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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 January 2020 (02.01.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/006132 A1

(51) International Patent Classification:
C12Q 1/68 (2018.01) *C07K 14/195* (2006.01)
C12N 9/22 (2006.01)

(21) International Application Number:
PCT/US2019/039326

(22) International Filing Date:
26 June 2019 (26.06.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

62/690,890	27 June 2018 (27.06.2018)	US
62/716,217	08 August 2018 (08.08.2018)	US
62/852,158	23 May 2019 (23.05.2019)	US

(71) Applicant: **ALTIUS INSTITUTE FOR BIOMEDICAL SCIENCES** [US/US]; 2211 Elliot avenue 4th Floor, Seattle, Washington 98121 (US).

(72) Inventors; and
(71) Applicants: **URNOV, Fyodor** [US/US]; 2211 Elliott Avenue, 4th Floor, Seattle, Washington 98121 (US). **STAMATOYANNOPOULOS, John A.** [US/US]; 2211 Elliott Avenue, 4th Floor, Seattle, Washington 98121 (US).

(74) Agent: **CHANDRA, Shweta**; Bozicevic, Field & Francis LLP, 201 Redwood Shores Pkwy., Suite 200, Redwood City, California 94065 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

(54) **Title:** GAPPED AND TUNABLE REPEAT UNITS FOR USE IN GENOME EDITING AND GENE REGULATION COMPOSITIONS

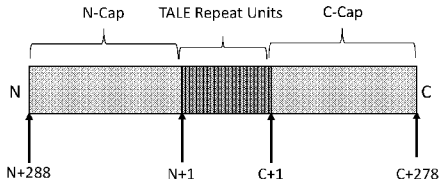

Fig. 1A

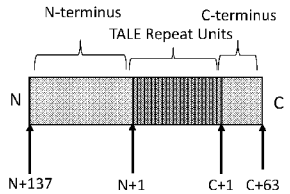

Fig. 1B

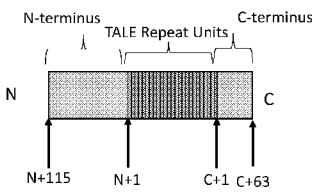

Fig. 1C

(57) **Abstract:** Provided herein are DNA binding domains comprising a plurality of repeat units, wherein each repeat unit is expanded or contracted in length. Also provided herein are DNA binding domains comprising a plurality of repeat units, wherein each repeat unit is separated from a neighboring repeat unit by a linker. In certain aspects, the linker includes a recognition site. Also disclosed are DNA binding proteins that include a fragment of N-cap sequence of a TALE protein. The TALE protein may be a Xanthomonas TALE protein.

WO 2020/006132 A1

[Continued on next page]

Published:

— *with international search report (Art. 21(3))*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
— *with sequence listing part of description (Rule 5.2(a))*

GAPPED AND TUNABLE REPEAT UNITS FOR USE IN GENOME EDITING AND GENE REGULATION COMPOSITIONS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority pursuant to 35 U.S.C. §119(e) to U.S. Provisional Application No. 62/690,890, filed June 27, 2018, U.S. Provisional Application No. 62/716,217, filed August 8, 2018, and U.S. Provisional Application No. 62/852,158, filed May 23, 2019, the disclosures of which are incorporated herein by reference in their entirety.

INCORPORATION BY REFERENCE OF SEQUENCE LISTING PROVIDED AS A TEXT FILE

[0002] A Sequence Listing is provided herewith as a text file, "ALTI-724WO Seq List_ST25.txt," created on June 26, 2019 and having a size of 443 KB. The contents of the text file are incorporated by reference herein in their entirety.

INTRODUCTION

[0003] Genome editing and gene regulation techniques require the development of nucleic acid binding domains having strong and specific binding to target genes. Provided herein are DNA binding domains with tunable binding activity. Additionally, genome editing and gene regulation compositions having functional linker regions are provided, yielding compositions that exhibit dual activities. Also provided herein are compositions and methods for genome editing and gene regulation, where the nucleic acid binding domain is derived from DNA binding proteins from bacteria from the genus *Xanthomonas*.

SUMMARY

[0004] In various aspects, the present disclosure provides a composition comprising a modular nucleic acid binding domain comprising a plurality of repeat units, wherein a repeat unit of the plurality of repeat units recognizes a target nucleic acid base and wherein the plurality of repeat units has one or more of the following characteristics: (a) at least one repeat unit comprising greater than 39 amino acid residues; (b) at least one repeat unit comprising greater than 35 amino acid residues derived from the genus of *Ralstonia*; (c) at least one repeat unit comprising less than 32 amino acid residues; and (d) each repeat unit of the plurality of repeat units is separated from a neighboring repeat unit by a linker comprising a recognition site.

[0005] In some aspects, the at least one repeat unit comprises an amino acid selected from glycine, alanine, threonine or histidine at a position after an amino acid residue at position 35. In some

aspects, the at least one repeat unit comprises an amino acid selected from glycine, alanine, threonine or histidine at a position after an amino acid residue at position 39. In some aspects, the recognition site is for a small molecule, a protease, or a kinase. In some aspects, the recognition site serves as a localization signal.

[0006] In further aspects, the composition further comprises a cleavage domain linked to the modular nucleic acid binding domain to form a non-naturally occurring fusion protein. In some aspects, the modular nucleic acid binding domain comprises a potency for a target site greater than 65% and a specificity ratio for the target site of 50:1; and a functional domain; wherein the modular nucleic acid binding domain comprises a plurality of repeat units, wherein at least one repeat unit of the plurality comprises a binding region configured to bind to a target nucleic acid base in the target site, wherein the potency comprises indel percentage at the target site, and wherein the specificity ratio comprises indel percentage at the target site over indel percentage at a top-ranked off-target site of the non-naturally occurring fusion protein.

[0007] In some aspects, the repeat unit comprises a sequence of $A_{1\text{-}11}X_1X_2B_{14\text{-}35}$, wherein each amino acid residue of $A_{1\text{-}11}$ comprises any amino acid residue; wherein X_1X_2 comprises the binding region; wherein each amino acid residue of $B_{14\text{-}35}$ comprises any amino acid; and wherein a first repeat unit of the plurality of repeat units comprises at least one residue in $A_{1\text{-}11}$, $B_{14\text{-}35}$, or a combination thereof that differs from a corresponding residue in a second repeat unit of the plurality of repeat units.

[0008] In some aspects, the binding region comprises an amino acid residue at position 13 or an amino acid residue at position 12 and the amino acid residue at position 13. In further aspects, the amino acid residue at position 13 binds to the target nucleic acid base. In still further aspects, the amino acid residue at position 12 stabilizes the configuration of the binding region. In some aspects, the indel percentage is measured by deep sequencing. In some aspects, the modular nucleic acid binding domain further comprises one or more properties selected from the following: (a) binds the target site, wherein the target site comprises a 5' guanine; (b) comprises from 7 repeat units to 25 repeat units; and (c) upon binding to the target site, the modular nucleic acid binding domain is separated from a second modular nucleic acid binding domain bound to a second target site by from 2 to 50 base pairs.

[0009] In some aspects, the plurality of repeat units comprises a *Ralstonia* repeat unit, a *Xanthomonas* repeat unit, a *Legionella* repeat unit, or any combination thereof. In further aspects, the *Ralstonia* repeat unit is a *Ralstonia solanacearum* repeat unit, the *Xanthomonas* repeat unit is a *Xanthomonas spp.* repeat unit, and the Legionella repeat unit is a *Legionella quateirensis* repeat unit.

In still further aspects, the B_{14-35} of at least one repeat unit of the plurality of repeat units has at least 92% sequence identity to GGKQALEAVRAQLLDLRAAPYG (SEQ ID NO: 280).

[0010] In some aspects, the binding region comprises HD binding to cytosine, NG binding to thymidine, NK binding to guanine, SI binding to adenosine, RS binding to adenosine, HN binding to guanine, or NT binds to adenosine. In some aspects, the at least one repeat unit comprises any one of SEQ ID NO: 267 – SEQ ID NO: 279. In some aspects, the at least one repeat unit comprises at least 80%, at least 85%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or a 100% sequence identity with any one of SEQ ID NO: 168 – SEQ ID NO: 263. In some aspects, the at least one repeat unit comprises at least 80%, at least 85%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or a 100% sequence identity with SEQ ID NO: 209, SEQ ID NO: 197, SEQ ID NO: 233, SEQ ID NO: 253, SEQ ID NO: 203, or SEQ ID NO: 218. In some aspects, the at least one repeat unit comprises any one of SEQ ID NO: 168 – SEQ ID NO: 263. In some aspects, the at least one repeat unit comprises SEQ ID NO: 209, SEQ ID NO: 197, SEQ ID NO: 233, SEQ ID NO: 253, SEQ ID NO: 203, or SEQ ID NO: 218.

[0011] In some aspects, the target nucleic acid base is cytosine, guanine, thymidine, adenosine, uracil, or a combination thereof. In some aspects, the modular nucleic acid binding domain comprises an N-terminus amino acid sequence, a C-terminus amino acid sequence, or a combination thereof. In further aspects, the N-terminus amino acid sequence is from *Xanthomonas spp.*, *Legionella quateirensis*, or *Ralstonia solanacearum*. In still further aspects, the N-terminus amino acid sequence comprises at least 80%, at least 85%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or a 100% sequence identity to SEQ ID NO: 264, SEQ ID NO: 300, SEQ ID NO: 335, SEQ ID NO: 303, SEQ ID NO: 301, SEQ ID NO: 304, or SEQ ID NO: 320, SEQ ID NO: 321, or SEQ ID NO: 322. In still further aspects, the N-terminus amino acid sequence comprises SEQ ID NO: 264, SEQ ID NO: 300, SEQ ID NO: 335, SEQ ID NO: 303, SEQ ID NO: 301, SEQ ID NO: 304, or SEQ ID NO: 320, SEQ ID NO: 321, or SEQ ID NO: 322.

[0012] In some aspects, the C-terminus amino acid sequence is from *Xanthomonas spp.*, *Legionella quateirensis*, or *Ralstonia solanacearum*. In further aspects, the C-terminus amino acid sequence comprises at least 80%, at least 85%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or a 100% sequence identity to SEQ ID NO: 266, SEQ ID NO: 298, or SEQ ID NO: 306. In still further aspects, the C-terminus amino acid sequence comprises SEQ ID NO: 266, SEQ ID NO: 298, or SEQ ID NO: 306. In some aspects, the C-terminus amino acid sequence serves as a linker between the modular nucleic acid binding domain and a functional domain.

[0013] In some aspects, the modular nucleic acid binding domain comprises a half repeat. In some aspects, the half repeat comprises at least 80%, at least 85%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or a 100% sequence identity to SEQ ID NO: 265, SEQ ID NO: 327 – SEQ ID NO: 334, or SEQ ID NO: 290. In some aspects, the functional domain is a cleavage domain or a repression domain. In some aspects, the cleavage domain comprises at least 33.3% divergence from SEQ ID NO: 163 and is immunologically orthogonal to SEQ ID NO: 163. In some aspects, the composition comprises one or more of the following characteristics: (a) induces greater than 1% indels at the target site; (b) the cleavage domain comprises a molecular weight of less than 23 kDa; (c) the cleavage domain comprises less than 196 amino acids; and (d) capable of cleaving across a spacer region greater than 24 base pairs.

[0014] In some aspects, the composition induces greater than 5%, greater than 10%, greater than 20%, greater than 30%, greater than 40%, greater than 50%, greater than 60%, greater than 70%, greater than 80%, or greater than 90% indels at the target site. In some aspects, the cleavage domain comprises at least 35%, at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, or at least 75% divergence from SEQ ID NO: 163. In some aspects, the cleavage domain comprises a sequence selected from SEQ ID NO: 316 – SEQ ID NO: 319.

[0015] In some aspects, the cleavage domain comprises a nucleic acid sequence encoding for a sequence having at least 80% sequence identity with SEQ ID NO: 1 – SEQ ID NO: 81. In some aspects, the cleavage domain comprises a nucleic acid sequence encoding for a sequence selected from SEQ ID NO: 1 – SEQ ID NO: 81. In some aspects, the nucleic acid sequence comprises at least 80% sequence identity with SEQ ID NO: 82 – SEQ ID NO: 162. In some aspects, the nucleotide sequence encoding for the sequence comprises any one of SEQ ID NO: 82 – SEQ ID NO: 162.

[0016] In some aspects, the repression domain comprises KRAB, Sin3a, LSD1, SUV39H1, G9A (EHMT2), DNMT1, DNMT3A-DNMT3L, DNMT3B, KOX, TGF-beta-inducible early gene (TIEG), v-erbA, SID, MBD2, MBD3, Rb, or MeCP2. In some aspects, the plurality of repeat units comprises 3 to 60 repeat units.

[0017] In some aspects, the target site is a nucleic acid sequence within a PDCD1 gene, a CTLA4 gene, a LAG3 gene, a TET2 gene, a BTLA gene, a HAVCR2 gene, a CCR5 gene, a CXCR4 gene, a TRA gene, a TRB gene, a B2M gene, an albumin gene, a HBB gene, a HBA1 gene, a TTR gene, a NR3C1 gene, a CD52 gene, an erythroid specific enhancer of the BCL11A gene, a CBLB gene, a TGFBR1 gene, a SERPINA1 gene, a HBV genomic DNA in infected cells, a CEP290 gene, a DMD gene, a CFTR gene, or an IL2RG gene.

[0018] In other aspects, a nucleic acid sequence encoding a chimeric antigen receptor (CAR), alpha-L iduronidase (IDUA), iduronate-2-sulfatase (IDS), or Factor 9 (F9), is inserted at the target site.

[0019] In various aspects, the present disclosure provides a method of genome editing, the method comprising: administering any of the above compositions and inducing a double stranded break.

[0020] Also provided herein is a non-naturally occurring DNA binding polypeptide that includes from N- to C-terminus: a N-terminus region comprising at least residues N+110 to N+1 of a TALE protein, where the N-terminus region does not include residues N+288 to N+116 of the TALE protein; a plurality of TALE repeat units derived from a TALE protein; and C-terminus region of a TALE protein. The N-terminus region may not include at least amino acids N+288 to N+116 of the TALE protein. The N-terminus region may not include amino acids N+288 to up to N+116 of the TALE protein. The N-terminus region may not include at least amino acids N+288 to up to N+111 of the TALE protein. The N-terminus region may include residues N+1 to up to N+115 of the TALE protein. The N-terminus region may include residues N+1 to up to N+110 of the TALE protein. The C-terminus region may include full length C-terminus region of a TALE protein or a fragment thereof, e.g., residues C+1 to C+63 of the TALE protein. The DNA binding polypeptide may be fused to a heterologous functional domain, such as, enzyme, a transcriptional activator, a transcriptional repressor, or a DNA nucleotide modifier. The N-terminus region, the TALE repeat units, and the C-terminus region may be derived from the same TALE protein or from different TALE proteins. The TALE proteins from which the N-terminus region, the TALE repeat units, and the C-terminus region may be derived include *Xanthomonas* TALE proteins, such as, AvrBs3, AVRHAH1, AvrXa7, AVRB6, or AvrXa10.

[0021] In various aspects, the present disclosure provides a method of genome editing, the method comprising: administering any of the above polypeptides or compositions thereof and inducing a double stranded break.

[0022] In various aspects, the present disclosure provides method of gene repression, the method comprising administering any of the above polypeptides or compositions thereof and repressing gene expression.

INCORPORATION BY REFERENCE

[0023] 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 DRAWINGS

[0024] FIGS. 1A-1C show schematics of the domain structure of DNA binding proteins (not drawn to scale).

FIG. 2 shows nuclease activity mediated by DNA binding protein dimers that each include from N-terminus to C-terminus: a N-terminus region of a TALE protein, TALE repeat units, C-terminus region of a TALE protein, and a Fok1 endonuclease.

DETAILED DESCRIPTION

[0025] The present disclosure provides compositions and methods for genome editing and gene regulation (including activation and repression) with DNA binding domains fused to functional domains via linkers that serve as recognition sites for further activity (e.g., a non-nuclease enzyme activity). The present disclosure also provides compositions and methods for genome editing and gene regulation with DNA binding domains that can have enhanced binding to a target nucleic acid sequence. Enhanced binding to a target nucleic acid sequence can be achieved with the DNA binding domains of the present disclosure in which repeat units can be varied in length to tune for binding activity.

Linkers Comprising Recognition Sites

[0026] In some embodiments, the present disclosure provides DNA binding domains with gapped repeat units for use as gene editing complexes. A DNA binding domain with gapped repeat can comprise of a plurality of repeat units in which each repeat unit of the plurality of repeat units is separated from a neighboring repeat unit by a linker. This linker can comprise a recognition site for additional functionality and activity. For example, the linker can comprise a recognition site for a small molecule. As another example, the linker can serve as a recognition site for a protease. In yet another example, the linker can serve as a recognition site for a kinase. In other embodiments, the recognition site can serve as a localization signal.

[0027] Each repeat unit of a DNA binding domain (*e.g.*, RNBDs, MAP-NBDs, TALEs) comprises a secondary structure in which the RVD interfaces with and binds to a target nucleic acid base on double stranded DNA, while the remainder of the repeat unit protrudes from the surface of the DNA. Thus, the linkers comprising a recognition site between each repeat unit are removed from the surface of the DNA and are solvent accessible. In some embodiments, these solvent accessible linkers comprising recognition sites can have extra activity while mediating gene editing.

[0028] Examples of a left and a right DNA binding domain comprising repeat units derived from *Xanthomonas spp.* are shown below in **TABLE 1** for AAVS1 and GA7. "X," shown in bold and

underlining, represents a linker comprising a recognition site and can comprise 1-40 amino acid residues. An amino acid residue of the linker can comprise a glycine, an alanine, a threonine, or a histidine.

[0029] In some embodiments, "derived" indicates that a protein is from a particular source (e.g., *Ralstonia*), is a variant of a protein from a particular source (e.g., *Ralstonia*), is a mutated or modified form of the protein from a particular source (e.g., *Ralstonia*), and shares at least 30% sequence identity with, at least 40% sequence identity with, at least 50% sequence identity with, at least 60% sequence identity with, at least 70% sequence identity with, at least 80% sequence identity with, or at least 90% sequence identity with a protein from a particular source (e.g., *Ralstonia*, *Xanthomonas*, or *Legionella*).

