating said host cell after said cultivation and lysing said host cell to produce a protein extract. In some embodiments, the method further comprises subjecting said protein extract to IMAC, or ion-affinity chromatography. In some embodiments, the method further comprises cleaving said IMAC affinity tag by contacting a protease corresponding to said protease cleavage site to said endonuclease. In some embodiments, the method further comprises performing subtractive IMAC affinity chromatography to remove said affinity tag from a composition comprising said endonuclease.

[0021] In some aspects, the present disclosure provides a method of disrupting a locus in a cell, comprising contacting to said cell a composition comprising: a class 2, type V Cas endonuclease having at least 75% identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and an engineered guide RNA, wherein said engineered guide RNA is configured to form a complex with said endonuclease and said engineered guide RNA comprises a spacer sequence configured to hybridize to a region of said locus, wherein said class 2, type V Cas endonuclease has at least equivalent cleavage activity to spCas9 in said cell. In some embodiments, said cleavage activity is measured in vitro by introducing said endonucleases alongside compatible guide RNAs to cells comprising said target nucleic acid and detecting cleavage of said target nucleic acid sequence in said cells. In some embodiments, said composition comprises 20 picomoles (pmol) or less of said class 2, type V Cas endonuclease. In some embodiments, said composition comprises 1 pmol or less of said class 2, type V Cas endonuclease.

[0022] In some aspects, the present disclosure provides for a method of disrupting an albumin locus in a cell, comprising contacting to said cell a composition comprising: an endonuclease

having at least 75% identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and an engineered guide RNA, wherein said engineered guide RNA is configured to form a complex with said endonuclease, and said engineered guide RNA comprises a spacer sequence configured to hybridize to a region of said locus, wherein said engineered guide RNA is configured to hybridize to the any one of the target sequences in Table 6. In some embodiments, said engineered guide RNA comprises a sequence having at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to at least 18 non-degenerate nucleotides of any one of SEQ ID NOs: 414-419432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474. In some embodiments, said engineered guide RNA comprises the modified nucleotides of any of the single guide RNA (sgRNA) sequences in Table 6. In some embodiments, said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629. In some embodiments, said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to SEQ ID NO: 57. In some embodiments, said region is 5' to a PAM sequence comprising any one of SEQ ID NOs: 433, 435, 437, 439, 441, 443, 445, 447, 449, 451, 453, 455, 457, 459, 461, 463, 465, 467, 469, 471, 473, and 475.

[0023] In some aspects, the present disclosure provides for an isolated RNA molecule comprising a sequence at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any sequence in Table 6. In some embodiments, the isolated RNA molecule further comprises the pattern of chemical modifications recited in any of the guide RNAs recited in Table 6.

[0024] In some aspects, the present disclosure provides for a use of any RNA molecule disclosed herein for modifying an albumin locus of a cell.

[0025] In some aspects, the present disclosure provides for an engineered nuclease system comprising, an endonuclease configured to be selective for a protospacer adjacent motif (PAM)

comprising any one of SEQ ID NOs: 433, 435, 437, 439, 441, 443, 445, 447, 449, 451, 453, 455, 457, 459, 461, 463, 465, 467, 469, 471, 473, and 475; and an engineered guide RNA, wherein said engineered guide RNA is configured to form a complex with said endonuclease, and said engineered guide RNA comprises a spacer sequence configured to hybridize to a target nucleic acid sequence. In some embodiments, said endonuclease is a class 2, type V Cas endonuclease. In some embodiments, said endonuclease is not a Cas12a nuclease. In some embodiments, said endonuclease is derived from an uncultivated organism. In some embodiments, said endonuclease further comprises a PAM interacting domain configured to interact with said PAM. In some embodiments, said endonuclease has at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof. In some embodiments, said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629.

[0026] In some aspects, the present disclosure provides an engineered nuclease system comprising: an endonuclease having at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and a DNA methyltransferase. In some embodiments, said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629. In some embodiments, said DNA methyltransferase binds non-covalently to said endonuclease. In some embodiments, said DNA methyltransferase is fused to said endonuclease in a single polypeptide. In some embodiments, said DNA methyltransferase comprises Dmnt3A or Dnmt3L. In some embodiments, said KRAB domain binds non-covalently to said endonuclease or said DNA methyltransferase.

[0027] In some embodiments, said KRAB domain is covalently linked to said endonuclease or said DNA methyltransferase. In some embodiments, said KRAB domain is fused to said endonuclease or said DNA methyltransferase in a single polypeptide. In some embodiments, said endonuclease is a nickase or is catalytically dead. In some embodiments, the engineered nuclease system further comprises an engineered guide RNA structure configured to form a complex with said endonuclease, and wherein said engineered guide RNA structure comprises a spacer

sequence configured to hybridize to a target nucleic acid sequence. In some embodiments, said target nucleic acid sequence is comprised in or proximal to a promoter of a target genome. In some embodiments, said engineered guide RNA structure comprises one or more: (a) 2'-O-methylnucleotide(s); (b) 2'-fluoronucleotide(s); or (c) phosphorothioate bond(s). In some embodiments, said engineered guide RNA structure comprises the pattern of chemically modified nucleotides of any of the single guide RNAs in Table 6.

[0028] In some aspects, the present disclosure provides for a method of modifying a target nucleic acid locus, said method comprising delivering to said target nucleic acid locus any engineered nuclease system disclosed herein, wherein said endonuclease is configured to forma complex with said engineered guide RNA structure, and wherein said complex is configured that upon binding of said complex to said target nucleic acid locus, said DNA methyltransferase modifies said target nucleic acid locus.

[0029] In some aspects, the present disclosure provides for a use any engineered nuclease system disclosed herein for modifying a nucleic acid locus. In some embodiments, modifying said nucleic acid locus comprises methylating or demethylating a nucleotide of said nucleic acid locus.

[0030] In some aspects, the present disclosure provides for an engineered nuclease system comprising: (a) an endonuclease comprising a RuvC domain, wherein the endonuclease is derived from an uncultivated microorganism, and wherein the endonuclease is not a Cas12a endonuclease; and (b) an engineered guide RNA, wherein the engineered guide RNA is configured to form a complex with the endonuclease and the engineered guide RNA comprises a spacer sequence configured to hybridize to a target nucleic acid sequence. In some aspects, the present disclosure provides an engineered nuclease system comprising: (a) an endonuclease having at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and (b) an engineered guide RNA, wherein the engineered guide RNA is configured to form a complex with the endonuclease and the engineered guide RNA comprises a spacer sequence configured to hybridize to a target nucleic acid sequence. In some embodiments, the endonuclease comprises a RuvCI, II, or III domain. In some embodiments, the endonuclease has at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99% identity to a RuvCI, II, or III domain of any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof. In some embodiments, the RuvCI domain comprises a D catalytic residue. In

some embodiments the RuvCII domain comprises an E catalytic residue. In some embodiments the RuvCIII domain comprises a D catalytic residue. In some embodiments, the RuvC domain does not have nuclease activity. In some embodiments, the endonuclease further comprises a WED II domain having at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99% identity to a WED II domain of any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof. In some embodiments, the guide RNA comprises a sequence with at least 80% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419. In some aspects, the present disclosure provides an engineered nuclease system comprising: (a) an engineered guide RNA comprising a sequence with at least 80% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419, and (b) a class 2, type V Cas endonuclease configured to bind to the engineered guide RNA. In some embodiments, the guide RNA comprises a sequence complementary to a eukaryotic, fungal, plant, mammalian, or human genomic polynucleotide sequence. In some embodiments, the guide RNA is 30-250 nucleotides in length. In some embodiments, the endonuclease comprises one or more nuclear localization sequences (NLSs) proximal to an N- or C-terminus of the endonuclease. In some embodiments, the NLS comprises a sequence at least 80% identical to a sequence from the group consisting of SEQ ID NO: 630-645.

[0031] In some embodiments, the engineered nuclease system further comprises a single- or double-stranded DNA repair template comprising from 5′ to 3′: a first homology arm comprising a sequence of at least 20 nucleotides 5′ to the target deoxyribonucleic acid sequence, a synthetic DNA sequence of at least 10 nucleotides, and a second homology arm comprising a sequence of at least 20 nucleotides 3′ to the target sequence. In some embodiments, the first or second homology arm comprises a sequence of at least 40, 80, 120, 150, 200, 300, 500, or 1,000 nucleotides. In some embodiments, the first and second homology arms are homologous to a genomic sequence of a prokaryote, bacteria, fungus, or eukaryote. In some embodiments, the single- or double-stranded DNA repair template comprises a transgene donor. In some embodiments, the engineered nuclease system further comprises a DNA repair template comprising a double-stranded DNA segment flanked by one or two single-stranded DNA segments. In some embodiments, the single-stranded DNA segments are conjugated to the 5′ ends of the double-stranded DNA segment. In some embodiments, the single stranded DNA segments are conjugated to the 3′ ends of the double-stranded DNA segment. In some

embodiments, the single-stranded DNA segments have a length from 4 to 10 nucleotide bases. In some embodiments, the single-stranded DNA segments have a nucleotide sequence complementary to a sequence within the spacer sequence. In some embodiments, the double-stranded DNA sequence comprises a barcode, an open reading frame, an enhancer, a promoter, a protein-coding sequence, a miRNA coding sequence, an RNA coding sequence, or a transgene. In some embodiments, the double-stranded DNA sequence is flanked by a nuclease cut site. In some embodiments, the nuclease cut site comprises a spacer and a PAM sequence. In some embodiments, the system further comprises a source of Mg^{2+}. In some embodiments, the guide RNA comprises a hairpin comprising at least 8, at least 10, or at least 12 base-paired ribonucleotides. In some embodiments, the hairpin comprises 10 base-paired ribonucleotides. In some embodiments, a) the endonuclease comprises a sequence at least 75%, 80%, or 90% identical to any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof; and b) the guide RNA structure comprises a sequence at least 80%, or 90% identical to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419. In some embodiments, the sequence identity is determined by a BLASTP, CLUSTALW, MUSCLE, MAFFT algorithm, or a CLUSTALW algorithm with the Smith-Waterman homology search algorithm parameters. In some embodiments, the sequence identity is determined by the BLASTP homology search algorithm using parameters of a wordlength (W) of 3, an expectation (E) of 10, and a BLOSUM62 scoring matrix setting gap costs at existence of 11, extension of 1, and using a conditional compositional score matrix adjustment.

[0032] In some aspects, the present disclosure provides an engineered guide RNA comprising: a) a DNA-targeting segment comprising a nucleotide sequence that is complementary to a target sequence in a target DNA molecule; and b) a protein-binding segment comprising two complementary stretches of nucleotides that hybridize to form a double-stranded RNA (dsRNA) duplex, wherein the two complementary stretches of nucleotides are covalently linked to one another with intervening nucleotides, and wherein the engineered guide ribonucleic acid polynucleotide is capable of forming a complex with an endonuclease having at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629, and targeting the complex to the target sequence of the target DNA molecule. In some embodiments, the DNA-targeting segment is positioned 3′ of both of the two complementary stretches of nucleotides. In some embodiments, the protein binding segment comprises a sequence having at least 70%, at least 80%, or at least 90% identity to the non-degenerate nucleotides of SEQ ID NO: 410-419. In some embodiments, the double-stranded RNA (dsRNA) duplex comprises at least 5, at least 8, at least 10, or at least 12 ribonucleotides.

[0033] In some aspects, the present disclosure provides a deoxyribonucleic acid polynucleotide

encoding an engineered guide ribonucleic acid polynucleotide described herein.

[0034] In some aspects, the present disclosure provides a nucleic acid comprising an engineered nucleic acid sequence optimized for expression in an organism, wherein the nucleic acid encodes a class 2, type V Cas endonuclease, and wherein the endonuclease is derived from an uncultivated microorganism, wherein the organism is not the uncultivated organism. In some embodiments, the endonuclease comprises a variant having at least 70% or at least 80% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629. In some embodiments, the endonuclease comprises a sequence encoding one or more nuclear localization sequences (NLSs) proximal to an N- or C-terminus of the endonuclease. In some embodiments, the NLS comprises a sequence selected from SEQ ID NOs: 630-645. In some embodiments, the NLS comprises SEQ ID NO: 631. In some embodiments, the NLS is proximal to the N-terminus of the endonuclease. In some embodiments, the NLS comprises SEQ ID NO: 630. In some embodiments, the NLS is proximal to the C-terminus of the endonuclease. In some embodiments, the organism is prokaryotic, bacterial, eukaryotic, fungal, plant, mammalian, rodent, or human.

[0035] In some aspects, the present disclosure provides an engineered vector comprising a nucleic acid sequence encoding a class 2, type V Cas endonuclease, wherein the endonuclease is derived from an uncultivated microorganism.

[0036] In some aspects, the present disclosure provides an engineered vector comprising a nucleic acid described herein.

[0037] In some aspects, the present disclosure provides an engineered vector comprising a deoxyribonucleic acid polynucleotide described herein. In some embodiments, the vector is a plasmid, a minicircle, a CELiD, an adeno-associated virus (AAV) derived virion, a lentivirus, or an adenovirus.

[0038] In some aspects, the present disclosure provides a cell comprising a vector described herein.

[0039] In some aspects, the present disclosure provides a method of manufacturing an endonuclease, comprising cultivating any of the host cells described herein.

[0040] In some aspects, the present disclosure provides a method for binding, cleaving, marking, or modifying a double-stranded deoxyribonucleic acid polynucleotide, comprising: (a) contacting the double-stranded deoxyribonucleic acid polynucleotide with a class 2, type V Cas endonuclease in complex with an engineered guide RNA configured to bind to the endonuclease and the double-stranded deoxyribonucleic acid polynucleotide; wherein the double-stranded deoxyribonucleic acid polynucleotide comprises a protospacer adjacent motif (PAM); and wherein the guide RNA structure comprises a sequence at least 80%, or 90% identical to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419. In some embodiments, the double-

stranded deoxyribonucleic acid polynucleotide comprises a first strand comprising a sequence complementary to a sequence of the engineered guide RNA and a second strand comprising the PAM. In some embodiments, the PAM is directly adjacent to the 5′ end of the sequence complementary to the sequence of the engineered guide RNA. In some embodiments, the class 2, type V Cas endonuclease is derived from an uncultivated microorganism. In some embodiments, the double-stranded deoxyribonucleic acid polynucleotide is a eukaryotic, plant, fungal, mammalian, rodent, or human double-stranded deoxyribonucleic acid polynucleotide.

[0041] In some aspects, the present disclosure provides a method of modifying a target nucleic acid locus, the method comprising delivering to the target nucleic acid locus the engineered nuclease system described herein, wherein the endonuclease is configured to form a complex with the engineered guide ribonucleic acid structure, and wherein the complex is configured such that upon binding of the complex to the target nucleic acid locus, the complex modifies the target nucleic acid locus. In some embodiments, modifying the target nucleic acid locus comprises binding, nicking, cleaving, or marking the target nucleic acid locus. In some embodiments, the target nucleic acid locus comprises deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). In some embodiments, the target nucleic acid comprises genomic DNA, viral DNA, viral RNA, or bacterial DNA. In some embodiments, the target nucleic acid locus is *in vitro*. In some embodiments, the target nucleic acid locus is within a cell. In some embodiments, the cell is a prokaryotic cell, a bacterial cell, a eukaryotic cell, a fungal cell, a plant cell, an animal cell, a mammalian cell, a rodent cell, a primate cell, a human cell, or a primary cell. In some embodiments, the cell is a primary cell. In some embodiments, the primary cell is a T cell. In some embodiments, the primary cell is a hematopoietic stem cell (HSC). In some embodiments, delivering the engineered nuclease system to the target nucleic acid locus comprises delivering a nucleic acid described herein or a vector described herein. In some embodiments, delivering the engineered nuclease system to the target nucleic acid locus comprises delivering a nucleic acid comprising an open reading frame encoding the endonuclease. In some embodiments, the nucleic acid comprises a promoter to which the open reading frame encoding the endonuclease is operably linked. In some embodiments, delivering the engineered nuclease system to the target nucleic acid locus comprises delivering a capped mRNA containing the open reading frame encoding the endonuclease. In some embodiments, delivering the engineered nuclease system to the target nucleic acid locus comprises delivering a translated polypeptide. In some embodiments, delivering the engineered nuclease system to the target nucleic acid locus comprises delivering a deoxyribonucleic acid (DNA) encoding the engineered guide RNA operably linked to a ribonucleic acid (RNA) pol III promoter. In some embodiments, the endonuclease induces a single-stranded break or a double-stranded break at or proximal to the

target locus. In some embodiments, the endonuclease induces a staggered single stranded break within or 3′ to the target locus.

[0042] In some aspects, the present disclosure provides a host cell comprising an open reading frame encoding a heterologous endonuclease having at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof. In some embodiments, the endonuclease has at least 75% sequence identity to any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof. In some embodiments, the host cell is an *E. coli* cell or a mammalian cell. In some embodiments, the host cell is an *E. coli* cell. In some embodiments, the *E. coli* cell is a λDE3 lysogen or the *E. coli* cell is a BL21(DE3) strain. In some embodiments, the *E. coli* cell has an *ompT lon* genotype. In some embodiments, the open reading frame is operably linked to a T7 promoter sequence, a T7-lac promoter sequence, a lac promoter sequence, a tac promoter sequence, a trc promoter sequence, a ParaBAD promoter sequence, a PrhaBAD promoter sequence, a T5 promoter sequence, a *cspA* promoter sequence, an $araP_{BAD}$ promoter, a strong leftward promoter from phage lambda (pL promoter), or any combination thereof. In some embodiments, the open reading frame comprises a sequence encoding an affinity tag linked in-frame to a sequence encoding the endonuclease. In some embodiments, the affinity tag is an immobilized metal affinity chromatography (IMAC) tag. In some embodiments, the IMAC tag is a polyhistidine tag. In some embodiments, the affinity tag is a myc tag, a human influenza hemagglutinin (HA) tag, a maltose binding protein (MBP) tag, a glutathione S-transferase (GST) tag, a streptavidin tag, a FLAG tag, or any combination thereof. In some embodiments, the affinity tag is linked in-frame to the sequence encoding the endonuclease via a linker sequence encoding a protease cleavage site. In some embodiments, the protease cleavage site is a tobacco etch virus (TEV) protease cleavage site, a PreScission® protease cleavage site, a Thrombin cleavage site, a Factor Xa cleavage site, an enterokinase cleavage site, or any combination thereof. In some embodiments, open reading frame is codon-optimized for expression in the host cell. In some embodiments, the open reading frame is provided on a vector. In some embodiments, the open reading frame is integrated into a genome of the host cell.

[0043] In some aspects, the present disclosure provides a culture comprising any of the host cells described herein in compatible liquid medium.

[0044] In some aspects, the present disclosure provides a method of producing an endonuclease, comprising cultivating any of the host cells described herein in compatible growth medium. In some embodiments, the method further comprises inducing expression of the endonuclease by addition of an additional chemical agent or an increased amount of a nutrient. In some embodiments, an additional chemical agent or an increased amount of a nutrient comprises Isopropyl β-D-1-thiogalactopyranoside (IPTG) or additional amounts of lactose. In some

embodiments, the method further comprises isolating the host cell after the cultivation and lysing the host cell to produce a protein extract. In some embodiments, the method further comprises subjecting the protein extract to IMAC, or ion-affinity chromatography. In some embodiments, the open reading frame comprises a sequence encoding an IMAC affinity tag linked in-frame to a sequence encoding the endonuclease. In some embodiments, the IMAC affinity tag is linked in-frame to the sequence encoding the endonuclease via a linker sequence encoding protease cleavage site. In some embodiments, the protease cleavage site comprises a tobacco etch virus (TEV) protease cleavage site, a PreScission® protease cleavage site, a Thrombin cleavage site, a Factor Xa cleavage site, an enterokinase cleavage site, or any combination thereof. In some embodiments, the method further comprises cleaving the IMAC affinity tag by contacting a protease corresponding to the protease cleavage site to the endonuclease. In some embodiments, the method further comprises performing subtractive IMAC affinity chromatography to remove the affinity tag from a composition comprising the endonuclease.

[0045] In some aspects, the present disclosure provides a method of disrupting a locus in a cell, comprising contacting to the cell a composition comprising: (a) a class 2, type V Cas endonuclease having at least 75% identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and (b) an engineered guide RNA, wherein the engineered guide RNA is configured to form a complex with the endonuclease and the engineered guide RNA comprises a spacer sequence configured to hybridize to a region of the locus, wherein the class 2, type V Cas endonuclease has at least equivalent cleavage activity to spCas9 in the cell. In some embodiments, the cleavage activity is measured *in vitro* by introducing the endonucleases alongside compatible guide RNAs to cells comprising the target nucleic acid and detecting cleavage of the target nucleic acid sequence in the cells. In some embodiments, the composition comprises 20 pmoles or less of the class 2, type V Cas endonuclease. In some embodiments, the composition comprises 1 pmol or less of the class 2, type V Cas endonuclease.

[0046] Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative embodiments of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

INCORPORATION BY REFERENCE

[0047] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or

patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0048] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0049] FIG. 1 depicts typical organizations of CRISPR/Cas loci of different classes and types that were previously described before this disclosure.

[0050] FIGs. 2A-2D depict an overview of the MG119 Family. **FIG. 2A** depicts a multiple alignment of MG119 effectors representatives showing domains compositions and conservation of the RuvC catalytic residues critical for function for a double stranded DNA cleavage activity. **FIG. 2B** depicts a representation of a CRISPR-containing contig with genomic context surrounding the CRISPR array and the Cas effector (example of MG119-1). **FIG. 2C** depicts folding of the Direct repeat of MG119-1. **FIG. 2D** depicts a single guide RNA designed for MG119-1.

[0051] FIGs. 3A-3C depict an overview of the MG90 Family. **FIG. 3A** depicts a multiple alignment of MG90 effectors representatives showing domains compositions and conservation of the RuvC catalytic residues critical for function for a double stranded DNA cleavage activity. **FIG. 3B** depicts a representation of a CRISPR-containing contig with genomic context surrounding the CRISPR array and the Cas effector (example of MG90-5). **FIG. 3C** depicts folding of the Direct repeat of MG90-5.

[0052] FIGs. 4A-4C depict an overview of the MG126 Family. **FIG. 4A** depicts a multiple alignment of MG126 effectors representatives showing domains compositions and conservation of the RuvC catalytic residues critical for function for a double stranded DNA cleavage activity. **FIG. 4B** depicts a representation of a CRISPR-containing contig with genomic context surrounding the CRISPR array and the Cas effector (example of MG126-4). **FIG. 4C** depicts folding of the Direct repeat of MG126-4.

[0053] FIGs. 5A-5C depict an overview of the MG118 Family. **FIG. 5A** depicts a multiple alignment of MG118 effectors representatives showing domains compositions and conservation of the RuvC catalytic residues critical for function for a double stranded DNA cleavage activity. **FIG. 5B** depicts a representation of a CRISPR-containing contig with genomic context surrounding the CRISPR array and the Cas effector (example of MG118-1). **FIG. 5C** depicts folding of the Direct repeat of MG118-1.

[0054] FIGs. 6A-6C depict an overview of the MG122 Family. **FIG. 6A** depicts a multiple alignment of MG122 effectors representatives showing domains compositions and conservation

of the RuvC catalytic residues critical for function for a double stranded DNA cleavage activity. **FIG. 6B** depicts a representation of a CRISPR-containing contig with genomic context surrounding the CRISPR array and the Cas effector (example of MG122-4). **FIG. 6C** depicts folding of the Direct repeat of MG122-4.

[0055] FIGs. **7A-7C** depict an overview of the MG120 Family. **FIG. 7A** depicts a multiple alignment of MG120 effectors representatives showing domains compositions and conservation of the RuvC catalytic residues critical for function for a double stranded DNA cleavage activity. **FIG. 7B** depicts a representation of a CRISPR-containing contig with genomic context surrounding the CRISPR array and the Cas effector (example of MG120-1). **FIG. 7C** depicts folding of the Direct repeat of MG120-1.

[0056] FIGs. **8A-8D** depict an overview of the MG91 Family. **FIG. 8A** depicts a representation of a CRISPR-containing contig with genomic context surrounding the CRISPR array and the Cas effector (example of MG91B-24). **FIG. 8B** depicts folding of the Direct repeats of MG91B-24. **FIG. 8C** depicts a representation of a CRISPR-containing contig with genomic context surrounding the CRISPR array and the Cas effector (example of MG91C-10).. **FIG. 8D** depicts folding of the Direct repeats of MG91C-10.

[0057] FIG. **9** depicts *in vitro* activity of MG119-2 using the TXTL assay. MG119-2 was tested for dsDNA cleavage with two intergenic sequences from the MG119-2 contig, minimal array (MA) sequences containing repeats in the forward or reverse orientation, and a PAM library target plasmid. Positive intergenic enrichment was observed in lane 1 as an amplified cleavage product with intergenic (IG) sequence 1 and the minimal array with repeats in the forward orientation. Lanes 3 and 7 are the negative controls where IGs were omitted, and lane 4 is a third negative control where both the arrays and IGs were omitted.

[0058] FIG. **10A** depicts a SeqLogo of the MG119-2 PAM (5'-nTnn-3') determined via next-generation sequencing (NGS) of the cleavage products obtained from the *in vitro* cleavage assay. **FIG**. **10B** depicts a histogram of the cutsite (23 bd away from the PAM).

[0059] FIGs. 11A and **11B** depict examples of active MG119 nuclease and their sgRNA designs. **FIG. 11A** depicts predicted folding for single guide RNA sequences without spacers. The blue circle represents the first 5' nucleotide of the tracrRNA and the red circle represents the 3' nucleotide of the repeat. TracrRNA and repeat sequences are looped with a GAAA tetraloop. The repeat anti-repeat fold is on the 3' end of each structure. Depicted are three different RNA structures of active guides within the same family. From left to right: the MG119-28 guide has four hairpins, three smaller ones on the 5' end and a very long hairpin with two bulges next to the repeat anti-repeat fold. The MG119-83 sgRNA has three small hairpins and the repeat anti-repeat has two bulges. The MG119-118 has four hairpins, the second hairpin from the 5' end branches

into three hairpins while the third hairpin and the repeat anti-repeat have one bulge. This guide also has some pairing nucleotides between the 5' end of the tracr and the 3' end of the repeat.

FIG. 11B depicts *in vitro* cleavage assay amplification products on 2% agarose gels. Low molecular weight DNA ladders (NEB) are in lanes 1, 7, and 11. Other lane contents from left to right: (2) MG119-28 nuclease only, MG119-28 nuclease plus (3) sgRNA1 with U67 spacer, (4) sgRNA1 with U40 spacer, (5) sgRNA2 with U67 spacer, and (6) sgRNA2 with U40 spacer; (8) MG119-83 nuclease only, MG119-83 nuclease plus (9) sgRNA1 with U67 spacer and (10) sgRNA1 with U40 spacer; (12) MG119-118 nuclease only, MG119-118 nuclease plus (13) sgRNA1 with U67 spacer and (14) sgRNA1 with U40 spacer. Resulting amplicon products are 188 bp with a U67 spacer carrying guide or 205 bp with a U40 spacer carrying guide.

[0060] FIG. 12 depicts sequence logos of protospacer adjacent motifs (PAMs) for active MG119 nucleases.

[0061] FIGs. 13A-13F depict example SDS-PAGE gels of protein purification steps and size exclusion chromatography (SEC) A280 traces. **FIG. 13A** depicts **MG119-28Δ purification with** samples recovered (1) post-sonication lysis, (2) post-clarification centrifugation, (3) Ni-NTA gravity column flow-through, (4) eluate from Ni-NTA resin, (5) concentrated sample. **FIG. 13B** depicts S200i 10 / 300 GL column SEC A280 trace. Peak fractions were pooled and concentrated. **FIGs. 13C** and **13D** depict **MBP-tagged/cleaved MG119-28Δ purification with** samples recovered (1) post-sonication lysis, (2) post-clarification centrifugation, (3) Ni-NTA gravity column flow-through, (4) eluate from Ni-NTA resin, (5) concentrated protein, (6) concentrated protein cleaved overnight with TEV protease, (7) and centrifuged (21,000 x g, 4 °C, 10min) to pellet aggregates, (8) Amylose column flow-through, (9) centrifuged flow-through (21,000 x g, 4 °C, 10 min) to pellet aggregate, and (10) concentrated flow-through. **FIG. 13E** depicts S200i 10 / 300 GL column SEC A280 trace. **FIG. 13F** depicts data demonstrating that of **the five MG119 candidates expressed in both the pMGB and pMGBΔ expression vectors, all showed higher yields in the pMGBΔ vector.**

[0062] FIGS. 14A and **14B** depict an example of *in vitro* cleavage efficiency with purified protein. **FIG. 14A** depicts an agarose gel showing RNP:substrate ratio titration and increasing substrate cleavage at higher ratios. **FIG. 14B** depicts the percent of substrate cleaved determined for each lane using densitometry. Cleavage fractions were plotted in Prism8, and the slope of the linear range of cleavage was used to calculate protein active fraction. This assay used MG119-28 **expressed in the pMGBΔ backbone.**

[0063] FIGS. 15A and **15B** depict examples of *in vitro* cleavage and editing efficiency of mouse Hepa1-6 cells DNA. **FIG. 15A** depicts percent cleavage of MG119-28 with four chemically modified guides targeting the mouse albumin gene at intron 1 (**Table 6**). Two concentrations of

nuclease were tested 15.6 nM (black bars) and 7.8 nM (white bars). Cleavage was normalized to the non-targeting control. MG119-28 can cleave Hepa 1-6 gDNA up to an average of 60% with sgRNA4 at 15.6 nM RNP and up to 33% at 7.8 nM RNP. **FIG. 15B** depicts percent INDEL generated by MG119-28 in Hepa 1-6 cells normalized to apo reactions. Each condition was performed in triplicate. An average of 25.12% of the sequenced reads were edited with sgRNA3. sgRNA3 is consistently active *in vitro* and in cells as shown here. The next best guide in cells is sgRNA4 with an average of 4.11% editing. The edits observed are largely a deletion between 4-24 bp.

BRIEF DESCRIPTION OF THE SEQUENCE LISTING

[0064] The Sequence Listing filed herewith provides exemplary polynucleotide and polypeptide sequences for use in methods, compositions, and systems according to the disclosure. Below are exemplary descriptions of sequences therein.

MG122

[0065] SEQ ID NOs: 1-5 show the full-length peptide sequences of MG122 nucleases.

MG120

[0066] SEQ ID NOs: 6-14 show the full-length peptide sequences of MG120 nucleases.

[0067] SEQ ID NOs: 333-335 and 355-357 show nucleotide sequences of MG120 tracrRNAs derived from the same loci as a MG120 Cas effector.

[0068] SEQ ID NOs: 374-375 and 389-390 show nucleotide sequences of MG120 minimal arrays.

MG118

[0069] SEQ ID NO: 15 shows the full-length peptide sequence of an MG118 nuclease.

[0070] SEQ ID NO: 376 shows a nucleotide sequence of an MG118 minimal array.

[0071] SEQ ID NO: 391 shows a nucleotide sequence of an MG118 minimal array.

[0072] SEQ ID NOs: 400-401 show nucleotide sequences of MG118 target CRISPR repeats.

[0073] SEQ ID NOs: 410-411 show nucleotide sequences of MG118 crRNAs.

MG90

[0074] SEQ ID NOs: 16-29 show the full-length peptide sequences of MG90 nucleases.

[0075] SEQ ID NOs: 346-347 and 368-369 show nucleotide sequences of MG90 tracrRNAs derived from the same loci as a MG90 Cas effector.

[0076] SEQ ID NOs: 383-384 and 398-399 show nucleotide sequences of MG90 minimal arrays.

[0077] SEQ ID NOs: 402-403 show nucleotide sequences of MG90 target CRISPR repeats.

[0078] SEQ ID NOs: 412-413 show nucleotide sequences of MG90 sgRNAs.

MG119

[0079] SEQ ID NOs: 30-150, 420-431, 476-624, and 629 show the full-length peptide sequences

of MG119 nucleases.

[0080] SEQ ID NOs: 326-332, 336-345, 348-354, and 358-367 show nucleotide sequences of MG119 tracrRNAs derived from the same loci as a MG119 Cas effector.

[0081] SEQ ID NOs: 370-373, 377-382, 385-388, and 392-397 show nucleotide sequences of MG119 minimal arrays.

[0082] SEQ ID NOs: 404-409 show nucleotide sequences of MG119 target CRISPR repeats.