TABLE 1: Exemplary Left or Right Gapped DNA Binding Domains

SEQ ID NO	Construct	Sequence
SEQ ID NO: 307	AAVS1_Left	LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVV AIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDG GKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDGGKQALET VQRLLPVLCQDHGXLTPDQVVAIASNGGGKQALETVQRLLPV LCQDHGXLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGX LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVV AIASNIGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDGG KQALETVQRLLPVLCQDHGXLTPDQVVAIASHDGGKQALETV QRLLPVLCQDHGXLTPDQVVAIASHDGGKQALETVQRLLPVL CQDHGXLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXL TPDQVVAIASNIGGKQALETVQRLLPVLCQDHGXLTPDQVVAI ASHDGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASNIGGK QALETVQRLLPVLCQDHGXLTPDQVVAIASNHGGKQALETVQ RLLPVLCQDHGXLTPDQVVAIASNGGG
SEQ ID NO: 308	AAVS1_Right	LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGXLTPDQVV AIASNGGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASNGG GKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDGGKQALET VQRLLPVLCQDHGXLTPDQVVAIASNGGGKQALETVQRLLPV LCQDHGXLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGX LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGXLTPDQVV AIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASNIGG KQALETVQRLLPVLCQDHGXLTPDQVVAIASHDGGKQALETV

SEQ ID NO	Construct	Sequence
		QRLLPVLCQDHGXLTPDQVVAIASHDGGKQALETVQRLLPVL CQDHGXLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGXL TPDQVVAIASNIGGKQALETVQRLLPVLCQDHGXLTPDQVVAI ASNGGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDGGK QALETVQRLLPVLCQDHGXLTPDQVVAIASHDGGKQALETVQ RLLPVLCQDHGXLTPDQVVAIASNGGGKQALESIVAQLSRPDP ALA
SEQ ID NO: 309	GA7.2 Left	LTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGXLTPDQVV AIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASNGG GKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDGGKQALET VQRLLPVLCQDHGXLTPDQVVAIASNIGGKQALETVQRLLPVL CQDHGXLTPDQVVAIASNHGGKQALETVQRLLPVLCQDHGXL TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVVA IASHDGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDGG KQALETVQRLLPVLCQDHGXLTPDQVVAIASNIGGKQALETV QRLLPVLCQDHGXLTPDQVVAIASNHGGKQALETVQRLLPVL CQDHGXLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXL TPDQVVAIASNGGGKQALETVQRLLPVLCQDHGXLTPDQVVA IASHDGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASNIGGK QALETVQRLLPVLCQDHGXLTPDQVVAIASNHGGKQALETVQ RLLPVLCQDHGXLTPDQVVAIASHDGGKQALETVQRLLPVLC QDHGXLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXLT PDQVVAIASNGGGK
SEQ ID NO: 310	GA7.2 Right	LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVV AIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDG GKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDGGKQALET VQRLLPVLCQDHGXLTPDQVVAIASHDGGKQALETVQRLLPV LCQDHGXLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGX LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVV AIASNGGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDG GKQALETVQRLLPVLCQDHGXLTPDQVVAIASNIGGKQALET VQRLLPVLCQDHGXLTPDQVVAIASNGGGKQALETVQRLLPV LCQDHGXLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGX LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVV

SEQ ID NO	Construct	Sequence
		AIASNGGGKQALETVQRLLPVLCQDHGXLTPDQVVAIASHDG GKQALETVQRLLPVLCQDHGXLTPDQVVAIASNGGGKQALET VQRLLPVLCQDHGXLTPDQVVAIASNIGGKQALETVQRLLPVL CQDHGXLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXL TPDQVVAIASHDGGKQALETVQRLLPVLCQDHGXLTPDQVVA IASNIGGKQALETVQRLLPVLCQDHGXLTPDQVVASASNGGG KQALESIVAQLSRPDPALA

Tunable Repeat Units

[0030] In some embodiments, the present disclosure provides DNA binding domains (*e.g.*, RNBDs, MAP-NBDs, TALEs) with expanded repeat units. For example, a DNA binding domain (*e.g.*, RNBDs, MAP-NBDs, TALEs) comprises a plurality of repeat units in which each repeat unit is usually 33-35 amino acid residues in length. The present disclosure provides repeat units, which are greater than 35 amino acid residues in length. In some embodiments, the present disclosure provides repeat units, which are greater than 39 amino acid residues in length. In some embodiments, the present disclosure provides repeat units which are 35 to 40 amino acid residues long, 39 to 40 amino acid residues long, 35 to 45 amino acid residues long, 39 to 45 amino acid residues long, 35 to 50 amino acid residues long, 39 to 50 amino acid residues long, 35 to 50 amino acid residues long, 35 to 60 amino acid residues long, 39 to 60 amino acid residues long, 35 to 70 amino acid residues long, 39 to 70 amino acid residues long, 35 to 79 amino acid residues long, or 39 to 79 amino acid residues long.

[0031] In other embodiments, the present disclosure provides DNA binding domains (*e.g.*, RNBDs, MAP-NBDs, TALEs) with contracted repeat units. For example, the present disclosure provides repeat units, which are less than 32 amino acid residues in length. In some embodiments, the present disclosure provides repeat units, which are 15 to 32 amino acid residues in length, 16 to 32 amino acid residues in length, 17 to 32 amino acid residues in length, 18 to 32 amino acid residues in length, 19 to 32 amino acid residues in length, 20 to 32 amino acid residues in length, 21 to 32 amino acid residues in length, 22 to 32 amino acid residues in length, 23 to 32 amino acid residues in length, 24 to 32 amino acid residues in length, 25 to 32 amino acid residues in length, 26 to 32 amino acid residues in length, 27 to 32 amino acid residues in length, 28 to 32 amino acid residues in length, 29 to 32 amino acid residues in length, 30 to 32 amino acid residues in length, or 31 to 32 amino acid residues in length.

[0032] In some embodiments, said expanded repeat units can be tuned to modulate binding of each repeat unit to its target nucleic acid, resulting in the ability to overall modulate binding of the DNA binding domain to a target gene of interest. For example, expanding repeat units can improve binding affinity of the repeat unit to its target nucleic acid base and thereby increase binding affinity of the DNA binding domain to a target gene. In some embodiments, expanding repeat units can improve specificity of the DNA binding domain for a target gene. In other embodiments, contracting repeat units can improve binding affinity of the repeat unit to its target nucleic acid base and thereby increase binding affinity of the DNA binding domain for a target gene.

[0033] Described in further detail below are DNA binding domains from the genus of *Ralstonia,* the genus of animal pathogens (e.g., *Legionella*, *Burkholderia*, *Paraburkholderia*, or *Francisella*), and the genus of *Xanthomonas*, which can comprise linkers comprising recognition sites, expanded repeat units, or contracted repeat units, as described in detail above.

[0034] In some embodiments, the present disclosure provides a composition comprising a modular nucleic acid binding domain comprising a plurality of repeat units, wherein a repeat unit of the plurality of repeat units recognizes a target nucleic acid base and wherein the plurality of repeat units has one or more of the following characteristics: (a) at least one repeat unit comprising greater than 39 amino acid residues; (b) at least one repeat unit comprising greater than 35 amino acid residues derived from the genus of *Ralstonia*; (c) at least one repeat unit comprising less than 32 amino acid residues; and (d) each repeat unit of the plurality of repeat units is separated from a neighboring repeat unit by a linker comprising a recognition site.

[0035] In some embodiments, the at least one repeat unit comprises an amino acid selected from glycine, alanine, threonine or histidine at a position after an amino acid residue at position 35. In some embodiments, the at least one repeat unit comprises an amino acid selected from glycine, alanine, threonine or histidine at a position after an amino acid residue at position 39. In some aspects, the recognition site is for a small molecule, a protease, or a kinase. In some aspects, the recognition site serves as a localization signal.

Ralstonia-Derived DNA Binding Domains

[0036] The present disclosure provides modular nucleic acid binding domains (NBDs) derived from the genus of bacteria. For example, in some embodiments, the present disclosure provides NBDs derived from bacteria that serve as plant pathogens, such as from the genus of *Xanthomonas spp.* and *Ralstonia*. In particular embodiments, the present disclosure provides NBDs from the genus of *Ralstonia*. Also provided herein are NBDs from the animal pathogen, *Legionella*, Provided herein

are sequences of repeat units derived from the genus of *Ralstonia*, which can be linked together to form non-naturally occurring modular nucleic acid binding domains (NBDs), capable of targeting and binding any target nucleic acid sequence (e.g., DNA sequence).

[0037] In some embodiments, "modular" indicates that a particular composition such as a nucleic acid binding domain, comprises a plurality of repeat units that can be switched and replaced with other repeat units. For example, any repeat unit in a modular nucleic acid binding domain can be switched with a different repeat unit. In some embodiments, modularity of the nucleic acid binding domains disclosed herein allows for switching the target nucleic acid base for a particular repeat unit by simply switching it out for another repeat unit. In some embodiments, modularity of the nucleic acid binding domains disclosed herein allows for swapping out a particular repeat unit for another repeat unit to increase the affinity of the repeat unit for a particular target nucleic acid. Overall, the modular nature of the nucleic acid binding domains disclosed herein enables the development of genome editing complexes that can precisely target any nucleic acid sequence of interest.

[0038] In particular embodiments, modular nucleic acid binding domains (NBDs), also referred to herein as "DNA binding polypeptides," are provided herein from the genus of *Ralstonia solanacearum*. In some embodiments, modular nucleic acid binding domains derived from *Ralstonia* (RNBDs) can be engineered to bind to a target gene of interest for purposes of gene editing or gene regulation. An RNBD can be engineered to target and bind a specific nucleic acid sequence. The nucleic acid sequence can be DNA or RNA.

[0039] In some embodiments, the RNBD can comprise a plurality of repeat units, wherein each repeat unit recognizes and binds to a single nucleotide (in DNA or RNA) or base pair. Each repeat unit in the plurality of repeat units can be specifically selected to target and bind to a specific nucleic acid sequence, thus contributing to the modular nature of the DNA binding polypeptide. A non-naturally occurring *Ralstonia*-derived modular nucleic acid binding domain can comprise a plurality of repeat units, wherein each repeat unit of the plurality of repeat units recognizes a single target nucleotide, base pair, or both.

[0040] In some embodiments, the repeat unit of a modular nucleic acid binding domain can be derived from a bacterial protein. For example, the bacterial protein can be a transcription activator like effector-like protein (TALE-like protein). The bacterial protein can be derived from *Ralstonia solanacearum*. Repeat units derived from *Ralstonia solanacearum* can be 33-35 amino acid residues in length. In some embodiments, the repeat can be derived from the naturally occurring *Ralstonia solanacearum* TALE-like protein.

[0041] TABLE 2 below shows exemplary repeat units derived from the genus of *Ralstonia*, which are capable of binding a target nucleic acid.

TABLE 2 – Exemplary *Ralstonia*-derived Repeat Units

SEQ ID NO	Sequence
SEQ ID NO: 168	LDTEQVVAIASHNGGKQALEAVKADLLDLLGAPYV
SEQ ID NO: 169	LDTEQVVAIASHNGGKQALEAVKADLLDLRGAPYA
SEQ ID NO: 170	LDTEQVVAIASHNGGKQALEAVKADLLELRGAPYA
SEQ ID NO: 171	LDTEQVVAIASHNGGKQALEAVKAHLLDLRGAPYA
SEQ ID NO: 172	LNTEQVVAIASHNGGKQALEAVKADLLDLRGAPYA
SEQ ID NO: 173	LNTEQVVAIASNNGGKQALEAVKTHLLDLRGARYA
SEQ ID NO: 174	LNTEQVVAIASNPGGKQALEAVRALFPDLRAAPYA
SEQ ID NO: 175	LNTEQVVAIASSHGGKQALEAVRALFPDLRAAPYA
SEQ ID NO: 176	LNTEQVVAVASNKGGKQALEAVGAQLLALRAVPYA
SEQ ID NO: 177	LNTEQVVAVASNKGGKQALEAVGAQLLALRAVPYE
SEQ ID NO: 178	LSAAQVVAIASHDGGKQALEAVGTQLVALRAAPYA
SEQ ID NO: 179	LSIAQVVAVASRSGGKQALEAVRAQLLALRAAPYG
SEQ ID NO: 180	LSPEQVVAIASNHGGKQALEAVRALFRGLRAAPYG
SEQ ID NO: 181	LSPEQVVAIASNNGGKQALEAVKAQLLELRAAPYE
SEQ ID NO: 182	LSTAQLVAIASNPGGKQALEAIRALFRELRAAPYA
SEQ ID NO: 183	LSTAQLVAIASNPGGKQALEAVRALFRELRAAPYA
SEQ ID NO: 184	LSTAQLVAIASNPGGKQALEAVRAPFREVRAAPYA
SEQ ID NO: 185	LSTAQLVSIASNPGGKQALEAVRALFRELRAAPYA
SEQ ID NO: 186	LSTAQVAAIASHDGGKQALEAVGTQLVVLRAAPYA
SEQ ID NO: 187	LSTAQVATIASSIGGRQALEALKVQLPVLRAAPYG
SEQ ID NO: 188	LSTAQVATIASSIGGRQALEAVKVQLPVLRAAPYG
SEQ ID NO: 189	LSTAQVVAIAANNGGKQALEAVRALLPVLRVAPYE
SEQ ID NO: 190	LSTAQVVAIAGNGGGKQALEGIGEQLLKLRTAPYG
SEQ ID NO: 191	LSTAQVVAIASHDGGKQALEAAGTQLVALRAAPYA
SEQ ID NO: 192	LSTAQVVAIASHDGGKQALEAVGAQLVELRAAPYA
SEQ ID NO: 193	LSTAQVVAIASHDGGKQALEAVGTQLVALRAAPYA
SEQ ID NO: 194	LSTAQVVAIASHDGGNQALEAVGTQLVALRAAPYA
SEQ ID NO: 195	LSTAQVVAIASHNGGKQALEAVKAQLLDLRGAPYA
SEQ ID NO: 196	LSTAQVVAIASNDGGKQALEEVEAQLLALRAAPYE
SEQ ID NO: 197	LSTAQVVAIASNGGGKQALEGIGEQLLKLRTAPYG
SEQ ID NO: 198	LSTAQVVAIASNGGGKQALEGIGEQLRKLRTAPYG
SEQ ID NO: 199	LSTAQVVAIASNPGGKQALEAVRALFRELRAAPYA
SEQ ID NO: 200	LSTAQVVAIASQNGGKQALEAVKAQLLDLRGAPYA
SEQ ID NO: 201	LSTAQVVAIASSHGGKQALEAVRALFRELRAAPYG
SEQ ID NO: 202	LSTAQVVAIASSNGGKQALEAVWALLPVLRATPYD
SEQ ID NO: 203	LSTAQVVAIATRSGGKQALEAVRAQLLDLRAAPYG

SEQ ID NO	Sequence
SEQ ID NO: 204	LSTAQVVAVAGRNGGKQALEAVRAQLPALRAAPYG
SEQ ID NO: 205	LSTAQVVAVASSNGGKQALEAVWALLPVLRATPYD
SEQ ID NO: 206	LSTAQVVTIASSNGGKQALEAVWALLPVLRATPYD
SEQ ID NO: 207	LSTEQVVAIAGHDGGKQALEAVGAQLVALRAAPYA
SEQ ID NO: 208	LSTEQVVAIASHDGGKQALEAVGAQLVALLAAPYA
SEQ ID NO: 209	LSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYA
SEQ ID NO: 210	LSTEQVVAIASHDGGKQALEAVGGQLVALRAAPYA
SEQ ID NO: 211	LSTEQVVAIASHDGGKQALEAVGTQLVALRAAPYA
SEQ ID NO: 212	LSTEQVVAIASHDGGKQALEAVGVQLVALRAAPYA
SEQ ID NO: 213	LSTEQVVAIASHDGGKQALEAVVAQLVALRAAPYA
SEQ ID NO: 214	LSTEQVVAIASHDGGKQPLEAVGAQLVALRAAPYA
SEQ ID NO: 215	LSTEQVVAIASHGGGKQVLEGIGEQLLKLRAAPYG
SEQ ID NO: 216	LSTEQVVAIASHKGGKQALEGIGEQLLKLRAAPYG
SEQ ID NO: 217	LSTEQVVAIASHNGGKQALEAVKADLLDLRGAPYA
SEQ ID NO: 218	LSTEQVVAIASHNGGKQALEAVKADLLELRGAPYA
SEQ ID NO: 219	LSTEQVVAIASHNGGKQALEAVKAHLLDLRGAPYA
SEQ ID NO: 220	LSTEQVVAIASHNGGKQALEAVKAHLLDLRGVPYA
SEQ ID NO: 221	LSTEQVVAIASHNGGKQALEAVKAHLLELRGAPYA
SEQ ID NO: 222	LSTEQVVAIASHNGGKQALEAVKAQLLDLRGAPYA
SEQ ID NO: 223	LSTEQVVAIASHNGGKQALEAVKAQLLELRGAPYA
SEQ ID NO: 224	LSTEQVVAIASHNGGKQALEAVKAQLPVLRRAPYG
SEQ ID NO: 225	LSTEQVVAIASHNGGKQALEAVKTQLLELRGAPYA
SEQ ID NO: 226	LSTEQVVAIASHNGGKQALEAVRAQLPALRAAPYG
SEQ ID NO: 227	LSTEQVVAIASHNGSKQALEAVKAQLLDLRGAPYA
SEQ ID NO: 228	LSTEQVVAIASNGGGKQALEGIGKQLQELRAAPHG
SEQ ID NO: 229	LSTEQVVAIASNGGGKQALEGIGKQLQELRAAPYG
SEQ ID NO: 230	LSTEQVVAIASNHGGKQALEAVRALFRELRAAPYA
SEQ ID NO: 231	LSTEQVVAIASNHGGKQALEAVRALFRGLRAAPYG
SEQ ID NO: 232	LSTEQVVAIASNKGGKQALEAVKADLLDLRGAPYV
SEQ ID NO: 233	LSTEQVVAIASNKGGKQALEAVKAHLLDLLGAPYV
SEQ ID NO: 234	LSTEQVVAIASNKGGKQALEAVKAQLLALRAAPYA
SEQ ID NO: 235	LSTEQVVAIASNKGGKQALEAVKAQLLELRGAPYA
SEQ ID NO: 236	LSTEQVVAIASNNGGKQALEAVKALLLELRAAPYE
SEQ ID NO: 237	LSTEQVVAIASNNGGKQALEAVKAQLLALRAAPYE
SEQ ID NO: 238	LSTEQVVAIASNNGGKQALEAVKAQLLDLRGAPYA
SEQ ID NO: 239	LSTEQVVAIASNNGGKQALEAVKAQLLVLRAAPYG
SEQ ID NO: 240	LSTEQVVAIASNNGGKQALEAVKAQLPALRAAPYE
SEQ ID NO: 241	LSTEQVVAIASNNGGKQALEAVKAQLPVLRRAPCG
SEQ ID NO: 242	LSTEQVVAIASNNGGKQALEAVKAQLPVLRRAPYG
SEQ ID NO: 243	LSTEQVVAIASNNGGKQALEAVKARLLDLRGAPYA