[0083] SEQ ID NOs: 414-419, 432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474 show nucleotide sequences of MG119 sgRNAs.

[0084] SEQ ID NOs: 433, 435, 437, 439, 441, 443, 445, 447, 449, 451, 453, 455, 457, 459, 461, 463, 465, 467, 469, 471, 473, and 475 show nucleotide sequences of MG119 PAMs.

MG91B

[0085] SEQ ID NOs: 151-291 show the full-length peptide sequences of MG91B nucleases.

MG91C

[0086] SEQ ID NOs: 292-318 show the full-length peptide sequences of MG91C nucleases.

MG91A

[0087] SEQ ID NO: 319 shows the full-length peptide sequence of an MG91A nuclease.

MG126

[0088] SEQ ID NOs: 320-325 show the full-length peptide sequences of MG126 nucleases.

DETAILED DESCRIPTION OF THE INVENTION

[0089] While various embodiments of the invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions may occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed.

[0090] The practice of some methods disclosed herein employ, unless otherwise indicated, techniques of immunology, biochemistry, chemistry, molecular biology, microbiology, cell biology, genomics, and recombinant DNA. See for example Sambrook and Green, Molecular Cloning: A Laboratory Manual, 4th Edition (2012); the series Current Protocols in Molecular Biology (F. M. Ausubel, et al. eds.); the series Methods In Enzymology (Academic Press, Inc.), PCR 2: A Practical Approach (M.J. MacPherson, B.D. Hames and G.R. Taylor eds. (1995)), Harlow and Lane, eds. (1988) Antibodies, A Laboratory Manual, and Culture of Animal Cells: A Manual of Basic Technique and Specialized Applications, 6th Edition (R.I. Freshney, ed. (2010)) (which is entirely incorporated by reference herein).

[0091] As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. Furthermore, to the extent that the terms "including", "includes", "having", "has", "with", or variants thereof are used in either the detailed description and/or the claims, such terms are intended to be inclusive in a manner similar to the term "comprising".

[0092] The term "about" or "approximately" means within an acceptable error range for the particular value as determined by one of ordinary skill in the art, which will depend in part on how the value is measured or determined, i.e., the limitations of the measurement system. For example, "about" can mean within one or more than one standard deviation, per the practice in the art. Alternatively, "about" can mean a range of up to 20%, up to 15%, up to 10%, up to 5%, or up to 1% of a given value.

[0093] As used herein, a "cell" generally refers to a biological cell. A cell may be the basic structural, functional and/or biological unit of a living organism. A cell may originate from any organism having one or more cells. Some non-limiting examples include: a prokaryotic cell, eukaryotic cell, a bacterial cell, an archaeal cell, a cell of a single-cell eukaryotic organism, a protozoa cell, a cell from a plant (e.g., cells from plant crops, fruits, vegetables, grains, soy bean, corn, maize, wheat, seeds, tomatoes, rice, cassava, sugarcane, pumpkin, hay, potatoes, cotton, cannabis, tobacco, flowering plants, conifers, gymnosperms, ferns, clubmosses, hornworts, liverworts, mosses), an algal cell, (e.g., *Botryococcus braunii*, *Chlamydomonas reinhardtii*, *Nannochloropsis gaditana*, *Chlorella pyrenoidosa*, *Sargassum patens C. Agardh*, and the like), seaweeds (e.g., kelp), a fungal cell (e.g.,, a yeast cell, a cell from a mushroom), an animal cell, a cell from an invertebrate animal (e.g., fruit fly, cnidarian, echinoderm, nematode, etc.), a cell from a vertebrate animal (e.g., fish, amphibian, reptile, bird, mammal), a cell from a mammal (e.g., a pig, a cow, a goat, a sheep, a rodent, a rat, a mouse, a non-human primate, a human, etc.), and etcetera. Sometimes a cell is not originating from a natural organism (e.g., a cell can be a synthetically made, sometimes termed an artificial cell).

[0094] The term "nucleotide," as used herein, generally refers to a base-sugar-phosphate combination. A nucleotide may comprise a synthetic nucleotide. A nucleotide may comprise a synthetic nucleotide analog. Nucleotides may be monomeric units of a nucleic acid sequence (e.g., deoxyribonucleic acid (DNA) and ribonucleic acid (RNA)). The term nucleotide may include ribonucleoside triphosphates adenosine triphosphate (ATP), uridine triphosphate (UTP), cytosine triphosphate (CTP), guanosine triphosphate (GTP) and deoxyribonucleoside triphosphates such as dATP, dCTP, dITP, dUTP, dGTP, dTTP, or derivatives thereof. Such derivatives may include, for example, [αS]dATP, 7-deaza-dGTP and 7-deaza-dATP, and nucleotide derivatives that confer nuclease resistance on the nucleic acid molecule containing

them. The term nucleotide as used herein may refer to dideoxyribonucleoside triphosphates (ddNTPs) and their derivatives. Illustrative examples of dideoxyribonucleoside triphosphates may include, but are not limited to, ddATP, ddCTP, ddGTP, ddITP, and ddTTP. A nucleotide may be unlabeled or detectably labeled, such as using moieties comprising optically detectable moieties (e.g., fluorophores). Labeling may also be carried out with quantum dots. Detectable labels may include, for example, radioactive isotopes, fluorescent labels, chemiluminescent labels, bioluminescent labels, and enzyme labels. Fluorescent labels of nucleotides may include but are not limited fluorescein, 5-carboxyfluorescein (FAM), 2′7′-dimethoxy-4′5-dichloro-6-carboxyfluorescein (JOE), rhodamine, 6-carboxyrhodamine (R6G), N,N,N′,N′-tetramethyl-6-carboxyrhodamine (TAMRA), 6-carboxy-X-rhodamine (ROX), 4-(4′dimethylaminophenylazo) benzoic acid (DABCYL), Cascade Blue, Oregon Green, Texas Red, Cyanine and 5-(2′-aminoethyl)aminonaphthalene-1-sulfonic acid (EDANS). Specific examples of fluorescently labeled nucleotides can include [R6G]dUTP, [TAMRA]dUTP, [R110]dCTP, [R6G]dCTP, [TAMRA]dCTP, [JOE]ddATP, [R6G]ddATP, [FAM]ddCTP, [R110]ddCTP, [TAMRA]ddGTP, [ROX]ddTTP, [dR6G]ddATP, [dR110]ddCTP, [dTAMRA]ddGTP, and [dROX]ddTTP available from Perkin Elmer, Foster City, Calif; FluoroLink DeoxyNucleotides, FluoroLink Cy3-dCTP, FluoroLink Cy5-dCTP, FluoroLink Fluor X-dCTP, FluoroLink Cy3-dUTP, and FluoroLink Cy5-dUTP available from Amersham, Arlington Heights, Il.; Fluorescein-15-dATP, Fluorescein-12-dUTP, Tetramethyl-rodamine-6-dUTP, IR770-9-dATP, Fluorescein-12-ddUTP, Fluorescein-12-UTP, and Fluorescein-15-2′-dATP available from Boehringer Mannheim, Indianapolis, Ind.; and Chromosome Labeled Nucleotides, BODIPY-FL-14-UTP, BODIPY-FL-4-UTP, BODIPY-TMR-14-UTP, BODIPY-TMR-14-dUTP, BODIPY-TR-14-UTP, BODIPY-TR-14-dUTP, Cascade Blue-7-UTP, Cascade Blue-7-dUTP, fluorescein-12-UTP, fluorescein-12-dUTP, Oregon Green 488-5-dUTP, Rhodamine Green-5-UTP, Rhodamine Green-5-dUTP, tetramethylrhodamine-6-UTP, tetramethylrhodamine-6-dUTP, Texas Red-5-UTP, Texas Red-5-dUTP, and Texas Red-12-dUTP available from Molecular Probes, Eugene, Oreg. Nucleotides can also be labeled or marked by chemical modification. A chemically-modified single nucleotide can be biotin-dNTP. Some non-limiting examples of biotinylated dNTPs can include, biotin-dATP (e.g., bio-N6-ddATP, biotin-14-dATP), biotin-dCTP (e.g., biotin-11-dCTP, biotin-14-dCTP), and biotin-dUTP (e.g., biotin-11-dUTP, biotin-16-dUTP, biotin-20-dUTP).

[0095] The terms “polynucleotide,” “oligonucleotide,” and “nucleic acid” are used interchangeably to generally refer to a polymeric form of nucleotides of any length, either deoxyribonucleotides or ribonucleotides, or analogs thereof, either in single-, double-, or multi-stranded form. A polynucleotide may be exogenous or endogenous to a cell. A polynucleotide may exist in a cell-free environment. A polynucleotide may be a gene or fragment thereof. A

polynucleotide may be DNA. A polynucleotide may be RNA. A polynucleotide may have any three-dimensional structure and may perform any function. A polynucleotide may comprise one or more analogs (e.g., altered backbone, sugar, or nucleobase). If present, modifications to the nucleotide structure may be imparted before or after assembly of the polymer. Some non-limiting examples of analogs include: 5-bromouracil, peptide nucleic acid, xeno nucleic acid, morpholinos, locked nucleic acids, glycol nucleic acids, threose nucleic acids, dideoxynucleotides, cordycepin, 7-deaza-GTP, fluorophores (e.g., rhodamine or fluorescein linked to the sugar), thiol-containing nucleotides, biotin-linked nucleotides, fluorescent base analogs, CpG islands, methyl-7-guanosine, methylated nucleotides, inosine, thiouridine, pseudouridine, dihydrouridine, queuosine, and wyosine. Non-limiting examples of polynucleotides include coding or non-coding regions of a gene or gene fragment, loci (locus) defined from linkage analysis, exons, introns, messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), short interfering RNA (siRNA), short-hairpin RNA (shRNA), micro-RNA (miRNA), ribozymes, cDNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, cell-free polynucleotides including cell-free DNA (cfDNA) and cell-free RNA (cfRNA), nucleic acid probes, and primers. The sequence of nucleotides may be interrupted by non-nucleotide components.

[0096] The terms “transfection” or “transfected” generally refer to introduction of a nucleic acid into a cell by non-viral or viral-based methods. The nucleic acid molecules may be gene sequences encoding complete proteins or functional portions thereof. See, e.g., Sambrook et al., 1989, Molecular Cloning: A Laboratory Manual, 18.1-18.88 (which is entirely incorporated by reference herein).

[0097] The terms “peptide,” “polypeptide,” and “protein” are used interchangeably herein to generally refer to a polymer of at least two amino acid residues joined by peptide bond(s). This term does not connote a specific length of polymer, nor is it intended to imply or distinguish whether the peptide is produced using recombinant techniques, chemical or enzymatic synthesis, or is naturally occurring. The terms apply to naturally occurring amino acid polymers as well as amino acid polymers comprising at least one modified amino acid. In some cases, the polymer may be interrupted by non-amino acids. The terms include amino acid chains of any length, including full length proteins, and proteins with or without secondary and/or tertiary structure (e.g., domains). The terms also encompass an amino acid polymer that has been modified, for example, by disulfide bond formation, glycosylation, lipidation, acetylation, phosphorylation, oxidation, and any other manipulation such as conjugation with a labeling component. The terms “amino acid” and “amino acids,” as used herein, generally refer to natural and non-natural amino

acids, including, but not limited to, modified amino acids and amino acid analogues. Modified amino acids may include natural amino acids and non-natural amino acids, which have been chemically modified to include a group or a chemical moiety not naturally present on the amino acid. Amino acid analogues may refer to amino acid derivatives. The term "amino acid" includes both D-amino acids and L-amino acids.

[0098] As used herein, the "non-native" can generally refer to a nucleic acid or polypeptide sequence that is not found in a native nucleic acid or protein. Non-native may refer to affinity tags. Non-native may refer to fusions. Non-native may refer to a naturally occurring nucleic acid or polypeptide sequence that comprises mutations, insertions and/or deletions. A non-native sequence may exhibit and/or encode for an activity (e.g., enzymatic activity, methyltransferase activity, acetyltransferase activity, kinase activity, ubiquitinating activity, etc.) that may also be exhibited by the nucleic acid and/or polypeptide sequence to which the non-native sequence is fused. A non-native nucleic acid or polypeptide sequence may be linked to a naturally-occurring nucleic acid or polypeptide sequence (or a variant thereof) by genetic engineering to generate a chimeric nucleic acid and/or polypeptide sequence encoding a chimeric nucleic acid and/or polypeptide.

[0099] The term "promoter", as used herein, generally refers to the regulatory DNA region which controls transcription or expression of a gene and which may be located adjacent to or overlapping a nucleotide or region of nucleotides at which RNA transcription is initiated. A promoter may contain specific DNA sequences which bind protein factors, often referred to as transcription factors, which facilitate binding of RNA polymerase to the DNA leading to gene transcription. A 'basal promoter', also referred to as a 'core promoter', may generally refer to a promoter that contains all the basic necessary elements to promote transcriptional expression of an operably linked polynucleotide. Eukaryotic basal promoters typically, though not necessarily, contain a TATA-box and/or a CAAT box.

[00100] The term "expression", as used herein, generally refers to the process by which a nucleic acid sequence or a polynucleotide is transcribed from a DNA template (such as into mRNA or other RNA transcript) and/or the process by which a transcribed mRNA is subsequently translated into peptides, polypeptides, or proteins. Transcripts and encoded polypeptides may be collectively referred to as "gene product." If the polynucleotide is derived from genomic DNA, expression may include splicing of the mRNA in a eukaryotic cell.

[00101] As used herein, "operably linked", "operable linkage", "operatively linked", or grammatical equivalents thereof generally refer to juxtaposition of genetic elements, e.g., a promoter, an enhancer, a polyadenylation sequence, etc., wherein the elements are in a relationship permitting them to operate in the expected manner. For instance, a regulatory

element, which may comprise promoter and/or enhancer sequences, is operatively linked to a coding region if the regulatory element helps initiate transcription of the coding sequence. There may be intervening residues between the regulatory element and coding region so long as this functional relationship is maintained.

[00102] A "vector" as used herein, generally refers to a macromolecule or association of macromolecules that comprises or associates with a polynucleotide and which may be used to mediate delivery of the polynucleotide to a cell. Examples of vectors include plasmids, viral vectors, liposomes, and other gene delivery vehicles. The vector generally comprises genetic elements, e.g., regulatory elements, operatively linked to a gene to facilitate expression of the gene in a target.

[00103] As used herein, "an expression cassette" and "a nucleic acid cassette" are used interchangeably generally to refer to a combination of nucleic acid sequences or elements that are expressed together or are operably linked for expression. In some cases, an expression cassette refers to the combination of regulatory elements and a gene or genes to which they are operably linked for expression.

[00104] A "functional fragment" of a DNA or protein sequence generally refers to a fragment that retains a biological activity (either functional or structural) that is substantially similar to a biological activity of the full-length DNA or protein sequence. A biological activity of a DNA sequence may be its ability to influence expression in a manner known to be attributed to the full-length sequence.

[00105] As used herein, an "engineered" object generally indicates that the object has been modified by human intervention. According to non-limiting examples: a nucleic acid may be modified by changing its sequence to a sequence that does not occur in nature; a nucleic acid may be modified by ligating it to a nucleic acid that it does not associate with in nature such that the ligated product possesses a function not present in the original nucleic acid; an engineered nucleic acid may synthesized in vitro with a sequence that does not exist in nature; a protein may be modified by changing its amino acid sequence to a sequence that does not exist in nature; an engineered protein may acquire a new function or property. An "engineered" system comprises at least one engineered component.

[00106] As used herein, "synthetic" and "artificial" can generally be used interchangeably to refer to a protein or a domain thereof that has low sequence identity (e.g., less than 50% sequence identity, less than 25% sequence identity, less than 10% sequence identity, less than 5% sequence identity, less than 1% sequence identity) to a naturally occurring human protein. For example, VPR and VP64 domains are synthetic transactivation domains.

[00107] As used herein, the term "Cas12a" generally refers to a family of Cas endonucleases that

are class 2, Type V-A Cas endonucleases and that (a) use a relatively small guide RNA (about 42-44 nucleotides) that is processed by the nuclease itself following transcription from the CRISPR array, and (b) cleave DNA to leave staggered cut sites. Further features of this family of enzymes can be found, e.g. in Zetsche B, Heidenreich M, Mohanraju P, et al. Nat Biotechnol 2017;35:31–34, and Zetsche B, Gootenberg JS, Abudayyeh OO, et al. Cell 2015;163:759–771, which are incorporated by reference herein.

[00108] As used herein, a "guide nucleic acid" can generally refer to a nucleic acid that may hybridize to another nucleic acid. A guide nucleic acid may be RNA. A guide nucleic acid may be DNA. The guide nucleic acid may be programmed to bind to a sequence of nucleic acid site-specifically. The nucleic acid to be targeted, or the target nucleic acid, may comprise nucleotides. The guide nucleic acid may comprise nucleotides. A portion of the target nucleic acid may be complementary to a portion of the guide nucleic acid. The strand of a double-stranded target polynucleotide that is complementary to and hybridizes with the guide nucleic acid may be called the complementary strand. The strand of the double-stranded target polynucleotide that is complementary to the complementary strand, and therefore may not be complementary to the guide nucleic acid may be called noncomplementary strand. A guide nucleic acid may comprise a polynucleotide chain and can be called a "single guide nucleic acid." A guide nucleic acid may comprise two polynucleotide chains and may be called a "double guide nucleic acid." If not otherwise specified, the term "guide nucleic acid" may be inclusive, referring to both single guide nucleic acids and double guide nucleic acids. A guide nucleic acid may comprise a segment that can be referred to as a "nucleic acid-targeting segment" or a "nucleic acid-targeting sequence" or "spacer sequence." A nucleic acid-targeting segment may comprise a sub-segment that may be referred to as a "protein binding segment" or "protein binding sequence" or "Cas protein binding segment".

[00109] The term "sequence identity" or "percent identity" in the context of two or more nucleic acids or polypeptide sequences, generally refers to two (e.g., in a pairwise alignment) or more (e.g., in a multiple sequence alignment) sequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence over a local or global comparison window, as measured using a sequence comparison algorithm. Suitable sequence comparison algorithms for polypeptide sequences include, e.g., BLASTP using parameters of a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix setting gap costs at existence of 11, extension of 1, and using a conditional compositional score matrix adjustment for polypeptide sequences longer than 30 residues; BLASTP using parameters of a wordlength (W) of 2, an expectation (E) of 1000000, and the PAM30 scoring matrix setting gap costs at 9 to open gaps and 1 to extend gaps

for sequences of less than 30 residues (these are the default parameters for BLASTP in the BLAST suite available at https://blast.ncbi.nlm.nih.gov); CLUSTALW with the Smith-Waterman homology search algorithm parameters with a match of 2, a mismatch of -1, and a gap of -1; MUSCLE with default parameters; MAFFT with parameters of a retree of 2 and max iterations of 1000; Novafold with default parameters; HMMER hmmalign with default parameters.

[00110] The term "optimally aligned" in the context of two or more nucleic acids or polypeptide sequences, generally refers to two (e.g., in a pairwise alignment) or more (e.g., in a multiple sequence alignment) sequences that have been aligned to maximal correspondence of amino acids residues or nucleotides, for example, as determined by the alignment producing a highest or "optimized" percent identity score.

[00111] Included in the current disclosure are variants of any of the enzymes described herein with one or more conservative amino acid substitutions. Such conservative substitutions can be made in the amino acid sequence of a polypeptide without disrupting the three-dimensional structure or function of the polypeptide. Conservative substitutions can be accomplished by substituting amino acids with similar hydrophobicity, polarity, and R chain length for one another. Additionally, or alternatively, by comparing aligned sequences of homologous proteins from different species, conservative substitutions can be identified by locating amino acid residues that have been mutated between species (e.g., non-conserved residues) without altering the basic functions of the encoded proteins. Such conservatively substituted variants may include variants with at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99% identity to any one of the endonuclease protein sequences described herein (e.g. MG90, MG91A, MG91B, MG91C, MG118, MG119, MG120, MG122, or MG126 family endonucleases described herein, or any other family nuclease described herein). In some embodiments, such conservatively substituted variants are functional variants. Such functional variants can encompass sequences with substitutions such that the activity of one or more critical active site residues or guide RNA binding residues of the endonuclease are not disrupted. In some embodiments, a functional variant of any of the proteins described herein lacks substitution of at least one of the conserved or functional residues called out in **FIGs. 2A**, **3A**, **4A**, **5A**, or **6A**. In some embodiments, a functional variant of any of the proteins described herein lacks substitution of all of the conserved or functional residues called out in **FIGs. 2A**, **3A**, **4A**, **5A**, or **6A**.

[00112] Also included in the current disclosure are variants of any of the enzymes described

herein with substitution of one or more catalytic residues to decrease or eliminate activity of the enzyme (e.g. decreased-activity variants). In some embodiments, a decreased activity variant as a protein described herein comprises a disrupting substitution of at least one, at least two, or all three catalytic residues called out in **FIGs. 2A**, **3A**, **4A**, **5A**, or **6A**.

[00113] Conservative substitution tables providing functionally similar amino acids are available from a variety of references (see, for e.g., Creighton, Proteins: Structures and Molecular Properties (W H Freeman & Co.; 2nd edition (December 1993)). The following eight groups each contain amino acids that are conservative substitutions for one another:

1) Alanine (A), Glycine (G);
2) Aspartic acid (D), Glutamic acid (E);
3) Asparagine (N), Glutamine (Q);
4) Arginine (R), Lysine (K);
5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V);
6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W);
7) Serine (S), Threonine (T); and
8) Cysteine (C), Methionine (M)

Overview

[00114] The discovery of new Cas enzymes with unique functionality and structure may offer the potential to further disrupt deoxyribonucleic acid (DNA) editing technologies, improving speed, specificity, functionality, and ease of use. Relative to the predicted prevalence of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) systems in microbes and the sheer diversity of microbial species, relatively few functionally characterized CRISPR/Cas enzymes exist in the literature. This is partly because a huge number of microbial species may not be readily cultivated in laboratory conditions. Metagenomic sequencing from natural environmental niches containing large numbers of microbial species may offer the potential to drastically increase the number of new CRISPR/Cas systems known and speed the discovery of new oligonucleotide editing functionalities. A recent example of the fruitfulness of such an approach is demonstrated by the 2016 discovery of CasX/CasY CRISPR systems from metagenomic analysis of natural microbial communities.

[00115] CRISPR/Cas systems are RNA-directed nuclease complexes that have been described to function as an adaptive immune system in microbes. In their natural context, CRISPR/Cas systems occur in CRISPR (clustered regularly interspaced short palindromic repeats) operons or loci, which generally comprise two parts: (i) an array of short repetitive sequences (30-40bp) separated by equally short spacer sequences, which encode the RNA-based targeting element;

and (ii) ORFs encoding the Cas encoding the nuclease polypeptide directed by the RNA-based targeting element alongside accessory proteins/enzymes. Efficient nuclease targeting of a particular target nucleic acid sequence generally requires both (i) complementary hybridization between the first 6-8 nucleic acids of the target (the target seed) and the crRNA guide; and (ii) the presence of a protospacer-adjacent motif (PAM) sequence within a defined vicinity of the target seed (the PAM usually being a sequence not commonly represented within the host genome). Depending on the exact function and organization of the system, CRISPR-Cas systems are commonly organized into 2 classes, 5 types, and 16 subtypes based on shared functional characteristics and evolutionary similarity (see **FIG. 1**).

[00116] Class I CRISPR-Cas systems have large, multi-subunit effector complexes, and comprise Types I, III, and IV. Class II CRISPR-Cas systems generally have single-polypeptide multidomain nuclease effectors, and comprise Types II, V and VI.

[00117] Type II CRISPR-Cas systems are considered the simplest in terms of components. In Type II CRISPR-Cas systems, the processing of the CRISPR array into mature crRNAs does not require the presence of a special endonuclease subunit, but rather a small trans-encoded crRNA (tracrRNA) with a region complementary to the array repeat sequence; the tracrRNA interacts with both its corresponding effector nuclease (e.g. Cas9) and the repeat sequence to form a precursor dsRNA structure, which is cleaved by endogenous RNAse III to generate a mature effector enzyme loaded with both tracrRNA and crRNA. Cas II nucleases are known as DNA nucleases. Type 2 effectors generally exhibit a structure consisting of a RuvC-like endonuclease domain that adopts the RNase H fold with an unrelated HNH nuclease domain inserted within the folds of the RuvC-like nuclease domain. The RuvC-like domain is responsible for the cleavage of the target (e.g., crRNA complementary) DNA strand, while the HNH domain is responsible for cleavage of the displaced DNA strand.

[00118] Type V CRISPR-Cas systems are characterized by a nuclease effector (e.g. Cas12) structure similar to that of Type II effectors, comprising a RuvC-like domain. Similar to Type II, most (but not all) Type V CRISPR systems use a tracrRNA to process pre-crRNAs into mature crRNAs; however, unlike Type II systems which requires RNAse III to cleave the pre-crRNA into multiple crRNAs, type V systems are capable of using the effector nuclease itself to cleave pre-crRNAs. Like Type-II CRISPR-Cas systems, Type V CRISPR-Cas systems are again known as DNA nucleases. Unlike Type II CRISPR-Cas systems, some Type V enzymes (e.g., Cas12a) appear to have a robust single-stranded nonspecific deoxyribonuclease activity that is activated by the first crRNA directed cleavage of a double-stranded target sequence.

[00119] CRISPR-Cas systems have emerged in recent years as the gene editing technology of choice due to their targetability and ease of use. The most commonly used systems are the Class

2 Type II SpCas9 and the Class 2 Type V-A Cas12a (previously Cpf1). The Type V-A systems in particular are becoming more widely used since their reported specificity in cells is higher than other nucleases, with fewer or no off-target effects. The V-A systems are also advantageous in that the guide RNA is small (42-44 nucleotides compared with approximately 100 nt for SpCas9) and is processed by the nuclease itself following transcription from the CRISPR array, simplifying multiplexed applications with multiple gene edits. Furthermore, the V-A systems have staggered cut sites, which may facilitate directed repair pathways, such as microhomology-dependent targeted integration (MITI).

[00120] The most commonly used Type V-A enzymes require a 5' protospacer adjacent motif (PAM) next to the chosen target site: 5'-TTTV-3' for *Lachnospiraceae* bacterium ND2006 LbCas12a and *Acidaminococcus sp.* AsCas12a; and 5'-TTV-3' for *Francisella novicida* FnCas12a. Recent exploration of orthologs has revealed proteins with less restrictive PAM sequences that are also active in mammalian cell culture, for example YTV, YYN or TTN. However, these enzymes do not fully encompass Type V biodiversity and targetability, and may not represent all possible activities and PAM sequence requirements. Here, thousands of genomic fragments were mined from numerous metagenomes for Type V nucleases. The known diversity of V enzymes may have been expanded and novel systems may have been developed into highly targetable, compact, and precise gene editing agents.

MG Enzymes

[00121] Type V CRISPR systems are quickly being adopted for use in a variety of genome editing applications. These programmable nucleases are part of adaptive microbial immune systems, the natural diversity of which has been largely unexplored. Novel families of Type V CRISPR enzymes were identified through a large-scale analysis of metagenomes collected from a variety of complex environments, and representatives of these were developed systems into gene-editing platforms. The majority of these systems come from uncultivated organisms, some of which encode a divergent Type V effector within the same CRISPR operon.

[00122] In some aspects, the present disclosure provides for novel Type V candidates. These candidates may represent one or more novel subtypes and some sub-families may have been identified. These nucleases are less than about 900 amino acids in length. These novel subtypes may be found in the same CRISPR locus as known Type V effectors. RuvC catalytic residues may have been identified for the novel Type V candidates, and these novel Type V candidates may not require tracrRNA.

[00123] In some aspects, the present disclosure provides for smaller Type V effectors. Such effectors may be small putative effectors. These effectors may simplify delivery and may extend

therapeutic applications.

[00124] In some aspects, the present disclosure provides for a novel type V effector. Such an effector may be MG90 as described herein (see **FIGs. 3A-3C**). Such an effector may be MG91 as described herein (see **FIGs. 8A-8B**). Such an effector may be MG118 as described herein (see **FIGs. 5A-5C**). Such an effector may be MG119 as described herein (see **FIGs. 2A-2D**). Such an effector may be MG120 as described herein (see **FIGs. 7A-7C**). Such an effector may be MG122 as described herein (see **FIGs. 6A-6C**). Such an effector may be MG126 as described herein (see **FIGs. 4A-4C**).

[00125] In one aspect, the present disclosure provides for an engineered nuclease system discovered through metagenomic sequencing. In some cases, the metagenomic sequencing is conducted on samples. In some cases, the samples may be collected from a variety of environments. Such environments may be a human microbiome, an animal microbiome, environments with high temperatures, environments with low temperatures. Such environments may include sediment.

[00126] In one aspect, the present disclosure provides for an engineered nuclease system comprising an endonuclease. In some cases, the endonuclease is a Cas endonuclease. In some cases, the endonuclease is a class 2, type V Cas endonuclease. In some cases, the endonuclease is a class 2, type V Cas endonuclease of a novel sub-type. In some cases, the endonuclease is derived from an uncultivated microorganism. The endonuclease may comprise a RuvC domain. In some cases, the engineered nuclease system comprises an engineered guide RNA. In some cases, the engineered guide RNA is configured to form a complex with the endonuclease. In some cases, the engineered guide RNA comprises a spacer sequence. In some cases, the spacer sequence is configured to hybridize to a target nucleic acid sequence.

[00127] In one aspect, the present disclosure provides for an engineered nuclease system comprising an endonuclease. In some cases, the endonuclease has at least about 70% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629. In some cases, the endonuclease has at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629.

[00128] In some cases, the endonuclease comprises a variant having at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least

about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629. In some cases, the endonuclease may be substantially identical to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629.

[00129] In some cases, the engineered nuclease system comprises an engineered guide RNA. In some cases, the engineered guide RNA is configured to form a complex with the endonuclease. In some cases, the engineered guide RNA comprises a spacer sequence. In some cases, the spacer sequence is configured to hybridize to a target nucleic acid sequence. In some cases, the endonuclease is configured to bind to a protospacer adjacent motif (PAM) sequence.

[00130] In some cases, the endonuclease is not a Cpf1 or Cms1 endonuclease.

[00131] In some cases, the guide RNA comprises a sequence with at least 80% sequence identity to the first 19 nucleotides or the non-degenerate nucleotides of SEQ ID NO: 410-419. In some cases, the guide RNA comprises a sequence with at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% identity to the first 19 nucleotides or the non-degenerate nucleotides of SEQ ID NO: 410-419. In some cases, the guide RNA comprises a variant having at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% identity to the first 19 nucleotides or the non-degenerate nucleotides of SEQ ID NO: 410-419. In some cases, the guide RNA comprises a sequence which is substantially identical to the first 19 nucleotides or the non-degenerate nucleotides of SEQ ID NO: 410-419.

[00132] In some cases, the guide RNA comprises a sequence with at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% identity to the first 19 nucleotides or the non-degenerate nucleotides of SEQ ID NO: 410-419. In some cases, the endonuclease is

configured to bind to the engineered guide RNA. In some cases, the Cas endonuclease is configured to bind to the engineered guide RNA. In some cases, the class 2 Cas endonuclease is configured to bind to the engineered guide RNA. In some cases, the class 2, type V Cas endonuclease is configured to bind to the engineered guide RNA. In some cases, the class 2, type V, novel subtype Cas endonuclease is configured to bind to the engineered guide RNA.

[00133] In some cases, the guide RNA comprises a sequence complementary to a eukaryotic, fungal, plant, mammalian, or human genomic polynucleotide sequence. In some cases, the guide RNA comprises a sequence complementary to a eukaryotic genomic polynucleotide sequence. In some cases, the guide RNA comprises a sequence complementary to a fungal genomic polynucleotide sequence. In some cases, the guide RNA comprises a sequence complementary to a plant genomic polynucleotide sequence. In some cases, the guide RNA comprises a sequence complementary to a mammalian genomic polynucleotide sequence. In some cases, the guide RNA comprises a sequence complementary to a human genomic polynucleotide sequence.

[00134] In some cases, the guide RNA is 30-250 nucleotides in length. In some cases, the guide RNA is 42-44 nucleotides in length. In some cases, the guide RNA is 42 nucleotides in length. In some cases, the guide RNA is 43 nucleotides in length. In some cases, the guide RNA is 44 nucleotides in length. In some cases, the guide RNA is 85-245 nucleotides in length. In some cases, the guide RNA is more than 90 nucleotides in length. In some cases, the guide RNA is less than 245 nucleotides in length.