SEQ ID NO	Sequence
SEQ ID NO: 244	LSTEQVVAIASNNGGKQALEAVKTQLLALRTAPYE
SEQ ID NO: 245	LSTEQVVAIASNPGGKQALEAVRALFPDLRAAPYA
SEQ ID NO: 246	LSTEQVVAIASSHGGKQALEAVRALFPDLRAAPYA
SEQ ID NO: 247	LSTEQVVAIASSHGGKQALEAVRALLPVLRATPYD
SEQ ID NO: 248	LSTEQVVAVASHNGGKQALEAVRAQLLDLRAAPYE
SEQ ID NO: 249	LSTEQVVAVASNKGGKQALAAVEAQLLRLRAAPYE
SEQ ID NO: 250	LSTEQVVAVASNKGGKQALEEVEAQLLRLRAAPYE
SEQ ID NO: 251	LSTEQVVAVASNKGGKQVLEAVGAQLLALRAVPYE
SEQ ID NO: 252	LSTEQVVAVASNNGGKQALKAVKAQLLALRAAPYE
SEQ ID NO: 253	LSTEQVVVIANSIGGKQALEAVKVQLPVLRAAPYE
SEQ ID NO: 254	LSTGQVVAIASNGGGRQALEAVREQLLALRAVPYE
SEQ ID NO: 255	LSVAQVVTIASHNGGKQALEAVRAQLLALRAAPYG
SEQ ID NO: 256	LTIAQVVAVASHNGGKQALEAIGAQLLALRAAPYA
SEQ ID NO: 257	LTIAQVVAVASHNGGKQALEVIGAQLLALRAAPYA
SEQ ID NO: 258	LTPQQVVAIAANTGGKQALGAITTQLPILRAAPYE
SEQ ID NO: 259	LTPQQVVAIASNTGGKQALEAVTVQLRVLRGARYG
SEQ ID NO: 260	LTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYR
SEQ ID NO: 261	LTPQQVVAIASNTGGKRALEAVRVQLPVLRAAPYE
SEQ ID NO: 262	LTTAQVVAIASNDGGKQALEAVGAQLLVLRAVPYE
SEQ ID NO: 263	LTTAQVVAIASNDGGKQTLEVAGAQLLALRAVPYE
SEQ ID NO: 336	LSTAQVVAVASGSGGKPALEAVRAQLLALRAAPYG
SEQ ID NO: 337	LSTAQVVAVASGSGGKPALEAVRAQLLALRAAPYG
SEQ ID NO: 338	LNTAQIVAIASHDGGKPALEAVWAKLPVLRGAPYA
SEQ ID NO: 339	LNTAQVVAIASHDGGKPALEAVRAKLPVLRGVPYA
SEQ ID NO: 340	LNTAQVVAIASHDGGKPALEAVWAKLPVLRGVPYA
SEQ ID NO: 341	LNTAQVVAIASHDGGKPALEAVWAKLPVLRGVPYE
SEQ ID NO: 342	LSTAQVVAIASHDGGKPALEAVWAKLPVLRGAPYA
SEQ ID NO: 343	LSTAQVVAVASHDGGKPALEAVRKQLPVLRGVPHQ
SEQ ID NO: 344	LSTAQVVAVASHDGGKPALEAVRKQLPVLRGVPHQ
SEQ ID NO: 345	LNTAQVVAIASHDGGKPALEAVWAKLPVLRGVPYA
SEQ ID NO: 346	LSTEQVVAIASHNGGKLALEAVKAHLLDLRGAPYA
SEQ ID NO: 347	LSTEQVVAIASHNGGKPALEAVKAHLLALRAAPYA
SEQ ID NO: 348	LNTAQVVAIASHYGGKPALEAVWAKLPVLRGVPYA
SEQ ID NO: 349	LNTEQVVAIASNNGGKPALEAVKAQLLELRAAPYE
SEQ ID NO: 350	LSPEQVVAIASNNGGKPALEAVKALLLALRAAPYE
SEQ ID NO: 351	LSPEQVVAIASNNGGKPALEAVKAQLLELRAAPYE
SEQ ID NO: 352	LSTEQVVAIASNNGGKPALEAVKALLLALRAAPYE
SEQ ID NO: 353	LSTEQVVAIASNNGGKPALEAVKALLLELRAAPYE
SEQ ID NO: 354	LSPEQVVAIASNNGGKPALEAVKALLLALRAAPYE
SEQ ID NO: 355	LSPEQVVAIASNNGGKPALEAVKAQLLELRAAPYE

SEQ ID NO	Sequence
SEQ ID NO: 356	LSTEQVVAIASNNGGKPALEAVKALLLELRAAPYE

[0042] In some embodiments, an RNBD of the present disclosure can comprise between 1 to 50 *Ralstonia solanacearum*-derived repeat units. In some embodiments, an RNBD of the present disclosure can comprise between 9 and 36 *Ralstonia solanacearum*-derived repeat units. Preferably, in some embodiments, an RNBD of the present disclosure can comprise between 12 and 30 *Ralstonia solanacearum*-derived repeat units. An RNBD described herein can comprise between 5 to 10 *Ralstonia solanacearum*-derived repeat units, between 10 to 15 *Ralstonia solanacearum*-derived repeat units, between 15 to 20 *Ralstonia solanacearum*-derived repeat units, between 20 to 25 *Ralstonia solanacearum*-derived repeat units, between 25 to 30 *Ralstonia solanacearum*-derived repeat units, or between 30 to 35 *Ralstonia solanacearum*-derived repeat units, between 35 to 40 *Ralstonia solanacearum*-derived repeat units. An RNBD described herein can comprise 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40 *Ralstonia solanacearum*-derived repeat units.

[0043] A *Ralstonia solanacearum*-derived repeat unit can be derived from a wild-type repeat unit, such as any one of SEQ ID NO: 168 – SEQ ID NO: 263 or SEQ ID NO: 336 – SEQ ID NO: 356. A *Ralstonia solanacearum*- repeat unit can have at least 80% sequence identity with any one of SEQ ID NO: 168 – SEQ ID NO: 263 or SEQ ID NO: 336 – SEQ ID NO: 356. A *Ralstonia solanacearum*-derived repeat unit can also comprise a modified *Ralstonia solanacearum*-derived repeat unit enhanced for specific recognition of a nucleotide or base pair. An RNBD described herein can comprise one or more wild-type *Ralstonia solanacearum*-derived repeat units, one or more modified *Ralstonia solanacearum*-derived repeat units, or a combination thereof. In some embodiments, a modified *Ralstonia solanacearum*-derived repeat unit can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, or 29 mutations that can enhance recognition of a specific nucleotide or base pair. In some embodiments, a modified *Ralstonia solanacearum*-derived repeat unit can comprise more than 1 modification, for example 1 to 5 modifications, 5 to 10 modifications, 10 to 15 modifications, 15 to 20 modifications, 20 to 25 modification, or 25-29 modifications. In some embodiments, An RNBD can comprise more than one modified *Ralstonia solanacearum*-derived repeat units, wherein each of the modified *Ralstonia solanacearum*-derived repeat units can have a different number of modifications.

[0044] The *Ralstonia solanacearum*-derived repeat units comprise amino acid residues at positions 12 and 13, what is referred to herein as, a repeat variable diresidue (RVD). The RVD can modulate binding affinity of the repeat unit for a particular nucleic acid base (e.g., adenosine, guanine,

cytosine, thymidine, or uracil (in RNA sequences)). In some embodiments, a single amino acid residue can modulate binding to the target nucleic acid base. In some embodiments, two amino acid residues (RVD) can modulate binding to the target nucleic acid base. In some embodiments, any repeat unit disclosed herein can have an RVD selected from HD, HG, HK, HN, ND, NG, NH, NK, NN, NP, NT, QN, RN, RS, SH, SI, or SN. In some embodiments, an RVD of HD can bind to cytosine. In some embodiments, an RVD of NG can bind to thymidine. In some embodiments, an RVD of NK can bind to guanine. In some embodiments, an RVD of SI can bind to adenosine. In some embodiments, an RVD of RS can bind to adenosine. In some embodiments, an RVD of HN can bind to guanine. In some embodiments, an RVD of NT can bind to adenosine.

[0045] In some embodiments, a repeat unit having at least 80% sequence identity with SEQ ID NO: 209 can be included in a DNA binding domain of the present disclosure to bind to cytosine. In some embodiments, a repeat unit having at least 80% sequence identity with SEQ ID NO: 197 can be included in a DNA binding domain of the present disclosure to bind to thymidine. In some embodiments, a repeat unit having at least 80% sequence identity with SEQ ID NO: 233 can be included in a DNA binding domain of the present disclosure to bind to guanine. In some embodiments, a repeat unit having at least 80% sequence identity with SEQ ID NO: 253 can be included in a DNA binding domain of the present disclosure to bind to adenosine. In some embodiments, a repeat unit having at least 80% sequence identity with SEQ ID NO: 203 can be included in a DNA binding domain of the present disclosure to bind to adenosine. In some embodiments, a repeat unit having at least 80% sequence identity with SEQ ID NO: 218 can be included in a DNA binding domain of the present disclosure to bind to guanine. In some embodiments, the repeat unit of SEQ ID NO: 209 can be included in a DNA binding domain of the present disclosure to bind to cytosine. In some embodiments, the repeat unit of SEQ ID NO: 197 can be included in a DNA binding domain of the present disclosure to bind to thymidine. In some embodiments, the repeat unit of SEQ ID NO: 233 can be included in a DNA binding domain of the present disclosure to bind to guanine. In some embodiments, the repeat unit of SEQ ID NO: 253 can be included in a DNA binding domain of the present disclosure to bind to adenosine. In some embodiments, the repeat unit of SEQ ID NO: 203 can be included in a DNA binding domain of the present disclosure to bind to adenosine. In some embodiments, the repeat unit of SEQ ID NO: 218 can be included in a DNA binding domain of the present disclosure to bind to guanine.

[0046] In some embodiments, the present disclosure provides repeat units as set forth in SEQ ID NO: 267 – SEQ ID NO: SEQ ID NO: 279. Unspecified amino acid residues in SEQ ID NO: 267 – SEQ ID NO: SEQ ID NO: 279 can be any amino acid residues. In particular embodiments, unspecified

amino acid residues in SEQ ID NO: 267 – SEQ ID NO: SEQ ID NO: 279 can be those set forth in the Variable Definition column of **TABLE 3**.

[0047] **TABLE 3** shows consensus sequences of *Ralstonia*-derived repeat units.

TABLE 3: Consensus Sequences of *Ralstonia*-derived Repeat Units

RVD	Consensus Sequence	Variable Definition
HN	$LX_1X_2X_3QVVX_4X_5ASHNGX_6KQALEX_7X_8X_9X_{10}X_{11}LX_{12}X_{13}LX_{14}X_{15}X_{16}PYX_{17}$ (SEQ ID NO: 267)	X_1: D|N|S|T, X_2: I|T|V, X_3: A|E, X_4: A|T, X_5: I|V, X_6: G|S, X_7: A|V, X_8: I|V, X_9: G|K|R, X_{10}: A|T, X_{11}: D|H|Q, X_{12}: L|P, X_{13}: A|D|E|V, X_{14}: L|R, X_{15}: A|G|R, X_{16}: A|V, X_{17}: A|E|G|V
NN	$LX_1X_2X_3QVVAX_4AX_5NNGGKQALX_6AVX_7X_8X_9LX_{10}X_{11}LRX_{12}AX_{13}X_{14}X_{15}$ (SEQ ID NO: 268)	X_1: N|S, X_2: P|T, X_3: A|E, X_4: I|V, X_5: A|S, X_6: E|K, X_7: K|R, X_8: A|T, X_9: H|L|Q|R, X_{10}: L|P, X_{11}: A|D|E|V, X_{12}: A|G|R|T|V, X_{13}: P|R, X_{14}: C|Y, X_{15}: A|E|G
NP	$LX_1TX_2QX_3VX_4IASNPGGKQALEAX_5RAX_6FX_7X_8X_9RAAPYA$ (SEQ ID NO: 269)	X_1: N|S, X_2: A|E, X_3: L|V, X_4: A|S, X_5: I|V, X_6: L|P, X_7: P|R, X_8: D|E, X_9: L|V
SH	$LX_1TX_2QVVAIASSHGGKQALEAVRALX_3X_4X_5LRAX_6PYX_7$ (SEQ ID NO: 270)	X_1: N|S, X_2: A|E, X_3: F|L, X_4: P|R, X_5: D|E|V, X_6: A|T, X_7: A|D|G
NK	$LX_1TEQVVAX_2ASNKGGKQX_3LX_4X_5VX_6AX_7LLX_8LX_9X_{10}X_{11}PYX_{12}$ (SEQ ID NO: 271)	X_1: N|S, X_{10}: A|G, X_{11}: A|V, X_{12}: A|E|V, X_2: I|V, X_3: A|V, X_4: A|E, X_5: A|E, X_6: E|G|K, X_7: D|H|Q, X_8: A|D|E|R, X_9: L|R
HD	$LSX_1X_2QVX_3AIAX_4HDGGX_5QX_6LEAX_7X_8X_9QLVX_{10}LX_{11}AAPYA$ (SEQ ID NO: 272)	X_1: A|T, X_2: A|E, X_3: A|V, X_4: G|S, X_5: K|N, X_6: A|P, X_7: A|V, X_8: G|V, X_9: A|G|T|V, X_{10}: A|E|V, X_{11}: L|R
RS	$LSX_1AQVVAX_2AX_3RSGGKQALEAVRAQLLX_4LRAAPYG$ (SEQ ID NO: 273)	X_1: I|T, X_2: I|V, X_3: S|T, X_4: A|D
NH	$LSX_1EQVVAIASNHGGKQALEAVRALFRX_2LRAAPYX_3$ (SEQ ID NO: 274)	X_1: P|T, X_2: E|G, X_3: A|G
SI	$LSTX_1QVX_2X_3IAX_4SIGGX_5QALEAX_6KVQLPVLRAAPYX_7$ (SEQ ID NO: 275)	X_1: A|E, X_2: A|V, X_3: T|V, X_4: N|S, X_5: K|R, X_6: L|V, X_7: E|G
ND	$LX_1TAQVVAIASNDGGKQX_2LEX_3X_4X_5AQLLX_6LRAX_7PYE$ (SEQ ID NO: 276)	X_1: S|T, X_2: A|T, X_3: A|E|V, X_4: A|V, X_5: E|G, X_6: A|V, X_7: A|V
SN	$LSTAQVVX_1X_2ASSNGGKQALEAVWALLPVLRATPYD$ (SEQ ID NO: 277)	X_1: A|T, X_2: I|V
NG	$LSTX_1QVVAIAX_2NGGGX_3QALEX_4X_5X_6X_7QLX_8X_9LRX_{10}X_{11}PX_{12}X_{13}$ (SEQ ID NO: 278)	X_1: A|E|G, X_2: G|S, X_3: K|R, X_4: A|G, X_5: I|V, X_6: G|R, X_7: E|K, X_8: L|Q|R, X_9: A|E|K, X_{10}: A|T, X_{11}: A|V, X_{12}: H|Y, X_{13}: E|G

RVD	Consensus Sequence	Variable Definition
NT	LTPQQVVAIAX$_1$NTGGKX$_2$ALX$_3$AX$_4$X$_5$X$_6$QL X$_7$X$_8$LRX$_9$AX$_{10}$YX$_{11}$ (SEQ ID NO: 279)	X$_1$: A|S, X$_{10}$: P|R, X$_{11}$: E|G|R, X$_2$: Q|R, X$_3$: E|G, X$_4$: I|V, X$_5$: C|R|T, X$_6$: T|V, X$_7$: P|R, X$_8$: I|V, X$_9$: A|G

[0048] In some aspects, the at least one repeat unit comprises any one of SEQ ID NO: 267 – SEQ ID NO: 279. In some embodiments, the present disclosure provides a modular nucleic acid binding domain (e.g., RNBD or MAP-NBD), wherein the modular nucleic acid binding domain comprises a repeat unit with a sequence of $A_{1\text{-}11}X_1X_2B_{14\text{-}35}$ (SEQ ID NO: 448), wherein $A_{1\text{-}11}$ comprises 11 amino acid residues and wherein each amino acid residue of $A_{1\text{-}11}$ can be any amino acid. In some embodiments, $A_{1\text{-}11}$ can be any amino acids in position 1 through position 11 of any one of SEQ ID NO: 168 – SEQ ID NO: 263 or SEQ ID NO: 336 – SEQ ID NO: 356. X_1X_2 comprises any repeat variable diresidue (RVD) disclosed herein and comprises at least one amino acid at position 12 or position 13. As described herein, this RVD contacts and binds to a target nucleic acid base of a target site. Said RVD can be the RVD of any repeat unit disclosed herein, such as position 12 and position 13 of any one of SEQ ID NO: 168 – SEQ ID NO: 263 or SEQ ID NO: 336 – SEQ ID NO: 356. $B_{14\text{-}35}$ can comprise 22 amino acid residues and each amino acid residue of $B_{14\text{-}35}$ can be any amino acid. In some embodiments, $B_{14\text{-}35}$ can be any amino acid in position 14 through position 35 of any one of SEQ ID NO: 168 – SEQ ID NO: 263 or SEQ ID NO: 336 – SEQ ID NO: 356. In particular embodiments, a modular nucleic acid binding domain (e.g., RNBD or MAP-NBD) having the above sequence of $A_{1\text{-}11}X_1X_2B_{14\text{-}35}$ (SEQ ID NO: 448) can have a first repeat unit with at least one residue in $A_{1\text{-}11}$, $B_{14\text{-}35}$, or a combination thereof that differs from a corresponding residue in a second repeat unit in the modular nucleic acid binding domain (e.g., RNBD or MAP-NBD). In other words, at least two repeat units in a modular nucleic acid binding domain (e.g., RNBD or MAP-NBD) described herein can have different amino acid residues with respect to each other, at the same position outside the RVD region. Thus, in some embodiments, a modular nucleic acid binding domain (e.g., RNBD or MAP-NBD) described herein can have variant backbones with respect to each repeat unit in the plurality of repeat units that make up the modular nucleic acid binding domain. In some embodiments, an RNBD of the present disclosure can have a sequence of GGKQALEAVRAQLLDLRAAPYG (SEQ ID NO: 280) at $B_{14\text{-}35}$.

[0049] In some embodiments, a modular nucleic acid binding sequence (e.g., RNBD) can comprise one or more of the following characteristics: the modular nucleic acid binding sequence (e.g.,

RNBD) can bind a nucleic acid sequence, wherein the target site comprises a 5' guanine, the modular nucleic acid binding domain (e.g., RNBD) can comprise 7 repeat units to 25 repeat units, a first modular nucleic acid binding sequence (e.g., RNBD) can bind a target nucleic acid sequence and be separated from a second modular nucleic acid binding domain (e.g., RNBD) from 2 to 50 base pairs, or any combination thereof.