[00135] In some cases, the endonuclease may comprise a variant having one or more nuclear localization sequences (NLSs). The NLS may be proximal to the N- or C-terminus of the endonuclease. The NLS may be appended N-terminal or C-terminal to any one of SEQ ID NOs: 630-645, or to a variant having at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% identity to any one of SEQ ID NOs: 630-645. In some cases, the NLS may comprise a sequence substantially identical to any one of SEQ ID NOs: 630-645.

Table 1: Example NLS Sequences that may be used with Cas Effectors according to the disclosure.

Source	NLS amino acid sequence	SEQ ID NO:
SV40	PKKKRKV	630
nucleoplasmin bipartite NLS	KRPAATKKAGQAKKKK	631

Source	NLS amino acid sequence	SEQ ID NO:
c-myc NLS	PAAKRVKLD	632
c-myc NLS	RQRRNELKRSP	633
hRNPA1 M9 NLS	NQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGY	634
Importin-alpha IBB domain	RMRIZFKNKGKDTAELRRRRVEVSVELRKAKKDEQILKRRNV	635
Myoma T protein	VSRKRPRP	636
Myoma T protein	PPKKARED	637
p53	PQPKKKPL	638
mouse c-abl IV	SALIKKKKKMAP	639
influenza virus NS1	DRLRR	640
influenza virus NS1	PKQKKRK	641
Hepatitis virus delta antigen	RKLKKKIKKL	642
mouse Mx1 protein	REKKKFLKRR	643
human poly(ADP-ribose) polymerase	KRKGDEVDGVDEVAKKKSKK	644
steroid hormone receptors (human) glucocorticoid	RKCLQAGMNLEARKTKK	645

[00136] In some cases, the engineered nuclease system further comprises a single- or double stranded DNA repair template. In some cases, the engineered nuclease system further comprises a single-stranded DNA repair template. In some cases, the engineered nuclease system further comprises a double-stranded DNA repair template. In some cases, the single- or double-stranded DNA repair template may comprise from 5' to 3': a first homology arm comprising a sequence of at least 20 nucleotides 5′ to said target deoxyribonucleic acid sequence, a synthetic DNA sequence of at least 10 nucleotides, and a second homology arm comprising a sequence of at least 20 nucleotides 3′ to said target sequence.

[00137] In some cases, the first homology arm comprises a sequence of at least 40, at least 50, at least 60, at least 70, at least 80, at least 90, at least 100, at least 110, at least 120, at least 130, at least 140, at least 150, at least 175, at least 200, at least 250, at least 300, at least 400, at least 500, at least 750, or at least 1000 nucleotides. In some cases, the second homology arm comprises a sequence of at least 40, at least 50, at least 60, at least 70, at least 80, at least 90, at least 100, at least 110, at least 120, at least 130, at least 140, at least 150, at least 175, at least 200, at least 250, at least 300, at least 400, at least 500, at least 750, or at least 1000 nucleotides.

[00138] In some cases, the first and second homology arms are homologous to a genomic sequence of a prokaryote. In some cases, the first and second homology arms are homologous to a genomic sequence of a bacteria. In some cases, the first and second homology arms are homologous to a genomic sequence of a fungus. In some cases, the first and second homology arms are homologous to a genomic sequence of a eukaryote.

[00139] In some cases, the engineered nuclease system further comprises a DNA repair

template. The DNA repair template may comprise a double-stranded DNA segment. The double-stranded DNA segment may be flanked by one single-stranded DNA segment. The double-stranded DNA segment may be flanked by two single-stranded DNA segments. In some cases, the single-stranded DNA segments are conjugated to the 5' ends of the double-stranded DNA segment. In some cases, the single stranded DNA segments are conjugated to the 3' ends of the double-stranded DNA segment.

[00140] In some cases, the single-stranded DNA segments have a length from 1 to 15 nucleotide bases. In some cases, the single-stranded DNA segments have a length from 4 to 10 nucleotide bases. In some cases, the single-stranded DNA segments have a length of 4 nucleotide bases. In some cases, the single-stranded DNA segments have a length of 5 nucleotide bases. In some cases, the single-stranded DNA segments have a length of 6 nucleotide bases. In some cases, the single-stranded DNA segments have a length of 7 nucleotide bases. In some cases, the single-stranded DNA segments have a length of 8 nucleotide bases. In some cases, the single-stranded DNA segments have a length of 9 nucleotide bases. In some cases, the single-stranded DNA segments have a length of 10 nucleotide bases.

[00141] In some cases, the single-stranded DNA segments have a nucleotide sequence complementary to a sequence within the spacer sequence. In some cases, the double-stranded DNA sequence comprises a barcode, an open reading frame, an enhancer, a promoter, a protein-coding sequence, a miRNA coding sequence, an RNA coding sequence, or a transgene.

[00142] In some cases, the engineered nuclease system further comprises a source of Mg^{2+}.

[00143] In some cases, the guide RNA comprises a hairpin comprising at least 8 base-paired ribonucleotides. In some cases, the guide RNA comprises a hairpin comprising at least 9 base-paired ribonucleotides. In some cases, the guide RNA comprises a hairpin comprising at least 10 base-paired ribonucleotides. In some cases, the guide RNA comprises a hairpin comprising at least 11 base-paired ribonucleotides. In some cases, the guide RNA comprises a hairpin comprising at least 12 base-paired ribonucleotides.

[00144] In some cases, the endonuclease comprises a sequence at least 70% identical to a variant of any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof. In some cases, the endonuclease comprises a sequence at least 75% identical to a variant of any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof. In some cases, the endonuclease comprises a sequence at least 80% identical to a variant of any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof. In some cases, the endonuclease comprises a sequence at least 85% identical to a variant of any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof. In some cases, the endonuclease comprises a sequence at least 90% identical to a variant of any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof. In some

cases, the endonuclease comprises a sequence at least 95% identical to a variant of any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof.

[00145] In some cases, sequence may be determined by a BLASTP, CLUSTALW, MUSCLE, or MAFFT algorithm, or a CLUSTALW algorithm with the Smith-Waterman homology search algorithm parameters. The sequence identity may be determined by said BLASTP homology search algorithm using parameters of a wordlength (W) of 3, an expectation (E) of 10, and a BLOSUM62 scoring matrix setting gap costs at existence of 11, extension of 1, and using a conditional compositional score matrix adjustment.

[00146] In one aspect, the present disclosure provides an engineered guide RNA comprising a DNA-targeting segment. In some cases, the DNA-targeting segment comprises a nucleotide sequence that is complementary to a target sequence. In some cases, the target sequence is in a target DNA molecule. In some cases, the engineered guide RNA comprises a protein-binding segment. In some cases, the protein-binding segment comprises two complementary stretches of nucleotides. In some cases, the two complementary stretches of nucleotides hybridize to form a double-stranded RNA (dsRNA) duplex. In some cases, the two complementary stretches of nucleotides are covalently linked to one another with intervening nucleotides. In some cases, the engineered guide ribonucleic acid polynucleotide is capable of forming a complex with an endonuclease. In some cases, the endonuclease has at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629. In some cases, the complex targets the target sequence of the target DNA molecule. In some cases, the DNA-targeting segment is positioned 3' of both of the two complementary stretches of nucleotides.

[00147] In some cases, the double-stranded RNA (dsRNA) duplex comprises at least 8 ribonucleotides. In some cases, the double-stranded RNA (dsRNA) duplex comprises at least 9 ribonucleotides. In some cases, the double-stranded RNA (dsRNA) duplex comprises at least 10 ribonucleotides. In some cases, the double-stranded RNA (dsRNA) duplex comprises at least 11 ribonucleotides. In some cases, the double-stranded RNA (dsRNA) duplex comprises at least 12 ribonucleotides.

[00148] In some cases, the deoxyribonucleic acid polynucleotide encodes the engineered guide ribonucleic acid polynucleotide.

[00149] In one aspect, the present disclosure provides a nucleic acid comprising an engineered

nucleic acid sequence. In some cases, the engineered nucleic acid sequence is optimized for expression in an organism. In some cases, the nucleic acid encodes an endonuclease. In some cases, the endonuclease is a Cas endonuclease. In some cases, the endonuclease is a class 2 endonuclease. In some cases, the endonuclease is a class2, type V Cas endonuclease. In some cases, the endonuclease is a class2, type V, novel subtype Cas endonuclease. In some cases, the endonuclease is derived from an uncultivated microorganism. In some cases, the organism is not the uncultivated organism.

[00150] In some cases, the endonuclease comprises a variant having at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629.

[00151] In some cases, the endonuclease may comprise a variant having one or more nuclear localization sequences (NLSs). The NLS may be proximal to the N- or C-terminus of the endonuclease. The NLS may be appended N-terminal or C-terminal to any one of SEQ ID NOs: 630-645, or to a variant having at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% sequence identity to any one of SEQ ID NOs: 630-645.

[00152] In some cases, the organism is prokaryotic. In some cases, the organism is bacterial. In some cases, the organism is eukaryotic. In some cases, the organism is fungal. In some cases, the organism is a plant. In some cases, the organism is mammalian. In some cases, the organism is a rodent. In some cases, the organism is human.

[00153] In one aspect, the present disclosure provides an engineered vector. In some cases, the engineered vector comprises a nucleic acid sequence encoding an endonuclease. In some cases, the endonuclease is a Cas endonuclease. In some cases, the endonuclease is a class 2 Cas endonuclease. In some cases, the endonuclease is a class 2, type V Cas endonuclease. In some cases, the endonuclease is a class2, type V, novel subtype Cas endonuclease. In some cases, the endonuclease is derived from an uncultivated microorganism.

[00154] In some cases, the engineered vector comprises a nucleic acid described herein. In some cases, the nucleic acid described herein is a deoxyribonucleic acid polynucleotide described

herein. In some cases, the vector is a plasmid, a minicircle, a CELiD, an adeno-associated virus (AAV) derived virion, or a lentivirus.

[00155] In one aspect, the present disclosure provides a cell comprising a vector described herein.

[00156] In one aspect, the present disclosure provides a method of manufacturing an endonuclease. In some cases, the method comprises cultivating the cell.

[00157] In one aspect, the present disclosure provides a method for binding, cleaving, marking, or modifying a double-stranded deoxyribonucleic acid polynucleotide. The method may comprise contacting the double-stranded deoxyribonucleic acid polynucleotide with an endonuclease. In some cases, the endonuclease is a Cas endonuclease. In some cases, the endonuclease is a class 2 Cas endonuclease. In some cases, the endonuclease is a class 2, type V Cas endonuclease. In some cases, the endonuclease is a class2, type V, novel subtype Cas endonuclease. In some cases, the endonuclease is in complex with an engineered guide RNA. In some cases, the engineered guide RNA is configured to bind to the endonuclease. In some cases, the engineered guide RNA is configured to bind to the double-stranded deoxyribonucleic acid polynucleotide. In some cases, the engineered guide RNA is configured to bind to the endonuclease and to the double-stranded deoxyribonucleic acid polynucleotide. In some cases, the double-stranded deoxyribonucleic acid polynucleotide comprises a protospacer adjacent motif (PAM).

[00158] In some cases, the double-stranded deoxyribonucleic acid polynucleotide comprises a first strand comprising a sequence complementary to a sequence of the engineered guide RNA and a second strand comprising the PAM. In some cases, the PAM is directly adjacent to the 5′ end of the sequence complementary to the sequence of the engineered guide RNA. In some cases, the endonuclease is not a Cpf1 endonuclease or a Cms1 endonuclease. In some cases, the endonuclease is derived from an uncultivated microorganism. In some cases, the double-stranded deoxyribonucleic acid polynucleotide is a eukaryotic, plant, fungal, mammalian, rodent, or human double-stranded deoxyribonucleic acid polynucleotide.

[00159] In one aspect, the present disclosure provides a method of modifying a target nucleic acid locus. The method may comprise delivering to the target nucleic acid locus the engineered nuclease system described herein. In some cases, the endonuclease is configured to form a complex with the engineered guide ribonucleic acid structure. In some cases, the complex is configured such that upon binding of the complex to the target nucleic acid locus, the complex modifies the target nucleic acid locus.

[00160] In some cases, modifying the target nucleic acid locus comprises binding, nicking, cleaving, or marking said target nucleic acid locus. In some cases, the target nucleic acid locus

comprises deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). In some cases, the target nucleic acid comprises genomic DNA, viral DNA, viral RNA, or bacterial DNA. In some cases, the target nucleic acid locus is *in vitro*. In some cases, the target nucleic acid locus is within a cell. In some cases, the cell is a prokaryotic cell, a bacterial cell, a eukaryotic cell, a fungal cell, a plant cell, an animal cell, a mammalian cell, a rodent cell, a primate cell, or a human cell.

[00161] In some cases, delivery of the engineered nuclease system to the target nucleic acid locus comprises delivering the nucleic acid described herein or the vector described herein. In some cases, delivery of engineered nuclease system to the target nucleic acid locus comprises delivering a nucleic acid comprising an open reading frame encoding the endonuclease. In some cases, the nucleic acid comprises a promoter. In some cases, the open reading frame encoding the endonuclease is operably linked to the promoter.

[00162] In some cases, delivery of the engineered nuclease system to the target nucleic acid locus comprises delivering a capped mRNA containing the open reading frame encoding the endonuclease. In some cases, delivery of the engineered nuclease system to the target nucleic acid locus comprises delivering a translated polypeptide. In some cases, delivery of the engineered nuclease system to the target nucleic acid locus comprises delivering a deoxyribonucleic acid (DNA) encoding the engineered guide RNA operably linked to a ribonucleic acid (RNA) pol III promoter.

[00163] In some cases, the endonuclease induces a single-stranded break or a double-stranded break at or proximal to the target locus. In some cases, the endonuclease induces a staggered **single stranded break within or 3′ to said target locus.**

[00164] In some cases, effector repeat motifs are used to inform guide design of MG nucleases. For example, the processed gRNA in Type V systems consists of the last 20-22 nucleotides of a CRISPR repeat. This sequence may be synthesized into a crRNA (along with a spacer) and tested *in vitro*, along with the synthesized nucleases, for cleavage on a library of possible targets. Using this method, the PAM may be determined. In some cases, Type V enzymes may use a "universal" gRNA. In some cases, Type V enzymes may need a unique gRNA.

[00165] Systems of the present disclosure may be used for various applications, such as, for example, nucleic acid editing (e.g., gene editing), binding to a nucleic acid molecule (e.g., sequence-specific binding). Such systems may be used, for example, for addressing (e.g., removing or replacing) a genetically inherited mutation that may cause a disease in a subject, inactivating a gene in order to ascertain its function in a cell, as a diagnostic tool to detect disease-causing genetic elements (e.g. via cleavage of reverse-transcribed viral RNA or an amplified DNA sequence encoding a disease-causing mutation), as deactivated enzymes in combination with a probe to target and detect a specific nucleotide sequence (e.g. sequence

encoding antibiotic resistance int bacteria), to render viruses inactive or incapable of infecting host cells by targeting viral genomes, to add genes or amend metabolic pathways to engineer organisms to produce valuable small molecules, macromolecules, or secondary metabolites, to establish a gene drive element for evolutionary selection, to detect cell perturbations by foreign small molecules and nucleotides as a biosensor.

EXAMPLES

[00166] In accordance with IUPAC conventions, the following abbreviations are used throughout the examples:

A = adenine
C = cytosine
G = guanine
T = thymine
R = adenine or guanine
Y = cytosine or thymine
S = guanine or cytosine
W = adenine or thymine
K = guanine or thymine
M = adenine or cytosine
B = C, G, or T
D = A, G, or T
H = A, C, or T
V = A, C, or G

Example 1 – A method of metagenomic analysis for new proteins

[00167] Metagenomic samples were collected from sediment, soil, and animals. Deoxyribonucleic acid (DNA) was extracted with a Zymobiomics DNA mini-prep kit and sequenced on an Illumina HiSeq® 2500. Samples were collected with consent of property owners. Additional raw sequence data from public sources included animal microbiomes, sediment, soil, hot springs, hydrothermal vents, marine, peat bogs, permafrost, and sewage sequences. Metagenomic sequence data was searched using Hidden Markov Models generated based on known Cas protein sequences including class II type V Cas effector proteins to identify new Cas effectors. Novel effector proteins identified by the search were aligned to known proteins to identify potential active sites. This metagenomic workflow resulted in the delineation of the MG90, MG91A, MG91B, MG91C, MG118, MG119, MG120, MG122, and MG126 families described herein.

Example 2 – Discovery of MG90, MG91A, MG91B, MG91C, MG118, MG119, MG120, MG122, and MG126 Families of CRISPR systems

[00168] Analysis of the data from the metagenomic analysis of Example 1 revealed new clusters of previously undescribed putative CRISPR systems comprising 9 families (MG90, MG91A,

MG91B, MG91C, MG118, MG119, MG120, MG122, and MG126). The corresponding protein and nucleic acid sequences for these new enzymes and their exemplary subdomains are presented as SEQ ID NOs: 1-325, 420-431, 476-624, or 629.

Example 3 – Template DNA for Transcription and Translation

[00169] E coli codon optimized sequences of all MG VU and CasPhi nucleases were ordered (Twist Biosciences) in a plasmid with a T7 promoter. Linear templates were amplified from the plasmids by PCR to include the T7 and nuclease sequence. Minimal array linear templates were amplified from sequences composed of a T7 promoter, native repeat, universal spacer, and native repeat, flanked by adapter sequences for amplification. The universal spacer matches the spacer in an 8N target library, where there are 8N mixed bases adjacent to the spacer for PAM determination. Three intergenic sequences near the ORF or CRISPR array were identified from the metagenomic contigs and ordered as gBlocks with flanking adapter sequences for amplification (Integrated DNA Technologies).

Example 4 – In vitro transcription of crRNA, Minimal Arrays, and sgRNA

[00170] RNA was produced by *in vitro* transcription using HiScribe™ T7 High Yield RNA Synthesis Kit and purified using the Monarch® RNA Cleanup Kit (New England Biolabs Inc.). Templates for T7 transcription varied. For crRNA, DNA oligos were designed with a T7 promoter, trimmed native repeat, and universal spacer. For minimal arrays the same templates as described above were used. For sgRNA, DNA ultramers were designed with a T7 promoter, trimmed tracrRNA, GAAA tetraloop, trimmed native repeat, and universal spacer. Minimal array templates were amplified with adapter primers. The crRNA and sgRNA templates were ordered as reverse complements and annealed with a primer with the T7 promoter sequence in 1X IDT duplex buffer at 95 °C for two minutes followed by cooling to 22 °C at 0.1 °C/second to produce a hybrid ds/ssDNA substrate suitable for transcription. After transcription, but prior to cleaning, each reaction was treated with DNAse I and incubated at 37 °C for 15 minutes. All transcription products were verified for yield and purity via RNA TapeStation or via a denaturing urea PAGE gel.

Example 5 – TXTL Expression

[00171] Nucleases, intergenic sequences, and minimal arrays were expressed in transcription-translation reaction mixtures using myTXTL®Sigma 70 Master Mix Kit (Arbor Biosciences). The final reaction mixtures contained 5 nM nuclease DNA template, 12 nM intergenic DNA template, 15 nM minimal array DNA template, 0.1 nM pTXTL-P70a-T7rnap, and 1X of

myTXTL®Sigma 70 Master Mix. The reactions were incubated at 29 °C for 16 hours then stored at 4 °C.

Example 6 – PURExpress Expressions

[00172] 10 nM of nuclease PCR templates were expressed at 37 °C for 3 hours with PURExpress® In Vitro Protein Synthesis Kit (New England Biolabs Inc.) for cleavage with *in vitro* transcribed RNA. These reactions were used to test in vitro cleavage with 50 nM sgRNA or minimal array RNA following the same procedure as described in the cleavage reactions section.

Example 7 – *E. coli* Expressions

[00173] Plasmids encoding the effector, intergenic sequence from the genomic contig, native repeat, and universal spacer sequences with a T7 promoter were transformed into BL21 DE3 or T7 Express lysY/Iq and cultured at 37 °C in 60 mL terrific broth media supplemented with 100 µg/mL of ampicillin. Expression was induced with 0.4 mM IPTG after cultures reached OD_{600nm} of 0.5 and incubated at 16 °C overnight. 25 mL of cells were pelleted by centrifugation and resuspended in 1.5 mL of lysis buffer (20 mM Tris-HCl, 500 mM NaCl, 1 mM TCEP, 5% glycerol, 10 mM MgCl2 pH 7.5 with Pierce Protease Inhibitor, (Thermo Scientific™)). Cells were then lysed by sonication. Supernatant and cell debris were separated by centrifugation.

Example 8 – Cleavage Reactions

[00174] Plasmid library DNA cleavage reactions were carried out by mixing 5 nM of the target library, a 5-fold dilution of the TXTL or PURExpress expressions, 10 nM Tris-HCl, 10 nM $MgCl_2$, and 100 mM NaCl at 37 °C for 2 hours. For reactions with E. coli expressions, 10 µL of the clarified lysate was added. Reactions were stopped and cleaned with HighPrep™ PCR clean up beads (MAGBIO Genomics, Inc.) and eluted in Tris EDTA pH 8.0 buffer. 3 nM of the cleavage product ends were blunted with 3.33 µM dNTPs, 1X T4 DNA ligase buffer, and 0.167 U/µL of Klenow Fragment (New England Biolabs Inc.) at 25 °C for 15 minutes. 1.5 nM of the cleavage products were ligated with 150 nM adapters, 1 X T4 DNA ligase buffer (New England Biolabs Inc.), 20 U/µL T4 DNA ligase (New England Biolabs Inc.) at room temperature for 20 minutes. The ligated products were amplified by PCR with NGS primers and sequenced by NGS to obtain the PAM. The *in vitro* activity of MG119-2 is depicted in **FIG. 9**, while the PAM determination for MG119-2 is depicted in **FIG. 10**.

Example 9 – RNAseq Library Prep of Intergenic Enrichment from TXTL and *E. coli* lysates

[00175] RNA is extracted from TXTL and cell lysate expressions following the Quick-RNA™

Miniprep Kit (Zymo Research) and eluted in 30-50 µL of water. The total concentration of the transcripts were measured on a Nanodrop, Tapestation, and Qubit.

[00176] 100ng-1ug of total RNA from each sample were prepped for RNA sequencing using the NEBNext Small RNA Library Prep Set for Illumina (New England Biolabs Inc.). Amplicons between 150-300 bp were quantified by Tapestation and Qubit and pooled to a final concentration of 4 nM. A final concentration of 12.5 pM was loaded into a MiSeq V3 kit and sequenced in a Miseq system (Illumina) for 176 total cycles. The RNAseq reads were used to identify the tracr sequence of the genes.

Example 10 – Predicted RNA folding

[00177] Predicted RNA folding of the active single RNA sequence was computed at 37 °C using the method of Andronescu 2007. The shading of the bases corresponds to the probability of base pairing of that base.

Example 11 –In vitro cleavage efficiency (Prophetic)

[00178] The protein is expressed in *E. coli* protease deficient B strain under T7 inducible promoter, the cells are lysed using sonication, and the His-tagged protein of interest is purified using HisTrap FF (GE Lifescience) Ni-NTA affinity chromatography on the AKTA Avant FPLC (GE Lifescience). Purity is determined using densitometry in ImageLab software (Bio-Rad) of the protein bands resolved on SDS-PAGE and InstantBlue Ultrafast (Sigma-Aldrich) coomassie stained acrylamide gels (Bio-Rad). The protein is desalted in a storage buffer composed of 50 mM Tris-HCl, 300 mM NaCl, 1 mM TCEP, 5% glycerol; pH 7.5 and stored at -80°C.

[00179] A target DNA is constructed that contains a spacer sequence and the PAM determined via NGS. In the case of degenerate bases in the PAM a single representative PAM is chosen for testing. The target DNA is 2200 bp of linear DNA derived from a plasmid via PCR amplification. The PAM and spacer are located 700 bp from one end. Successful cleavage results in fragments of 700 and 1500 bp.

[00180] The target DNA, in vitro transcribed single RNA, and purified recombinant protein are combined in cleavage buffer (10 mM Tris, 100 mM NaCl, 10 mM MgCl2) with an excess of protein and RNA and incubated for 5' to 3 hours, usually 1 hr. The reaction is stopped via addition of RNAse A and incubation at 60°. The reaction is resolved on a 1.2% TAE agarose gel and the fraction of cleaved target DNA is quantified in ImageLab software.

Example 12 –Activity in *E. coli* (Prophetic)

[00181] For testing of nuclease activity in bacterial cells, strains are constructed with genome sequences containing the target spacer and corresponding PAM sequence specific to the enzyme

of interest. Engineered strains are then transformed with the nuclease of interest and transformants are then subsequently made chemocompetent and transformed with 50 ng of single guides either specific to the target sequence, on target, or non specific to the target, off target. After heat shock, transformations are recovered in SOC for 2 hrs at 37 °C, and nuclease efficiency is determined by a 5-fold dilution series grown on induction media. Colonies are quantified from the dilution series in triplicate.

Example 13 – Activity in mammalian cells (Prophetic)

[00182] To show targeting and cleavage activity in mammalian cells, the protein sequences are cloned into 2 mammalian expression vectors, one with a C-terminal SV40 NLS and a 2A-GFP tag and one with no GFP tag and 2 NLS sequences, one on the N-terminus and one on the C-terminus. Alternative NLS sequences that can also be used. The DNA sequence for the protein can be the native sequence, the *E. coli* codon optimized sequence, or the mammalian codon optimized sequence. The single guide RNA sequence with a gene target of interest is also cloned into a mammalian expression vector. The two plasmids are cotransfected into HEK293T cells. 72 hr after co-transfection of the expression plasmid and a sgRNA targeting plasmid into HEK293T cells, the DNA is extracted and used for the preparation of an NGS-library. Percent NHEJ is measured via indels in the sequencing of the target site to demonstrate the targeting efficiency of the enzyme in mammalian cells. At least 10 different target sites are chosen for testing each protein's activity.

Example 14 – Characterization of compact Type V nucleases in the MG119 family

[00183] ***In silico identification of novel compact type V nucleases in the MG119 family***

[00184] Discovery of predicted proteins related to nuclease sequences in the MG119 family of compact type V nucleases was based on homology searches. Searches were performed using HMMER software (http://hmmer.org/). Type V nuclease sequence hits were retained if they met the following criteria: (i) the hmmsearch e-value was $\leq 10^{-5}$, (ii) the genes encoding the nuclease were within 1 kb from a CRISPR array, and (iii) the amino acid sequence length ranged between 350 and 700 aa. MMSeqs2 (https://github.com/soedinglab/MMseqs2) was used to cluster sequences at 100% amino acid identity, with coverage mode 1 and 80% coverage of the target sequence (parameters --cov-mode 1 -c 0.8 --min-seq-id 1.0). Sequence representatives were chosen to build a multiple sequence alignment using MAFFT (https://mafft.cbrc.jp/alignment/software/) with the Needleman-Wunsch algorithm for global alignment, and FastTree (https://doi.org/10.1371/journal.pone.0009490) was used to build a phylogenetic tree. Careful examination of individual clades on the phylogenetic tree, including the nuclease gene's genomic context, led to the identification of several novel compact type V

nuclease sequences in the MG119 family (SEQ ID NOs: 476-624 and 629).

[00185] ***In vitro characterization to identify putative tracrRNAs***

[00186] To identify putative tracrRNA sequences, e.g., for nuclease MG119-2, adjacent intergenic sequences and a minimal array were expressed in transcription-translation reaction mixtures using myTXTL®Sigma 70 Master Mix Kit (Arbor Biosciences). The final reaction mixtures contained 5 nM nuclease DNA template, 12 nM intergenic DNA template, 15 nM minimal array DNA template, 0.1 nM pTXTL-P70a-T7rnap, and 1X of myTXTL®Sigma 70 Master Mix. The reactions were incubated at 29 °C for 16 hours, then stored at 4 °C.

[00187] Ribonucleoprotein complexes were tested via *in vitro* cleavage reactions. Plasmid DNA library cleavage reactions were carried out by mixing 5 nM of the target plasmid DNA library representing all possible 8N PAMs, a 5-fold dilution of the TXTL expressions, 10 nM Tris-HCl, 10 nM $MgCl_2$ and, 100 mM NaCl at 37 °C for 2 hours. Reactions were stopped and cleaned with HighPrep™ PCR clean up beads (MAGBIO Genomics, Inc.) and eluted in Tris EDTA pH 8.0 buffer.

[00188] To obtain the PAM sequences, 3 nM of the cleavage product ends were blunted with 3.33 µM dNTPs, 1X T4 DNA ligase buffer, and 0.167 U/µL of Klenow Fragment (New England Biolabs Inc.) at 25 °C for 15 minutes. 1.5 nM of the cleavage products were ligated with 150 nM adapters, 1X T4 DNA ligase buffer (New England Biolabs Inc.), and 20 U/µL T4 DNA ligase (New England Biolabs Inc.) at room temperature for 20 minutes. The ligated products were amplified by PCR with NGS primers and sequenced by NGS.

[00189] To obtain the sequence of the tracrRNA and the crRNA, RNA was extracted from TXTL lysate following the Quick-RNA™ Miniprep Kit (Zymo Research) and eluted in 30-50 µL of water. 100 ng – 1 µg of total RNA from each sample were prepped for RNA sequencing using the NEBNext Small RNA Library Prep Set for Illumina (New England Biolabs Inc.). Amplicons between 150-300 bp were quantified by Tapestation and Qubit and pooled to a final concentration of 4 nM. A final concentration of 12.5 pM was loaded into a MiSeq V3 kit and sequenced in a Miseq system (Illumina) for 176 total cycles. The RNAseq reads were used to identify the tracr sequence of the genes by mapping back to the original sequences.

[00190] ***In silico search for novel tracrRNA sequences***

[00191] To identify additional non-coding regions containing potential tracrRNAs, the sequence of the active tracrRNA was mapped to other contigs containing nucleases in the same nuclease family (e.g. MG119-1 and MG119-3). The newly identified sequences were used to generate covariance models to predict additional tracrRNAs. Covariance models were built from a multiple sequence alignment (MSA) of the active and predicted tracrRNA sequences. The secondary structure of the MSA was obtained with RNAalifold (Vienna Package), and the

covariance models were built with Infernal packages (http://eddylab.org/infernal/). Other contigs containing candidate nucleases were searched using the covariance models with the Infernal command 'cmsearch'. TracrRNA candidates were tested *in vitro* (see below), and in an iterative process, sequences from active candidates were used to improve the covariance models and search for additional tracrRNAs in the intergenic regions associated with other nuclease candidates.

[00192] ***sgRNA design***

[00193] Predicted tracrRNAs obtained from the covariance models and their associated CRISPR repeat sequence were modified to generate sgRNAs (**FIG. 11A**) as follows: the 3' end of the predicted tracrRNA sequence as well as the 5' end of the repeat sequence were trimmed, and then connected with a GAAA tetraloop.

[00194] ***In vitro cleavage reactions to confirm nuclease activity and enable PAM determination***

[00195] 5 nM of nuclease amplified DNA templates and 25 nM sgRNA amplified DNA templates (including one of the spacer sequences listed in **Table 2**) were expressed at 37 °C for 3 hours with PURExpress® In Vitro Protein Synthesis Kit (New England Biolabs Inc.). Plasmid library DNA cleavage reactions were carried out by mixing 5 nM of the target library representing all possible 8N PAMs, a 5-fold dilution of PURExpress expressions, 10 mM Tris-HCl pH 7.9, 10 mM $MgCl_2$, 100 µg/mL BSA, and 50 mM NaCl (NEB 2.1 Buffer, NEB Inc.) at 37 °C for 2 hours. Reactions were stopped and cleaned with HighPrep™ PCR clean up beads (MAGBIO Genomics, Inc.) and eluted in Tris EDTA pH 8.0 buffer. 3 nM of the cleavage product ends were blunted with 3.33 µM dNTPs, 1X T4 DNA ligase buffer, and 0.167 U/µL of Klenow Fragment (New England Biolabs Inc.) at 25 °C for 15 minutes. 1.5 nM of the cleavage products were ligated with 150 nM adapters, 1 X T4 DNA ligase buffer (New England Biolabs Inc.), and 20 U/µL T4 DNA ligase (New England Biolabs Inc.) at room temperature for 20 minutes. The ligated products were amplified by PCR with NGS primers and sequenced by NGS to obtain the PAM. Active proteins that successfully cleaved the PAM library yielded a band around 188 or 205 bp in an agarose gel, depending on which target site was encoded in the sgRNA (**FIG. 11B**).

Table 2: Spacer sequences for tested guides

Code	Sequence
U67 spacer	GTCGAGGCTTGCGACGTGGT
U40 spacer	TGGAGATATCTTGAACCTTG

[00196] The PAMs recognized by MG119 nucleases are shown as sequence logos made with Seqlog maker (**FIG. 12**). The preferred cut position on target strand of the protospacer sequence complementary to the U40 spacer is listed in **Table 3**.

Table 3: MG119 nucleases preferred cutsites in the protospacer sequence

Nuclease	sgRNA	Cutsite
119-1	MG119-1_sgRNA1	20 & 23
119-2	MG119-2_sgRNA1_Mutant1	22
119-3	MG119-3_sgRNA1_Mutant1	22-23
119-4	MG119-4_sgRNA1	22-23
119-10	MG119-10_sgRNA1	22-23
119-19	MG119-19_sgRNA1	23
119-27	MG119-27_sgRNA2_Mutant2	22-23
119-28	MG119-28_sgRNA2	22-23
119-32	MG119-32_sgRNA1	23
119-54	MG119-54_sgRNA1	22
119-64	MG119-64_sgRNA2	20
119-72	MG119-72_sgRNA1	23
119-83	MG119-83_sgRNA1	23
119-97	MG119-97_sgRNA1_Mutant1	22
119-109	MG119-109_sgRNA1	24-25
119-118	MG119-118_sgRNA1_Mutant2	23
119-121	MG119-121_sgRNA1_Mutant1	20 & 22
119-125	MG119-125_sgRNA1	22-23
119-128	MG119-128_sgRNA2_Mutant1	22
119-129	MG119-129_sgRNA1_Mutant1	22-23
119-133	MG119-133_sgRNA1_Mutant1	22
119-136	MG119-136_sgRNA1_Mutant2	23
119-137	MG119-137_sgRNA1	22-23

[00197] ***Protein expression and purification***

[00198] Isolating pure and functional proteins is essential for extensive *in vitro* analysis of biochemical properties and mechanistic studies. The expression and purification of MG119 candidates was optimized to obtain proteins of sufficient quantity and quality for such characterizations. All constructs were expressed in *E. coli* (NEBExpress I^q Competent *E. coli*, NEB C3037I). Constructs were expressed in either the pMGB expression vector (MBP-fused), the pMGBΔ expression vector (no fusion protein), or both.

[00199] ***Protein expression***

[00200] Protein expression protocols for pMGB and pMGBΔ constructs are identical. Cultures were grown at 37 °C in 2xYT media (1.6 % tryptone, 1 % yeast extract, 0.5 % NaCl) or TB media (Teknova T0690) with 100 μg / L Carbenicillin. At OD600 ≈ 0.8 – 1.2, cultures were induced with 0.5 mM IPTG (GoldBio I2481) and incubated at 18 °C overnight or 24 °C for 4-6 hrs, depending on construct. Cultures were then harvested by centrifugation at 6,000 x g for 10 min, and pellets were resuspended in Nickel_A Buffer (50 mM Tris pH 7.5, 750 mM NaCl, 10 mM $MgCl_2$, 20 mM imidazole, 0.5 mM EDTA, 5 % glycerol, 0.5 mM TCEP) + protease inhibitors (Pierce Protease Inhibitor Tablets, EDTA-free, ThermoFisher A32965) and stored at -80 °C.

[00201] ***Protein purification – pMGBΔ expression vector***

[00202] Proteins expressed in this vector have the following sequence architecture: 6xHis-(GS)2-PSP-nucleoplasmin bipartite NLS-(GGS)1-(GS)1-MG119-X-(GGS)3-SV40 NLS (**Table 5**). Proteins expressed in this vector are denoted MG119-XΔ. Cell pellets were thawed and the volume supplemented to 120 mL with Cf = 0.5 % n-Octyl-ß-D-glucoside detergent (P212121, CI-00234). Samples were sonicated in an ice-water bath at 75% amplitude for a total processing time of 3 min using a 15 s on / 45 s off cycle. Lysates were clarified by centrifugation at 30,000 x g for 25 min, and supernatants batch bound to 5 mL Ni-NTA resin (HisPur Ni-NTA Resin, ThermoFisher 88223) for ≥ 20 min. Samples were loaded onto a gravity column and washed with 30 CV Nickel_A Buffer, then eluted in 4 CV Nickel_B Buffer (Nickel_A Buffer + 250 mM imidazole) before concentrating in a 50 kDa MWCO concentrator (Amicon Ultra-15, MilliporeSigma UFC9050). Samples were taken throughout the purification process and run on an SDS-PAGE protein gel (BioRad #4568126), which was imaged on a ChemiDoc in the stain-free channel following 5 min UV activation (**FIG. 13A**). ΔMBP constructs were then loaded onto an S200i 10 / 300 GL column (Cytiva 28-9909-44) and run into Nickel_A buffer (**FIG. 13B**). Peak fractions were pooled and concentrated in a 50 kDa MWCO concentrator. Purification of proteins expressed in the pMGBΔ vector typically yielded 25 – 125 nmol protein per L expression culture (**FIG. 13F**).

[00203] ***Protein purification – pMGB expression vector***

[00204] Proteins expressed in this vector have the following sequence architecture: 6xHis-(GS)1-MBP-(GS)1-TEV- nucleoplasmin bipartite NLS-(GGGGS)3-(GS)1-MG119-X-(GGS)3-SV40 NLS (**Table 5**). MBP-fused constructs were purified identically to pMGBΔ proteins through lysis, clarification, affinity purification, and elution in Nickel_B (**FIG. 13C**). Following protein concentration in a 50 kDa MWCO concentrator, TEV protease (GenScript Z03030) was added to each sample (Cf = 1 UI/µL) and incubated at 4 °C overnight, gently rotating end-over-end. Samples were centrifuged (21,000 x g, 4 °C, 10min) to pellet aggregates, and the supernatant was then batch-bound to 3 mL amylose resin (NEB E8021L) for 30 min at 4 °C, then loaded onto a gravity column. The flow-through was collected and concentrated in a 50 kDa MWCO concentrator (**FIG. 13D**). Again, samples were centrifuged (21,000 x g, 4 °C, 10min) to pellet aggregates before loading on an S200i 10 / 300 GL column and run into Nickel_A buffer (**FIG. 13E**). Peak fractions were pooled and concentrated in a 50 kDa MWCO concentrator. Samples were taken throughout the purification process and run on an SDS-PAGE protein gel (BioRad #4568126), which was imaged on a ChemiDoc in the stain-free channel following 5 min UV activation (**FIG. 13D**).

[00205] A select few MG119 candidates were purified both from the pMGB and pMGBΔ expression vectors. A comparison of final protein yield, normalized to the initial expression culture volume, shows a trend of higher expression yields from the pMGBΔ vector (**FIG. 13E**). Purification of proteins expressed in the pMGBΔ vector typically yielded 2 – 15 nmol protein per L expression culture (**FIG. 13E**). Protein purification yields are shown in **Table 4**.

Table 4: Protein purification yields

Nuclease	Expression Vector	Expression Media	Yield (nmol)
MG119-1	pMGB	TB	8.8
MG119-1	pMGBΔ	TB	105.2
MG119-2	pMGB	2xYT	5.0
MG119-2	pMGBΔ	2xYT	95.0
MG119-3	pMGB	2xYT	7.6
MG119-3	pMGBΔ	2xYT	78.1
MG119-27	pMGB	2xYT	11.9
MG119-28	pMGB	2xYT	11.6
MG119-28	pMGBΔ	2xYT	102.2

Nuclease	Expression Vector	Expression Media	Yield (nmol)
MG119-32	pMGB	2xYT	4.3
MG119-54	pMGB	2xYT	5.6
MG119-64	pMGB	2xYT	2.1
MG119-97	pMGB	2xYT	4.3
MG119-109	pMGB	2xYT	4.1
MG119-121	pMGB	2xYT	8.2
MG119-128	pMGB	2xYT	9.9
MG119-128	pMGBΔ	2xYT	37.0
MG119-129	pMGB	2xYT	2.8
MG119-136	pMGB	2xYT	14.3
MG119-137	pMGB	2xYT	10.4

Table 5: Sequence element glossary

Element name	Element amino acid sequence
6xHis	HHHHHH
$(GS)_n$	GS
$(GGS)_n$	GGS
$(GGGGS)_n$	GGGGS
PSP	LEVQFQGP
TEV	ENLYFQG
Nucleoplasmin bipartite NLS	KRPAATKKAGQAKKKK
SV40 NLS	PKKKRKV

[00206] ***In vitro cleavage efficiency with purified protein***

[00207] The active fraction of protein aliquots was determined in a linear DNA substrate cleavage assay. Effector proteins were preincubated with a 2-fold molar excess of sgRNA for 20 min at room temperature to form the ribonucleoprotein complex (RNP). Reactions were set up using 25 nM DNA substrate and a titration of RNP from 0.25X to 10X molar excess over substrate. The reaction buffer composition was 10 mM Tris pH 7.5, 10 mM $MgCl_2$, and 100 mM NaCl. The DNA substrate is 522 bp long. Successful cleavage results in fragments of 172 and 350 bp. The reaction was incubated at 37 °C for 60 min, then incubated at 75 °C for 10 min.

RNase (NEB T3018) was added to each reaction (Cf = 0.33 µg/µL), and samples were incubated at 37 °C for 10 min. Proteinase K (NEB P8107) was added to each reaction (Cf = 60 units / mL), and samples were incubated at 55 °C for 15 min. The entirety of each reaction was then run on a 1.5 % agarose gel with GelGreen dye (Biotium, #41005) (FIG. 14A) and imaged on a ChemiDoc in the GelGreen channel. Percent cleaved substrate was calculated for each lane through densitometry analysis using BioRad's Image Lab software (Version 6.1.0 build 7). Active fraction was determined by the slope of the linear range of cleavage (**FIG. 14B**).

[00208] ***In vitro cleavage of purified Hepa1-6 genomic DNA with purified protein***

[00209] To assess cleavage of purified mouse Hepa1-6 genomic DNA (gDNA), the mouse albumin gene was targeted at intron 1 (**Table 6**). gDNA was extracted from Hepa1-6 cell pellets with 8 million cells following the PurelinkTM Genomic DNA Mini kit (Invitrogen) and eluted in 10 mM TrisHCl at pH 8. sgRNAs were ordered from Integrated DNA technologies (IDT) at 2 nmol then resuspended in 10 mM Tris EDTA Buffer at 20 µM (**Table 6**). Ribonucleoproteins (RNPs) were made by pre-incubating nucleases with targeting or non-targeting guides at a 1:2 molar ratio for 30 minutes at room temperature in 1X effector buffer (100 mM NaCl, 10 mM $MgCl_2$, 10 mM Tris HCl, at pH 7.5). All reactions were done in replicates of three including negative controls with no sgRNA. Following RNP formation, RNP was added to a digest reaction containing 20 ng/µL of the purified gDNA in 1X effector buffer and incubated at 37 °C for 1 hour. The nuclease was tested at two final concentrations, 7.8 and 15.6 nM. These concentrations were normalized by dividing the targeted concentrations with the active fractions for each nuclease. Following incubation, these reactions were immediately moved to 4 °C, diluted 30X in water, then prepared for qPCR in a mastermix containing 1X PrimeTime® Gene Expression Master Mix, 10 µM forward primer, 10 µM reverse primer, and 5 µM 5'-FAM and ZEN / Iowa Black fluorescence quencher Taqman probe (IDT) (**Table 7**). The AriaMx Real-Time PCR System (Agilent) was used with the following cycles 1) at 95 °C for 15 minutes, 2) at 95 °C for 5 seconds, and 3) at 60 °C for 1 min, where steps 2-3 were repeated 40X. Cq values were used to calculate the gDNA percent cleavage of each reaction following the Percent Cleavage Equation (below). All were normalized to the non-targeting control reactions. **FIG. 15A** illustrates an example of an average 60% gDNA cleavage by MG119-28 and sgRNA3 and 21% cleavage with sgRNA2 at the higher concentration of protein used.

Percent Cleavage Equation

$$\% \text{ Cleavage} = 100 - (2^{-(Cq(\text{experimental}) - Cq(\text{non targeting control}))} \times 100)$$

Table 6: Targeting sequences in mouse albumin intron 1 and chemically modified sgRNAs (IDT)

sgRNA Name	Sequence (5'-3')	Mouse Albumin Target (5'-3')
119-28 sgRNA1_Mouse_Alb	mU*mU*mG*rArArArUrArArArArUrGrArArUrUrUrCrArArArCrCrCrCrUrUrCrGrGrGrGrGrArGrGrGrCrGrCrGrUrUrGrGrArGrCrGrCrCrUrUrArGrUrUrUrGrArGrGrUrGrCrArGrArArUrCrArArArArArArArCrUrGrCrGrArCrGrArUrGrGrArGrGrUrCrGrUrUrUrCrArGrUrCrUrCrUrGrUrArUrArArArArArArArArUrUrUrCrArCrUrUrGrArGrArArArUrArUrCrArArArGrUrGrArArUrArUrCrCrArArCrArArGrArUrUrGrArUrGrArArGrArCrArA*mC*mU*mA	AAGATTGATGAAGACAACTA
119-28 sgRNA2_Mouse_Alb	mU*mU*mG*rArArArUrArArArArUrGrArArUrUrUrCrArArArCrCrCrCrUrUrCrGrGrGrGrGrArGrGrGrCrGrCrGrUrUrGrGrArGrCrGrCrCrUrUrArGrUrUrUrGrArGrGrUrGrCrArGrArArUrCrArArArArArArArArCrUrGrCrGrArCrGrArUrGrGrArGrGrUrCrGrUrUrUrCrArGrUrCrUrCrUrGrUrArUrArArArArArArArArUrUrUrCrArCrUrUrGrArGrArArArUrArUrCrArArArGrUrGrArArUrArUrCrCrArArCrArArGrArUrUrGrArUrGrArArGrArCrArA*mA*mC*mA	GGTCAGTGAAGAGAAGAACA
119-28 sgRNA3_Mouse_Alb	mU*mU*mG*rArArArUrArArArArUrGrArArUrUrUrCrArArArCrCrCrCrUrUrCrGrGrGrGrGrArGrGrGrCrGrCrGrUrUrGrGrArGrCrGrCrCrUrUrArGrUrUrUrGrArGrGrUrGrCrArGrArArUrCrArArArArArArArArCrUrGrCrGrArCrGrArUrGrGrArGrGrUrCrGrUrUrUrCrArGrUrCrUrCrUrGrUrArUrArArArArArArArArUrUrUrCrArCrUrUrGrArGrArArArUrArUrCrArArArGrUrGrArArUrArUrCrCrArArCrArArGrArUrUrGrArUrGrArArGrArCrArA*m	AGTGTAGCAGAGAGGAACCA

sgRNA Name	Sequence (5'-3')	Mouse Albumin Target (5'-3')
	C*mC*mA	
119-28 sgRNA4_Mouse_Alb	mU*mU*mG*rArArArUrArAr ArArUrGrArArUrUrUrCrArAr ArCrCrCrCrUrUrCrGrGrGrGrG rArGrGrGrCrGrCrGrUrUrGrGr ArGrCrGrCrCrUrUrArGrUrUrU rGrArGrGrUrGrCrArGrArArUr CrArArArArArArArCrUrGrCrG rArCrGrArUrGrGrArGrGrUrCr GrUrUrUrCrArGrUrCrUrCrUrG rUrArCrArCrUrCrArArArArAr ArUrUrCrArCrUrUrGrArGrAr ArArUrCrArArGrUrGrArArUr ArUrCrCrArArCrUrCrUrGrUrG rGrArArArCrArGrGrGrArG*m A*mG*mA	TCTGTGGAAACAGGGAGAG A

Table 7: DNA oligos used for qPCR

Oligo Names	Oligo sequences
611F_HE	TGCACAGATATAAACACTTAACGGG
869R_HE	GGGCGATCTCACTCTTGTCT
680_HE Taqman Probe	5'-FAM-AGCAGAGAGGAACCATTGCCACCTTCAG

[00210] ***In vivo cleavage of genomic DNA in Hepa 1-6 cells with purified protein***

[00211] In cell editing was demonstrated with RNP complexes of nucleases and guides targeting the mouse albumin gene at intron 1 (**Table 6**). Hepa1-6 cells were thawed, washed, and resuspended in Dulbecco's Modified Eagle Medium (DMEM, 10% FBS, and 1% Pen-strep). Cells were seeded at a density of 4 x 10^6 cells per 15 cm dish in 30 mL of media at 37 °C. After two days when the cells reached 70-80% confluency, the cells were split. Cells were trypsinized with 0.25% trypsin, then incubated at 37 °C for 30 seconds. DMEM was added then split into 3 mL and further diluted with 27 mL of media. Split cells were incubated for another two days. Prior to nucleofection, the media was aspirated from the plates, and cells were washed with 1X phosphate buffer saline (PBS, Gibco™) pH 7.2 before trypsinizing. Trypsin was neutralized and cells were resuspended with DMEM. Cells in the cell suspension were counted with a Countess 3 FL (Invitrogen) to calculate the volume of cells to pellet. Each treatment downstream required a total of 100,000 cells. Cells were centrifuged at 300 x g for 7 minutes in a sorvall X Pro Series Centrifuge (Thermo Fisher), then washed in PBS pH 7.2 before resuspending in Nucleofector™ Solution from the Amaxa™ 4D-Nucleofector™ Kit (Lonza).

[00212] RNP complexes were individually prepared by incubating 120 pmol of the nucleases with 120 pmol of the guides for 90 min at room temperature. 20 µL of the prepared cells were added to the RNPs. Nucleofections were done as recommended by the Amaxa™ 4D-Nucleofector™ Protocol in a 4D-Nucleofector™ System (Lonza). The nucleofected cells were transferred from the nucleofection cassettes to the 24 well plates, each well containing 500 µL of media. Following a two day incubation, gDNA from all treatments was extracted with QuickExtract (Lucigen) using the following cycles 1) at 65 °C for 15 min, 2) at 68 °C for 15 min, and 3) at 98 °C for 10 min and then held at 4 °C until use. The targeting window of 317 bp was amplified from the resulting extracted gDNA with Phusion Flash High-Fidelity PCR Master Mix (Thermo Fisher) using the following cycles 1) at 98 °C for 10 sec, 2) at 98 °C for 1 sec, 3) at 63 °C for 5 sec, 4) at 72 °C for 15 sec, and 5) at 72 °C for 1 min repeating steps 2-5 for 30 cycles then held at 4 °C. Amplicons were visualized on 2% agarose gels before cleaning and concentrating with HighPrep Magnetic Beads (MagBio Genomics Inc.) with 1.8X bead volume to sample. Samples were eluted in water. INDELs were sequenced by NGS on a MiSeq with a v3 reagent kit (600-cycles; **Table 8**) and 5% phiX for 2 x 301 bp paired-end reads, with a minimum of 20,000 reads per sample. INDEL analysis was performed with a modified CRISPResso2 program (Clement et al., 2019; https://doi.org/10.1038/s41587-019-0032-3), and results are shown in **Table 9** and **FIG. 15B**.

Table 8: Oligos used for NGS PCR1

Oligo Name	Oligo Sequence (5'-3')
611F_NGS	GCTCTTCCGATCTNNNNNTGCACAGATATAAACACTTAACGGG
927R_NGS	GCTCTTCCGATCTNNNNNTTCAGCATTATAACTTACAGGCCT

Table 9: Percent INDELs normalized to Apo conditions

RNP	Replicate1	Replicate2	Replicate3	Average
	% INDEL	% INDEL	% INDEL	% INDEL
119-28 sgRNA1_Mouse_Alb	0.70	0.23	0.87	0.60
119-28 sgRNA2_Mouse_Alb	0.32	0.48	0.38	0.39
119-28 sgRNA3_Mouse_Alb	50.57	13.12	11.68	25.12
119-28 sgRNA4_Mouse_Alb	9.23	2.52	0.60	4.12

Example 15 – Buffer optimization for MG119 protein purification (prophetic)

[00213] Thus far, MG119 proteins have been purified in Nickel_A buffer. Nickel_A buffer is incompatible with downstream *in vivo* assays due to its high salinity, and rapid dilution into low-salt solutions induces protein precipitation. To optimize buffers for protein stability and downstream assay compatibility, MG119 nucleases are purified initially in high-salt buffers (750 mM NaCl) and gradually washed into a Nickel_A buffer variant with 200 mM NaCl and the zwitterionic amino acids L-arginine (50 mM) and L-glutamate (50 mM). On an empirical basis, various stabilizing sugars (ribose, sorbitol, mannitol, xylitol) are also added to the buffers to enhance protein stability in low salt buffers.

[00214] Example 16 – Fluorescence-based measurement of nuclease activity (prophetic)

[00215] ***Novel cell line engineering***

[00216] Current assays used to measure *in vivo* (i.e., in mammalian cell lines) nuclease activity require extensive data analysis and turnaround times of up to a week. To expedite evaluation of *in vivo* nuclease activity, an immortalized mammalian cell line is engineered to provide immediate data on editing of genomic DNA. K562 mammalian cells, grown in IMDM (Gibco #12440053) + 10 % FBS (Corning™ Regular Fetal Bovine Serum, MT35011CV), are used for this assay. K562 mammalian cells are transfected with 12 pmol Cas9 protein (IDT #1081058), 60 pmol sgRNA (Mali et al. Science, 2013 Feb 15;339(6121):823-6.), and 1200 ng plasmid (pUC backbone) containing an expression sequence for an mMBP-(GGS)3-eGFP protein. Genomic integration of this construct results in constitutive expression under the synthetic MND promoter. Cells are left to grow for 6 days, passaging every 3 days. Monogenic cell lines are isolated from single cells by sorting individual GFP-expressing cells into a 96-well plate using a Sony MA900 Cell Sorter.

[00217] ***Fluorescence-based in vivo nuclease activity screen***

[00218] Appropriate sgRNAs are designed to direct nuclease cleavage along the mMBP and eGFP genes, such that indel formation produces a frameshift mutation resulting in loss of fluorescence. MG119 RNP complexes are formed by combining 100 pmol protein and 200 pmol sgRNA and incubating at room temperature for ≥ 20 min in a final volume of 5 µL. K562 cells are washed in 1x PBS and resuspended in Nucleofector Solution (SF Cell Line 96-well Nucleofector™ Solution) with approximately 200,000 cells per well. Cells and RNP are combined in a Lonza 96-well nucleofection plate (SF Cell Line 96-well Nucleofector™ Kit, V4SC-2096) in a final volume of 25 µL, nucleofected (K562 cells, FF-120), and recovered in IMDM + 10 % FBS media. Cells are left to recover for 2 – 3 days at 37 °C. To analyze, cells are washed twice with 1x PBS, then stained with 1x PBS + LIVE/DEAD Fixable Near-IR Dead Cell Stain Kit dye (ThermoFisher L10119) for 20 min at room temperature. Cells are washed once

more with 1x PBS before being resuspended in 1x PBS and loaded into an Attune NxT, Acoustic Focusing Flow Cytometer (model AFC2) for fluorescence analysis. Positive unedited controls (nucleofected without RNP) and negative controls (non-fluorescent K562 cells) are used to establish positive and negative fluorescence gates, and cell populations are analyzed for loss-of-fluorescence in the GFP channel to evaluate *in vivo* nuclease activity.

[00219] Example 17 – Use for epigenome editing (prophetic)

[00220] Epigenome editing is a gene modulation technique that comprises turning genes on or off constitutively or temporarily. Such techniques may use catalytically dead Cas9 (dCas9) fused to 3 proteins: Dnmt3A, Dnmt3L, and KRAB (*e.g.*, as described in Nuñez *et al. Cell* **2021**, *184*(9), 2503-2519, which is incorporated herein by this reference in its entirety). Dnmt3A and Dnmt3L are DNA methyltransferases. The KRAB domain mediates histone methylation. The methylation of DNA and histones in the promoter region mediates constitutive gene repression. dCas9 and a guide RNA may recruit the DNA and histone methylation complex to the promoter region, requiring no nuclease activity. Together, Dnmt3A, Dnmt3L, and KRAB are 579 aa, and dCas9 is 1,368 aa. The fusion protein consists of 1,947 aa or 5,841 nucleotides, exceeding the adeno-associated virus vector (AAV) packaging limit (4.7 Kb). Therefore, there is a need to create more compact epigenome editors. Compact Type V nucleases from the MG119 family represent great candidates for use as the dead nuclease partners in technologies for epigenome editing. Due to their small size, ranging from 350 to 700 aa, when fused to DNA and histone methylation complexes, the size of the fusion proteins may range from, for example, about 929 to about 1,279 aa, or about 2787 to about 3837 nucleotides, allowing easy packaging in AAVs.

[00221] To test MG119 fusion proteins as epigenome editors, HEK293T cells expressing GFP under a chimeric promoter (GAPDH-Srnpn) are generated by lentiviral transduction. MG119 family guide RNAs targeting the chimeric promoter are designed. Guides are ordered from IDT, modifying the 5' and 3' nucleotides with 3 2'-O-methyl substituents and 3 phosphorothioate bonds for stability. Dead versions of MG119 nucleases are fused to DNA and histone methylation complexes (MG119 epigenome editors). The fusion proteins are cloned in mammalian expression plasmids under the CMV promoter. GFP expressing HEK293T cells are transfected with chemically synthesized guides and plasmids expressing MG119 epigenome editors. Transfected cells are analyzed by flow cytometry. Successful MG119 epigenome editors are determined by the loss of GFP fluorescence in transfected cells. MG119 epigenome editors are then used to target genes of therapeutic interest.

Table 10 – Protein and nucleic acid sequences referred to herein

Cat.	SEQ ID NO:	Description	Type
MG122 effectors	1	MG122-1 Effector	protein
MG122 effectors	2	MG122-2 Effector	protein
MG122 effectors	3	MG122-3 Effector	protein
MG122 effectors	4	MG122-4 Effector	protein
MG122 effectors	5	MG122-5 Effector	protein
MG120 effectors	6	MG120-1 Effector	protein
MG120 effectors	7	MG120-2 Effector	protein
MG120 effectors	8	MG120-3 Effector	protein
MG120 effectors	9	MG120-4 Effector	protein
MG120 effectors	10	MG120-5 Effector	protein
MG120 effectors	11	MG120-6 Effector	protein
MG120 effectors	12	MG120-7 Effector	protein
MG120 effectors	13	MG120-8 Effector	protein
MG120 effectors	14	MG120-9 Effector	protein
MG118 Effectors	15	MG118-1 Effector	protein
MG90 Effectors	16	MG90-3 Effector	protein
MG90 Effectors	17	MG90-5 Effector	protein
MG90 Effectors	18	MG90-6 Effector	protein
MG90 Effectors	19	MG90-7 Effector	protein
MG90 Effectors	20	MG90-8 Effector	protein
MG90 Effectors	21	MG90-16 Effector	protein
MG90 Effectors	22	MG90-17 Effector	protein
MG90 Effectors	23	MG90-18 Effector	protein
MG90 Effectors	24	MG90-19 Effector	protein
MG90 Effectors	25	MG90-20 Effector	protein
MG90 Effectors	26	MG90-21 Effector	protein
MG90 Effectors	27	MG90-22 Effector	protein
MG90 Effectors	28	MG90-23 Effector	protein
MG90 Effectors	29	MG90-24 Effector	protein
MG119 Effectors	30	MG119-1 Effector	protein
MG119 Effectors	31	MG119-2 Effector	protein
MG119 Effectors	32	MG119-3 Effector	protein
MG119 Effectors	33	MG119-4 Effector	protein
MG119 Effectors	34	MG119-5 Effector	protein
MG119 Effectors	35	MG119-6 Effector	protein
MG119 Effectors	36	MG119-7 Effector	protein
MG119 Effectors	37	MG119-8 Effector	protein
MG119 Effectors	38	MG119-9 Effector	protein
MG119 Effectors	39	MG119-10 Effector	protein
MG119 Effectors	40	MG119-11 Effector	protein
MG119 Effectors	41	MG119-12 Effector	protein
MG119 Effectors	42	MG119-13 Effector	protein
MG119 Effectors	43	MG119-14 Effector	protein
MG119 Effectors	44	MG119-15 Effector	protein
MG119 Effectors	45	MG119-16 Effector	protein
MG119 Effectors	46	MG119-17 Effector	protein
MG119 Effectors	47	MG119-18 Effector	protein
MG119 Effectors	48	MG119-19 Effector	protein
MG119 Effectors	49	MG119-20 Effector	protein
MG119 Effectors	50	MG119-21 Effector	protein
MG119 Effectors	51	MG119-22 Effector	protein
MG119 Effectors	52	MG119-23 Effector	protein
MG119 Effectors	53	MG119-24 Effector	protein
MG119 Effectors	54	MG119-25 Effector	protein
MG119 Effectors	55	MG119-26 Effector	protein
MG119 Effectors	56	MG119-27 Effector	protein

Cat.	SEQ ID NO:	Description	Type
MG119 Effectors	57	MG119-28 Effector	protein
MG119 Effectors	58	MG119-29 Effector	protein
MG119 Effectors	59	MG119-30 Effector	protein
MG119 Effectors	60	MG119-31 Effector	protein
MG119 Effectors	61	MG119-32 Effector	protein
MG119 Effectors	62	MG119-33 Effector	protein
MG119 Effectors	63	MG119-34 Effector	protein
MG119 Effectors	64	MG119-35 Effector	protein
MG119 Effectors	65	MG119-36 Effector	protein
MG119 Effectors	66	MG119-37 Effector	protein
MG119 Effectors	67	MG119-38 Effector	protein
MG119 Effectors	68	MG119-39 Effector	protein
MG119 Effectors	69	MG119-40 Effector	protein
MG119 Effectors	70	MG119-41 Effector	protein
MG119 Effectors	71	MG119-42 Effector	protein
MG119 Effectors	72	MG119-43 Effector	protein
MG119 Effectors	73	MG119-44 Effector	protein
MG119 Effectors	74	MG119-45 Effector	protein
MG119 Effectors	75	MG119-46 Effector	protein
MG119 Effectors	76	MG119-47 Effector	protein
MG119 Effectors	77	MG119-48 Effector	protein
MG119 Effectors	78	MG119-49 Effector	protein
MG119 Effectors	79	MG119-50 Effector	protein
MG119 Effectors	80	MG119-51 Effector	protein
MG119 Effectors	81	MG119-52 Effector	protein
MG119 Effectors	82	MG119-53 Effector	protein
MG119 Effectors	83	MG119-54 Effector	protein
MG119 Effectors	84	MG119-55 Effector	protein
MG119 Effectors	85	MG119-56 Effector	protein
MG119 Effectors	86	MG119-57 Effector	protein
MG119 Effectors	87	MG119-58 Effector	protein
MG119 Effectors	88	MG119-59 Effector	protein
MG119 Effectors	89	MG119-61 Effector	protein
MG119 Effectors	90	MG119-62 Effector	protein
MG119 Effectors	91	MG119-63 Effector	protein
MG119 Effectors	92	MG119-64 Effector	protein
MG119 Effectors	93	MG119-65 Effector	protein
MG119 Effectors	94	MG119-66 Effector	protein
MG119 Effectors	95	MG119-67 Effector	protein
MG119 Effectors	96	MG119-68 Effector	protein
MG119 Effectors	97	MG119-69 Effector	protein
MG119 Effectors	98	MG119-70 Effector	protein
MG119 Effectors	99	MG119-71 Effector	protein
MG119 Effectors	100	MG119-72 Effector	protein
MG119 Effectors	101	MG119-73 Effector	protein
MG119 Effectors	102	MG119-74 Effector	protein
MG119 Effectors	103	MG119-75 Effector	protein
MG119 Effectors	104	MG119-76 Effector	protein
MG119 Effectors	105	MG119-77 Effector	protein
MG119 Effectors	106	MG119-78 Effector	protein
MG119 Effectors	107	MG119-79 Effector	protein
MG119 Effectors	108	MG119-80 Effector	protein
MG119 Effectors	109	MG119-81 Effector	protein
MG119 Effectors	110	MG119-83 Effector	protein
MG119 Effectors	111	MG119-84 Effector	protein
MG119 Effectors	112	MG119-85 Effector	protein
MG119 Effectors	113	MG119-86 Effector	protein
MG119 Effectors	114	MG119-87 Effector	protein
MG119 Effectors	115	MG119-88 Effector	protein