[0050] In some embodiments, an RNBD of the present disclosure can have the full length naturally occurring N-terminus of a naturally occurring *Ralstonia solanacearum*-derived protein. In some embodiments, any truncation of the full length naturally occurring N-terminus of a naturally occurring *Ralstonia solanacearum*-derived protein can be used at the N-terminus of an RNBD of the present disclosure. For example, in some embodiments, amino acid residues at positions 1 (H) to position 137 (F) of the naturally occurring *Ralstonia solanacearum*-derived protein N-terminus can be used. In particular embodiments, said truncated N-terminus from position 1 (H) to position 137 (F) can have a sequence as follows:
FGKLVALGYSREQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPE LTRAHIVDIARQRSGDLALQALLPVATALTAAPLRLSASQIATVAQYGERPAIQALYRLRRKLTRAPL H (SEQ ID NO: 264). In some embodiments, the naturally occurring N-terminus of *Ralstonia solanacearum* can be truncated to any length and used at the N-terminus of the engineered DNA binding domain. For example, the naturally occurring N-terminus of *Ralstonia solanacearum* can be truncated to amino acid residues at position 1 (H) to position 120 (K) as follows:
KQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELAAALPELTRAHIVDIARQ RSGDLALQALLPVATALTAAPLRLSASQIATVAQYGERPAIQALYRLRRKLTRAPLH (SEQ ID NO: 303) and used at the N-terminus of the RNBD. The naturally occurring N-terminus of *Ralstonia solanacearum* can be truncated amino acid residues at positions 1 to 115 and used at the N-terminus of the engineered DNA binding domain as set forth in SEQ ID NO: 320. The naturally occurring N-terminus of *Ralstonia solanacearum* can be truncated to amino acid residues at positions 1 to 50, 1 to 70, 1 to 100, 1 to 120, 1 to 130, 10 to 40, 60 to 100, or 100 to 120 and used at the N-terminus of the engineered DNA binding domain. Truncation of the N-termini can be particularly advantageous for obtaining DNA binding domains, which are smaller in size including number of amino acids and overall molecular weight. A reduced number of amino acids can allow for more efficient packaging into a viral vector and a smaller molecular weight can result in more efficient loading of the DNA binding domains in non-viral vectors for delivery.

[0051] In some embodiments, the N-terminus, referred to as the amino terminus or the "NH2" domain, can recognize a guanine. In some embodiments, the N-terminus can be engineered to bind a cytosine, adenosine, thymidine, guanine, or uracil.

[0052] In some embodiments, an RNBD of the present disclosure can have a DNA binding domain, in which the final full length repeat unit of 33-35 amino acid residues is followed by a half-repeat also derived from *Ralstonia solanacearum*. The half repeat can have 15 to 23 amino acid residues, for example, the half repeat can have 19 amino acid residues. In particular embodiments, the half-repeat can have a sequence as follows: LSTAQVVAIACISGQQALE (SEQ ID NO: 265).

[0053] In some embodiments, an RNBD of the present disclosure can have the full length naturally occurring C-terminus of a naturally occurring *Ralstonia solanacearum*-derived protein. In some embodiments, any truncation of the full length naturally occurring C-terminus of a naturally occurring *Ralstonia solanacearum*-derived protein can be used at the C-terminus of an RNBD of the present disclosure. For example, in some embodiments, the RNBD can comprise amino acid residues at position 1 (A) to position 63 (S) as follows: AIEAHMPTLRQASHSLSPERVAAIACIGGRSAVEAVRQGLPVKAIRRIRREKAPVAGPPPAS (SEQ ID NO: 266) of the naturally occurring *Ralstonia solanacearum*-derived protein C-terminus. In some embodiments, the naturally occurring C-terminus of *Ralstonia solanacearum* can be truncated to any length and used at the C-terminus of the RNBD. For example, the naturally occurring C-terminus of *Ralstonia solanacearum* can be truncated to amino acid residues at positions 1 to 63 and used at the C-terminus of the RNBD. The naturally occurring C-terminus of *Ralstonia solanacearum* can be truncated amino acid residues at positions 1 to 50 and used at the C-terminus of the RNBD. The naturally occurring C-terminus of *Ralstonia solanacearum* can be truncated to amino acid residues at positions 1 to 63, 1 to 50, 1 to 70, 1 to 100, 1 to 120, 1 to 130, 10 to 40, 60 to 100, or 100 to 120 and used at the C-terminus of the RNBD.

[0054] **TABLE 4** shows N-termini, C-termini, and half-repeats derived from *Ralstonia*.

TABLE 4: *Ralstonia*-Derived N-terminus, C-terminus, and Half-Repeat

SEQ ID NO	Description	Sequence
SEQ ID NO: 320	Truncated N-terminus; positions 1 (H) to 115 (S) of the naturally occurring *Ralstonia solanacearum*-derived protein N-terminus	SEIAKYHTTLTGQGFTHADICRISRRRQSLRV VARNYPELAAALPELTRAHIVDIARQRSGDL ALQALLPVATALTAAPLRLSASQIATVAQYG ERPAIQALYRLRRKLTRAPLH
SEQ ID NO: 264	Truncated N-terminus; positions 1 (H) to 137 (F) of the naturally occurring *Ralstonia solanacearum*-derived protein N-terminus	FGKLVALGYSREQIRKLKQESLSEIAKYHTT LTGQGFTHADICRISRRRQSLRVVARNYPEL AAALPELTRAHIVDIARQRSGDLALQALLPV ATALTAAPLRLSASQIATVAQYGERPAIQAL

SEQ ID NO	Description	Sequence
		YRLRRKLTRAPLH
SEQ ID NO: 303	Truncated N-terminus; positions 1 (H) to 120 (K) of the naturally occurring *Ralstonia solanacearum*-derived protein N-terminus	KQESLSEIAKYHTTLTGQGFTHADICRISRRR QSLRVVARNYPELAAALPELTRAHIVDIARQ RSGDLALQALLPVATALTAAPLRLSASQIAT VAQYGERPAIQALYRLRRKLTRAPLH
SEQ ID NO: 265	Half-repeat	LSTAQVVAIACISGQQALE
SEQ ID NO: 266	Truncated C-terminus; positions 1 (A) to 63 (S) of the naturally occurring *Ralstonia solanacearum*-derived protein C-terminus	AIEAHMPTLRQASHSLSPERVAAIACIGGRS AVEAVRQGLPVKAIRRIRREKAPVAGPPPAS

[0055] In some embodiments, an RNBD can be engineered to target and bind to a site in the PDCD1 gene. For example, an RNBD with the sequence FGKLVALGYSREQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELA AALPELTRAHIVDIARQRSGDLALQALLPVATALTAAPLRLSASQIATVAQYGERPAIQALY RLRRKLTRAPLHLTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYRLSTEQVVAIASHDG GKQALEAVGAQLVALRAAPYALSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYALST AQVVAIASNGGGKQALEGIGEQLLKLRTAPYGLSTEQVVAIASNKGGKQALEAVKAHLLDL LGAPYVLSTEQVVAIASNKGGKQALEAVKAHLLDLLGAPYVLSTEQVVAIASNKGGKQAL EAVKAHLLDLLGAPYVLSTEQVVVIANSIGGKQALEAVKVQLPVLRAAPYELSTEQVVAIA SHDGGKQALEAVGAQLVALRAAPYALSTEQVVVIANSIGGKQALEAVKVQLPVLRAAPYE LSTEQVVAIASNKGGKQALEAVKAHLLDLLGAPYVLSTAQVVAIASNGGGKQALEGIGEQL LKLRTAPYGLSTAQVVAIASNGGGKQALEGIGEQLLKLRTAPYGLSTAQVVAIASNGGGKQ ALEGIGEQLLKLRTAPYGLSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYALSTEQVVA IASHDGGKQALEAVGAQLVALRAAPYALSTEQVVAIASHDGGKQALEAVGAQLVALRAAP YALSTAQVVAIASNGGGKQALEGIGEQLLKLRTAPYGLSTAQVVAIASNGGGKQALEGIGE QLLKLRTAPYGLSTAQVVAIACISGQQALEAIEAHMPTLRQASHSLSPERVAAIACIGGRSAV EAVRQGLPVKAIRRIRREKAPVAGPPPAS (SEQ ID NO: 311) can bind to the GACCTGGGACAGTTTCCCTT (SEQ ID NO: 312) nucleic acid sequence in the PDCD1 gene. As another example, an RNBD with the sequence FGKLVALGYSREQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELA AALPELTRAHIVDIARQRSGDLALQALLPVATALTAAPLRLSASQIATVAQYGERPAIQALY RLRRKLTRAPLHLTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYRLSTAQVVAIASNGG GKQALEGIGEQLLKLRTAPYGLSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYALSTA QVVAIASNGGGKQALEGIGEQLLKLRTAPYGLSTEQVVAIASHNGGKQALEAVKADLLELR

GAPYALSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYALSTEQVVVIANSIGGKQALEA
VKVQLPVLRAAPYELSTAQVVAIASNGGGKQALEGIGEQLLKLRTAPYGLSTEQVVAIASH
NGGKQALEAVKADLLELRGAPYALSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYALS
TEQVVAIASHDGGKQALEAVGAQLVALRAAPYALSTAQVVAIASNGGGKQALEGIGEQLL
KLRTAPYGLSTEQVVAIASHNGGKQALEAVKADLLELRGAPYALSTEQVVAIASHNGGKQ
ALEAVKADLLELRGAPYALSTEQVVVIANSIGGKQALEAVKVQLPVLRAAPYELSTEQVVA
IASHNGGKQALEAVKADLLELRGAPYALSTEQVVAIASHDGGKQALEAVGAQLVALRAAP
YALSTAQVVAIACISGQQALEAIEAHMPTLRQASHSLSPERVAAIACIGGRSAVEAVRQGLP
VKAIRRIRREKAPVAGPPPAS (SEQ ID NO: 313) can bind to the GATCTGCATGCCTGGAGC (SEQ ID NO: 314) nucleic acid sequence in the PDCD1 gene. As yet another example, an RNBD with the sequence
FGKLVALGYSREQIRKLKQESLSEIAKYHTTLTGQGFTHADICRISRRRQSLRVVARNYPELA
AALPELTRAHIVDIARQRSGDLALQALLPVATALTAAPLRLSASQIATVAQYGERPAIQALY
RLRRKLTRAPLHLTPQQVVAIASNTGGKRALEAVCVQLPVLRAAPYRLSTAQVVAIASNGG
GKQALEGIGEQLLKLRTAPYGLSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYALSTA
QVVAIASNGGGKQALEGIGEQLLKLRTAPYGLSTEQVVAIASHNGGKQALEAVKADLLELR
GAPYALSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYALSTAQVVAIATRSGGKQALE
AVRAQLLDLRAAPYGLSTAQVVAIASNGGGKQALEGIGEQLLKLRTAPYGLSTEQVVAIAS
HNGGKQALEAVKADLLELRGAPYALSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYA
LSTEQVVAIASHDGGKQALEAVGAQLVALRAAPYALSTAQVVAIASNGGGKQALEGIGEQ
LLKLRTAPYGLSTEQVVAIASHNGGKQALEAVKADLLELRGAPYALSTEQVVAIASHNGGK
QALEAVKADLLELRGAPYALSTAQVVAIATRSGGKQALEAVRAQLLDLRAAPYGLSTEQV
VAIASHNGGKQALEAVKADLLELRGAPYALSTEQVVAIASHDGGKQALEAVGAQLVALRA
APYALSTAQVVAIACISGQQALEAIEAHMPTLRQASHSLSPERVAAIACIGGRSAVEAVRQG
LPVKAIRRIRREKAPVAGPPPAS (SEQ ID NO: 315) can bind to the GATCTGCATGCCTGGAGC (SEQ ID NO: 314) nucleic acid sequence in the PDCD1 gene. Any one of SEQ ID NO: 311, SEQ ID NO; 313, or SEQ ID NO: 315 can be fused to any repression domain described herein (e.g., KRAB) to yield a gene repressor capable of repressing expression of the target gene.

Xanthomonas Derived Transcription Activator Like Effector (TALE)

[0056] The present disclosure provides a modular nucleic acid binding domain derived from *Xanthomonas spp.*, also referred to herein as a transcription activator-like effector (TALE) protein,

can comprise a plurality of repeat units. A repeat unit of the plurality of repeat units recognizes a single target nucleotide, base pair, or both. A repeat unit from *Xanthomonas spp.* can comprise 33-35 amino acid residues. In some embodiments, a repeat unit can be from *Xanthomonas spp.* and have a sequence of

MDPIRSRTPSPARELLPGPQPDGVQPTADRGVSPPAGGPLDGLPARRTMSRTRLPSPPAPSPA FSAGSFSDLLRQFDPSLFNTSLFDSLPPFGAHHTEAATGEWDEVQSGLRAADAPPPTMRVAV TAARPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVG HGFTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVA GELRGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLNLTPEQVVAIASHDGGKQ ALETVQRLLPVLCQAHGLTPQQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPQQVVAIA SNSGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTP EQVVAIASNIGGKQALETVQALLPVLCQAHGLTPEQVVAIASNIGGKQALETVQALLPVLCQ AHGLTPEQVVAIASNIGGKQALETVQALLPVLCQAHGLTPEQVVAIASHDGGKQALETVQR LLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPQQVVAIASNGGGK QALETVQRLLPVLCQAHGLTPEQVVAIASNSGGKQALETVQALLPVLCQAHGLTPEQVVAI ASNSGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGL TPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPV LCQAHGLTPQQVVAIASNGGGRPALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALET VQRLLPVLCQAHGLTPQQVVAIASNGGGRPALE*SIVAQLSRPDPALAALTNDHLVALACLGGR PALDAVKKGLPHAPALIKRTNRRIPERTSHRVADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGM SRHGLLQLFRRVGVTELEARSGTLPPASQRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADSL ERDLDAPSPMHEGDQTRASSRKRSRSDRAVTGPSAQQSFEVRVPEQRDALHLPLSWRVKRPRTSI GGGLPDPGTPTAADLAASSTVMREQDEDPFAGAADDFPAFNEEELAWLMELLPQ* (SEQ ID NO: 299).

[0057] In some embodiments, a TALE of the present disclosure can comprise between 1 to 50 *Xanthomonas spp.*-derived repeat units. In some embodiments, a TALE of the present disclosure can comprise between 9 and 36 *Xanthomonas spp.*-derived repeat units. Preferably, in some embodiments, a TALE of the present disclosure can comprise between 12 and 30 *Xanthomonas spp.*-derived repeat units. A TALE described herein can comprise between 5 to 10 *Xanthomonas spp.*-derived repeat units, between 10 to 15 *Xanthomonas spp.*-derived repeat units, between 15 to 20 *Xanthomonas spp.*-derived repeat units, between 20 to 25 *Xanthomonas spp.*-derived repeat units, between 25 to 30 *Xanthomonas spp.*-derived repeat units, or between 30 to 35 *Xanthomonas spp.*-derived repeat units, between 35 to 40 *Xanthomonas spp.*-derived repeat units. A TALE described

herein can comprise at least 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 , 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40, or more *Xanthomonas spp.*-derived repeat units, such as, repeat units derived from *Xanthomonas spp.* protein having the amino acid sequence set forth in SEQ ID NO:299.

[0058] A *Xanthomonas spp.*-derived repeat units can be derived from a wild-type repeat unit, such as any one of SEQ ID NO: 323 – SEQ ID NO: 326. For example, a *Xanthomonas spp.*-derived repeat units can have a sequence of LTPDQVVAIASNHGGKQALETVQRLLPVLCQDHG (SEQ ID NO: 323) comprising an RVD of NH, which recognizes guanine. A *Xanthomonas spp.*-derived repeat units can have a sequence of LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG (SEQ ID NO: 324) comprising an RVD of NG, which recognizes thymidine. A *Xanthomonas spp.*-derived repeat units can have a sequence of LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG (SEQ ID NO: 325) comprising an RVD of NI, which recognizes adenosine. A *Xanthomonas spp.*-derived repeat units can have a sequence of LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG (SEQ ID NO: 326) comprising an RVD of HD, which recognizes cytosine.

[0059] A *Xanthomonas spp.*-derived repeat unit can also comprise a modified *Xanthomonas spp.*-derived repeat units enhanced for specific recognition of a nucleotide or base pair. A TALE described herein can comprise one or more wild-type *Xanthomonas spp.*-derived repeat units, one or more modified *Xanthomonas spp.*-derived repeat units, or a combination thereof. In some embodiments, a modified *Xanthomonas spp.*-derived repeat units can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, or 29 mutations that can enhance recognition of a specific nucleotide or base pair. In some embodiments, a modified *Xanthomonas spp.*-derived repeat unit can comprise more than 1 modification, for example 1 to 5 modifications, 5 to 10 modifications, 10 to 15 modifications, 15 to 20 modifications, 20 to 25 modification, or 25-29 modifications. In some embodiments, A TALE can comprise more than one modified *Xanthomonas spp.*-derived repeat units, wherein each of the modified *Xanthomonas spp.*-derived repeat units can have a different number of modifications.

[0060] In some embodiments, a TALE of the present disclosure can have the full length naturally occurring N-terminus of a naturally occurring *Xanthomonas spp.*-derived protein, such as the N-terminus of SEQID NO: 299. The N-terminus sequence in SEQ ID NO:299 is indicated by underlining.

[0061] In some embodiments, a TALE of the present disclosure can comprise the amino acid residues at position 1 (N) through position 137 (M) of the naturally occurring *Xanthomonas spp.*-derived protein as follows:

MVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQ DMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTA VEAVHAWRNALTGAPLN (SEQ ID NO: 300).

[0062] The amino acid sequence set forth in SEQ ID NO:300 includes a M added to the N-terminus which is not present in the wild type N-terminus region of a TALE protein. The N-terminus fragment sequence set out in SEQ ID NO:300 is generated by deleting amino acids N+288 through N+137 of the N-terminus region of a TALE protein, adding a M, such that amino acids N+136 through N+1 of the N-terminus region of the TALE protein are present.

[0063] In some embodiments, the N-terminus can be truncated such that the fragment of the N-terminus includes amino acids from position 1 (N) through position 120 (K) of the naturally occurring *Xanthomonas spp.*-derived protein as follows:
KPKVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGV GKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPL N (SEQ ID NO: 301).

[0064] In some embodiments, the N-terminus can be truncated such that the fragment of the N-terminus includes amino acids from position 1 (N) through position 115 (S) of the naturally occurring *Xanthomonas spp.*-derived protein as follows:
STVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWS GARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLN (SEQ ID NO: 321).

[0065] In some embodiments, the N-terminus can be truncated such that the fragment of the N-terminus includes amino acids from position 1 (N) through position 110 (H) of the naturally occurring *Xanthomonas spp.*-derived protein as follows:
HHEALVGHGFTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARAL EALLTVAGELRGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLN (SEQ ID NO: 447).

[0066] In some embodiments, any truncation of the naturally occurring *Xanthomonas spp.*-derived protein can be used at the N-terminus of a TALE disclosed herein. The naturally occurring N-terminus of *Xanthomonas spp.* can be truncated to amino acid residues at positions 1 to 50, 1 to 70, 1 to 100, 1 to 120, 1 to 130, 10 to 40, 60 to 100, or 100 to 120 and used at the N-terminus of the TALE.