Cat.	SEQ ID NO:	Description	Type
MG119 Effectors	116	MG119-89 Effector	protein
MG119 Effectors	117	MG119-90 Effector	protein
MG119 Effectors	118	MG119-91 Effector	protein
MG119 Effectors	119	MG119-92 Effector	protein
MG119 Effectors	120	MG119-93 Effector	protein
MG119 Effectors	121	MG119-94 Effector	protein
MG119 Effectors	122	MG119-95 Effector	protein
MG119 Effectors	123	MG119-96 Effector	protein
MG119 Effectors	124	MG119-97 Effector	protein
MG119 Effectors	125	MG119-98 Effector	protein
MG119 Effectors	126	MG119-99 Effector	protein
MG119 Effectors	127	MG119-100 Effector	protein
MG119 Effectors	128	MG119-101 Effector	protein
MG119 Effectors	129	MG119-102 Effector	protein
MG119 Effectors	130	MG119-103 Effector	protein
MG119 Effectors	131	MG119-104 Effector	protein
MG119 Effectors	132	MG119-105 Effector	protein
MG119 Effectors	133	MG119-106 Effector	protein
MG119 Effectors	134	MG119-107 Effector	protein
MG119 Effectors	135	MG119-108 Effector	protein
MG119 Effectors	136	MG119-109 Effector	protein
MG119 Effectors	137	MG119-110 Effector	protein
MG119 Effectors	138	MG119-111 Effector	protein
MG119 Effectors	139	MG119-112 Effector	protein
MG119 Effectors	140	MG119-113 Effector	protein
MG119 Effectors	141	MG119-114 Effector	protein
MG119 Effectors	142	MG119-115 Effector	protein
MG119 Effectors	143	MG119-116 Effector	protein
MG119 Effectors	144	MG119-117 Effector	protein
MG119 Effectors	145	MG119-118 Effector	protein
MG119 Effectors	146	MG119-119 Effector	protein
MG119 Effectors	147	MG119-120 Effector	protein
MG119 Effectors	148	MG119-121 Effector	protein
MG119 Effectors	149	MG119-122 Effector	protein
MG119 Effectors	150	MG119-123 Effector	protein
MG91B Effectors	151	MG91B-1 Effector	protein
MG91B Effectors	152	MG91B-2 Effector	protein
MG91B Effectors	153	MG91B-3 Effector	protein
MG91B Effectors	154	MG91B-4 Effector	protein
MG91B Effectors	155	MG91B-5 Effector	protein
MG91B Effectors	156	MG91B-6 Effector	protein
MG91B Effectors	157	MG91B-7 Effector	protein
MG91B Effectors	158	MG91B-8 Effector	protein
MG91B Effectors	159	MG91B-9 Effector	protein
MG91B Effectors	160	MG91B-10 Effector	protein
MG91B Effectors	161	MG91B-11 Effector	protein
MG91B Effectors	162	MG91B-12 Effector	protein
MG91B Effectors	163	MG91B-13 Effector	protein
MG91B Effectors	164	MG91B-14 Effector	protein
MG91B Effectors	165	MG91B-15 Effector	protein
MG91B Effectors	166	MG91B-16 Effector	protein
MG91B Effectors	167	MG91B-17 Effector	protein
MG91B Effectors	168	MG91B-18 Effector	protein
MG91B Effectors	169	MG91B-19 Effector	protein
MG91B Effectors	170	MG91B-20 Effector	protein
MG91B Effectors	171	MG91B-21 Effector	protein
MG91B Effectors	172	MG91B-22 Effector	protein
MG91B Effectors	173	MG91B-23 Effector	protein
MG91B Effectors	174	MG91B-24 Effector	protein

Cat.	SEQ ID NO:	Description	Type
MG91B Effectors	175	MG91B-25 Effector	protein
MG91B Effectors	176	MG91B-26 Effector	protein
MG91B Effectors	177	MG91B-27 Effector	protein
MG91B Effectors	178	MG91B-28 Effector	protein
MG91B Effectors	179	MG91B-29 Effector	protein
MG91B Effectors	180	MG91B-30 Effector	protein
MG91B Effectors	181	MG91B-31 Effector	protein
MG91B Effectors	182	MG91B-32 Effector	protein
MG91B Effectors	183	MG91B-33 Effector	protein
MG91B Effectors	184	MG91B-34 Effector	protein
MG91B Effectors	185	MG91B-35 Effector	protein
MG91B Effectors	186	MG91B-36 Effector	protein
MG91B Effectors	187	MG91B-37 Effector	protein
MG91B Effectors	188	MG91B-38 Effector	protein
MG91B Effectors	189	MG91B-39 Effector	protein
MG91B Effectors	190	MG91B-40 Effector	protein
MG91B Effectors	191	MG91B-41 Effector	protein
MG91B Effectors	192	MG91B-42 Effector	protein
MG91B Effectors	193	MG91B-43 Effector	protein
MG91B Effectors	194	MG91B-44 Effector	protein
MG91B Effectors	195	MG91B-45 Effector	protein
MG91B Effectors	196	MG91B-46 Effector	protein
MG91B Effectors	197	MG91B-47 Effector	protein
MG91B Effectors	198	MG91B-48 Effector	protein
MG91B Effectors	199	MG91B-49 Effector	protein
MG91B Effectors	200	MG91B-50 Effector	protein
MG91B Effectors	201	MG91B-51 Effector	protein
MG91B Effectors	202	MG91B-52 Effector	protein
MG91B Effectors	203	MG91B-53 Effector	protein
MG91B Effectors	204	MG91B-54 Effector	protein
MG91B Effectors	205	MG91B-55 Effector	protein
MG91B Effectors	206	MG91B-56 Effector	protein
MG91B Effectors	207	MG91B-57 Effector	protein
MG91B Effectors	208	MG91B-58 Effector	protein
MG91B Effectors	209	MG91B-59 Effector	protein
MG91B Effectors	210	MG91B-60 Effector	protein
MG91B Effectors	211	MG91B-61 Effector	protein
MG91B Effectors	212	MG91B-62 Effector	protein
MG91B Effectors	213	MG91B-63 Effector	protein
MG91B Effectors	214	MG91B-64 Effector	protein
MG91B Effectors	215	MG91B-65 Effector	protein
MG91B Effectors	216	MG91B-66 Effector	protein
MG91B Effectors	217	MG91B-67 Effector	protein
MG91B Effectors	218	MG91B-68 Effector	protein
MG91B Effectors	219	MG91B-69 Effector	protein
MG91B Effectors	220	MG91B-70 Effector	protein
MG91B Effectors	221	MG91B-71 Effector	protein
MG91B Effectors	222	MG91B-72 Effector	protein
MG91B Effectors	223	MG91B-73 Effector	protein
MG91B Effectors	224	MG91B-74 Effector	protein
MG91B Effectors	225	MG91B-75 Effector	protein
MG91B Effectors	226	MG91B-76 Effector	protein
MG91B Effectors	227	MG91B-77 Effector	protein
MG91B Effectors	228	MG91B-78 Effector	protein
MG91B Effectors	229	MG91B-79 Effector	protein
MG91B Effectors	230	MG91B-80 Effector	protein
MG91B Effectors	231	MG91B-81 Effector	protein
MG91B Effectors	232	MG91B-82 Effector	protein
MG91B Effectors	233	MG91B-83 Effector	protein

Cat.	SEQ ID NO:	Description	Type
MG91B Effectors	234	MG91B-84 Effector	protein
MG91B Effectors	235	MG91B-85 Effector	protein
MG91B Effectors	236	MG91B-86 Effector	protein
MG91B Effectors	237	MG91B-87 Effector	protein
MG91B Effectors	238	MG91B-88 Effector	protein
MG91B Effectors	239	MG91B-89 Effector	protein
MG91B Effectors	240	MG91B-90 Effector	protein
MG91B Effectors	241	MG91B-91 Effector	protein
MG91B Effectors	242	MG91B-92 Effector	protein
MG91B Effectors	243	MG91B-93 Effector	protein
MG91B Effectors	244	MG91B-94 Effector	protein
MG91B Effectors	245	MG91B-95 Effector	protein
MG91B Effectors	246	MG91B-96 Effector	protein
MG91B Effectors	247	MG91B-97 Effector	protein
MG91B Effectors	248	MG91B-98 Effector	protein
MG91B Effectors	249	MG91B-99 Effector	protein
MG91B Effectors	250	MG91B-100 Effector	protein
MG91B Effectors	251	MG91B-101 Effector	protein
MG91B Effectors	252	MG91B-102 Effector	protein
MG91B Effectors	253	MG91B-103 Effector	protein
MG91B Effectors	254	MG91B-104 Effector	protein
MG91B Effectors	255	MG91B-105 Effector	protein
MG91B Effectors	256	MG91B-106 Effector	protein
MG91B Effectors	257	MG91B-107 Effector	protein
MG91B Effectors	258	MG91B-108 Effector	protein
MG91B Effectors	259	MG91B-109 Effector	protein
MG91B Effectors	260	MG91B-110 Effector	protein
MG91B Effectors	261	MG91B-111 Effector	protein
MG91B Effectors	262	MG91B-112 Effector	protein
MG91B Effectors	263	MG91B-113 Effector	protein
MG91B Effectors	264	MG91B-114 Effector	protein
MG91B Effectors	265	MG91B-115 Effector	protein
MG91B Effectors	266	MG91B-116 Effector	protein
MG91B Effectors	267	MG91B-117 Effector	protein
MG91B Effectors	268	MG91B-118 Effector	protein
MG91B Effectors	269	MG91B-119 Effector	protein
MG91B Effectors	270	MG91B-120 Effector	protein
MG91B Effectors	271	MG91B-121 Effector	protein
MG91B Effectors	272	MG91B-122 Effector	protein
MG91B Effectors	273	MG91B-123 Effector	protein
MG91B Effectors	274	MG91B-124 Effector	protein
MG91B Effectors	275	MG91B-125 Effector	protein
MG91B Effectors	276	MG91B-126 Effector	protein
MG91B Effectors	277	MG91B-127 Effector	protein
MG91B Effectors	278	MG91B-128 Effector	protein
MG91B Effectors	279	MG91B-129 Effector	protein
MG91B Effectors	280	MG91B-130 Effector	protein
MG91B Effectors	281	MG91B-131 Effector	protein
MG91B Effectors	282	MG91B-132 Effector	protein
MG91B Effectors	283	MG91B-133 Effector	protein
MG91B Effectors	284	MG91B-134 Effector	protein
MG91B Effectors	285	MG91B-135 Effector	protein
MG91B Effectors	286	MG91B-136 Effector	protein
MG91B Effectors	287	MG91B-137 Effector	protein
MG91B Effectors	288	MG91B-138 Effector	protein
MG91B Effectors	289	MG91B-139 Effector	protein
MG91B Effectors	290	MG91B-140 Effector	protein
MG91B Effectors	291	MG91B-141 Effector	protein
MG91C Effectors	292	MG91C-1 Effector	protein

Cat.	SEQ ID NO:	Description	Type
MG91C Effectors	293	MG91C-2 Effector	protein
MG91C Effectors	294	MG91C-3 Effector	protein
MG91C Effectors	295	MG91C-4 Effector	protein
MG91C Effectors	296	MG91C-5 Effector	protein
MG91C Effectors	297	MG91C-6 Effector	protein
MG91C Effectors	298	MG91C-7 Effector	protein
MG91C Effectors	299	MG91C-8 Effector	protein
MG91C Effectors	300	MG91C-9 Effector	protein
MG91C Effectors	301	MG91C-10 Effector	protein
MG91C Effectors	302	MG91C-11 Effector	protein
MG91C Effectors	303	MG91C-12 Effector	protein
MG91C Effectors	304	MG91C-13 Effector	protein
MG91C Effectors	305	MG91C-14 Effector	protein
MG91C Effectors	306	MG91C-15 Effector	protein
MG91C Effectors	307	MG91C-16 Effector	protein
MG91C Effectors	308	MG91C-17 Effector	protein
MG91C Effectors	309	MG91C-18 Effector	protein
MG91C Effectors	310	MG91C-19 Effector	protein
MG91C Effectors	311	MG91C-20 Effector	protein
MG91C Effectors	312	MG91C-21 Effector	protein
MG91C Effectors	313	MG91C-22 Effector	protein
MG91C Effectors	314	MG91C-23 Effector	protein
MG91C Effectors	315	MG91C-24 Effector	protein
MG91C Effectors	316	MG91C-25 Effector	protein
MG91C Effectors	317	MG91C-26 Effector	protein
MG91C Effectors	318	MG91C-27 Effector	protein
MG91A Effectors	319	MG91A-1 Effector	protein
MG126 Effectors	320	MG126-3 Effector	protein
MG126 Effectors	321	MG126-4 Effector	protein
MG126 Effectors	322	MG126-5 Effector	protein
MG126 Effectors	323	MG126-6 Effector	protein
MG126 Effectors	324	MG126-7 Effector	protein
MG126 Effectors	325	MG126-8 Effector	protein
MG119-3 effector intergenic region encoding potential tracrRNA	326	MG119-3_IG1	nucleotide
MG119-3 effector intergenic region encoding potential tracrRNA	327	MG119-3_IG2	nucleotide
MG119-3 effector intergenic region encoding potential tracrRNA	328	MG119-3_IG3	nucleotide
MG119-4 effector intergenic region encoding potential tracrRNA	329	MG119-4_IG1	nucleotide
MG119-4 effector intergenic region encoding potential tracrRNA	330	MG119-4_IG2	nucleotide
MG119-4 effector intergenic region encoding potential tracrRNA	331	MG119-4_IG3	nucleotide
MG119-4 effector intergenic region encoding potential tracrRNA	332	MG119-4_IG4	nucleotide
MG120-1 effector intergenic region encoding potential tracrRNA	333	MG120-1_IG1	nucleotide
MG120-1 effector intergenic region encoding potential tracrRNA	334	MG120-1_IG2	nucleotide

Cat.	SEQ ID NO:	Description	Type
MG120-1 effector intergenic region encoding potential tracrRNA	335	MG120-1_IG3	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA	336	MG119-1_IG1	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA	337	MG119-1_IG2	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA	338	MG119-1_IG3	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA	339	MG119-1_IG4	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA	340	MG119-1_IG5	nucleotide
MG119-2 effector intergenic region encoding potential tracrRNA	341	MG119-2_IG1	nucleotide
MG119-2 effector intergenic region encoding potential tracrRNA	342	MG119-2_IG2	nucleotide
MG119-5 effector intergenic region encoding potential tracrRNA	343	MG119-5_IG1	nucleotide
MG119-5 effector intergenic region encoding potential tracrRNA	344	MG119-5_IG2	nucleotide
MG119-5 effector intergenic region encoding potential tracrRNA	345	MG119-5_IG3	nucleotide
MG90-3 effector intergenic region encoding potential tracrRNA	346	MG90-3_IG1	nucleotide
MG90-3 effector intergenic region encoding potential tracrRNA	347	MG90-3_IG2	nucleotide
MG119-3 effector intergenic region encoding potential tracrRNA plus adapters	348	MG119-3_IG1_adapters	nucleotide
MG119-3 effector intergenic region encoding potential tracrRNA plus adapters	349	MG119-3_IG2_adapters	nucleotide
MG119-3 effector intergenic region encoding potential tracrRNA plus adapters	350	MG119-3_IG3_adapters	nucleotide
MG119-4 effector intergenic region encoding potential tracrRNA plus adapters	351	MG119-4_IG1_adapters	nucleotide
MG119-4 effector intergenic region encoding potential tracrRNA plus adapters	352	MG119-4_IG2_adapters	nucleotide
MG119-4 effector intergenic region encoding potential tracrRNA plus adapters	353	MG119-4_IG3_adapters	nucleotide
MG119-4 effector intergenic region encoding potential tracrRNA plus adapters	354	MG119-4_IG4_adapters	nucleotide

Cat.	SEQ ID NO:	Description	Type
MG120-1 effector intergenic region encoding potential tracrRNA plus adapters	355	MG120-1_IG1_adapters	nucleotide
MG120-1 effector intergenic region encoding potential tracrRNA plus adapters	356	MG120-1_IG2_adapters	nucleotide
MG120-1 effector intergenic region encoding potential tracrRNA plus adapters	357	MG120-1_IG3_adapters	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA plus adapters	358	MG119-1_IG1_adapters	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA plus adapters	359	MG119-1_IG2_adapters	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA plus adapters	360	MG119-1_IG3_adapters	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA plus adapters	361	MG119-1_IG4_adapters	nucleotide
MG119-1 effector intergenic region encoding potential tracrRNA plus adapters	362	MG119-1_IG5_adapters	nucleotide
MG119-2 effector intergenic region encoding potential tracrRNA plus adapters	363	MG119-2_IG1_adapters	nucleotide
MG119-2 effector intergenic region encoding potential tracrRNA plus adapters	364	MG119-2_IG2_adapters	nucleotide
MG119-5 effector intergenic region encoding potential tracrRNA plus adapters	365	MG119-5_IG1_adapters	nucleotide
MG119-5 effector intergenic region encoding potential tracrRNA plus adapters	366	MG119-5_IG2_adapters	nucleotide
MG119-5 effector intergenic region encoding potential tracrRNA plus adapters	367	MG119-5_IG3_adapters	nucleotide
MG90-3 effector intergenic region encoding potential tracrRNA plus adapters	368	MG90-3_IG1_adapters	nucleotide
MG90-3 effector intergenic region encoding potential tracrRNA plus adapters	369	MG90-3_IG2_adapters	nucleotide
MG119-3 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer	370	119-3_5U40_31_F	nucleotide
MG119-3 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer	371	119-3_5U40_31_R	nucleotide
MG119-4 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer	372	119-4_5U40_31_F	nucleotide
MG119-4 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer	373	119-4_5U40_31_R	nucleotide

Cat.	SEQ ID NO:	Description	Type
MG120-1 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer	374	120-1_5U40_37_F	nucleotide
MG120-1 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer	375	120-1_5U40_37_R	nucleotide
MG118-1 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer	376	118-1_5U40_38_R	nucleotide
MG119-1 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer	377	119-1_5U67_32_F	nucleotide
MG119-1 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer	378	119-1_5U67_32_R	nucleotide
MG119-2 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer	379	119-2_5U40_28_F	nucleotide
MG119-2 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer	380	119-2_5U40_28_R	nucleotide
MG119-5 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer	381	119-5_5U40_31_F	nucleotide
MG119-5 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer	382	119-5_5U40_31_R	nucleotide
MG90-3 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer	383	90-3_5U67_37_F	nucleotide
MG90-3 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer	384	90-3_5U67_37_R	nucleotide
MG119-3 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer plus adapters	385	119-3_5U40_31_F_adapters	nucleotide
MG119-3 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer plus adapters	386	119-3_5U40_31_R_adapters	nucleotide
MG119-4 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer plus adapters	387	119-4_5U40_31_F_adapters	nucleotide
MG119-4 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer plus adapters	388	119-4_5U40_31_R_adapters	nucleotide

Cat.	SEQ ID NO:	Description	Type
MG120-1 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer plus adapters	389	120-1_5U40_37_F_adapters	nucleotide
MG120-1 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer plus adapters	390	120-1_5U40_37_R_adapters	nucleotide
MG118-1 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer plus adapters	391	118-1_5U40_38_R_adapters	nucleotide
MG119-1 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer plus adapters	392	119-1_5U67_32_F_adapters	nucleotide
MG119-1 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer plus adapters	393	119-1_5U67_32_R_adapters	nucleotide
MG119-2 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer plus adapters	394	119-2_5U40_28_F_adapters	nucleotide
MG119-2 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer plus adapters	395	119-2_5U40_28_R_adapters	nucleotide
MG119-5 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer plus adapters	396	119-5_5U40_31_F_adapters	nucleotide
MG119-5 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer plus adapters	397	119-5_5U40_31_R_adapters	nucleotide
MG90-3 minimal array with T7 promoter, two repeats in the forward orientation, and one spacer plus adapters	398	90-3_5U67_37_F_adapters	nucleotide
MG90-3 minimal array with T7 promoter, two repeats in the reverse orientation, and one spacer plus adapters	399	90-3_5U67_37_R_adapters	nucleotide
MG118-1 crRNA with trimmed repeat and 18 nt universal spacer, target sequence	400	MG118-1_U40_18nt_target	nucleotide
MG118-1 crRNA with trimmed repeat and 18 nt universal spacer, target sequence	401	MG118-1_U67_18nt_target	nucleotide
MG90-3 sgRNA with 10 RAR and 24 nt universal spacer, target sequence	402	90-3_sgRNA_10bp_RAR_U40_24_target	nucleotide
MG90-3 sgRNA with 16 RAR and 24 nt universal spacer, target sequence	403	90-3_sgRNA_16bp_RAR_U40_24_target	nucleotide
MG119-1 sgRNA with 10 RAR and 24 nt universal spacer, target sequence	404	119-1_sgRNA_10bp_RAR_U40_24_target	nucleotide

Cat.	SEQ ID NO:	Description	Type
MG119-1 sgRNA with 15 RAR and 24 nt universal spacer, target sequence	405	119-1_sgRNA_15bp_RAR_U40_24_target	nucleotide
MG119-2 sgRNA with 10 RAR and 24 nt universal spacer, target sequence	406	119-2_sgRNA_10bp_RAR_U40_24_target	nucleotide
MG119-2 sgRNA with 16 RAR and 24 nt universal spacer, target sequence	407	119-2_sgRNA_16bp_RAR_U40_24_target	nucleotide
MG119-5 sgRNA with 9 RAR and 24 nt universal spacer, target sequence	408	119-5_sgRNA_9bp_RAR_U40_24_target	nucleotide
MG119-5 sgRNA with 16 RAR and 24 nt universal spacer, target sequence	409	119-5_sgRNA_16bp_RAR_U40_24_target	nucleotide
MG118-1 crRNA with trimmed repeat and 18 nt universal spacer	410	MG118-1_U40_18nt	nucleotide
MG118-1 crRNA with trimmed repeat and 18 nt universal spacer	411	MG118-1_U67_18nt	nucleotide
MG90-3 sgRNA with 10 RAR and 24 nt universal spacer	412	90-3_sgRNA_10bp_RAR_U40_24	nucleotide
MG90-3 sgRNA with 16 RAR and 24 nt universal spacer	413	90-3_sgRNA_16bp_RAR_U40_24	nucleotide
MG119-1 sgRNA with 10 RAR and 24 nt universal spacer	414	119-1_sgRNA_10bp_RAR_U40_24	nucleotide
MG119-1 sgRNA with 15 RAR and 24 nt universal spacer	415	119-1_sgRNA_15bp_RAR_U40_24	nucleotide
MG119-2 sgRNA with 10 RAR and 24 nt universal spacer	416	119-2_sgRNA_10bp_RAR_U40_24	nucleotide
MG119-2 sgRNA with 16 RAR and 24 nt universal spacer	417	119-2_sgRNA_16bp_RAR_U40_24	nucleotide
MG119-5 sgRNA with 9 RAR and 24 nt universal spacer	418	119-5_sgRNA_9bp_RAR_U40_24	nucleotide
MG119-5 sgRNA with 16 RAR and 24 nt universal spacer	419	119-5_sgRNA_16bp_RAR_U40_24	nucleotide
MG119 Effectors	420	MG119-124 Effector	protein
MG119 Effectors	421	MG119-125 Effector	protein
MG119 Effectors	422	MG119-126 Effector	protein
MG119 Effectors	423	MG119-127 Effector	protein
MG119 Effectors	424	MG119-128 Effector	protein
MG119 Effectors	425	MG119-129 Effector	protein
MG119 Effectors	426	MG119-130 Effector	protein
MG119 Effectors	427	MG119-131 Effector	protein
MG119 Effectors	428	MG119-132 Effector	protein
MG119 Effectors	429	MG119-133 Effector	protein
MG119 Effectors	430	MG119-134 Effector	protein
MG119 Effectors	431	MG119-135 Effector	protein
MG119-1 effectors sgRNA1	432	MG119-1_sgRNA1	Nucleotide
MG119-1 effectors PAM (5')	433	MG119-1 PAM (5')	Nucleotide
MG119-2 effectors sgRNA1	434	MG119-2_sgRNA1_Mutant1	Nucleotide
MG119-2 effectors PAM (5')	435	MG119-2 PAM (5')	Nucleotide
MG119-3 effectors sgRNA1	436	MG119-3_sgRNA1_Mutant1	Nucleotide
MG119-3 effectors PAM (5')	437	MG119-3 PAM (5')	Nucleotide
MG119-4 effectors sgRNA1	438	MG119-4_sgRNA1	Nucleotide
MG119-4 effectors PAM (5')	439	MG119-4 PAM (5')	Nucleotide
MG119-10 effectors sgRNA1	440	MG119-10_sgRNA1	Nucleotide
MG119-10 effectors PAM (5')	441	MG119-10 PAM (5')	Nucleotide
MG119-19 effectors sgRNA1	442	MG119-19_sgRNA1	Nucleotide
MG119-19 effectors PAM (5')	443	MG119-19 PAM (5')	Nucleotide
MG119-27 effectors sgRNA2	444	MG119-27_sgRNA2_Mutant2	Nucleotide

Cat.	SEQ ID NO:	Description	Type
MG119-27 effectors PAM (5')	445	MG119-27 PAM (5')	Nucleotide
MG119-28 effectors sgRNA2	446	MG119-28_sgRNA2	Nucleotide
MG119-28 effectors PAM (5')	447	MG119-28 PAM (5')	Nucleotide
MG119-32 effectors sgRNA1	448	MG119-32_sgRNA1	Nucleotide
MG119-32 effectors PAM (5')	449	MG119-32 PAM (5')	Nucleotide
MG119-54 effectors sgRNA1	450	MG119-54_sgRNA1	Nucleotide
MG119-54 effectors PAM (5')	451	MG119-54 PAM (5')	Nucleotide
MG119-64 effectors sgRNA2	452	MG119-64_sgRNA2	Nucleotide
MG119-64 effectors PAM (5')	453	MG119-64 PAM (5')	Nucleotide
MG119-72 effectors sgRNA1	454	MG119-72_sgRNA1	Nucleotide
MG119-72 effectors PAM (5')	455	MG119-72 PAM (5')	Nucleotide
MG119-83 effectors sgRNA1	456	MG119-83_sgRNA1	Nucleotide
MG119-83 effectors PAM (5')	457	MG119-83 PAM (5')	Nucleotide
MG119-97 effectors sgRNA1	458	MG119-97_sgRNA1_Mutant1	Nucleotide
MG119-97 effectors PAM (5')	459	MG119-97 PAM (5')	Nucleotide
MG119-109 effectors sgRNA1	460	MG119-109_sgRNA1	Nucleotide
MG119-109 effectors PAM (5')	461	MG119-109 PAM (5')	Nucleotide
MG119-118 effectors sgRNA1	462	MG119-118_sgRNA1_Mutant2	Nucleotide
MG119-118 effectors PAM (5')	463	MG119-118 PAM (5')	Nucleotide
MG119-121 effectors sgRNA1	464	MG119-121_sgRNA1_Mutant1	Nucleotide
MG119-121 effectors PAM (5')	465	MG119-121 PAM (5')	Nucleotide
MG119-125 effectors sgRNA1	466	MG119-125_sgRNA1	Nucleotide
MG119-125 effectors PAM (5')	467	MG119-125 PAM (5')	Nucleotide
MG119-128 effectors sgRNA1	468	MG119-128_sgRNA2_Mutant1	Nucleotide
MG119-128 effectors PAM (5')	469	MG119-128 PAM (5')	Nucleotide
MG119-129 effectors sgRNA1	470	MG119-129_sgRNA1_Mutant1	Nucleotide
MG119-129 effectors PAM (5')	471	MG119-129 PAM (5')	Nucleotide
MG119-133 effectors sgRNA1	472	MG119-133_sgRNA1_Mutant1	Nucleotide
MG119-133 effectors PAM (5')	473	MG119-133 PAM (5')	Nucleotide
MG119-136 effectors sgRNA1	474	MG119-136_sgRNA1_Mutant2	Nucleotide
MG119-136 effectors PAM (5')	475	MG119-136 PAM (5')	Nucleotide
MG119-136 active effectors	476	MG119-136 effector	Protein
MG119-139 effectors	477	MG119-139 effector	Protein
MG119-140 effectors	478	MG119-140 effector	Protein
MG119-141 effectors	479	MG119-141 effector	Protein
MG119-142 effectors	480	MG119-142 effector	Protein
MG119-143 effectors	481	MG119-143 effector	Protein
MG119-144 effectors	482	MG119-144 effector	Protein
MG119-145 effectors	483	MG119-145 effector	Protein
MG119-146 effectors	484	MG119-146 effector	Protein
MG119-147 effectors	485	MG119-147 effector	Protein
MG119-148 effectors	486	MG119-148 effector	Protein
MG119-149 effectors	487	MG119-149 effector	Protein
MG119-150 effectors	488	MG119-150 effector	Protein
MG119-151 effectors	489	MG119-151 effector	Protein
MG119-152 effectors	490	MG119-152 effector	Protein
MG119-153 effectors	491	MG119-153 effector	Protein
MG119-154 effectors	492	MG119-154 effector	Protein
MG119-155 effectors	493	MG119-155 effector	Protein
MG119-156 effectors	494	MG119-156 effector	Protein
MG119-157 effectors	495	MG119-157 effector	Protein
MG119-158 effectors	496	MG119-158 effector	Protein
MG119-159 effectors	497	MG119-159 effector	Protein
MG119-160 effectors	498	MG119-160 effector	Protein
MG119-161 effectors	499	MG119-161 effector	Protein
MG119-162 effectors	500	MG119-162 effector	Protein
MG119-163 effectors	501	MG119-163 effector	Protein
MG119-164 effectors	502	MG119-164 effector	Protein
MG119-165 effectors	503	MG119-165 effector	Protein

Cat.	SEQ ID NO:	Description	Type
MG119-166 effectors	504	MG119-166 effector	Protein
MG119-167 effectors	505	MG119-167 effector	Protein
MG119-168 effectors	506	MG119-168 effector	Protein
MG119-169 effectors	507	MG119-169 effector	Protein
MG119-170 effectors	508	MG119-170 effector	Protein
MG119-171 effectors	509	MG119-171 effector	Protein
MG119-172 effectors	510	MG119-172 effector	Protein
MG119-173 effectors	511	MG119-173 effector	Protein
MG119-174 effectors	512	MG119-174 effector	Protein
MG119-175 effectors	513	MG119-175 effector	Protein
MG119-176 effectors	514	MG119-176 effector	Protein
MG119-177 effectors	515	MG119-177 effector	Protein
MG119-178 effectors	516	MG119-178 effector	Protein
MG119-179 effectors	517	MG119-179 effector	Protein
MG119-180 effectors	518	MG119-180 effector	Protein
MG119-181 effectors	519	MG119-181 effector	Protein
MG119-182 effectors	520	MG119-182 effector	Protein
MG119-183 effectors	521	MG119-183 effector	Protein
MG119-184 effectors	522	MG119-184 effector	Protein
MG119-185 effectors	523	MG119-185 effector	Protein
MG119-186 effectors	524	MG119-186 effector	Protein
MG119-187 effectors	525	MG119-187 effector	Protein
MG119-188 effectors	526	MG119-188 effector	Protein
MG119-189 effectors	527	MG119-189 effector	Protein
MG119-190 effectors	528	MG119-190 effector	Protein
MG119-191 effectors	529	MG119-191 effector	Protein
MG119-192 effectors	530	MG119-192 effector	Protein
MG119-193 effectors	531	MG119-193 effector	Protein
MG119-194 effectors	532	MG119-194 effector	Protein
MG119-195 effectors	533	MG119-195 effector	Protein
MG119-196 effectors	534	MG119-196 effector	Protein
MG119-197 effectors	535	MG119-197 effector	Protein
MG119-198 effectors	536	MG119-198 effector	Protein
MG119-199 effectors	537	MG119-199 effector	Protein
MG119-200 effectors	538	MG119-200 effector	Protein
MG119-201 effectors	539	MG119-201 effector	Protein
MG119-202 effectors	540	MG119-202 effector	Protein
MG119-203 effectors	541	MG119-203 effector	Protein
MG119-204 effectors	542	MG119-204 effector	Protein
MG119-205 effectors	543	MG119-205 effector	Protein
MG119-206 effectors	544	MG119-206 effector	Protein
MG119-207 effectors	545	MG119-207 effector	Protein
MG119-208 effectors	546	MG119-208 effector	Protein
MG119-209 effectors	547	MG119-209 effector	Protein
MG119-210 effectors	548	MG119-210 effector	Protein
MG119-211 effectors	549	MG119-211 effector	Protein
MG119-212 effectors	550	MG119-212 effector	Protein
MG119-213 effectors	551	MG119-213 effector	Protein
MG119-214 effectors	552	MG119-214 effector	Protein
MG119-215 effectors	553	MG119-215 effector	Protein
MG119-216 effectors	554	MG119-216 effector	Protein
MG119-217 effectors	555	MG119-217 effector	Protein
MG119-218 effectors	556	MG119-218 effector	Protein
MG119-219 effectors	557	MG119-219 effector	Protein
MG119-220 effectors	558	MG119-220 effector	Protein
MG119-221 effectors	559	MG119-221 effector	Protein
MG119-222 effectors	560	MG119-222 effector	Protein
MG119-223 effectors	561	MG119-223 effector	Protein
MG119-224 effectors	562	MG119-224 effector	Protein

Cat.	SEQ ID NO:	Description	Type
MG119-225 effectors	563	MG119-225 effector	Protein
MG119-226 effectors	564	MG119-226 effector	Protein
MG119-227 effectors	565	MG119-227 effector	Protein
MG119-228 effectors	566	MG119-228 effector	Protein
MG119-229 effectors	567	MG119-229 effector	Protein
MG119-230 effectors	568	MG119-230 effector	Protein
MG119-231 effectors	569	MG119-231 effector	Protein
MG119-232 effectors	570	MG119-232 effector	Protein
MG119-233 effectors	571	MG119-233 effector	Protein
MG119-234 effectors	572	MG119-234 effector	Protein
MG119-235 effectors	573	MG119-235 effector	Protein
MG119-236 effectors	574	MG119-236 effector	Protein
MG119-237 effectors	575	MG119-237 effector	Protein
MG119-238 effectors	576	MG119-238 effector	Protein
MG119-239 effectors	577	MG119-239 effector	Protein
MG119-240 effectors	578	MG119-240 effector	Protein
MG119-241 effectors	579	MG119-241 effector	Protein
MG119-242 effectors	580	MG119-242 effector	Protein
MG119-243 effectors	581	MG119-243 effector	Protein
MG119-244 effectors	582	MG119-244 effector	Protein
MG119-245 effectors	583	MG119-245 effector	Protein
MG119-246 effectors	584	MG119-246 effector	Protein
MG119-247 effectors	585	MG119-247 effector	Protein
MG119-248 effectors	586	MG119-248 effector	Protein
MG119-249 effectors	587	MG119-249 effector	Protein
MG119-250 effectors	588	MG119-250 effector	Protein
MG119-251 effectors	589	MG119-251 effector	Protein
MG119-252 effectors	590	MG119-252 effector	Protein
MG119-253 effectors	591	MG119-253 effector	Protein
MG119-254 effectors	592	MG119-254 effector	Protein
MG119-255 effectors	593	MG119-255 effector	Protein
MG119-256 effectors	594	MG119-256 effector	Protein
MG119-257 effectors	595	MG119-257 effector	Protein
MG119-258 effectors	596	MG119-258 effector	Protein
MG119-259 effectors	597	MG119-259 effector	Protein
MG119-260 effectors	598	MG119-260 effector	Protein
MG119-261 effectors	599	MG119-261 effector	Protein
MG119-262 effectors	600	MG119-262 effector	Protein
MG119-263 effectors	601	MG119-263 effector	Protein
MG119-264 effectors	602	MG119-264 effector	Protein
MG119-265 effectors	603	MG119-265 effector	Protein
MG119-266 effectors	604	MG119-266 effector	Protein
MG119-267 effectors	605	MG119-267 effector	Protein
MG119-268 effectors	606	MG119-268 effector	Protein
MG119-269 effectors	607	MG119-269 effector	Protein
MG119-270 effectors	608	MG119-270 effector	Protein
MG119-271 effectors	609	MG119-271 effector	Protein
MG119-272 effectors	610	MG119-272 effector	Protein
MG119-273 effectors	611	MG119-273 effector	Protein
MG119-274 effectors	612	MG119-274 effector	Protein
MG119-275 effectors	613	MG119-275 effector	Protein
MG119-276 effectors	614	MG119-276 effector	Protein
MG119-277 effectors	615	MG119-277 effector	Protein
MG119-278 effectors	616	MG119-278 effector	Protein
MG119-279 effectors	617	MG119-279 effector	Protein
MG119-280 effectors	618	MG119-280 effector	Protein
MG119-281 effectors	619	MG119-281 effector	Protein
MG119-282 effectors	620	MG119-282 effector	Protein
MG119-283 effectors	621	MG119-283 effector	Protein

Cat.	SEQ ID NO:	Description	Type
MG119-284 effectors	622	MG119-284 effector	Protein
MG119-285 effectors	623	MG119-285 effector	Protein
MG119-286 effectors	624	MG119-286 effector	Protein
sgRNAs	625	119-28 sgRNA1_Mouse_Alb	Nucleotide (RNA)
sgRNAs	626	119-28 sgRNA2_Mouse_Alb	Nucleotide (RNA)
sgRNAs	627	119-28 sgRNA3_Mouse_Alb	Nucleotide (RNA)
sgRNAs	628	119-28 sgRNA4_Mouse_Alb	Nucleotide (RNA)
MG119 effectors	629	MG119-137 Effector	protein

Table 11– Protein and nucleic acid sequences referred to herein

Cat.	SEQ ID NO:	Description	Type	Organism	Other Information	Sequence
MG119 effectors	629	MG119-137 Effector	protein	unknown	uncultivated organism	MSKDKYVITRKIKLLPVGG ENEVDRVYDFIRNGQYSQY QALNLLMGQLASKYYDCK KDLSSAEFKDAQKSILSNSN PNLCDIEFVKGCDTKSAVV QKVRQDFSTAIKNGLPRGE RNITNYKRTVPLITRGRDLV FVHGYENYTEFLDNLYTDR NLKVFIKWVNKIQFKIVFG NPYKSAELRSVVQNIFEERY KINGSSICIDDDDIILNLSLT MPKEIKELDESKVVGVDLG IAIPAVCALNTNSYSRKSIGS ADDFLRVRTKIRAQRRRLQ KSLSQTSGGHGRKKKLRAL DKFSEYEKHWVQNYNHYV SKQVVDFAIKNNAKYINLE DLEGYGEEEKNKFILSNWS YYQLQQYIAYKAEKYGIEV RKINPYHTSQVCSCCGHWE SGQRVNQKTFICKNPECEN FGEEVNADFNAARNIALST NWSDIDEKKNKKNKKK

[00222] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. It is not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the embodiments herein are not meant to be construed in a limiting sense. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is therefore contemplated that the invention shall also cover any such alternatives, modifications, variations or equivalents. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

CLAIMS

WHAT IS CLAIMED IS:

1. An engineered nuclease system comprising:

(a) an endonuclease having at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and

(b) an engineered guide RNA, wherein said engineered guide RNA is configured to form a complex with said endonuclease, and said engineered guide RNA comprises a spacer sequence configured to hybridize to a target nucleic acid sequence.

2. The engineered nuclease system of claim 1, wherein said guide RNA comprises a sequence with at least 80% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419, 432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474.

3. The engineered nuclease system of any one of claims 1-2, wherein said endonuclease has at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629.

4. The engineered nuclease system of any one of claims 1-3, wherein said guide RNA comprises a sequence with at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 414-419, 432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474.

5. An engineered nuclease system comprising:

(a) an engineered guide RNA comprising a sequence with at least 80% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419, 432, 434, 436,

438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474, and

(b) a class 2, type V Cas endonuclease configured to bind to said engineered guide RNA.

6. The engineered nuclease system of any one of claims 1-5, wherein said guide RNA comprises a sequence complementary to a eukaryotic, fungal, plant, mammalian, or human genomic polynucleotide sequence.

7. The engineered nuclease system of any one of claims 1-6, wherein said guide RNA is 30-250 nucleotides in length.

8. The engineered nuclease system of any one of claims 1-7, wherein said endonuclease comprises one or more nuclear localization sequences (NLSs) proximal to an N- or C-terminus of said endonuclease.

9. The engineered nuclease system of any one of claims 1-8, wherein said NLS comprises a sequence at least 80% identical to a sequence from the group consisting of SEQ ID NO: 630-645.

10. The engineered nuclease system of any one of claims 1-9, further comprising a single- or double-stranded DNA repair template comprising from 5′ to 3′: a first homology arm comprising a sequence of at least 20 nucleotides 5′ to said target deoxyribonucleic acid sequence, a synthetic DNA sequence of at least 10 nucleotides, and a second homology arm comprising a sequence of at least 20 nucleotides 3′ to said target sequence.

11. The engineered nuclease system of claim 10, wherein said first or second homology arm comprises a sequence of at least 40, 80, 120, 150, 200, 300, 500, or 1,000 nucleotides.

12. The engineered nuclease system of claim 10 or claim 11, wherein said first and second homology arms are homologous to a genomic sequence of a prokaryote, bacteria, fungus, or eukaryote.

13. The engineered nuclease system of claims 10-12, wherein said single- or double-stranded DNA repair template comprises a transgene donor.

14. The engineered nuclease system of any one of claims 1-13, further comprising a DNA repair template comprising a double-stranded DNA segment flanked by one or two single-stranded DNA segments.

15. The engineered nuclease system of claim 14, wherein said single-stranded DNA segments are conjugated to the 5′ ends of said double-stranded DNA segment.

16. The engineered nuclease system of claim 14, wherein said single stranded DNA segments are conjugated to the 3′ ends of said double-stranded DNA segment.

17. The engineered nuclease system of any one of claims 14-16, wherein said single-stranded DNA segments have a length from 4 to 10 nucleotide bases.

18. The engineered nuclease system of any one of claims 14-17, wherein said single-stranded DNA segments have a nucleotide sequence complementary to a sequence within said spacer sequence.

19. The engineered nuclease system of any one of claims 14-18, wherein said double-stranded DNA sequence comprises a barcode, an open reading frame, an enhancer, a promoter, a protein-coding sequence, a miRNA coding sequence, an RNA coding sequence, or a transgene.

20. The engineered nuclease system of any one of claims 14-18, wherein said double-stranded DNA sequence is flanked by a nuclease cut site.

21. The engineered nuclease system of claim 20, wherein said nuclease cut site comprises a spacer and a PAM sequence.

22. The engineered nuclease system of claim 21, wherein said PAM comprises a sequence of any one of SEQ ID NOs: 433, 435, 437, 439, 441, 443, 445, 447, 449, 451, 453, 455, 457, 459, 461, 463, 465, 467, 469, 471, 473, and 475.

23. The engineered nuclease system of any one of claims 1-22, wherein said system further comprises a source of Mg^{2+}.

24. The engineered nuclease system of any one of claims 1-23, wherein said guide RNA comprises a hairpin comprising at least 8, at least 10, or at least 12 base-paired ribonucleotides.

25. The engineered nuclease system of claim 24, wherein said hairpin comprises 10 base-paired ribonucleotides.

26. The engineered nuclease system of any one of claims 1-25, wherein:

a) said endonuclease comprises a sequence at least 75%, 80%, or 90% identical to any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof; and

b) said guide RNA structure comprises a sequence at least 80%, or 90% identical to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419.

27. The engineered nuclease system of any one of claims 1-25, wherein

a) said endonuclease comprises a sequence at least about 75%, at least about 80%, at least about 85%, at least about at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629; and

b) said guide RNA structure comprises a sequence at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to the non-degenerate nucleotides of any one of SEQ ID NOs: 414-419, 432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474.

28. The engineered nuclease system of any one of claims 1-27, wherein said sequence identity is determined by a BLASTP, CLUSTALW, MUSCLE, MAFFT algorithm, or a CLUSTALW algorithm with the Smith-Waterman homology search algorithm parameters.

29. The engineered nuclease system of claim 28, wherein said sequence identity is determined by said BLASTP homology search algorithm using parameters of a wordlength (W) of 3, an expectation (E) of 10, and a BLOSUM62 scoring matrix setting gap costs at existence of 11, extension of 1, and using a conditional compositional score matrix adjustment.

30. An engineered guide ribonucleic acid (RNA) polynucleotide comprising:

a) a DNA-targeting segment comprising a nucleotide sequence that is complementary to a target sequence in a target DNA molecule; and

b) a protein-binding segment comprising two complementary stretches of nucleotides that hybridize to form a double-stranded RNA (dsRNA) duplex,

wherein said two complementary stretches of nucleotides are covalently linked to one another with intervening nucleotides, and

wherein said engineered guide ribonucleic acid polynucleotide is capable of forming a complex with a type 2, class V Cas endonuclease.

31. The engineered guide RNA of claim 30, wherein said type 2, class V Cas endonuclease is derived from an uncultivated organism.

32. The engineered guide ribonucleic acid polynucleotide of claim 30 or claim 31, wherein said Cas endonuclease has at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629, and targeting said complex to said target sequence of said target DNA molecule.

33. The engineered guide ribonucleic acid polynucleotide of any one of claims 30-32, wherein said DNA-targeting segment is positioned 3′ of both of said two complementary stretches of nucleotides.

34. The engineered guide ribonucleic acid polynucleotide of any one of claims 30-33, wherein said protein binding segment comprises a sequence having at least 70%, at least 80%, or at least 90% identity to the non-degenerate nucleotides of SEQ ID NO: 410-419.

35. The engineered guide ribonucleic acid polynucleotide of any one of claims 30-34, wherein said double-stranded RNA (dsRNA) duplex comprises at least 5, at least 8, at least 10, or at least 12 ribonucleotides.

36. A deoxyribonucleic acid polynucleotide encoding the engineered guide ribonucleic acid polynucleotide of any one of claims 30-35.

37. A nucleic acid comprising an engineered nucleic acid sequence optimized for expression in an organism, wherein said nucleic acid encodes a class 2, type V Cas endonuclease, and wherein said endonuclease is derived from an uncultivated microorganism, wherein the organism is not said uncultivated organism.

38. The nucleic acid of claim 37, wherein said endonuclease comprises a variant having at least 70% or at least 80% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629.

39. The nucleic acid of claim 37 or 38, wherein said endonuclease comprises a sequence encoding one or more nuclear localization sequences (NLSs) proximal to an N- or C-terminus of said endonuclease.

40. The nucleic acid of claim 39, wherein said NLS comprises a sequence selected from SEQ ID NOs: 630-645.

41. The nucleic acid of claim 39 or 40, wherein said NLS comprises SEQ ID NO: 631.

42. The nucleic acid of claim 41, wherein said NLS is proximal to said N-terminus of said endonuclease.

43. The nucleic acid of claim 39 or 40, wherein said NLS comprises SEQ ID NO: 630.

44. The nucleic acid of claim 43, wherein said NLS is proximal to said C-terminus of said endonuclease.

45. The nucleic acid of any one of claims 37-44, wherein said organism is prokaryotic, bacterial, eukaryotic, fungal, plant, mammalian, rodent, or human.

46. An engineered vector comprising a nucleic acid sequence encoding a class 2, type V Cas endonuclease, wherein said endonuclease is derived from an uncultivated microorganism.

47. An engineered vector comprising the nucleic acid of any of claims 37-45.

48. An engineered vector comprising the deoxyribonucleic acid polynucleotide of claim 36.

49. The engineered vector of any of claims 46-48, wherein the vector is a plasmid, a minicircle, a CELiD, an adeno-associated virus (AAV) derived virion, a lentivirus, or an adenovirus.

50. A cell comprising the engineered vector of any of claims 46-49.

51. A method of manufacturing an endonuclease, comprising cultivating said cell of claim 50.

52. A method for binding, cleaving, marking, or modifying a double-stranded deoxyribonucleic acid polynucleotide, comprising:

(a) contacting said double-stranded deoxyribonucleic acid polynucleotide with a class 2, type V Cas endonuclease in complex with an engineered guide RNA configured to bind to said endonuclease and said double-stranded deoxyribonucleic acid polynucleotide;

wherein said double-stranded deoxyribonucleic acid polynucleotide comprises a protospacer adjacent motif (PAM); and

wherein said guide RNA structure comprises a sequence at least 80%, or 90% identical to the non-degenerate nucleotides of any one of SEQ ID NOs: 410-419.

53. The method of claim 52, wherein said double-stranded deoxyribonucleic acid polynucleotide comprises a first strand comprising a sequence complementary to a sequence of said engineered guide RNA and a second strand comprising said PAM.

54. The method of claim 53, wherein said PAM is directly adjacent to the 5′ end of said sequence complementary to said sequence of said engineered guide RNA.

55. The method of any one of claims 52-54, wherein said PAM comprises a sequence of any one of SEQ ID NOs: 433, 435, 437, 439, 441, 443, 445, 447, 449, 451, 453, 455, 457, 459, 461, 463, 465, 467, 469, 471, 473, and 475.

56. The method of any one of claims 52-55, wherein said class 2, type V Cas endonuclease is derived from an uncultivated microorganism.

57. The method of any one of claims 52-56, wherein said class 2, type V Cas endonuclease further comprises a PAM interacting domain.

58. The method of any one of claims 52-57, wherein said double-stranded deoxyribonucleic acid polynucleotide is a eukaryotic, plant, fungal, mammalian, rodent, or human double-stranded deoxyribonucleic acid polynucleotide.

59. A method of modifying a target nucleic acid locus, said method comprising delivering to said target nucleic acid locus said engineered nuclease system of any one of claims 1-29, wherein said endonuclease is configured to form a complex with said engineered guide

ribonucleic acid structure, and wherein said complex is configured such that upon binding of said complex to said target nucleic acid locus, said complex modifies said target nucleic acid locus.

60. The method of claim 59, wherein modifying said target nucleic acid locus comprises binding, nicking, cleaving, or marking said target nucleic acid locus.

61. The method of claim 59 or 60, wherein said target nucleic acid locus comprises deoxyribonucleic acid (DNA) or ribonucleic acid (RNA).

62. The method of claim 59, wherein said target nucleic acid comprises genomic DNA, viral DNA, viral RNA, or bacterial DNA.

63. The method of any one of claims 59-62, wherein said target nucleic acid locus is *in vitro*.

64. The method of any one of claims 59-62, wherein said target nucleic acid locus is within a cell.

65. The method of claim 64, wherein said cell is a prokaryotic cell, a bacterial cell, a eukaryotic cell, a fungal cell, a plant cell, an animal cell, a mammalian cell, a rodent cell, a primate cell, a human cell, or a primary cell.

66. The method of claim 64 or 65, wherein said cell is a primary cell.

67. The method of claim 66, wherein said primary cell is a T cell.

68. The method of claim 66, wherein said primary cell is a hematopoietic stem cell (HSC).

69. The method of any one of claims 59-68, wherein delivering said engineered nuclease system to said target nucleic acid locus comprises delivering the nucleic acid of any of claims 37-45 or the engineered vector of any of claims 46-49.

70. The method of any one of claims 59-69, wherein delivering said engineered nuclease system to said target nucleic acid locus comprises delivering a nucleic acid comprising an open reading frame encoding said endonuclease.

71. The method of claim 70, wherein said nucleic acid comprises a promoter to which said open reading frame encoding said endonuclease is operably linked.

72. The method of any one of claims 59-71, wherein delivering said engineered nuclease system to said target nucleic acid locus comprises delivering a capped mRNA containing said open reading frame encoding said endonuclease.

73. The method of any one of claims 59-72, wherein delivering said engineered nuclease system to said target nucleic acid locus comprises delivering a translated polypeptide.

74. The method of any one of claims 59-72, wherein delivering said engineered nuclease system to said target nucleic acid locus comprises delivering a deoxyribonucleic acid (DNA) encoding said engineered guide RNA operably linked to a ribonucleic acid (RNA) pol III promoter.

75. The method of any one of claims 59-74, wherein said endonuclease induces a single-stranded break or a double-stranded break at or proximal to said target locus.

76. The method of claim 75, wherein said endonuclease induces a staggered single stranded break within or 3′ to said target locus.

77. A host cell comprising an open reading frame encoding a heterologous endonuclease having at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof.

78. The host cell of claim 77, wherein said endonuclease has at least 75% sequence identity to any one of SEQ ID NOs: 1, 6, 15, 30, 151, 292, or 319, or a variant thereof.

79. The host cell of claim 77, wherein said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629.

80. The host cell of any one of claims 77-79, wherein said host cell is an *E. coli* cell.

81. The host cell of claim 80, wherein said *E. coli* cell is a λDE3 lysogen or said *E. coli* cell is a BL21(DE3) strain.

82. The host cell of claim 80 or 81, wherein said *E. coli* cell has an *ompT lon* genotype.

83. The host cell of any one of claims 77-82, wherein said open reading frame is operably linked to a T7 promoter sequence, a T7-lac promoter sequence, a lac promoter sequence, a tac promoter sequence, a trc promoter sequence, a ParaBAD promoter sequence, a PrhaBAD promoter sequence, a T5 promoter sequence, a *cspA* promoter sequence, an $araP_{BAD}$ promoter, a strong leftward promoter from phage lambda (pL promoter), or any combination thereof.

84. The host cell of any one of claims 77-83, wherein said open reading frame comprises a sequence encoding an affinity tag linked in-frame to a sequence encoding said endonuclease.

85. The host cell of claim 84, wherein said affinity tag is an immobilized metal affinity chromatography (IMAC) tag.

86. The host cell of claim 85, wherein said IMAC tag is a polyhistidine tag.

87. The host cell of claim 84, wherein said affinity tag is a myc tag, a human influenza hemagglutinin (HA) tag, a maltose binding protein (MBP) tag, a glutathione S-transferase (GST) tag, a streptavidin tag, a FLAG tag, or any combination thereof.

88. The host cell of any one of claims 84-87, wherein said affinity tag is linked in-frame to said sequence encoding said endonuclease via a linker sequence encoding a protease cleavage site.

89. The host cell of claim 88, wherein said protease cleavage site is a tobacco etch virus (TEV) protease cleavage site, a PreScission® protease (PSP) cleavage site, a Thrombin cleavage site, a Factor Xa cleavage site, an enterokinase cleavage site, or any combination thereof.

90. The host cell of any one of claims 77-89, wherein said open reading frame is codon-optimized for expression in said host cell.

91. The host cell of any one of claims 77-90, wherein said open reading frame is provided on a vector.

92. The host cell of any one of claims 77-90, wherein said open reading frame is integrated into a genome of said host cell.

93. A culture comprising the host cell of any one of claims 77-92 in compatible liquid medium.

94. A method of producing an endonuclease, comprising cultivating the host cell of any one of claims 77-92 in compatible growth medium.

95. The method of claim 94, further comprising inducing expression of said endonuclease by addition of an additional chemical agent or an increased amount of a nutrient.

96. The method of claim 95, wherein an additional chemical agent or an increased amount of a nutrient comprises Isopropyl β-D-1-thiogalactopyranoside (IPTG) or additional amounts of lactose.

97. The method of any one of claims 94-96, further comprising isolating said host cell after said cultivation and lysing said host cell to produce a protein extract.

98. The method of claim 97, further comprising subjecting said protein extract to IMAC, or ion-affinity chromatography.

99. The method of claim 98, wherein said open reading frame comprises a sequence encoding an IMAC affinity tag linked in-frame to a sequence encoding said endonuclease.

100. The method of claim 99, wherein said IMAC affinity tag is linked in-frame to said sequence encoding said endonuclease via a linker sequence encoding protease cleavage site.

101. The method of claim 100, wherein said protease cleavage site comprises a tobacco etch virus (TEV) protease cleavage site, a PreScission® protease cleavage site, a Thrombin cleavage site, a Factor Xa cleavage site, an enterokinase cleavage site, or any combination thereof.

102. The method of any one of claims 100-101, further comprising cleaving said IMAC affinity tag by contacting a protease corresponding to said protease cleavage site to said endonuclease.

103. The method of claim 102, further comprising performing subtractive IMAC affinity chromatography to remove said affinity tag from a composition comprising said endonuclease.

104. A method of disrupting a locus in a cell, comprising contacting to said cell a composition comprising:

(a) a class 2, type V Cas endonuclease having at least 75% identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and

(b) an engineered guide RNA, wherein said engineered guide RNA is configured to form a complex with said endonuclease and said engineered guide RNA comprises a spacer sequence configured to hybridize to a region of said locus,

wherein said class 2, type V Cas endonuclease has at least equivalent cleavage activity to spCas9 in said cell.

105. The method of claim 104, wherein said cleavage activity is measured *in vitro* by introducing said endonucleases alongside compatible guide RNAs to cells comprising said target nucleic acid and detecting cleavage of said target nucleic acid sequence in said cells.

106. The method of claim 104 or claim 105, wherein said composition comprises 20 picomoles (pmol) or less of said class 2, type V Cas endonuclease.

107. The method of claim 106, wherein said composition comprises 1 pmol or less of said class 2, type V Cas endonuclease.

108. A method of disrupting an albumin locus in a cell, comprising contacting to said cell a composition comprising:

(a) an endonuclease having at least 75% identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and

(b) an engineered guide RNA, wherein said engineered guide RNA is configured to form a complex with said endonuclease, and said engineered guide RNA comprises a spacer sequence configured to hybridize to a region of said locus,

wherein said engineered guide RNA is configured to hybridize to the any one of the target sequences in Table 6.

109. The method of claim 108, wherein said engineered guide RNA comprises a sequence having at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to at least 18 non-degenerate nucleotides of any one of SEQ ID NOs: 414-419432, 434, 436, 438, 440, 442, 444, 446, 448, 450, 452, 454, 456, 458, 460, 462, 464, 466, 468, 470, 472, and 474.

110. The method of claim 108 or claim 109, wherein said engineered guide RNA comprises the modified nucleotides of any of the single guide RNA (sgRNA) sequences in Table 6.

111. The method of claim any one of claims 108-110, wherein said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629.

112. The method of claim 111, wherein said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to SEQ ID NO: 57.

113. The method of any one of claims 108-112, wherein said region is 5' to a PAM sequence comprising any one of SEQ ID NOs: 433, 435, 437, 439, 441, 443, 445, 447, 449, 451, 453, 455, 457, 459, 461, 463, 465, 467, 469, 471, 473, and 475.

114. An isolated RNA molecule comprising a sequence at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any sequence in Table 6.

115. The isolated RNA molecule of claim 114, further comprising the pattern of chemical modifications recited in any of the guide RNAs recited in Table 6.

116. Use of the RNA molecule of claim 114 or claim 115 for modifying an albumin locus of a cell.

117. An engineered nuclease system comprising,

(a) an endonuclease configured to be selective for a protospacer adjacent motif (PAM) comprising any one of SEQ ID NOs: 433, 435, 437, 439, 441, 443, 445, 447, 449, 451, 453, 455, 457, 459, 461, 463, 465, 467, 469, 471, 473, and 475; and

(b) an engineered guide RNA, wherein said engineered guide RNA is configured to form a complex with said endonuclease, and said engineered guide RNA comprises a spacer sequence configured to hybridize to a target nucleic acid sequence.

118. The engineered nuclease system of claim 117, wherein said endonuclease is a class 2, type V Cas endonuclease.

119. The engineered nuclease system of claim 117 or claim 118, wherein said endonuclease is not a Cas12a nuclease.

120. The engineered nuclease system of any one of claims 117-119, wherein said endonuclease is derived from an uncultivated organism.

121. The engineered nuclease system of any one of claims 117-120, wherein said endonuclease further comprises a PAM interacting domain configured to interact with said PAM.

122. The engineered nuclease system of any one of claims 117-121, wherein said endonuclease has at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof.

123. The engineered nuclease system of claim 122, wherein said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629.

124. An engineered nuclease system comprising:

(a) an endonuclease having at least 75% sequence identity to any one of SEQ ID NOs: 1-325, 420-431, 476-624, or 629 or a variant thereof; and

(b) a DNA methyltransferase.

125. The engineered nuclease system of claim 124, wherein said endonuclease has at least about 75%, at least about 80%, at least about 85%, at least about 86%, at least about 87%, at least about 88%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at

least about 98%, at least about 99%, or 100% sequence identity to any one of SEQ ID NOs: 30-33, 39, 48, 56, 57, 61, 83, 92, 100, 110, 124, 136, 145, 148, 424, 425, 429, 476, or 629.

126. The engineered nuclease system of claim 124 or claim 125, wherein said DNA methyltransferase binds non-covalently to said endonuclease.

127. The engineered nuclease system of claim 124 or claim 125, wherein said DNA methyltransferase is fused to said endonuclease in a single polypeptide.

128. The engineered nuclease system of any one of claims 124-127, wherein said DNA methyltransferase comprises Dmnt3A or Dnmt3L.

129. The engineered nuclease system of any one of claims 124-128, further comprising a KRAB domain.

130. The engineered nuclease system of claim 129, wherein said KRAB domain binds non-covalently to said endonuclease or said DNA methyltransferase.

131. The engineered nuclease system of claim 129, wherein said KRAB domain is covalently linked to said endonuclease or said DNA methyltransferase.

132. The engineered nuclease system of claim 131, wherein said KRAB domain is fused to said endonuclease or said DNA methyltransferase in a single polypeptide.

133. The engineered nuclease system of any one of claims 124-132, wherein said endonuclease is a nickase or is catalytically dead.

134. The engineered nuclease system of any one of claims 124-133, further comprising an engineered guide RNA structure configured to form a complex with said endonuclease, and wherein said engineered guide RNA structure comprises a spacer sequence configured to hybridize to a target nucleic acid sequence.

135. The engineered nuclease system of claim 134, wherein said target nucleic acid sequence is comprised in or proximal to a promoter of a target genome.

136. The engineered nuclease system of claim 134 or claim 135, wherein said engineered guide RNA structure comprises one or more: (a) 2'-O-methylnucleotide(s); (b) 2'-fluoronucleotide(s); or (c) phosphorothioate bond(s).

137. The engineered nuclease system of claim 134 or claim 135, wherein said engineered guide RNA structure comprises the pattern of chemically modified nucleotides of any of the single guide RNAs in Table 6.

138. A method of modifying a target nucleic acid locus, said method comprising delivering to said target nucleic acid locus said engineered nuclease system of any one of claims 124-137, wherein said endonuclease is configured to forma complex with said engineered guide RNA structure, and wherein said complex is configured that upon binding of said complex to said target nucleic acid locus, said DNA methyltransferase modifies said target nucleic acid locus.

139. Use of the engineered nuclease system of any one of claims 124-137 for modifying a nucleic acid locus.

140. The use of claim 139, wherein modifying said nucleic acid locus comprises methylating or demethylating a nucleotide of said nucleic acid locus.