[0067] FIGS. 1A-1C show schematics of the domain structure of a TALE protein (not drawn to scale). `N` and `C` indicate the amino and carboxy termini, respectively. The TALE repeat domain

comprising TALE repeat units, N-Cap and C-Cap regions are labeled and the residue numbering scheme for the N-Cap and C-Cap regions and the N- terminus and C- terminus fragments are indicated. FIG. 1A includes the full-length N-cap region that extends from amino acid position N+1 to N+288 and full-length C-cap region that extends from amino acid position C+1 through C+278. FIG. 1B provides a schematic of a DNA binding protein comprising TALE repeat units and a truncated N-terminus that extends from amino acid position N+1 to N+136 (the notation N+137 indicates that a methionine added to the N-terminus increases the length to 137) and a truncated C-terminus that extends from amino acid position C+1 through C+63. FIG. 1C provides a schematic of a DNA binding protein comprising TALE repeat units and a truncated N-terminus that extends from amino acid position N+1 to N+115 and a truncated C-terminus that extends from amino acid position C+1 through C+63. In certain cases, the last repeat domain may be a half-repeat or a partial repeat as disclosed herein.

[0068] In some embodiments, a TALE of the present disclosure can have a DNA binding domain, in which the final full length repeat unit of 33-35 amino acid residues is followed by a half-repeat also derived from *Xanthomonas spp*. The half repeat can have 15 to 23 amino acid residues, for example, the half repeat can have 19 amino acid residues. In particular embodiments, the half-repeat can have a sequence as set forth in LTPQQVVAIASNGGGRPALE (SEQ ID NO: 297). In some embodiments, the half-repeat can have a sequence as set forth in SEQ ID NO: 327, 328, 329, 330, 331, 332, 333, or 334.

TABLE 5: *Xanthomonas* Repeat Sequences

SEQ ID NO	Amino Acid Sequence	Description
323	LTPDQVVAIASNHGGKQALETVQRLLPVLCQDHG	RVD of NH recognizing guanine
324	LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG	RVD of NG recognizing thymidine
325	LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG	RVD of NI recognizing adenosine
SEQ ID NO: 326	LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG	RVD of HD recognizing cytosine
SEQ ID NO: 297	LTPQQVVAIASNGGGRPALE	Half repeat
SEQ ID NO: 327	LTPEQVVAIASNGGGRPALE	Half repeat

SEQ ID NO: 328	LTPDQVVAIASNGGGRPALE	Half repeat
SEQ ID NO: 329	LTPEQVVAIASNIGGRPALE	Half repeat
SEQ ID NO: 330	LTPDQVVAIASNIGGRPALE	Half repeat
SEQ ID NO: 331	LTPEQVVAIASHDGGRPALE	Half repeat
SEQ ID NO: 332	LTPDQVVAIASHDGGRPALE	Half repeat
SEQ ID NO: 333	LTPEQVVAIASNHGGRPALE	Half repeat
SEQ ID NO: 334	LTPDQVVAIASNHGGRPALE	Half repeat

[0069] In some embodiments, a TALE of the present disclosure can have the full length naturally occurring C-terminus of a naturally occurring *Xanthomonas spp.*-derived protein, such as the C-terminus of SEQ ID NO: 299. The C-terminus of the TALE protein sequence set forth in SEQ ID NO:299 is italicized. In some embodiments, the C-terminus can be a fragment of the full length naturally occurring C-terminus of a naturally occurring *Xanthomonas spp.*-derived protein. In some embodiments, the C-terminus can be less than 250 amino acids long. In some embodiments, the C-terminus can be positions 1 (S) through position 278 (Q) of the naturally occurring *Xanthomonas spp.*-derived protein as follows:
SIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRV
ADHAQVVRVLGFFQCHSHPAQAFDDAMTQFGMSRHGLLQLFRRVGVTELEARSGTLPPAS
QRWDRILQASGMKRAKPSPTSTQTPDQASLHAFADSLERDLDAPSPTHEGDQRRASSRKRS
RSDRAVTGPSAQQSFEVRAPEQRDALHLPLSWRVKRPRTSIGGGLPDPGTPTAADLAASSTV
MREQDEDPFAGAADDFPAFNEEELAWLMELLPQ (SEQ ID NO: 302). In some embodiments, any truncation of the full length naturally occurring C-terminus of a naturally occurring *Xanthomonas spp.* -derived protein can be used at the C-terminus of a TALE of the present disclosure. For example, in some embodiments, the naturally occurring N-terminus of *Xanthomonas spp.* can be truncated to amino acid residues at position 1 (S) to position 63 (X) as follows:
SIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRV
A (SEQ ID NO: 298). The naturally occurring C-terminus of *Xanthomonas spp.* can be truncated amino acid residues at positions 1 to 50 and used at the C-terminus of the engineered DNA binding domain. The naturally occurring C-terminus of *Xanthomonas spp.* can be truncated to amino acid residues at positions 1 to 63, 1 to 50, 1 to 70, 1 to 100, 1 to 120, 1 to 130, 10 to 40, 60 to 100, or 100 to 120 and used at the C-terminus of the engineered DNA binding domain.

[0070] The terms "N-cap" polypeptide and "N-terminal sequence" are used to refer to an amino acid sequence (polypeptide) that flanks the N-terminal portion of the first TALE repeat unit. The N-cap sequence can be of any length (including no amino acids), so long as the TALE-repeat unit(s) function to bind DNA. An N-terminal fragment and grammatical equivalents thereof refers to a shortened sequence of an N-terminal sequence which fragment is sufficient for the TALE repeat units to bind to DNA.

[0071] The term "C-cap" or "C-terminal region" refers to optionally present amino acid sequences that may be flanking the C-terminal portion of the last TALE repeat unit. The C-cap can also comprise any part of a terminal C-terminal TALE repeat, including 0 residues, truncations of a TALE repeat or a full TALE repeat. A C-terminal fragment and grammatical equivalents thereof refers to a shortened sequence of a C-terminal sequence which fragment is sufficient for the TALE repeat units to bind to DNA.

Animal Pathogen Derived Modular Nucleic Acid Binding Domains

[0072] The present disclosure provides a modular nucleic acid binding domain derived from an animal pathogen protein (MAP-NBD) can comprise a plurality of repeat units, wherein a repeat unit of the plurality of repeat units recognizes a single target nucleotide, base pair, or both.

[0073] In some embodiments, the repeat unit can be derived from an animal pathogen, and can be referred to as a non-naturally occurring modular nucleic acid binding domain derived from an animal pathogen protein (MAP-NBD), or "modular animal pathogen-nucleic acid binding domain" (MAP-NBD). For example, in some cases, the animal pathogen can be from the Gram-negative bacterium genus, Legionella. In other cases, the animal pathogen can be from *Burkholderia*. In some cases, the animal pathogen can be from *Paraburkholderia*. In other cases, the animal pathogen can be from *Francisella*.

[0074] In particular embodiments, the repeat unit can be derived from a species of the genus of Legionella, such as *Legionella quateirensis*, the genus of *Burkholderia*, the genus of *Paraburkholderia*, or the genus of *Francisella*. In some embodiments, the repeat unit can comprise from 19 amino acid residues to 35 amino acid residues. In particular embodiments, the repeat unit can comprise 33 amino acid residues. In other embodiments, the repeat unit can comprise 35 amino acid residues. In some embodiments, the MAP-NBD is non-naturally occurring, and comprises a plurality of repeat units and wherein a repeat unit of the plurality of repeat units recognizes a single target nucleic acid.

[0075] In some embodiments, a repeat unit can be derived from a *Legionella quateirensis* protein with the following sequence:
MPDLELNFAIPLHLFDDETVFTHDATNDNSQASSSYSSKSSPASANARKRTSRKEMSGPPSK
EPANTKSRRANSQNNKLSLADRLTKYNIDEEFYQTRSDSLLSLNYTKKQIERLILYKGRTSA
VQQLLCKHEELLNLISPDGLGHKELIKIAARNGGGNNLIAVLSCYAKLKEMGFSSQQIIRMV
SHAGGANNLKAVTANHDDLQNMGFNVEQIVRMVSHNGGSKNLKAVTDNHDDLKNMGFN
AEQIVRMVSHGGGSKNLKAVTDNHDDLKNMGFNAEQIVSMVSNNGGSKNLKAVTDNHDD
LKNMGFNAEQIVSMVSNGGGSLNLKAVKKYHDALKDRGFNTEQIVRMVSHDGGSLNLKA
VKKYHDALRERKFNVEQIVSIVSHGGGSLNLKAVKKYHDVLKDREFNAEQIVRMVSHDGG
SLNLKAVTDNHDDLKNMGFNAEQIVRMVSHKGGSKNLALVKEYFPVFSSFHFTADQIVALI
CQSKQCFRNLKKNHQQWKNKGLSAEQIVDLILQETPPKPNFNNTSSSTPSPSAPSFFQGPSTP
IPTPVLDNSPAPIFSNPVCFFSSRSENNTEQYLQDSTLDLDSQLGDPTKNFNVNNFWSLFPFD
DVGYHPHSNDVGYHLHSDEESPFFDF (SEQ ID NO: 281).

[0076] In some embodiments, a repeat from a *Legionella quateirensis* protein can comprise a repeat with a canonical RVD or a non-canonical RVD. In some embodiments, a canonical RVD can comprise NN, NG, HD, or HD. In some embodiments, a non-canonical RVD can comprise RN, HA, HN, HG, HG, or HK.

[0077] In some embodiments, a repeat of SEQ ID NO: 282 comprises an RVD of HA and primarily recognizes a base of adenine (A). In some embodiments, a repeat of SEQ ID NO: 283 comprises an RVD of HN and recognizes a base comprising guanine (G). In some embodiments, a repeat of S SEQ ID NO: 284 comprises an RVD of HG and recognizes a base comprising thymine (T). In some embodiments, a repeat of SEQ ID NO: 285 comprises an RVD of NN and recognizes a base comprising guanine (G). In some embodiments, a repeat of SEQ ID NO: 286 comprises an RVD of NG and recognizes a base comprising thymine (T). In some embodiments, a repeat of SEQ ID NO: 287 comprises an RVD of HD and recognizes a base comprising cytosine (C). In some embodiments, a repeat of SEQ ID NO: 288 comprises an RVD of HG and recognizes a base comprising thymine (T). In some embodiments, a repeat of SEQ ID NO: 289 comprises an RVD of HD and recognizes a base comprising cytosine (C). In some embodiments, a half-repeat of SEQ ID NO: 290 comprises an RVD of HK and recognizes a base comprising guanine (G). In some embodiments, a repeat of SEQ ID NO: 357 comprises an RVD of RN and recognizes a base comprising guanine (G).

[0078] TABLE 6 illustrates exemplary repeats from *Legionella quateirensis*, *Burkholderia*, *Paraburkholderia*, or *Francisella* that can make up a MAP-NBD of the present disclosure and the RVD at position 12 and 13 of the particular repeat. A MAP-NBD of the present disclosure can

comprise at least one of the repeats disclosed in **TABLE 5** including any one of SEQ ID NO: 357, SEQ ID NO: 282 – SEQ ID NO: 290, or SEQ ID NO: 358 – SEQ ID NO: 446. A MAP-NBD of the present disclosure can comprise any combination of repeats disclosed in **TABLE 5** including any one of SEQ ID NO: 357, SEQ ID NO: 282 – SEQ ID NO: 290, or SEQ ID NO: 358 – SEQ ID NO: 446.

TABLE 6 – Animal Pathogen Derived Repeat Units

SEQ ID NO	Organism	Repeat Unit Sequence	RVD
SEQ ID NO: 357	*L. quateirensis*	LGHKELIKIAARNGGGNNLIAVLSCYAKLKEMG	RN
SEQ ID NO: 282	*L. quateirensis*	FSSQQIIRMVSHAGGANNLKAVTANHDDLQNMG	HA
SEQ ID NO: 283	*L. quateirensis*	FNVEQIVRMVSHNGGSKNLKAVTDNHDDLKNMG	HN
SEQ ID NO: 284	*L. quateirensis*	FNAEQIVRMVSHGGGSKNLKAVTDNHDDLKNMG	HG
SEQ ID NO: 285	*L. quateirensis*	FNAEQIVSMVSNNGGSKNLKAVTDNHDDLKNMG	NN
SEQ ID NO: 286	*L. quateirensis*	FNAEQIVSMVSNGGGSLNLKAVKKYHDALKDRG	NG
SEQ ID NO: 287	*L. quateirensis*	FNTEQIVRMVSHDGGSLNLKAVKKYHDALRERK	HD
SEQ ID NO: 288	*L. quateirensis*	FNVEQIVSIVSHGGGSLNLKAVKKYHDVLKDRE	HG
SEQ ID NO: 289	*L. quateirensis*	FNAEQIVRMVSHDGGSLNLKAVTDNHDDLKNMG	HD
SEQ ID NO: 290 (half-repeat)	*L. quateirensis*	FNAEQIVRMVSHKGGSKNL	HK
SEQ ID NO: 358	*L. quateirensis*	FSAEQIVRIAAHDGGSRNIEAVQQAQHVLKELG	HD
SEQ ID NO: 359	*L. quateirensis*	FSAEQIVSIVAHDGGSRNIEAVQQAQHILKELG	HD
SEQ ID NO: 360	*L. quateirensis*	FSRQQILRIASHDGGSKNIAAVQKFLPKLMNFGFN	HD
SEQ ID NO: 361	*L. quateirensis*	FSAEQIVRIAAHDGGSLNIDAVQQAQQALKELG	HD
SEQ ID NO: 362	*L. quateirensis*	FSTEQIVCIAGHGGGSLNIKAVLLAQQALKDLG	HG
SEQ ID NO: 363	*L. quateirensis*	FSSEQIVRVAAHGGGSLNIKAVLQAHQALKELD	HG
SEQ ID NO: 364	*L. quateirensis*	FSAEQIVHIAAHGGGSLNIKAILQAHQTLKELN	HG
SEQ ID NO: 365	*L. quateirensis*	FSAEQIVRIAAHIGGSRNIEAIQQAHHALKELG	HI
SEQ ID NO: 366	*L. quateirensis*	FSAEQIVRIAAHIGGSHNLKAVLQAQQALKELD	HI
SEQ ID NO: 367	*L. quateirensis*	FSAKHIVRIAAHIGGSLNIKAVQQAQQALKELG	HI
SEQ ID NO: 368	*L. quateirensis*	FNAEQIVRMVSHKGGSKNLALVKEYFPVFSSFH	HK
SEQ ID NO: 369	*L. quateirensis*	FNAEQIVRMVSHKGGSKNLALVKEYFPVFSSFHFT	HK
SEQ ID NO: 370	*L. quateirensis*	FSADQIVRIAAHKGGSHNIVAVQQAQQALKELD	HK
SEQ ID NO: 371	*L. quateirensis*	FNVEQIVRMVSHNGGSKNLKAVTDNHDDLKNMGFN	HN
SEQ ID NO: 372	*L. quateirensis*	FSADQVVKIAGHSGGSNNIAVMLAVFPRLRDFGFK	HS

SEQ ID NO	Organism	Repeat Unit Sequence	RVD
SEQ ID NO: 373	*L. quateirensis*	FSAEQIVSIAAHVGGSHNIEAVQKAHQALKELD	HV
SEQ ID NO: 374	*L. quateirensis*	FNAEQIVSMVSNNGGSKNLKAVTDNHDDLKNMGFN	NN
SEQ ID NO: 375	*L. quateirensis*	FSHKELIKIAARNGGGNNLIAVLSCYAKLKEMG	RN
SEQ ID NO: 376	*L. quateirensis*	FSHKELIKIAARNGGGNNLIAVLSCYAKLKEMGFS	RN
SEQ ID NO: 377	*Burkholderia*	FSSGETVGATVGAGGTETVAQGGTASNTTVSSGGY	GA
SEQ ID NO: 378	*Burkholderia*	FSGGMATSTTVGSGGTQDVLAGGAAVGGTVGTGGV	GS
SEQ ID NO: 379	*Burkholderia*	FSAADIVKIAGKIGGAQALQAFITHRAALIQAGFS	KI
SEQ ID NO: 380	*Burkholderia*	FNPTDIVKIAGNDGGAQALQAVLELEPALRERGFS	ND
SEQ ID NO: 381	*Burkholderia*	FNPTDIVRMAGNDGGAQALQAVFELEPAFRERSFS	ND
SEQ ID NO: 382	*Burkholderia*	FNPTDIVRMAGNDGGAQALQAVLELEPAFRERGFS	ND
SEQ ID NO: 383	*Burkholderia*	FSQVDIVKIASNDGGAQALYSVLDVEPTFRERGFS	ND
SEQ ID NO: 384	*Burkholderia*	FSRADIVKIAGNDGGAQALYSVLDVEPPLRERGFS	ND
SEQ ID NO: 385	*Burkholderia*	FSRGDIVKIAGNDGGAQALYSVLDVEPPLRERGFS	ND
SEQ ID NO: 386	*Burkholderia*	FNRADIVRIAGNGGGAQALYSVRDAGPTLGKRGFS	NG
SEQ ID NO: 387	*Burkholderia*	FRQADIVKIASNGGSAQALNAVIKLGPTLRQRGFS	NG
SEQ ID NO: 388	*Burkholderia*	FRQADIVKMASNGGSAQALNAVIKLGPTLRQRGFS	NG
SEQ ID NO: 389	*Burkholderia*	FSRADIVKIAGNGGGAQALQAVLELEPTFRERGFS	NG
SEQ ID NO: 390	*Burkholderia*	FSRADIVRIAGNGGGAQALYSVLDVGPTLGKRGFS	NG
SEQ ID NO: 391	*Burkholderia*	FSRGDIVRIAGNGGGAQALQAVLELEPTLGERGFS	NG
SEQ ID NO: 392	*Burkholderia*	FSRADIVKIAGNGGGAQALQAVITHRAALTQAGFS	NG
SEQ ID NO: 393	*Burkholderia*	FSRGDTVKIAGNIGGAQALQAVLELEPTLRERGFS	NI
SEQ ID NO: 394	*Burkholderia*	FNPTDIVKIAGNIGGAQALQAVLELEPAFRERGFS	NI
SEQ ID NO: 395	*Burkholderia*	FSAADIVKIAGNIGGAQALQAIFTHRAALIQAGFS	NI
SEQ ID NO: 396	*Burkholderia*	FSAADIVKIAGNIGGAQALQAVITHRATLTQAGFS	NI
SEQ ID NO: 397	*Burkholderia*	FSATDIVKIASNIGGAQALQAVISRRAALIQAGFS	NI
SEQ ID NO: 398	*Burkholderia*	FSQPDIVKIAGNIGGAQALQAVLELEPAFRERGFS	NI
SEQ ID NO: 399	*Burkholderia*	FSRADIVKIAGNIGGAQALQAVLELESTFRERSFN	NI
SEQ ID NO: 400	*Burkholderia*	FSRADIVKIAGNIGGAQALQAVLELESTLRERSFN	NI
SEQ ID NO: 401	*Burkholderia*	FSRGDIVKMAGNIGGAQALQAGLELEPAFRERGFS	NI
SEQ ID NO: 402	*Burkholderia*	FSRGDIVKMAGNIGGAQALQAVLELEPAFHERSFC	NI
SEQ ID NO: 403	*Burkholderia*	FTLTDIVKMAGNIGGAQALKAVLEHGPTLRQRDLS	NI
SEQ ID NO: 404	*Burkholderia*	FTLTDIVKMAGNIGGAQALKVVLEHGPTLRQRDLS	NI
SEQ ID NO: 405	*Burkholderia*	FNPTDIVKIAGNNGGAQALQAVLELEPALRERGFS	NN
SEQ ID NO: 406	*Burkholderia*	FNPTDIVKIAGNNGGAQALQAVLELEPALRERSFS	NN
SEQ ID NO: 407	*Burkholderia*	FNPTDMVKIAGNNGGAQALQAVLELEPALRERGFS	NN
SEQ ID NO: 408	*Burkholderia*	FSAADIVKIASNNGGAQALQALIDHWSTLSGKTKA	NN
SEQ ID NO: 409	*Burkholderia*	FSAADIVKIASNNGGAQALQAVISRRAALIQAGFS	NN
SEQ ID NO: 410	*Burkholderia*	FSAADIVKIASNNGGAQALQAVITHRAALAQAGFS	NN
SEQ ID NO: 411	*Burkholderia*	FSAADIVKIASNNGGARALQALIDHWSTLSGKTKA	NN
SEQ ID NO: 412	*Burkholderia*	FTLTDIVEMAGNNGGAQALKAVLEHGSTLDERGFT	NN
SEQ ID NO: 413	*Burkholderia*	FTLTDIVKMAGNNGGAQALKAVLEHGPTLDERGFT	NN
SEQ ID NO: 414	*Burkholderia*	FTLTDIVKMAGNNGGAQALKVVLEHGPTLRQRGFS	NN
SEQ ID NO: 415	*Burkholderia*	FTLTDIVKMASNNGGAQALKAVLEHGPTLDERGFT	NN
SEQ ID NO: 416	*Burkholderia*	FSAADIVKIAGNSGGAQALQAVISHRAALTQAGFS	NS

SEQ ID NO	Organism	Repeat Unit Sequence	RVD
SEQ ID NO: 417	*Burkholderia*	FSGGDAVSTVVRSGGAQSVASGGTASGTTVSAGAT	RS
SEQ ID NO: 418	*Burkholderia*	FRQTDIVKMAGSGGSAQALNAVIKHGPTLRQRGFS	SG
SEQ ID NO: 419	*Burkholderia*	FSLIDIVEIASNGGAQALKAVLKYGPVLTQAGRS	SN
SEQ ID NO: 420	*Burkholderia*	FSGGDAAGTVVSSGGAQNVTGGLASGTTVASGGAA	SS
SEQ ID NO: 421	*Paraburkholderia*	FNLTDIVEMAANSGGAQALKAVLEHGPTLRQRGLS	NS
SEQ ID NO: 422	*Paraburkholderia*	FNRASIVKIAGNSGGAQALQAVLKHGPTLDERGFN	NS
SEQ ID NO: 423	*Paraburkholderia*	FSQANIVKMAGNSGGAQALQAVLDLELVFRERGFS	NS
SEQ ID NO: 424	*Paraburkholderia*	FSQPDIVKMAGNSGGAQALQAVLDLELAFRERGFS	NS
SEQ ID NO: 425	*Paraburkholderia*	FSLIDIVEIASNGGAQALKAVLKYGPVLMQAGRS	SN
SEQ ID NO: 426	*Francisella*	YKSEDIIRLASHDGGSVNLEAVLRLHSQLTRLG	HD
SEQ ID NO: 427	*Francisella*	YKPEDIIRLASHGGGSVNLEAVLRLNPQLIGLG	HG
SEQ ID NO: 428	*Francisella*	YKSEDIIRLASHGGGSVNLEAVLRLHSQLTRLG	HG
SEQ ID NO: 429	*Francisella*	YKSEDIIRLASHGGGSVNLEAVLRLNPQLIGLG	HG
SEQ ID NO: 430	*Paraburkholderia*	FNLTDIVEMAGKGGGAQALKAVLEHGPTLRQRGFN	KG
SEQ ID NO: 431	*Paraburkholderia*	FRQADIIKIAGNDGGAQALQAVIEHGPTLRQHGFN	ND
SEQ ID NO: 432	*Paraburkholderia*	FSQADIVKIAGNDGGTQALHAVLDLERMLGERGFS	ND
SEQ ID NO: 433	*Paraburkholderia*	FSRADIVKIAGNGGGAQALKAVLEHEATLDERGFS	NG
SEQ ID NO: 434	*Paraburkholderia*	FSRADIVRIAGNGGGAQALYSVLDVEPTLGKRGFS	NG
SEQ ID NO: 435	*Paraburkholderia*	FSQPDIVKMASNIGGAQALQAVLELEPALRERGFS	NI
SEQ ID NO: 436	*Paraburkholderia*	FSQPDIVKMAGNIGGAQALQAVLSLGPALRERGFS	NI
SEQ ID NO: 437	*Paraburkholderia*	FSQPEIVKIAGNIGGAQALHTVLELEPTLHKRGFN	NI
SEQ ID NO: 438	*Paraburkholderia*	FSQSDIVKIAGNIGGAQALQAVLDLESMLGKRGFS	NI
SEQ ID NO: 439	*Paraburkholderia*	FSQSDIVKIAGNIGGAQALQAVLELEPTLRESDFR	NI
SEQ ID NO: 440	*Paraburkholderia*	FNPTDIVKIAGNKGGAQALQAVLELEPALRERGFN	NK
SEQ ID NO: 441	*Paraburkholderia*	FSPTDIIKIAGNNGGAQALQAVLDLELMLRERGFS	NN
SEQ ID NO: 442	*Paraburkholderia*	FSQADIVKIAGNNGGAQALYSVLDVEPTLGKRGFS	NN
SEQ ID NO: 443	*Paraburkholderia*	FSRGDIVTIAGNNGGAQALQAVLELEPTLRERGFN	NN
SEQ ID NO: 444	*Paraburkholderia*	FSRIDIVKIAANNGGAQALHAVLDLGPTLRECGFS	NN
SEQ ID NO: 445	*Paraburkholderia*	FSQADIVKIVGNNGGAQALQAVFELEPTLRERGFN	NN
SEQ ID NO: 446	*Paraburkholderia*	FSQPDIVRITGNRGGAQALQAVLALELTLRERGFS	NR

[0079] In any one of the animal pathogen-derived repeat domains of SEQ ID NO: 357, SEQ ID NO: 282 – SEQ ID NO: 290, or SEQ ID NO: 358 – SEQ ID NO: 446, there can be considerable sequence divergence between repeats of a MAP-NBD outside of the RVD.

[0080] In some embodiments, a MAP-NBD of the present disclosure can comprise between 1 to 50 animal pathogen-derived repeat units. In some embodiments, a MAP-NBD of the present disclosure can comprise between 9 and 36 animal pathogen-derived repeat units. Preferably, in some embodiments, a MAP-NBD of the present disclosure can comprise between 12 and 30 animal pathogen-derived repeat units. A MAP-NBD described herein can comprise between 5 to 10, 10 to 15, 15-20, 20 to 25, 25 to 30, 30 to 35, or 35 to 40, e.g., 15-25 animal pathogen-derived repeat units. A MAP-NBD described herein can comprise 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,

21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39 or 40 animal pathogen-derived repeat units, e.g., .

[0081] A MAP-NBD described herein can comprise 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39 or 40 animal pathogen-derived repeat units.

[0082] An animal pathogen-derived repeat units can be derived from a wild-type repeat unit, such as any one of SEQ ID NO: 357, SEQ ID NO: 282 – SEQ ID NO: 290, or SEQ ID NO: 358 – SEQ ID NO: 446. An animal pathogen-derived repeat unit can also comprise a modified animal pathogen-derived repeat units enhanced for specific recognition of a nucleotide or base pair. A MAP-NBD described herein can comprise one or more wild-type animal pathogen-derived repeat units, one or more modified animal pathogen-derived repeat units, or a combination thereof. In some embodiments, a modified animal pathogen-derived repeat units can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, or 29 mutations that can enhance recognition of a specific nucleotide or base pair. In some embodiments, a modified animal pathogen-derived repeat unit can comprise more than 1 modification, for example 1 to 5 modifications, 5 to 10 modifications, 10 to 15 modifications, 15 to 20 modifications, 20 to 25 modification, or 25-29 modifications. In some embodiments, A MAP-NBD can comprise more than one modified animal pathogen-derived repeat units, wherein each of the modified animal pathogen-derived repeat units can have a different number of modifications.

[0083] In some embodiments, a MAP-NBD of the present disclosure can have the full length naturally occurring N-terminus of a naturally occurring *Legionella quateirensis*-derived protein, such as the N-terminus of SEQ ID NO: 281. A N-terminus can be the full length N-terminus sequence and can have a sequence of MPDLELNFAIPLHLFDDETVFTHDATNDNSQASSSYSSKSSPASANARKRTSRKEMSGPPSKEPANTKSRRANSQNNKLSLADRLTKYNIDEEFYQTRSDSLLSLNYTKKQIERLILYKGRTSAVQQLLCKHEELLNLISPDG (SEQ ID NO: 291). In some embodiments, any truncation of SEQ ID NO: 291 can be used as the N-terminus in a MAP-NBD of the present disclosure. For example, in some embodiments, a MAP-NBD comprises a truncated N-terminus including amino acid residues at position 1 (G) to position 137 (S) of the naturally occurring *Legionella quateirensis* N-terminus as follows: NFAIPLHLFDDETVFTHDATNDNSQASSSYSSKSSPASANARKRTSRKEMSGPPSKEPANTKSRRANSQNNKLSLADRLTKYNIDEEFYQTRSDSLLSLNYTKKQIERLILYKGRTSAVQQLLCKHEELLNLISPDG (SEQ ID NO: 335). For example, in some embodiments, a MAP-NBD

comprises a truncated N-terminus including amino acid residues at position 1 (G) to position 120 (S) of the naturally occurring *Legionella quateirensis* N-terminus as follows: DATNDNSQASSSYSSKSSPASANARKRTSRKEMSGPPSKEPANTKSRRANSQNNKLSLADR LTKYNIDEEFYQTRSDSLLSLNYTKKQIERLILYKGRTSAVQQLLCKHEELLNLISPDG (SEQ ID NO: 304). In some embodiments, a MAP-NBD comprises a truncated N-terminus including amino acid residues at position 1 (G) to position 115 (K) of the naturally occurring *Legionella quateirensis* N-terminus as follows: NSQASSSYSSKSSPASANARKRTSRKEMSGPPSKEPANTKSRRANSQNNKLSLADRLTKYNI DEEFYQTRSDSLLSLNYTKKQIERLILYKGRTSAVQQLLCKHEELLNLISPDG (SEQ ID NO: 322). In some embodiments, any truncation of the naturally occurring *Legionella quateirensis*-derived protein can be used at the N-terminus of a DNA binding domain disclosed herein. The naturally occurring N-terminus of *Legionella quateirensis* can be truncated to amino acid residues at positions 1 to 50, 1 to 70, 1 to 100, 1 to 120, 1 to 130, 10 to 40, 60 to 100, or 100 to 120 and used at the N-terminus of the MAP-NBD.

[0084] In some embodiments, a MAP-NBD of the present disclosure can have the full length naturally occurring C-terminus of a naturally occurring *Legionella quateirensis*-derived protein. In some embodiments, A MAP-NBD of the present disclosure can have at its C-terminus amino acid residues at position 1 (A) to position 176 (F) of the naturally occurring *Legionella quateirensis*-derived protein as follows: ALVKEYFPVFSSFHFTADQIVALICQSKQCFRNLKKNHQQWKNKGLSAEQIVDLILQETPPK PNFNNTSSSTPSPSAPSFFQGPSTPIPTPVLDNSPAPIFSNPVCFFSSRSENNTEQYLQDSTLDLD SQLGDPTKNFNVNNFWSLFPFDDVGYHPHSNDVGYHLHSDEESPFFDF (SEQ ID NO: 305). In some embodiments, a MAP-NBD of the present disclosure can have at its C-terminus amino acid residues at position 1 (A) to position 63 (P) of the naturally occurring *Legionella quateirensis*-derived protein as follows: ALVKEYFPVFSSFHFTADQIVALICQSKQCFRNLKKNHQQWKNKGLSAEQIVDLILQETPPK P (SEQ ID NO: 306).

[0085] In some embodiments, the present disclosure provides methods for identifying an animal pathogen-derived repeat unit. For example, a consensus sequence can be defined comprising a first repeat motif, a spacer, and a second repeat motif. The consensus sequence can be 1xxx211x1xxx33x2x1xxxxxxxxx1xxxx1xxx211x1xxx33x2x1xxxxxxxxx1 (SEQ ID NO: 292), 1xxx211x1xxx33x2x1xxxxxxxxx1xxxxx1xxx211x1xxx33x2x1xxxxxxxxx1 (SEQ ID NO: 293), 1xxx211x1xxx33x2x1xxxxxxxxx1xxxxxx1xxx211x1xxx33x2x1xxxxxxxxx1 (SEQ ID NO: 294),

1xxx211x1xxx33x2x1xxxxxxxxx1xxxxxxx1xxx211x1xxx33x2x1xxxxxxxxx1 (SEQ ID NO: 295), 1xxx211x1xxx33x2x1xxxxxxxxx1xxxxxxxx1xxx211x1xxx33x2x1xxxxxxxxx1 (SEQ ID NO: 296). For any one of SEQ ID NO: 292 – SEQ ID NO: 296, x can be any amino acid residue, 1, 2, and 3 are flexible residues that are defined as follows: 1 can be selected from any one of A, F, I, L, M, T, or V, 2 can be selected from any one of D, E, K, N, M, S, R, or Q, and 3 can be selected from any one of A, G, N, or S. Thus, in some embodiments, a MAP-NBD can be derived from an animal pathogen comprising the consensus sequence of SEQ ID NO: 292, SEQ ID NO: 293, SEQ ID NO: 294, SEQ ID NO: 295, or SEQ ID NO: 296. Any one of consensus sequences of SEQ ID NO: 292 – SEQ ID NO: 296 can be compared against all sequences downloaded from NCBI, MGRast, JGI, and EBI databases to identify matches corresponding to animal pathogen proteins containing repeat units of a DNA-binding repeat unit.

[0086] In some embodiments, a MAP-NBD repeat unit can itself have a consensus sequence of 1xxx211x1xxx33x2x1xxxxxxxxx1 (SEQ ID NO: 293), wherein x can be any amino acid residue, 1, 2, and 3 are flexible residues that are defined as follows: 1 can be selected from any one of A, F, I, L, M, T, or V, 2 can be selected from any one of D, E, K, N, M, S, R, or Q, and 3 can be selected from any one of A, G, N, or S.

Mixed DNA Binding Domains

[0087] In some embodiments, the present disclosure provides DNA binding domains in which the repeat units, the N-terminus, and the C-terminus can be derived from any one of *Ralstonia solanacearum, Xanthomonas spp.*, *Legionella quateirensis*, *Burkholderia*, *Paraburkholderia*, or *Francisella*. For example, the present disclosure provides a DNA binding domain wherein the plurality of repeat units are selected from any one of SEQ ID NO: 168 – SEQ ID NO: 263 or SEQ ID NO: 336 – SEQ ID NO 356 and can further comprise an N-terminus and/or C-terminus from *Xanthomonas spp.*, (N-termini: SEQ ID NO: 298, SEQ ID NO: 300, SEQ ID NO: 301, and SEQ ID NO: 321; C-termini: SEQ ID NO: 302 and SEQ ID NO: 298) or *Legionella quateirensis* (N-termini: SEQ ID NO: 304 or SEQ ID NO: 322; C-termini: SEQ ID NO: 305 and SEQ ID NO: 306). In some embodiments, the present disclosure provides modular DNA binding domains in which the repeat units can be from *Ralstonia solanacearum* (e.g., any one of SEQ ID NO: 168 – SEQ ID NO: 263 or SEQ ID NO: 336 – SEQ ID NO 356), *Xanthomonas spp.* (e.g., any one of SEQ ID NO: 323 – SEQ ID NO: 334), an animal pathogen such as *Legionella quateirensis*, *Burkholderia*, *Paraburkholderia*, or *Francisella* (e.g., any one of SEQ ID NO: 357, SEQ ID NO: 282 – SEQ ID NO: 290, or SEQ ID NO: 358 – SEQ ID NO: 446), or any combination thereof.

Nucleases for Genome Editing

[0088] Genome editing can include the process of modifying a DNA of a cell in order to introduce or knock out a target gene or a target gene region. In some instances, a subject may have a disease in which a protein is aberrantly expressed or completely lacking. One therapeutic strategy for treating this disease can be introduction of a target gene or a target gene region to correct the aberrant or missing protein. For example, genome editing can be used to modify the DNA of a cell in the subject in order to introduce a functional gene, which gives rise to a functional protein. Introduction of this functional gene and expression of the functional protein can relieve the disease state of the subject.

[0089] In other instances, a subject may have a disease in which protein is overexpressed or is targeted by a virus for infection of a cell. Alternatively, a therapy such as a cell therapy for cancer can be ineffective due to repression of certain processes by tumor cells (*e.g.*, checkpoint inhibition). Still alternatively, it may be desirable to eliminate a particular protein expressed at the surface of a cell in order to generate a universal, off-the-shelf cell therapy for a subject in need thereof (*e.g.*, TCR). In such cases, it can be desirable to partially or completely knock out the gene encoding for such a protein. Genome editing can be used to modify the DNA of a cell in the subject in order to partially or completely knock out the target gene, thus reducing or eliminating expression of the protein of interest.

[0090] Genome editing can include the use of any nuclease as described herein in combination with any DNA binding domain disclosed herein in order to bind to a target gene or target gene region and induce a double strand break, mediated by the nuclease. Genes can be introduced during this process, or DNA binding domains can be designed to cut at regions of the DNA such that after non-homologous end joining, the target gene or target gene region is removed. Genome editing systems that are further disclosed and described in detail herein can include DNA binding domains from *Xanthomonas, Ralstonia,* or *Legionella* fused to nucleases.

[0091] The specificity and efficiency of genome editing can be dependent on the nuclease responsible for cleavage. More than 3,000 type II restriction endonucleases have been identified. They recognize short, usually palindromic, sequences of 4–8 bp and, in the presence of Mg2+, cleave the DNA within or in close proximity to the recognition sequence. Naturally, type IIs restriction enzymes themselves have a DNA recognition domain that can be separated from the catalytic, or cleavage, domain. As such, since cleavage occurs at a site adjacent to the DNA sequence bound by the recognition domain, these enzymes can be referred to as exhibiting "shifted" cleavage. These type IIs restriction enzymes having both the recognition domain and the cleavage domain can be 400-600 amino acids. The main criterion for classifying a restriction endonuclease as

a type II enzyme is that it cleaves specifically within or close to its recognition site and that it does not require ATP hydrolysis for its nucleolytic activity. An example of a type II restriction endonucleases is FokI, which consists of a DNA recognition domain and a non-specific DNA cleavage domain. FokI cleaves DNA nine and thirteen bases downstream of an asymmetric sequence (recognizing a DNA sequence of GGATG).