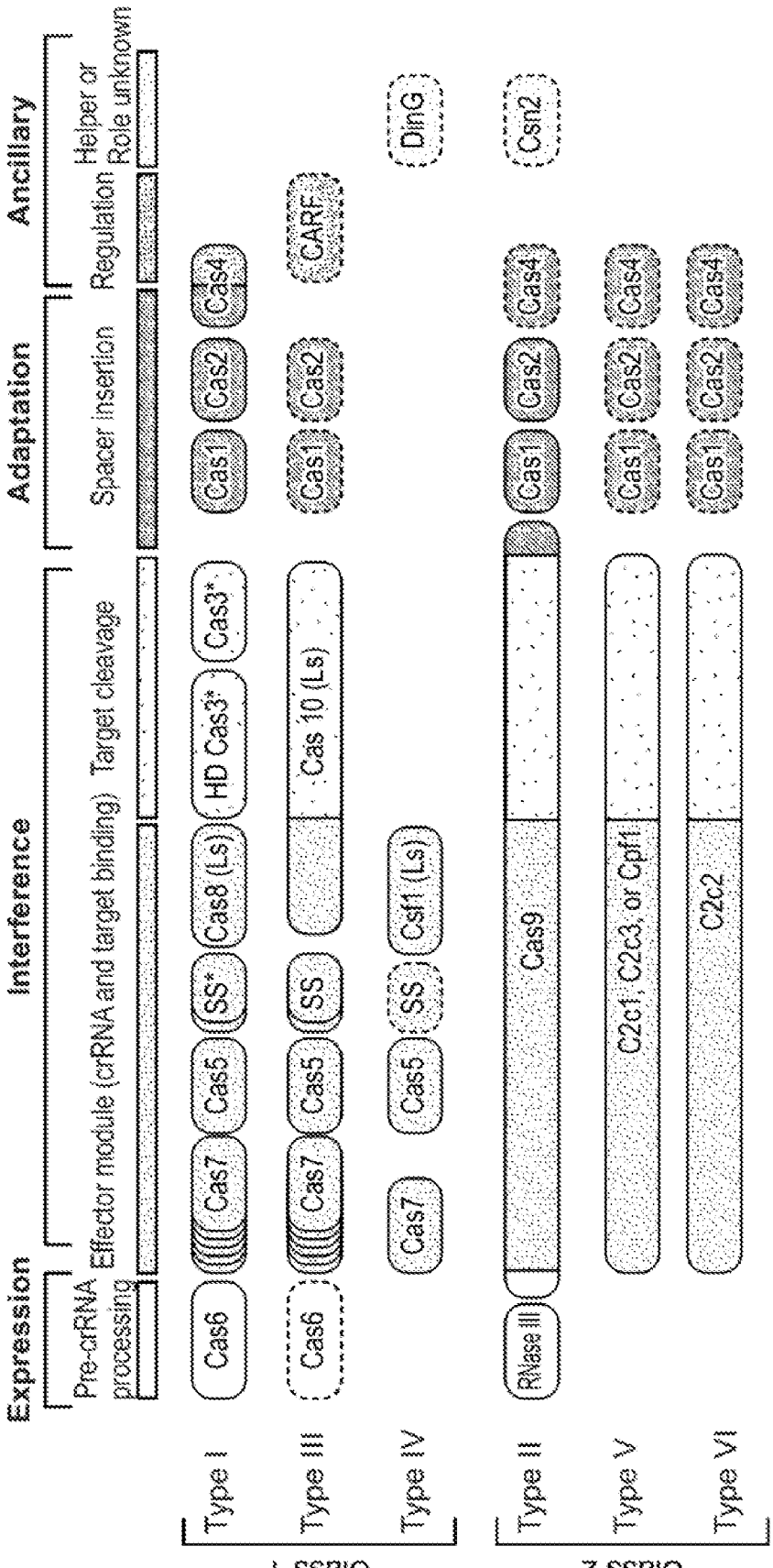

Expression
Interference
Adaptation
Ancillary
Pre-crRNA processing
Effector module (crRNA and target binding)
Target cleavage
Spacer insertion
Regulation
Helper or Role unknown
Class 1
Type I
Type III
Type IV
Class 2
Type II
Type V
Type VI
Cas6
Cas7
Cas5
SS*
Cas8 (Ls)
HD Cas3*
Cas3*
SS
Cas 10 (Ls)
Csf1 (Ls)
Cas1
Cas2
Cas4
CARF
DinG
RNase III
Cas9
C2c1, C2c3, or Cpf1
C2c2
Csn2

FIG. 1

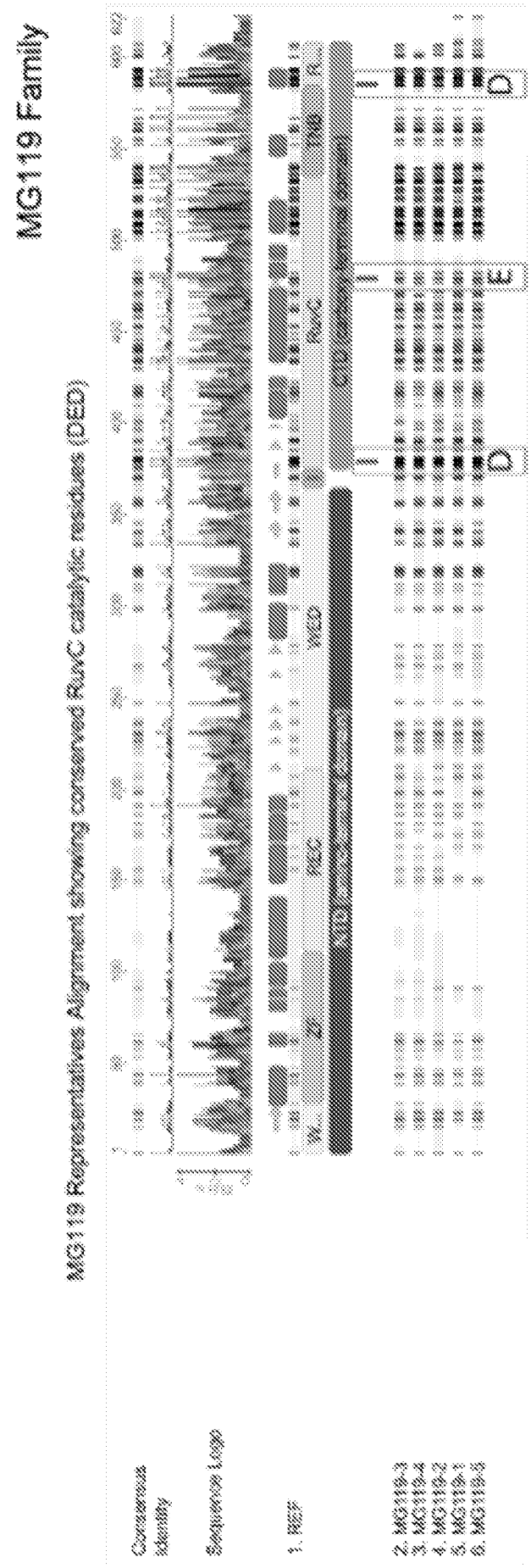

MG119 Family
MG119 Representative Alignment showing conserved RuvC catalytic residues (DED)
Consensus Identity
Sequence Logo
1. REF
2. MG119-3
3. MG119-4
4. MG119-2
5. MG119-1
6. MG119-5
RuvC
REC
WED
D
E
D

FIG. 2A

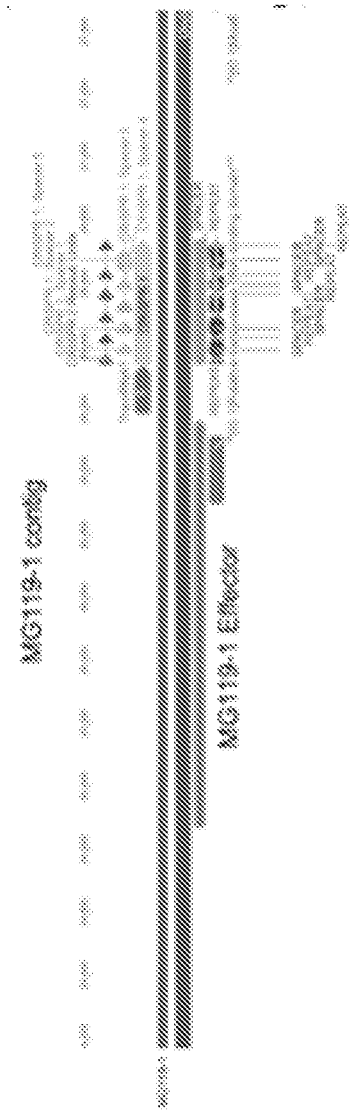
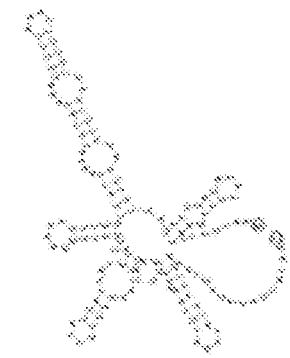


FIG. 2B



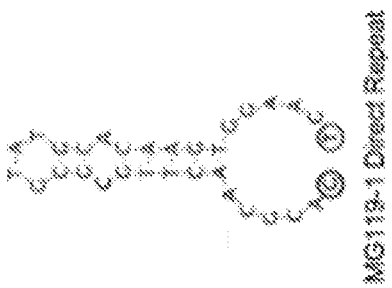


FIG. 2C

FIG. 2D

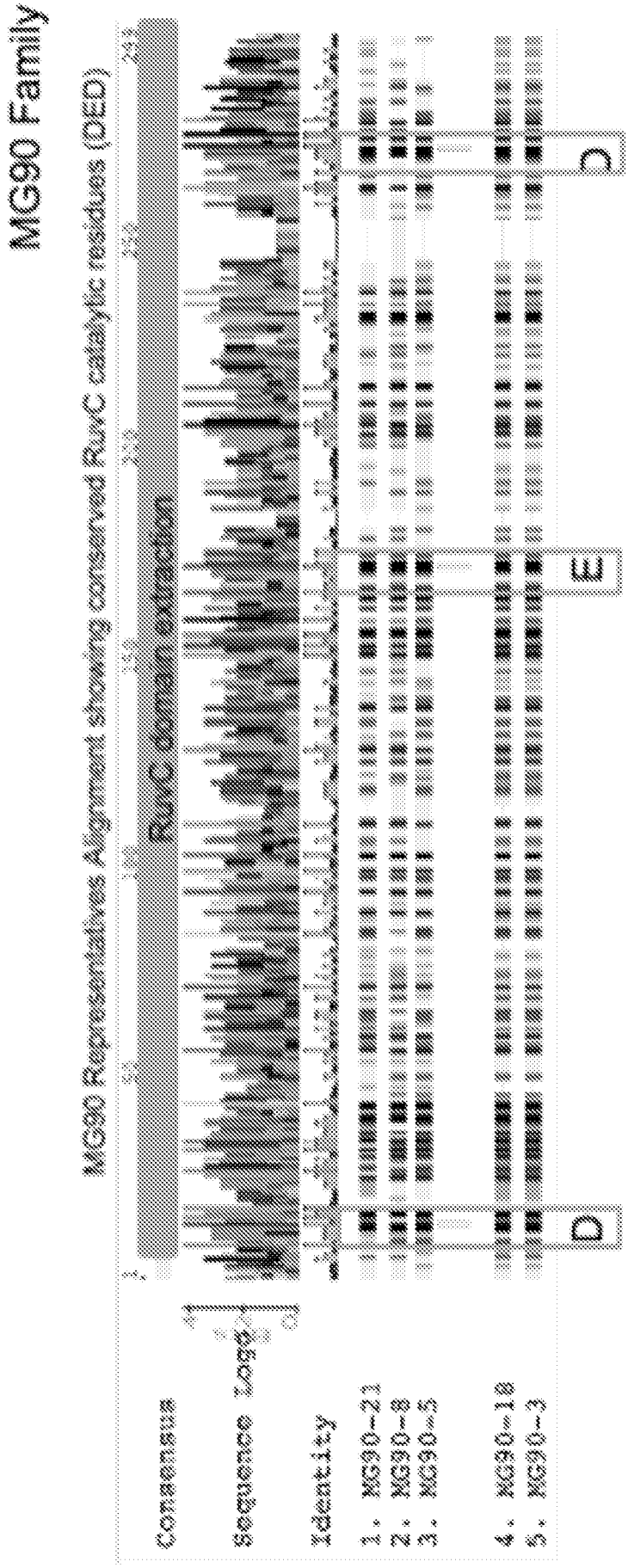

MG90 Family
MG90 Representatives Alignment showing conserved RuvC catalytic residues (DED)
RuvC domain extraction
Consensus
Sequence Logo
Identity
1. MG90-21
2. MG90-8
3. MG90-5
4. MG90-10
5. MG90-3
D
E
D

FIG. 3A

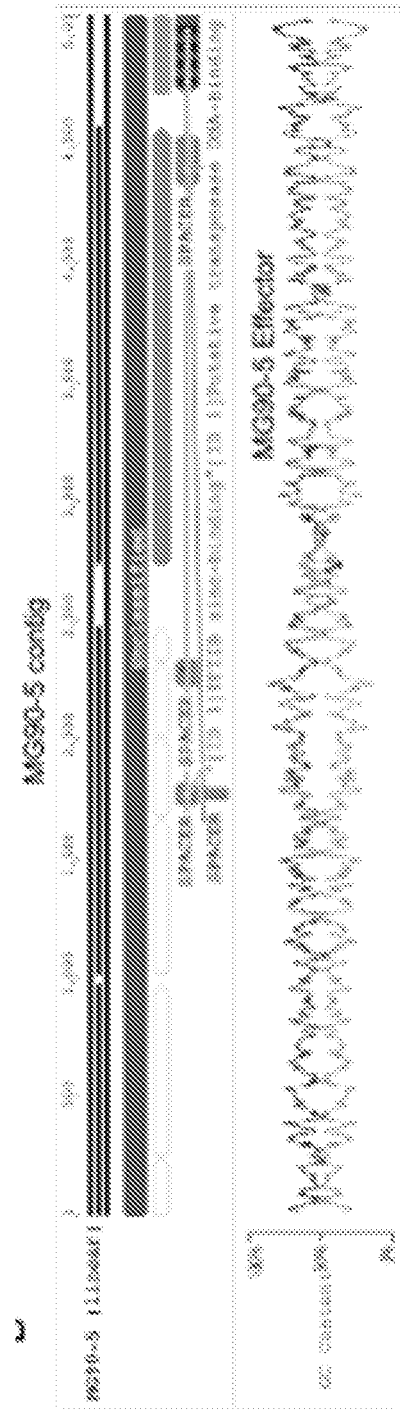


FIG. 3B

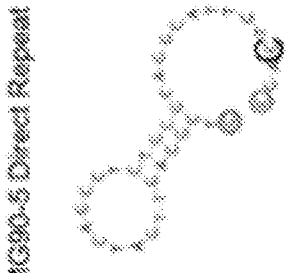


FIG. 3C

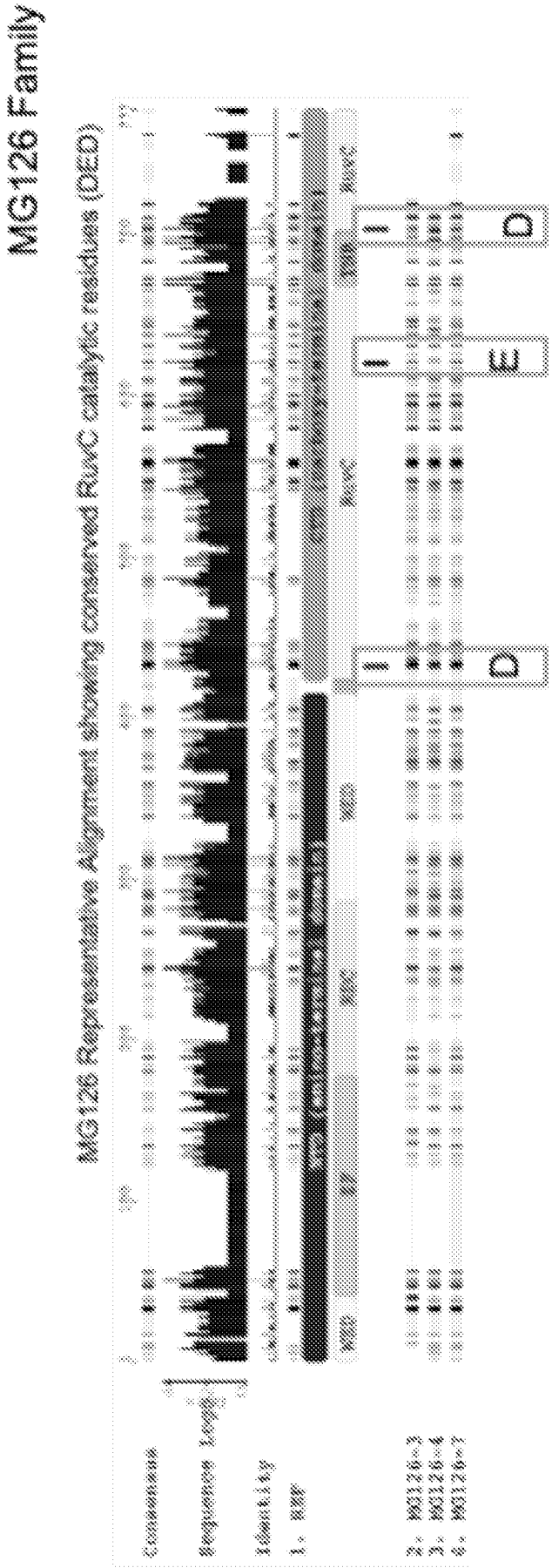

MG126 Family
MG126 Representative Alignment showing conserved RuvC catalytic residues (DED)
Consensus
Sequence Logo
Identity
1. MG126-1
2. MG126-3
3. MG126-4
4. MG126-7
RuvC
I
D
E
I
I
D

FIG. 4A

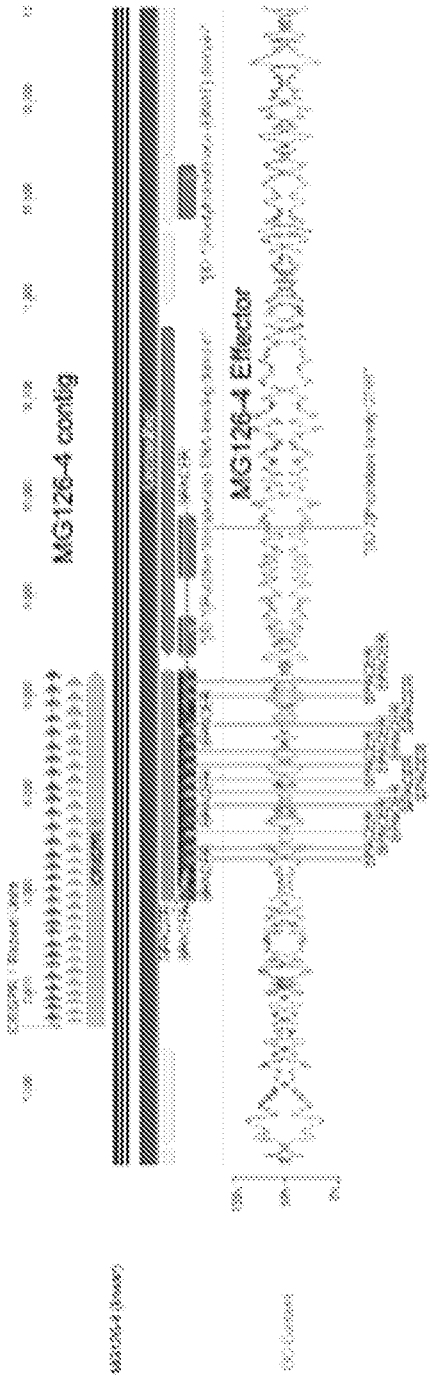


FIG. 4B

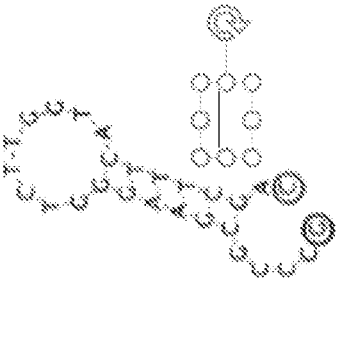


FIG. 4C

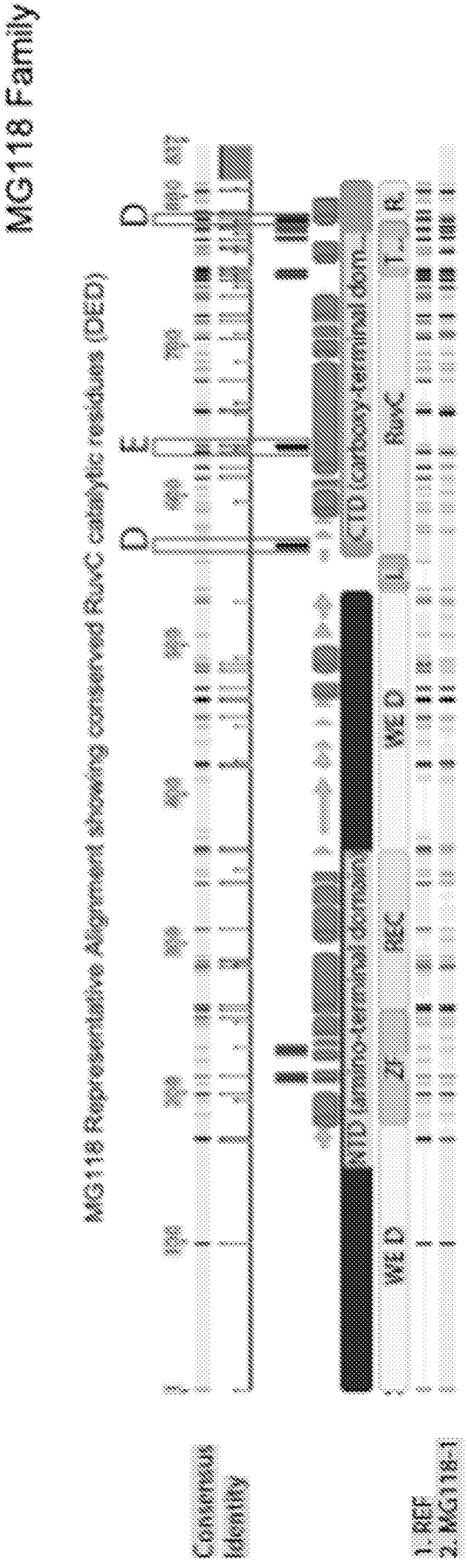

MG118 Family
MG118 Representative Alignment showing conserved RuvC catalytic residues (DED)
D
E
D
Consensus Identity
1. REF
2. MG118-1
WED
Zf
REC
RuvC
WED
RuvC
TNB
RuvC-terminal domain
CTD (carboxy-terminal dom.)

FIG. 5A

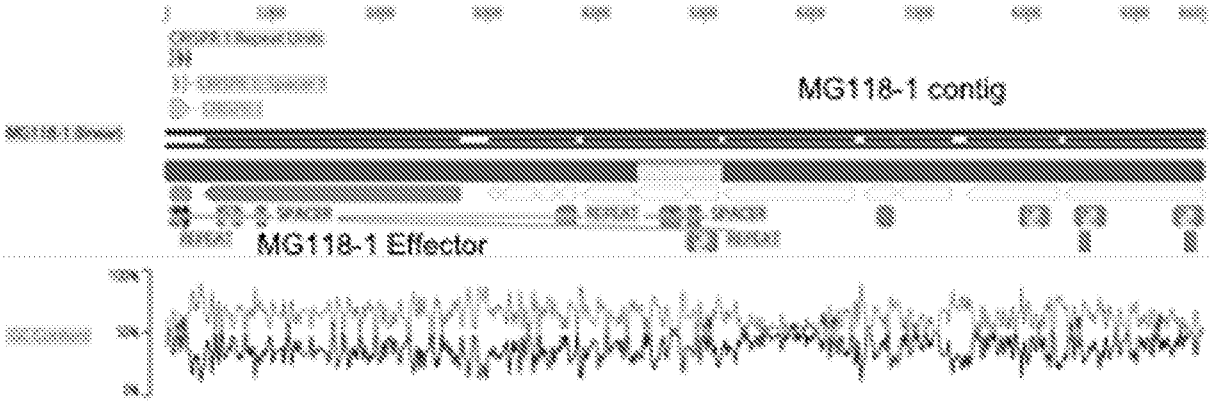


FIG. 5B

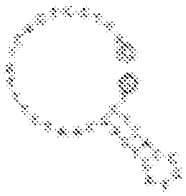


FIG. 5C

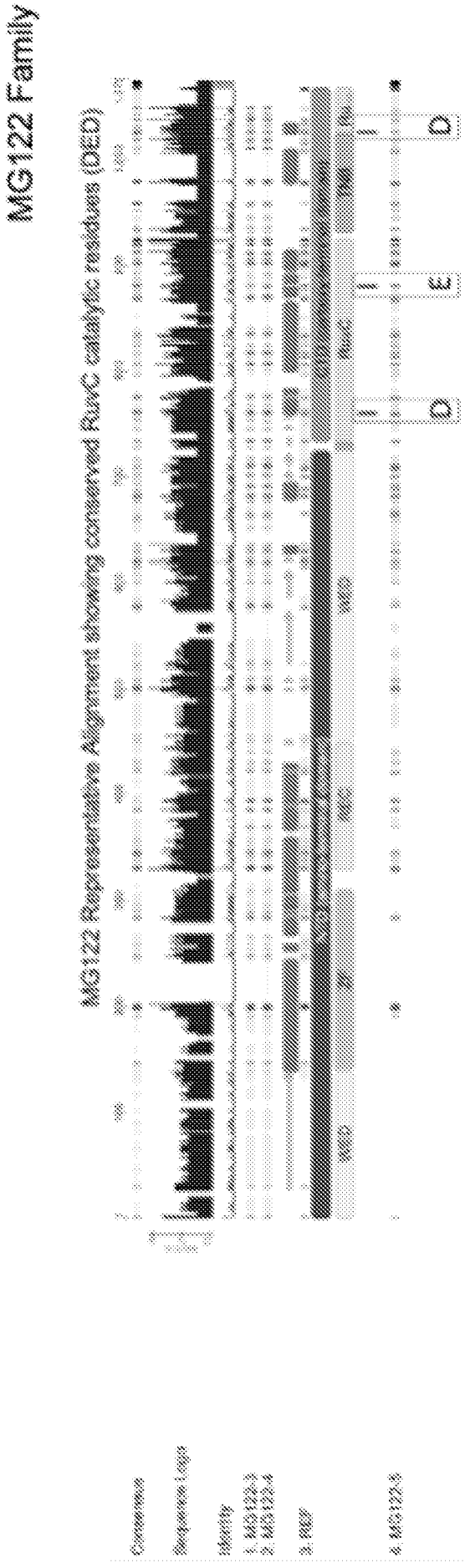

MG122 Family
MG122 Representative Alignment showing conserved RuvC catalytic residues (DED)
Consensus
Sequence Logo
Identity
1. MG122-3
2. MG122-4
3. REF
4. MG122-5
RuvC
WED
PI
REC
D
E
D

FIG. 6A

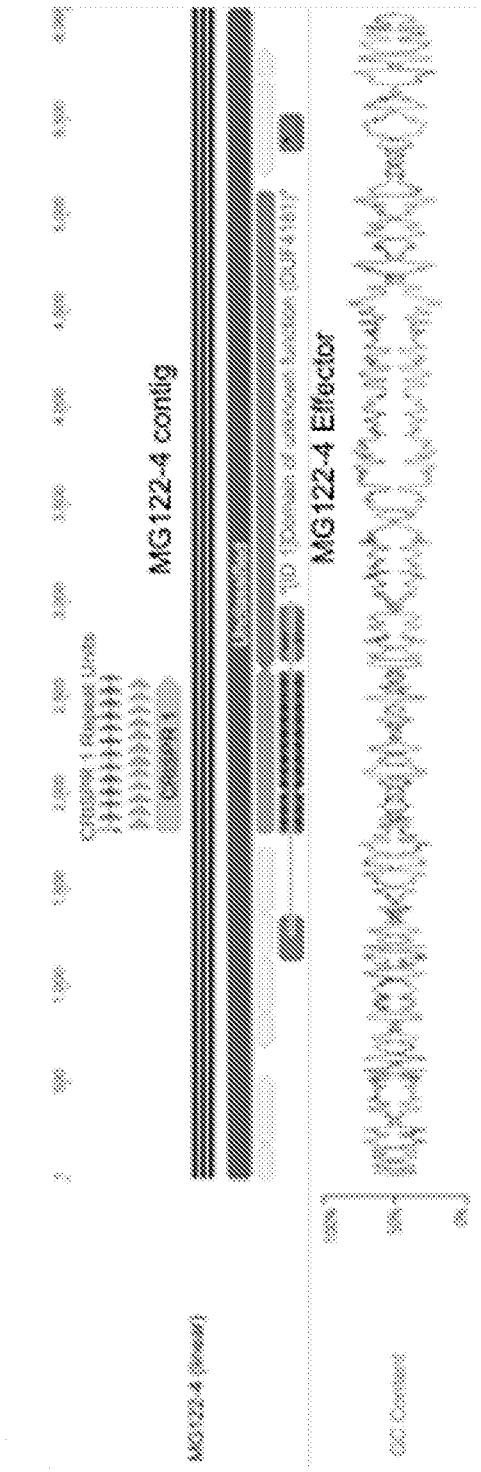


FIG. 6B

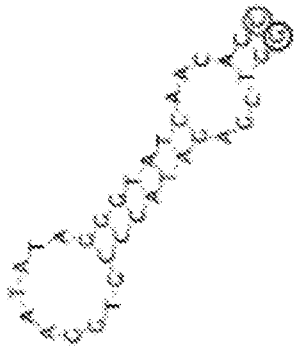


FIG. 6C

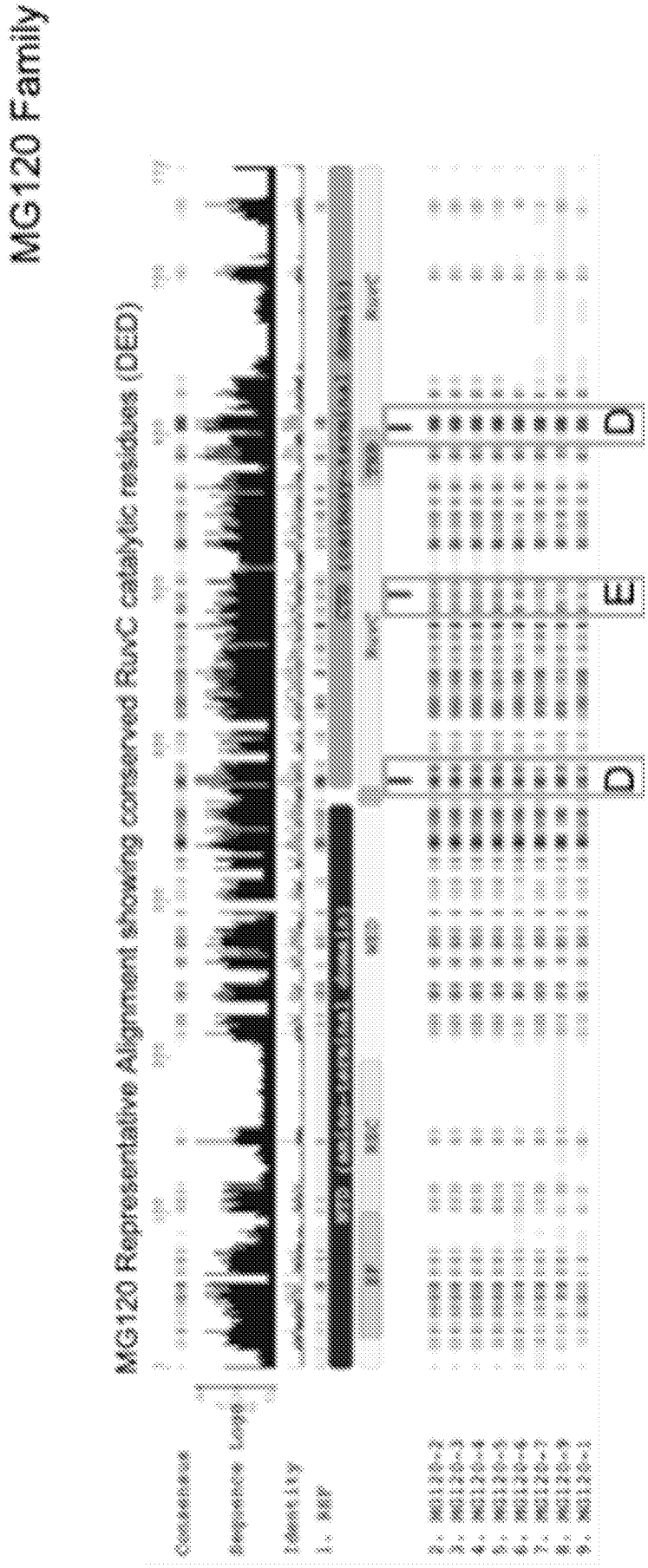

MG120 Family
MG120 Representative Alignment showing conserved RuvC catalytic residues (DED)
Consensus
Sequence Logo
Identity
1 - 100
MG120-2
MG120-3
MG120-4
MG120-5
MG120-6
MG120-1
MG120-7
MG120-8
MG120-9
MG120-10
D
E
D

FIG. 7A

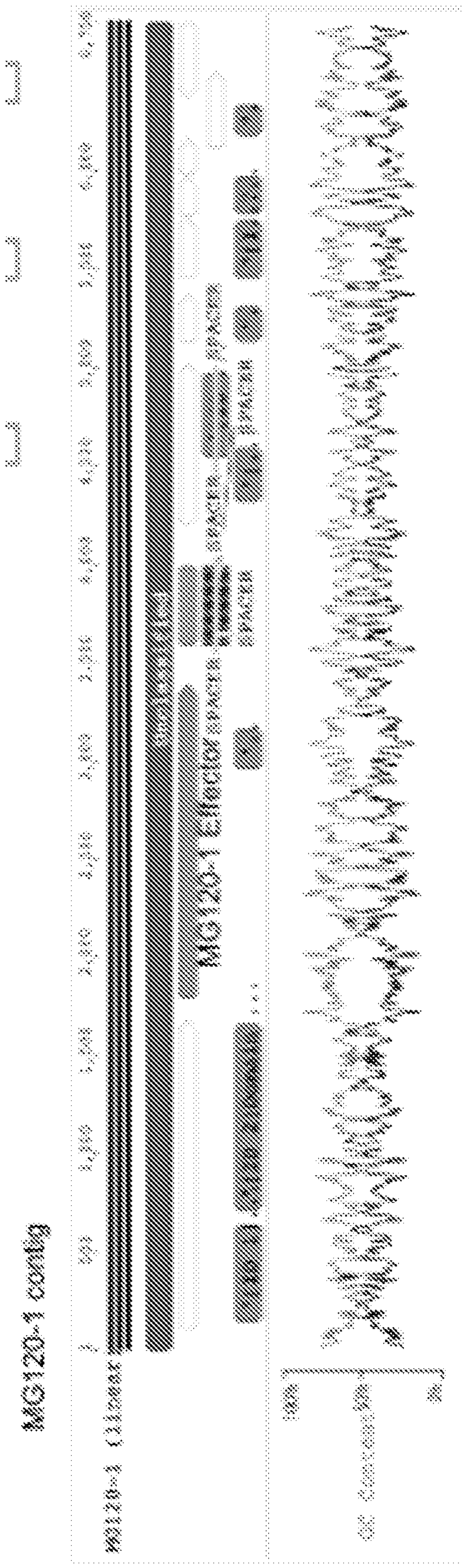

MG120-1 contig
MG120-1 Effector
SPACER

FIG. 7B

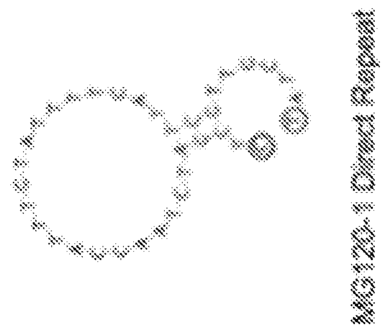

MG120-1 Direct Repeat

FIG. 7C

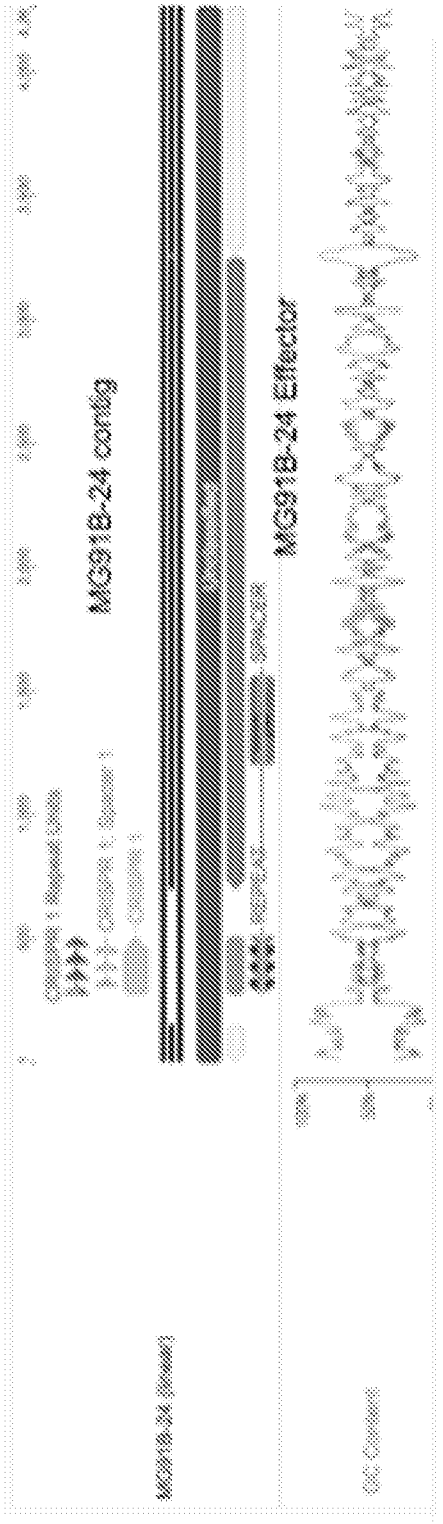
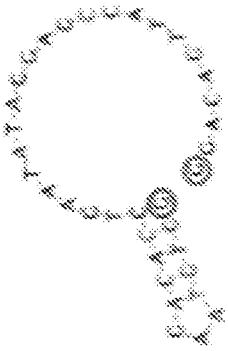


FIG. 8A



MG918-24 Direct Repeat

FIG. 8B

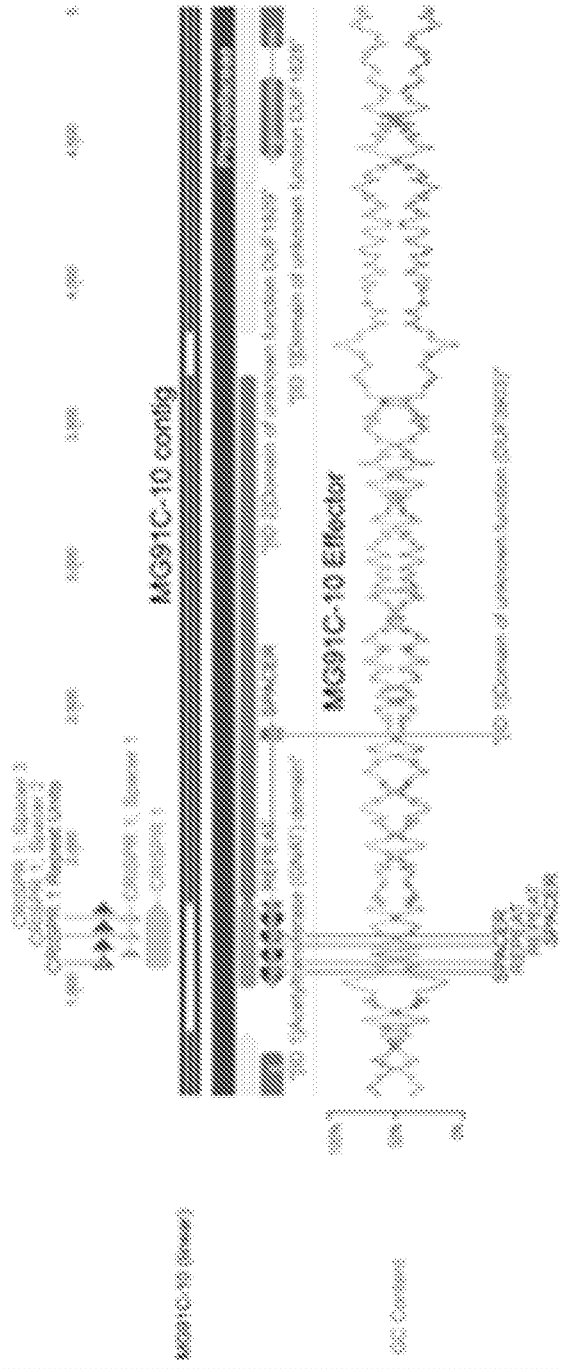
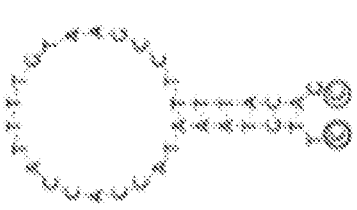


FIG. 8C



MG91C-10 Direct Repeat

FIG. 8D

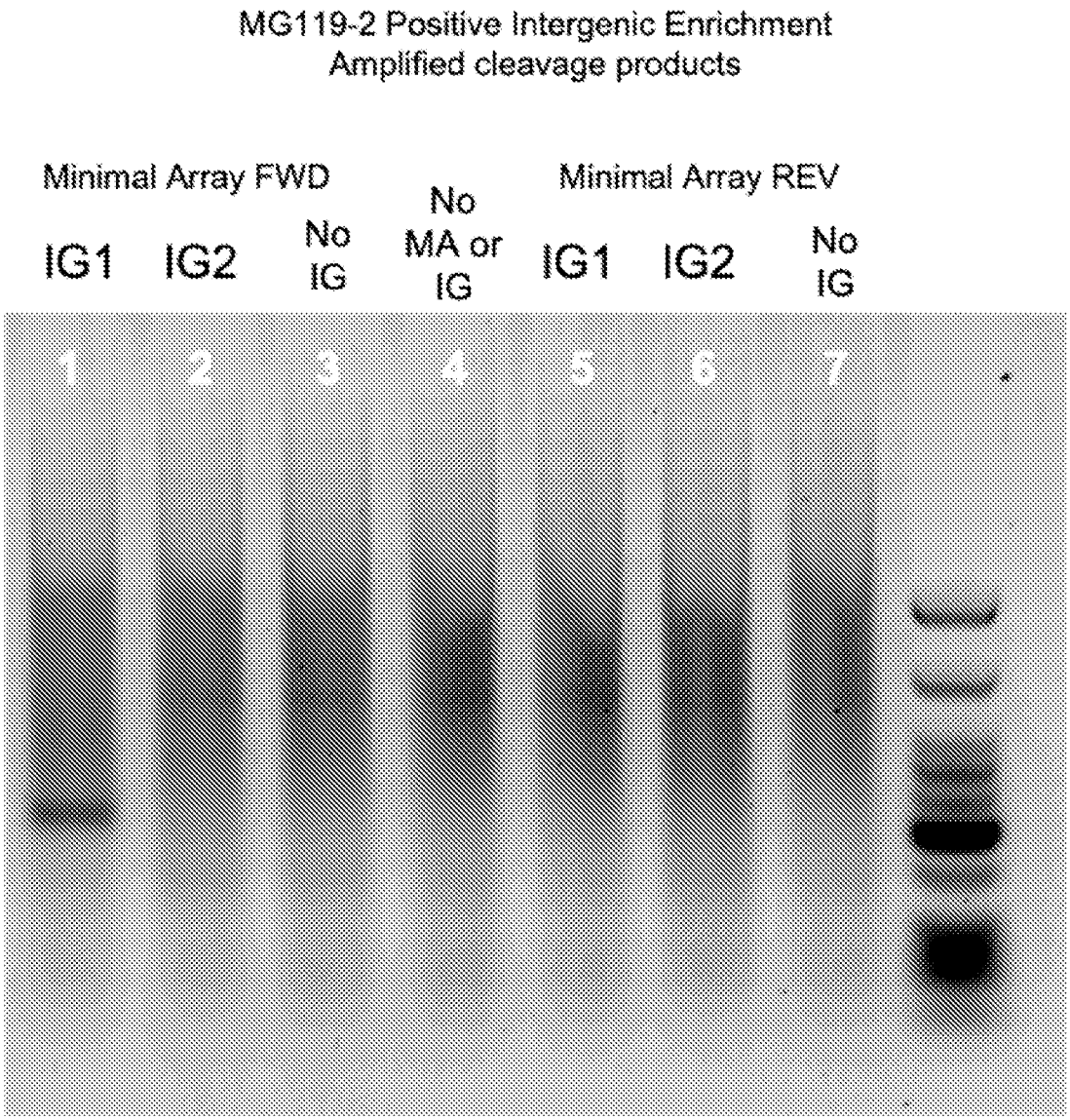

MG119-2 Positive Intergenic Enrichment
Amplified cleavage products
Minimal Array FWD
Minimal Array REV
IG1
IG2
No IG
No MA or IG
IG1
IG2
No IG
1
2
3
4
5
6
7

FIG. 9

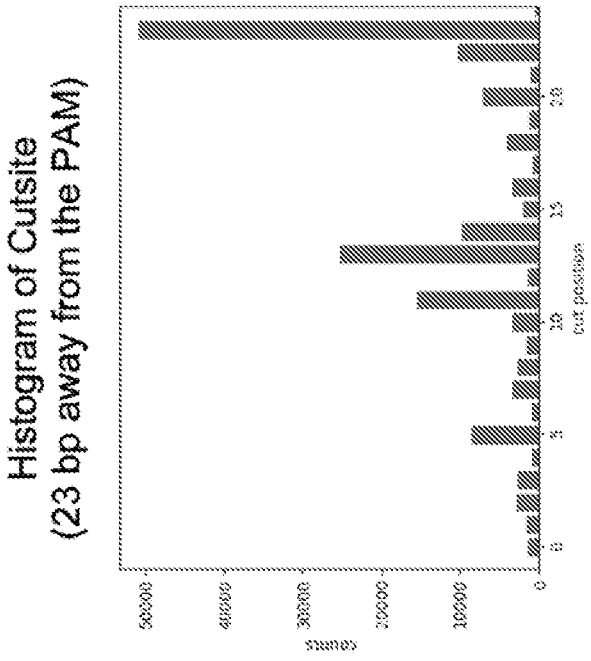


FIG. 10B

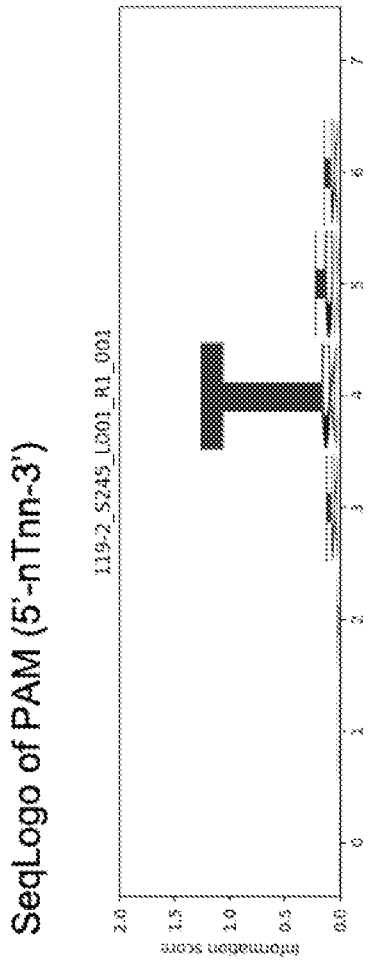


FIG. 10A

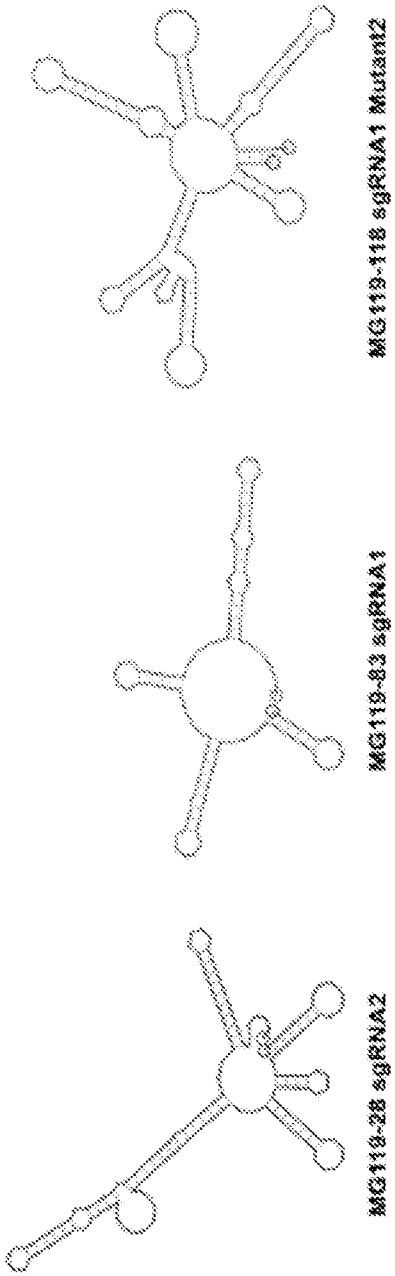


FIG. 11A

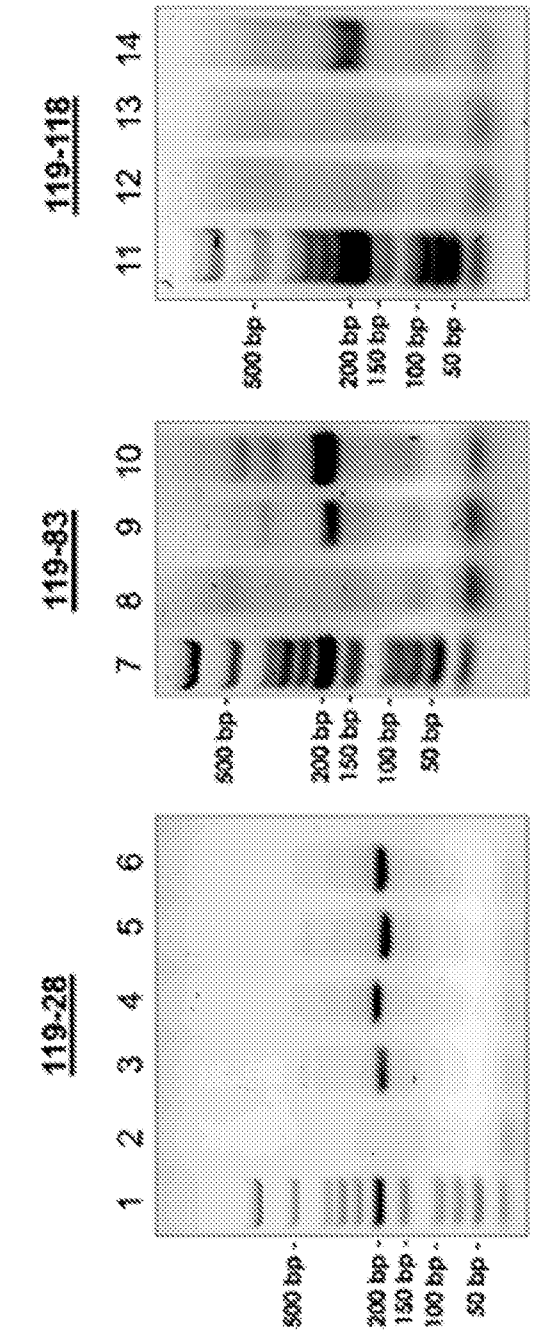


FIG. 11B

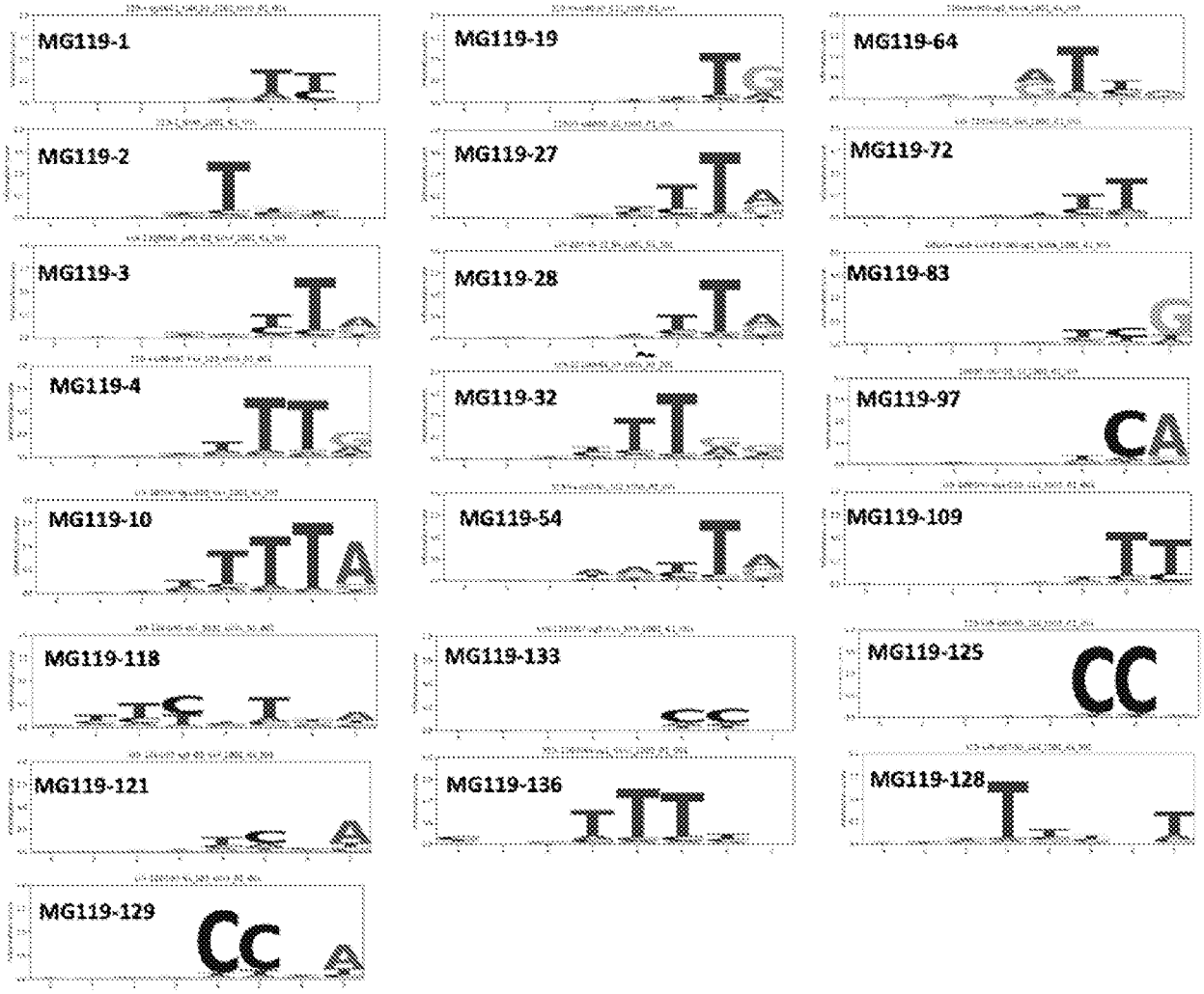

MG119-1
MG119-19
MG119-64
MG119-2
MG119-27
MG119-72
MG119-3
MG119-28
MG119-83
MG119-4
MG119-32
MG119-97
MG119-10
MG119-54
MG119-109
MG119-118
MG119-133
MG119-125
MG119-121
MG119-136
MG119-128
MG119-129

FIG. 12

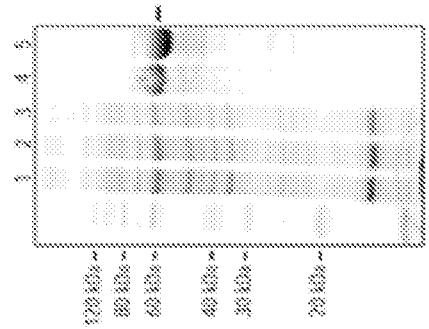


FIG. 13A

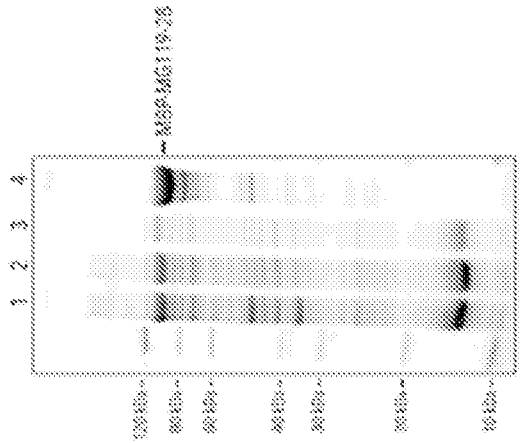


FIG. 13C

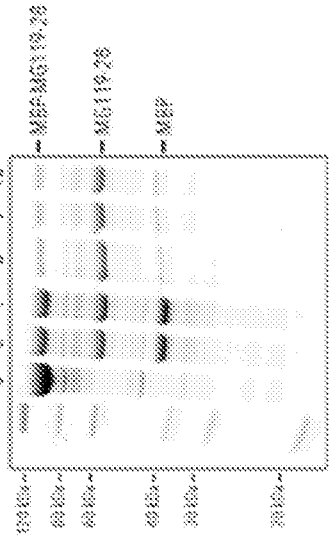


FIG. 13D

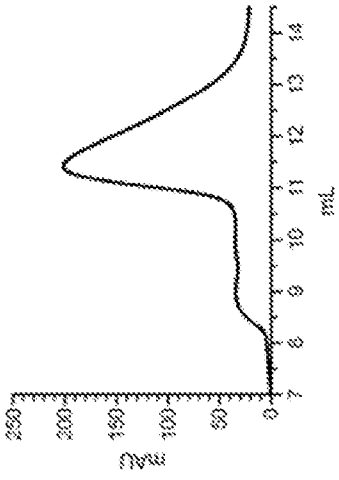


FIG. 13E

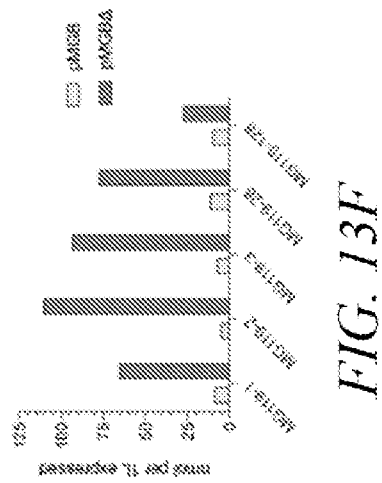


FIG. 13F

FIG. 13B

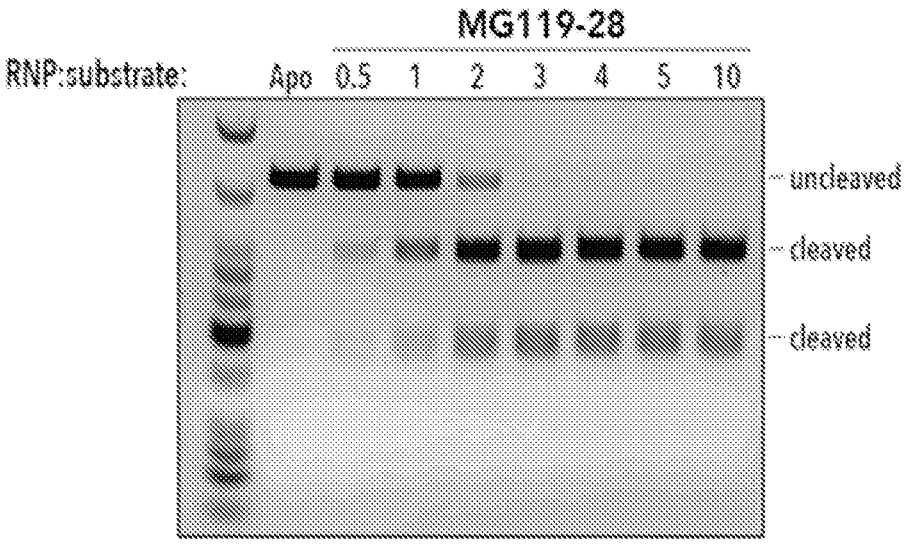


FIG. 14A

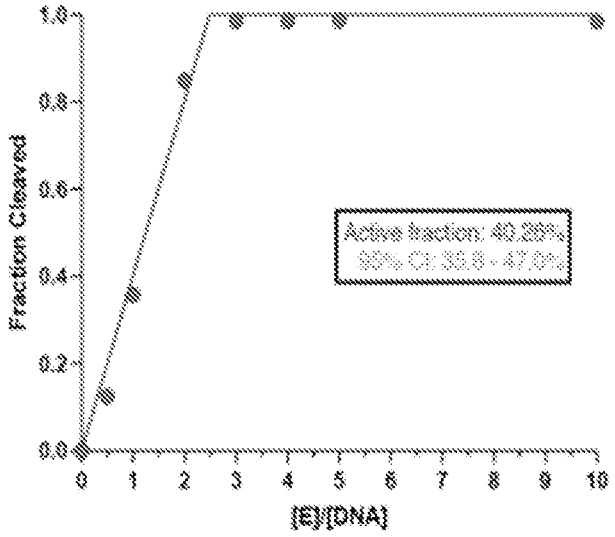


FIG. 14B

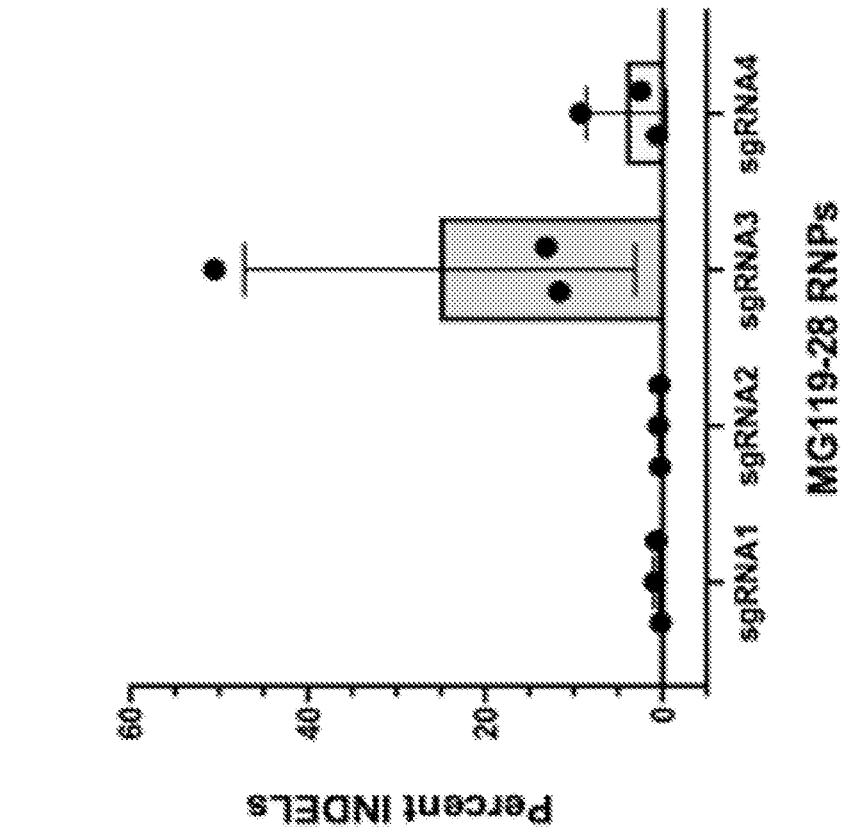


FIG. 15B

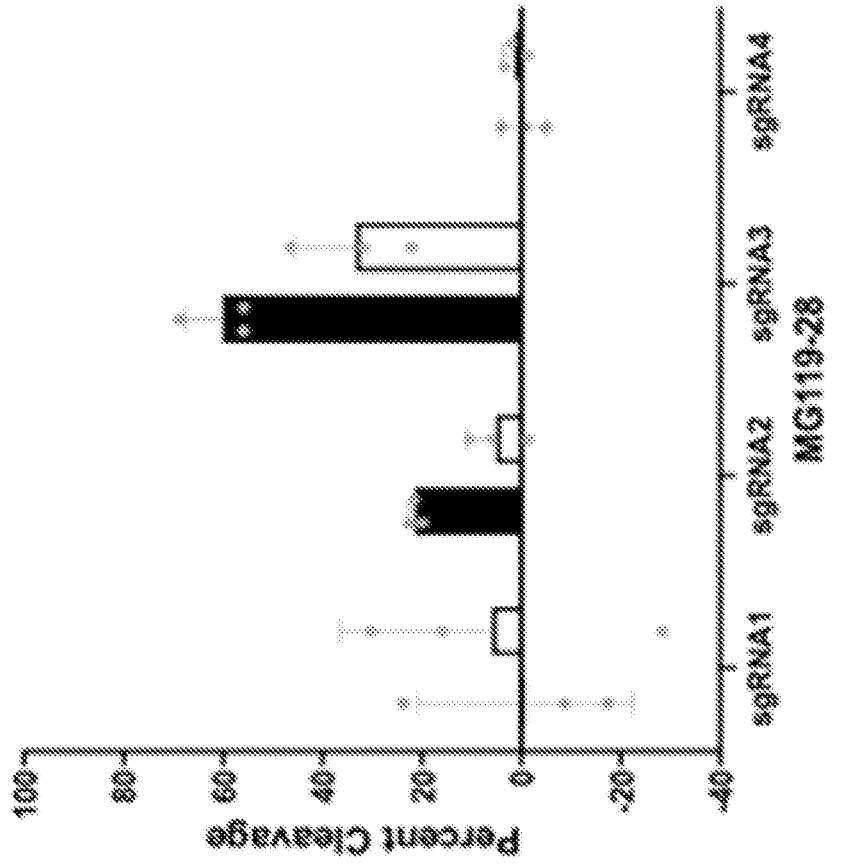


FIG. 15A