[0092] In some embodiments, the DNA cleavage domain at the C-terminus of FokI itself can be combined with a variety of DNA-binding domains (e.g., RNBDs, TALEs, MAP-NBDs) of other molecules for genome editing purposes. This cleavage domain can be 180 amino acids in length and can be directly linked to a DNA binding domain (e.g., RNBDs, TALEs, MAP-NBDs). In some embodiments, the FokI cleavage domain only comprises a single catalytic site. Thus, in order to cleave phosphodiester bonds, these enzymes form transient homodimers, providing two catalytic sites capable of cleaving double stranded DNA. In some embodiments, a single DNA-binding domains (e.g., RNBDs, TALEs, MAP-NBDs) linked to a Type IIS cleaving domain may not nick the double stranded DNA at the targeted site. In some embodiments, cleaving of target DNA only occurs when a pair of DNA-binding domains (e.g., RNBDs, TALEs, MAP-NBDs), each linked to a Type IIS cleaving domain (e.g., any one of SEQ ID NO: 1 – SEQ ID NO: 81 (nucleotide sequences of SEQ ID NO: 82 – SEQ ID NO: 162)) bind to opposing strands of DNA and allow for formation of a transient homodimer in the spacer region (the base pairs between the C-terminus of the DNA binding domain on a top strand of DNA and the C-terminus of the DNA binding domain on a bottom strand of DNA). Said spacer region can be greater than 2 base pairs, greater than 5 base pairs, greater than 10 base pairs, greater than 15 base pairs, greater than 24 base pairs, greater than 25 base pairs, greater than 30 base pairs, greater than 35 base pairs, greater than 40 base pairs, greater than 45 base pairs, or greater than 50 base pairs. In some embodiments, the spacer region can be anywhere from 2 to 50 base pairs, 5 to 40 base pairs, 10 to 30 base pairs, 14 to 40 base pairs, 24 to 30 base pairs, 24 to 40 base pairs, or 24 to 50 base pairs. In some embodiments, the nuclease disclosed herein (e.g., any one of SEQ ID NO: 1 – SEQ ID NO: 81 (nucleotide sequences of SEQ ID NO: 82 – SEQ ID NO: 162) can be capable of cleaving over a spacer region of greater than 24 base pairs upon formation of a transient homodimer.

[0093] In some instances, such enzymes can comprise one or more mutations relative to SEQ ID NO: 1 – SEQ ID NO: 81 (nucleotide sequences of SEQ ID NO: 82 – SEQ ID NO: 162). In some cases, the non-naturally occurring enzymes described herein can comprise about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more mutations. A mutation can be engineered to enhance cleavage efficiency. A mutation can abolish cleavage activity. In some cases, a mutation can enhance homodimerization. For

example, FokI can have a mutation at one or more amino acid residue positions 446, 447, 479, 483, 484, 486, 487, 490, 491, 496, 498, 499, 500, 531, 534, 537, and 538 to modulate homodimerization, and similar mutations can be designed based on the phylogenetic analysis of SEQ ID NO: 1 – SEQ ID NO: 81 (nucleotide sequences of SEQ ID NO: 82 – SEQ ID NO: 162).

[0094] TABLE 7 shows exemplary amino acid sequences (SEQ ID NO: 1 – SEQ ID NO: 81) of endonucleases for genome editing and the corresponding back-translated nucleic acid sequences (SEQ ID NO: 82 – SEQ ID NO: 162) of the endonucleases, which were obtained using Geneious software and selecting for human codon optimization.

TABLE 7 – Amino Acid and Nucleic Acid Sequences of Endonucleases

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
1	FLVKGAMEIKKSEL RHKLRHVPHEYIELI EIAQDSKQNRLLEFK VVEFFKKIYGYRGK HLGGSRKPDGALFT DGLVLNHGIILDTKA YKDGYRLPISQADE MQRYVDENNKRSQ VINPNEWWEIYPTSI TDFKFLFVSGFFQGD YRKQLERVSHLTKC QGAVMSVEQLLLGG EKIKEGSLTLEEVGK KFKNDEIVF	82	TTCCTGGTGAAGGGCGCCATGGAGATCAAGAAGAGCGAGCTGAG GCACAAGCTGAGGCACGTGCCCCACGAGTACATCGAGCTGATCG AGATCGCCCAGGACAGCAAGCAGAACAGGCTGCTGGAGTTCAAG GTGGTGGAGTTCTTCAAGAAGATCTACGGCTACAGGGGCAAGCAC CTGGGCGGCAGCAGGAAGCCCGACGGCGCCCTGTTCACCGACGG CCTGGTGCTGAACCACGGCATCATCCTGGACACCAAGGCCTACAA GGACGGCTACAGGCTGCCCATCAGCCAGGCCGACGAGATGCAGA GGTACGTGGACGAGAACAACAAGAGGAGCCAGGTGATCAACCCC AACGAGTGGTGGGAGATCTACCCCACCAGCATCACCGACTTCAAG TTCCTGTTCGTGAGCGGCTTCTTCCAGGGCGACTACAGGAAGCAG CTGGAGAGGGTGAGCCACCTGACCAAGTGCCAGGGCGCCGTGAT GAGCGTGGAGCAGCTGCTGCTGGGCGGCGAGAAGATCAAGGAGG GCAGCCTGACCCTGGAGGAGGTGGGCAAGAAGTTCAAGAACGAC GAGATCGTGTTC
2	QIVKSSIEMSKANM RDNLQMLPHDYIELI EISQDPYQNRIFEMK VMDLFINEYGFSGS HLGGSRKPDGAMY AHGFGVIVDTKAYK DGYNLPISQADEME RYVRENIDRNEHVN SNRWWNIFPEDTNE YKFLFVSGFFKGNFE KQLERISIDTGVQGG ALSVEHLLLGAEYIK RGILTLYDFKNSFLN KEIQF	83	CAGATCGTGAAGAGCAGCATCGAGATGAGCAAGGCCAACATGAG GGACAACCTGCAGATGCTGCCCCACGACTACATCGAGCTGATCGA GATCAGCCAGGACCCCTACCAGAACAGGATCTTCGAGATGAAGG TGATGGACCTGTTCATCAACGAGTACGGCTTCAGCGGCAGCCACC TGGGCGGCAGCAGGAAGCCCGACGGCGCCATGTACGCCCACGGC TTCGGCGTGATCGTGGACACCAAGGCCTACAAGGACGGCTACAAC CTGCCCATCAGCCAGGCCGACGAGATGGAGAGGTACGTGAGGGA GAACATCGACAGGAACGAGCACGTGAACAGCAACAGGTGGTGGA ACATCTTCCCCGAGGACACCAACGAGTACAAGTTCCTGTTCGTGA GCGGCTTCTTCAAGGGCAACTTCGAGAAGCAGCTGGAGAGGATC AGCATCGACACCGGCGTGCAGGGCGGCGCCCTGAGCGTGGAGCA CCTGCTGCTGGGCGCCGAGTACATCAAGAGGGGCATCCTGACCCT GTACGACTTCAAGAACAGCTTCCTGAACAAGGAGATCCAGTTC
3	QTIKSSIEELKSELRT QLNVISHDYLQLVDI SQDSQQNRLFEMKV MDLFINEFGYNGSH LGGSRKPDGILYTEG LSKDYGIIVDTKAYK DGYNLPIAQADEME RYIRENIDRNEVVNP NRWWEVFPSKINDY KFLFVSAYFKGNFK	84	CAGACCATCAAGAGCAGCATCGAGGAGCTGAAGAGCGAGCTGAG GACCCAGCTGAACGTGATCAGCCACGACTACCTGCAGCTGGTGGA CATCAGCCAGGACAGCCAGCAGAACAGGCTGTTCGAGATGAAGG TGATGGACCTGTTCATCAACGAGTTCGGCTACAACGGCAGCCACC TGGGCGGCAGCAGGAAGCCCGACGGCATCCTGTACACCGAGGGC CTGAGCAAGGACTACGGCATCATCGTGGACACCAAGGCCTACAA GGACGGCTACAACCTGCCCATCGCCCAGGCCGACGAGATGGAGA GGTACATCAGGGAGAACATCGACAGGAACGAGGTGGTGAACCCC AACAGGTGGTGGGAGGTGTTCCCCAGCAAGATCAACGACTACAA GTTCCTGTTCGTGAGCGCCTACTTCAAGGGCAACTTCAAGGAGCA

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
	EQLERISINTGILGGA ISVEHLLLGAEYFKR GILSLEDVRDKFCNT EIEF		GCTGGAGAGGATCAGCATCAACACCGGCATCCTGGGCGGCGCCA TCAGCGTGGAGCACCTGCTGCTGGGCGCCGAGTACTTCAAGAGGG GCATCCTGAGCCTGGAGGACGTGAGGGACAAGTTCTGCAACACC GAGATCGAGTTC
4	GKSEVETIKEQMRG ELTHLSHEYLGLLDL AYDSKQNRLFELKT MQLLTEECGFEGLH LGGSRKPDGIVYTK DENEQVGKENYGIII DTKAYSGGYSLPISQ ADEMERYIGENQTR DIRINPNEWWKNFG DGVTEYYYLFVAGH FKGKYQEQIDRINCN KNIKGAAVSIQQLLR IVNDYKAGKLTHED MKLKIFHY	85	GGCAAGAGCGAGGTGGAGACCATCAAGGAGCAGATGAGGGGCG AGCTGACCCACCTGAGCCACGAGTACCTGGGCCTGCTGGACCTGG CCTACGACAGCAAGCAGAACAGGCTGTTCGAGCTGAAGACCATG CAGCTGCTGACCGAGGAGTGCGGCTTCGAGGGCCTGCACCTGGGC GGCAGCAGGAAGCCCGACGGCATCGTGTACACCAAGGACGAGAA CGAGCAGGTGGGCAAGGAGAACTACGGCATCATCATCGACACCA AGGCCTACAGCGGCGGCTACAGCCTGCCCATCAGCCAGGCCGAC GAGATGGAGAGGTACATCGGCGAGAACCAGACCAGGGACATCAG GATCAACCCCAACGAGTGGTGGAAGAACTTCGGCGACGGCGTGA CCGAGTACTACTACCTGTTCGTGGCCGGCCACTTCAAGGGCAAGT ACCAGGAGCAGATCGACAGGATCAACTGCAACAAGAACATCAAG GGCGCCGCCGTGAGCATCCAGCAGCTGCTGAGGATCGTGAACGA CTACAAGGCCGGCAAGCTGACCCACGAGGACATGAAGCTGAAGA TCTTCCACTAC
5	MKILELLINECGYKG LHLGGARKPDGIIYT EKEKYNYGVIIDTK AYSKGYNLPIGQIDE MIRYIIENNERNIKR NTNCWWNNFEKNV NEFYFSFISGEFTGNI EEKLNRIFISTNIKGN AMSVKTLLYLANEI KANRISYIELLNYFD NKV	86	ATGAAGATCCTGGAGCTGCTGATCAACGAGTGCGGCTACAAGGG CCTGCACCTGGGCGGCGCCAGGAAGCCCGACGGCATCATCTACAC CGAGAAGGAGAAGTACAACTACGGCGTGATCATCGACACCAAGG CCTACAGCAAGGGCTACAACCTGCCCATCGGCCAGATCGACGAG ATGATCAGGTACATCATCGAGAACAACGAGAGGAACATCAAGAG GAACACCAACTGCTGGTGGAACAACTTCGAGAAGAACGTGAACG AGTTCTACTTCAGCTTCATCAGCGGCGAGTTCACCGGCAACATCG AGGAGAAGCTGAACAGGATCTTCATCAGCACCAACATCAAGGGC AACGCCATGAGCGTGAAGACCCTGCTGTACCTGGCCAACGAGATC AAGGCCAACAGGATCAGCTACATCGAGCTGCTGAACTACTTCGAC AACAAGGTG
6	AKSSQSETKEKLRE KLRNLPHEYLSLVD LAYDSKQNRLFEMK VIELLTEECGFQGLH LGGSRRPDGVLYTA GLTDNYGIILDTKAY SSGYSLPIAQADEME RYVRENQTRDELVN PNQWWENFENGLG TFYFLFVAGHFNGN VQAQLERISRNTGV LGAAASISQLLLLAD AIRGGRMDRERLRH LMFQNEEFL	87	GCCAAGAGCAGCCAGAGCGAGACCAAGGAGAAGCTGAGGGAGA AGCTGAGGAACCTGCCCCACGAGTACCTGAGCCTGGTGGACCTGG CCTACGACAGCAAGCAGAACAGGCTGTTCGAGATGAAGGTGATC GAGCTGCTGACCGAGGAGTGCGGCTTCCAGGGCCTGCACCTGGGC GGCAGCAGGAGGCCCGACGGCGTGCTGTACACCGCCGGCCTGAC CGACAACTACGGCATCATCCTGGACACCAAGGCCTACAGCAGCG GCTACAGCCTGCCCATCGCCCAGGCCGACGAGATGGAGAGGTAC GTGAGGGAGAACCAGACCAGGGACGAGCTGGTGAACCCCAACCA GTGGTGGGAGAACTTCGAGAACGGCCTGGGCACCTTCTACTTCCT GTTCGTGGCCGGCCACTTCAACGGCAACGTGCAGGCCCAGCTGGA GAGGATCAGCAGGAACACCGGCGTGCTGGGCGCCGCCGCCAGCA TCAGCCAGCTGCTGCTGCTGGCCGACGCCATCAGGGGCGGCAGGA TGGACAGGGAGAGGCTGAGGCACCTGATGTTCCAGAACGAGGAG TTCCTG

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
7	NSEKSEFTQEKDNL REKLDTLSHEYLSLV DLAFDSQQNRLFEM KTVELLTKECNYKG VHLGGSRKPDGIIYT ENSTDNYGVIIDTKA YSNGYNLPISQVDE MVRYVEENNKREK ERNSNEWWKEFGD NINKFYFSFISGKFIG NIEEKLQRITIFTNVY GNAMTIITLLYLANE IKANRLKTMEVVKY FDNKV	88	AACAGCGAGAAGAGCGAGTTCACCCAGGAGAAGGACAACCTGAG GGAGAAGCTGGACACCCTGAGCCACGAGTACCTGAGCCTGGTGG ACCTGGCCTTCGACAGCCAGCAGAACAGGCTGTTCGAGATGAAG ACCGTGGAGCTGCTGACCAAGGAGTGCAACTACAAGGGCGTGCA CCTGGGCGGCAGCAGGAAGCCCGACGGCATCATCTACACCGAGA ACAGCACCGACAACTACGGCGTGATCATCGACACCAAGGCCTAC AGCAACGGCTACAACCTGCCCATCAGCCAGGTGGACGAGATGGT GAGGTACGTGGAGGAGAACAACAAGAGGGAGAAGGAGAGGAAC AGCAACGAGTGGTGGAAGGAGTTCGGCGACAACATCAACAAGTT CTACTTCAGCTTCATCAGCGGCAAGTTCATCGGCAACATCGAGGA GAAGCTGCAGAGGATCACCATCTTCACCAACGTGTACGGCAACGC CATGACCATCATCACCCTGCTGTACCTGGCCAACGAGATCAAGGC CAACAGGCTGAAGACCATGGAGGTGGTGAAGTACTTCGACAACA AGGTG
8	NLTCSDLTEIKEEVR NALTHLSHEYLALID LAYDSTQNRLFEMK TLQLLVEECGYQGT HLGGSRKPDGICYSE EAKSEGLEANYGIII DTKSYSGGYGLPISQ ADEMERYIRENQTR DAEVNRNKWWEAF PETIDIFYFMFVAGH FKGNYFNQLERLQR STGIKGAAVDIKTLL LTANRCKTGELDHA GIESCFFNNCRL	89	AACCTGACCTGCAGCGACCTGACCGAGATCAAGGAGGAGGTGAG GAACGCCCTGACCCACCTGAGCCACGAGTACCTGGCCCTGATCGA CCTGGCCTACGACAGCACCCAGAACAGGCTGTTCGAGATGAAGA CCCTGCAGCTGCTGGTGGAGGAGTGCGGCTACCAGGGCACCCACC TGGGCGGCAGCAGGAAGCCCGACGGCATCTGCTACAGCGAGGAG GCCAAGAGCGAGGGCCTGGAGGCCAACTACGGCATCATCATCGA CACCAAGAGCTACAGCGGCGGCTACGGCCTGCCCATCAGCCAGG CCGACGAGATGGAGAGGTACATCAGGGAGAACCAGACCAGGGAC GCCGAGGTGAACAGGAACAAGTGGTGGGAGGCCTTCCCCGAGAC CATCGACATCTTCTACTTCATGTTCGTGGCCGGCCACTTCAAGGGC AACTACTTCAACCAGCTGGAGAGGCTGCAGAGGAGCACCGGCAT CAAGGGCGCCGCCGTGGACATCAAGACCCTGCTGCTGACCGCCAA CAGGTGCAAGACCGGCGAGCTGGACCACGCCGGCATCGAGAGCT GCTTCTTCAACAACTGCAGGCTG
9	DNVKSNFNQEKDEL REKLDTLSHEYLYL LDLAYDSKQNKLFE MKILELLINECGYRG LHLGGVRKPDGIIYT EKEKYNYGVIIDTK AYSKGYNLPIGQIDE MIRYIIENNERNIKR NTNCWWNNFEKNV NEFYFSFISGEFTGNI EEKLNRIFISTNIKGN AMSVKTLLYLANEI KANRISFLEMEKYF DNKV	90	GACAACGTGAAGAGCAACTTCAACCAGGAGAAGGACGAGCTGAG GGAGAAGCTGGACACCCTGAGCCACGAGTACCTGTACCTGCTGGA CCTGGCCTACGACAGCAAGCAGAACAAGCTGTTCGAGATGAAGA TCCTGGAGCTGCTGATCAACGAGTGCGGCTACAGGGGCCTGCACC TGGGCGGCGTGAGGAAGCCCGACGGCATCATCTACACCGAGAAG GAGAAGTACAACTACGGCGTGATCATCGACACCAAGGCCTACAG CAAGGGCTACAACCTGCCCATCGGCCAGATCGACGAGATGATCA GGTACATCATCGAGAACAACGAGAGGAACATCAAGAGGAACACC AACTGCTGGTGGAACAACTTCGAGAAGAACGTGAACGAGTTCTAC TTCAGCTTCATCAGCGGCGAGTTCACCGGCAACATCGAGGAGAAG CTGAACAGGATCTTCATCAGCACCAACATCAAGGGCAACGCCATG AGCGTGAAGACCCTGCTGTACCTGGCCAACGAGATCAAGGCCAA CAGGATCAGCTTCCTGGAGATGGAGAAGTACTTCGACAACAAGGT G
10	EGIKSNISLLKDELR GQISHISHEYLSLIDL AFDSKQNRLFEMKV LELLVNEYGFKGRH LGGSRKPDGIVYSTT LEDNFGIIVDTKAYS EGYSLPISQADEMER YVRENSNRDEEVNP NKWWENFSEEVKK YYFVFISGSFKGKFE EQLRRLSMTTGVNG SAVNVVNLLLGAEK IRSGEMTIEELERAM	91	GAGGGCATCAAGAGCAACATCAGCCTGCTGAAGGACGAGCTGAG GGGCCAGATCAGCCACATCAGCCACGAGTACCTGAGCCTGATCGA CCTGGCCTTCGACAGCAAGCAGAACAGGCTGTTCGAGATGAAGGT GCTGGAGCTGCTGGTGAACGAGTACGGCTTCAAGGGCAGGCACCT GGGCGGCAGCAGGAAGCCCGACGGCATCGTGTACAGCACCACCC TGGAGGACAACTTCGGCATCATCGTGGACACCAAGGCCTACAGCG AGGGCTACAGCCTGCCCATCAGCCAGGCCGACGAGATGGAGAGG TACGTGAGGGAGAACAGCAACAGGGACGAGGAGGTGAACCCCAA CAAGTGGTGGGAGAACTTCAGCGAGGAGGTGAAGAAGTACTACT TCGTGTTCATCAGCGGCAGCTTCAAGGGCAAGTTCGAGGAGCAGC TGAGGAGGCTGAGCATGACCACCGGCGTGAACGGCAGCGCCGTG AACGTGGTGAACCTGCTGCTGGGCGCCGAGAAGATCAGGAGCGG CGAGATGACCATCGAGGAGCTGGAGAGGGCCATGTTCAACAACA

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
	FNNSEFI		GCGAGTTCATC
11	ISKTNVLELKDKVR DKLKYVDNRYLALI DLAYDGTANRDFEI QTIDLLINELKFKGV RLGESRKPDGIISYDI NGVIIDNKAYSSGY NLPINQADEMIRYIE ENQTRDKKINPNKW WESFDDKVKDFNYL FVSSFFKGNFKNNL KHIANRTGVNGGVI NVENLLYFAEELKS GRLSYVDLFKMYDN DEINI	92	ATCAGCAAGACCAACGTGCTGGAGCTGAAGGACAAGGTGAGGGA CAAGCTGAAGTACGTGGACAACAGGTACCTGGCCCTGATCGACCT GGCCTACGACGGCACCGCCAACAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGAAGTTCAAGGGCGTGAGGCTGG GCGAGAGCAGGAAGCCCGACGGCATCATCAGCTACGACATCAAC GGCGTGATCATCGACAACAAGGCCTACAGCAGCGGCTACAACCT GCCCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGAGA ACCAGACCAGGGACAAGAAGATCAACCCCAACAAGTGGTGGGAG AGCTTCGACGACAAGGTGAAGGACTTCAACTACCTGTTCGTGAGC AGCTTCTTCAAGGGCAACTTCAAGAACAACCTGAAGCACATCGCC AACAGGACCGGCGTGAACGGCGGCGTGATCAACGTGGAGAACCT GCTGTACTTCGCCGAGGAGCTGAAGAGCGGCAGGCTGAGCTACGT GGACCTGTTCAAGATGTACGACAACGACGAGATCAACATC
12	ISKTNVLELKDKVR DKLKYVDHRYLALI DLAYDGTANRDFEI QTIDLLINELKFKGV RLGESRKPDGIISYDI NGVIIDNKAYSTGY NLPINQADEMIRYIE ENQTRDKKINSNKW WESFDDKVKNFNYL FVSSFFKGNFKNNL KHIANRTGVNGGAI NVENLLYFAEELKA GRLSYVDSFTMYDN DEIYV	93	ATCAGCAAGACCAACGTGCTGGAGCTGAAGGACAAGGTGAGGGA CAAGCTGAAGTACGTGGACCACAGGTACCTGGCCCTGATCGACCT GGCCTACGACGGCACCGCCAACAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGAAGTTCAAGGGCGTGAGGCTGG GCGAGAGCAGGAAGCCCGACGGCATCATCAGCTACGACATCAAC GGCGTGATCATCGACAACAAGGCCTACAGCACCGGCTACAACCTG CCCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGAGAA CCAGACCAGGGACAAGAAGATCAACAGCAACAAGTGGTGGGAGA GCTTCGACGACAAGGTGAAGAACTTCAACTACCTGTTCGTGAGCA GCTTCTTCAAGGGCAACTTCAAGAACAACCTGAAGCACATCGCCA ACAGGACCGGCGTGAACGGCGGCGCCATCAACGTGGAGAACCTG CTGTACTTCGCCGAGGAGCTGAAGGCCGGCAGGCTGAGCTACGTG GACAGCTTCACCATGTACGACAACGACGAGATCTACGTG
13	KAEKSEFLIEKDKLR EKLDTLPHDYLSMV DLAYDSKQNRLFEM KTIELLINECNYKGL HLGGTRKPDGIVYT NNEVENYGIIIDTKA YSKGYNLPISQVDE MTRYVEENNKREK KRNPNEWWNNFDS NVKKFYFSFISGKFV GNIEEKLQRITLFTEI YGNAITVTTLLYIAN EIKANRMKKSDIME YFNDKV	94	AAGGCCGAGAAGAGCGAGTTCCTGATCGAGAAGGACAAGCTGAG GGAGAAGCTGGACACCCTGCCCCACGACTACCTGAGCATGGTGG ACCTGGCCTACGACAGCAAGCAGAACAGGCTGTTCGAGATGAAG ACCATCGAGCTGCTGATCAACGAGTGCAACTACAAGGGCCTGCAC CTGGGCGGCACCAGGAAGCCCGACGGCATCGTGTACACCAACAA CGAGGTGGAGAACTACGGCATCATCATCGACACCAAGGCCTACA GCAAGGGCTACAACCTGCCCATCAGCCAGGTGGACGAGATGACC AGGTACGTGGAGGAGAACAACAAGAGGGAGAAGAAGAGGAACC CCAACGAGTGGTGGAACAACTTCGACAGCAACGTGAAGAAGTTC TACTTCAGCTTCATCAGCGGCAAGTTCGTGGGCAACATCGAGGAG AAGCTGCAGAGGATCACCCTGTTCACCGAGATCTACGGCAACGCC ATCACCGTGACCACCCTGCTGTACATCGCCAACGAGATCAAGGCC AACAGGATGAAGAAGAGCGACATCATGGAGTACTTCAACGACAA GGTG

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
14	ISKTNVLELKDKVR DKLKYVDHRYLALI DLAYDGTANRDFEI QTIDLLINELKFKGV RLGESRKPDGIISYNI NGVIIDNKAYSTGY NLPINQADEMIRYIE ENQTRDEKINSNKW WESFDDEVKDFNYL FVSSFFKGNFKNNL KHIANRTGVNGGAI NVENLLYFAEELKA GRLSYVDSFTMYDN DEIYV	95	ATCAGCAAGACCAACGTGCTGGAGCTGAAGGACAAGGTGAGGGA CAAGCTGAAGTACGTGGACCACAGGTACCTGGCCCTGATCGACCT GGCCTACGACGGCACCGCCAACAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGAAGTTCAAGGGCGTGAGGCTGG GCGAGAGCAGGAAGCCCGACGGCATCATCAGCTACAACATCAAC GGCGTGATCATCGACAACAAGGCCTACAGCACCGGCTACAACCTG CCCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGAGAA CCAGACCAGGGACGAGAAGATCAACAGCAACAAGTGGTGGGAGA GCTTCGACGACGAGGTGAAGGACTTCAACTACCTGTTCGTGAGCA GCTTCTTCAAGGGCAACTTCAAGAACAACCTGAAGCACATCGCCA ACAGGACCGGCGTGAACGGCGGCGCCATCAACGTGGAGAACCTG CTGTACTTCGCCGAGGAGCTGAAGGCCGGCAGGCTGAGCTACGTG GACAGCTTCACCATGTACGACAACGACGAGATCTACGTG
15	ISKTNILELKDKVRD KLKYVDHRYLALID LAYDGTANRDFEIQ TIDLLINELKFKGVR LGESRKPDGIISYNIN GVIIDNKAYSTGYNL PINQADEMIRYIEEN QTRDEKINSNKWWE SFDEKVKDFNYLFV SSFFKGNFKNNLKHI ANRTGVNGGAINVE NLLYFAEELKAGRIS YLDSFKMYNNDEIY L	96	ATCAGCAAGACCAACATCCTGGAGCTGAAGGACAAGGTGAGGGA CAAGCTGAAGTACGTGGACCACAGGTACCTGGCCCTGATCGACCT GGCCTACGACGGCACCGCCAACAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGAAGTTCAAGGGCGTGAGGCTGG GCGAGAGCAGGAAGCCCGACGGCATCATCAGCTACAACATCAAC GGCGTGATCATCGACAACAAGGCCTACAGCACCGGCTACAACCTG CCCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGAGAA CCAGACCAGGGACGAGAAGATCAACAGCAACAAGTGGTGGGAGA GCTTCGACGAGAAGGTGAAGGACTTCAACTACCTGTTCGTGAGCA GCTTCTTCAAGGGCAACTTCAAGAACAACCTGAAGCACATCGCCA ACAGGACCGGCGTGAACGGCGGCGCCATCAACGTGGAGAACCTG CTGTACTTCGCCGAGGAGCTGAAGGCCGGCAGGATCAGCTACCTG GACAGCTTCAAGATGTACAACAACGACGAGATCTACCTG
16	ISKTNVLELKDKVR DKLKYVDHRYLALI DLAYDGTANRDFEI QTIDLLINELKFKGV RLGESRKPDGIISYNI NGVIIDNKAYSTGY NLPINQADEMIRYIE ENQTRDEKINSNKW WESFDDKVKDFNYL FVSSFFKGNFKNNL KHIANRTGVSGGAI NVENLLYFAEELKA GRLSYVDSFKMYDN DEIYV	97	ATCAGCAAGACCAACGTGCTGGAGCTGAAGGACAAGGTGAGGGA CAAGCTGAAGTACGTGGACCACAGGTACCTGGCCCTGATCGACCT GGCCTACGACGGCACCGCCAACAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGAAGTTCAAGGGCGTGAGGCTGG GCGAGAGCAGGAAGCCCGACGGCATCATCAGCTACAACATCAAC GGCGTGATCATCGACAACAAGGCCTACAGCACCGGCTACAACCTG CCCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGAGAA CCAGACCAGGGACGAGAAGATCAACAGCAACAAGTGGTGGGAGA GCTTCGACGACAAGGTGAAGGACTTCAACTACCTGTTCGTGAGCA GCTTCTTCAAGGGCAACTTCAAGAACAACCTGAAGCACATCGCCA ACAGGACCGGCGTGAGCGGCGGCGCCATCAACGTGGAGAACCTG CTGTACTTCGCCGAGGAGCTGAAGGCCGGCAGGCTGAGCTACGTG GACAGCTTCAAGATGTACGACAACGACGAGATCTACGTG
17	ISKTNVLELKDKVR NKLKYVDHRYLALI DLAYDGTANRDFEI QTIDLLINELKFKGV RLGESRKPDGIISYDI NGVIIDNKSYSTGYN LPINQADEMIRYIEE NQTRDEKINSNKW WESFDEKVKDFNYL FVSSFFKGNFKNNL KHIANRTGVNGGAI NVENLLYFAEELKS GRLSYVDSFTMYDN	98	ATCAGCAAGACCAACGTGCTGGAGCTGAAGGACAAGGTGAGGAA CAAGCTGAAGTACGTGGACCACAGGTACCTGGCCCTGATCGACCT GGCCTACGACGGCACCGCCAACAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGAAGTTCAAGGGCGTGAGGCTGG GCGAGAGCAGGAAGCCCGACGGCATCATCAGCTACGACATCAAC GGCGTGATCATCGACAACAAGAGCTACAGCACCGGCTACAACCT GCCCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGAGA ACCAGACCAGGGACGAGAAGATCAACAGCAACAAGTGGTGGGAG AGCTTCGACGAGAAGGTGAAGGACTTCAACTACCTGTTCGTGAGC AGCTTCTTCAAGGGCAACTTCAAGAACAACCTGAAGCACATCGCC AACAGGACCGGCGTGAACGGCGGCGCCATCAACGTGGAGAACCT GCTGTACTTCGCCGAGGAGCTGAAGAGCGGCAGGCTGAGCTACGT GGACAGCTTCACCATGTACGACAACGACGAGATCTACGTG

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
	DEIYV		
18	ISKTNVLELKDKVR DKLKYVDHRYLSLI DLAYDGNANRDFEI QTIDLLINELNFKGV RLGESRKPDGIISYNI NGVIIDNKAYSTGY NLPINQADEMIRYIE ENQTRDEKINSNKW WESFDDKVKDFNYL FVSSFFKGNFKNNL KHIANRTGVSGGAI NVENLLYFAEELKA GRLSYADSFTMYDN DEIYV	99	ATCAGCAAGACCAACGTGCTGGAGCTGAAGGACAAGGTGAGGGA CAAGCTGAAGTACGTGGACCACAGGTACCTGAGCCTGATCGACCT GGCCTACGACGGCAACGCCAACAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGAACTTCAAGGGCGTGAGGCTGG GCGAGAGCAGGAAGCCCGACGGCATCATCAGCTACAACATCAAC GGCGTGATCATCGACAACAAGGCCTACAGCACCGGCTACAACCTG CCCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGAGAA CCAGACCAGGGACGAGAAGATCAACAGCAACAAGTGGTGGGAGA GCTTCGACGACAAGGTGAAGGACTTCAACTACCTGTTCGTGAGCA GCTTCTTCAAGGGCAACTTCAAGAACAACCTGAAGCACATCGCCA ACAGGACCGGCGTGAGCGGCGGCGCCATCAACGTGGAGAACCTG CTGTACTTCGCCGAGGAGCTGAAGGCCGGCAGGCTGAGCTACGCC GACAGCTTCACCATGTACGACAACGACGAGATCTACGTG
19	IAKTNVLGLKDKVR DRLKYVDHRYLALI DLAYDGTANRDFEI QTIDLLINELKFKGV RLGESRKPDGIISYN VNGVIIDNKAYSKG YNLPINQADEMIRYI EENQTRDEKINANK WWESFDDKVEEFSY LFVSSFFKGNFKNNL KHIANRTGVNGGAI NVENLLYFAEELKS GRLSYMDSFSLYDN DEICV	100	ATCGCCAAGACCAACGTGCTGGGCCTGAAGGACAAGGTGAGGGA CAGGCTGAAGTACGTGGACCACAGGTACCTGGCCCTGATCGACCT GGCCTACGACGGCACCGCCAACAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGAAGTTCAAGGGCGTGAGGCTGG GCGAGAGCAGGAAGCCCGACGGCATCATCAGCTACAACGTGAAC GGCGTGATCATCGACAACAAGGCCTACAGCAAGGGCTACAACCT GCCCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGAGA ACCAGACCAGGGACGAGAAGATCAACGCCAACAAGTGGTGGGAG AGCTTCGACGACAAGGTGGAGGAGTTCAGCTACCTGTTCGTGAGC AGCTTCTTCAAGGGCAACTTCAAGAACAACCTGAAGCACATCGCC AACAGGACCGGCGTGAACGGCGGCGCCATCAACGTGGAGAACCT GCTGTACTTCGCCGAGGAGCTGAAGAGCGGCAGGCTGAGCTACAT GGACAGCTTCAGCCTGTACGACAACGACGAGATCTGCGTG
20	ELKDEQSEKRKAKF LKETKLPMKYIELLD IAYDGKRNRDFEIVT MELFREVYRLNSKL LGGGRKPDGLIYTD DFGVIVDTKAYGEG YSKSINQADEMIRYI EDNKRRDEKRNPIK WWESFPSSISQNNFY FLWVSSKFVGKFQE QLAYTANETQTKGG AINVEQILIGADLIM QKMLDINTIPSFFEN QEIIF	101	GAGCTGAAGGACGAGCAGAGCGAGAAGAGGAAGGCCAAGTTCCT GAAGGAGACCAAGCTGCCCATGAAGTACATCGAGCTGCTGGACA TCGCCTACGACGGCAAGAGGAACAGGGACTTCGAGATCGTGACC ATGGAGCTGTTCAGGGAGGTGTACAGGCTGAACAGCAAGCTGCT GGGCGGCGGCAGGAAGCCCGACGGCCTGATCTACACCGACGACT TCGGCGTGATCGTGGACACCAAGGCCTACGGCGAGGGCTACAGC AAGAGCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGA CAACAAGAGGAGGGACGAGAAGAGGAACCCCATCAAGTGGTGGG AGAGCTTCCCCAGCAGCATCAGCCAGAACAACTTCTACTTCCTGT GGGTGAGCAGCAAGTTCGTGGGCAAGTTCCAGGAGCAGCTGGCC TACACCGCCAACGAGACCCAGACCAAGGGCGGCGCCATCAACGT GGAGCAGATCCTGATCGGCGCCGACCTGATCATGCAGAAGATGCT GGACATCAACACCATCCCCAGCTTCTTCGAGAACCAGGAGATCAT CTTC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
21	IFKTNVLELKDSIRE KLDYIDHRYLSLVD LAYDSKANRDFEIQ TIDLLINELDFKGLR LGESRKPDGIISYDIN GVIIDNKAYSKGYN LPINQADEMIRYIQE NQSRNEKINPNKWW ENFEDKVIKFNYLFI SSLFVGGFKKNLQHI ANRTGVNGGAIDVE NLLYFAEEIKSGRLT YKDSFSRYINDEIKM	102	ATCTTCAAGACCAACGTGCTGGAGCTGAAGGACAGCATCAGGGA GAAGCTGGACTACATCGACCACAGGTACCTGAGCCTGGTGGACCT GGCCTACGACAGCAAGGCCAACAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGGACTTCAAGGGCCTGAGGCTGG GCGAGAGCAGGAAGCCCGACGGCATCATCAGCTACGACATCAAC GGCGTGATCATCGACAACAAGGCCTACAGCAAGGGCTACAACCT GCCCATCAACCAGGCCGACGAGATGATCAGGTACATCCAGGAGA ACCAGAGCAGGAACGAGAAGATCAACCCCAACAAGTGGTGGGAG AACTTCGAGGACAAGGTGATCAAGTTCAACTACCTGTTCATCAGC AGCCTGTTCGTGGGCGGCTTCAAGAAGAACCTGCAGCACATCGCC AACAGGACCGGCGTGAACGGCGGCGCCATCGACGTGGAGAACCT GCTGTACTTCGCCGAGGAGATCAAGAGCGGCAGGCTGACCTACA AGGACAGCTTCAGCAGGTACATCAACGACGAGATCAAGATG
22	LPVKSEVSVFKDYL RTHLTHVDHRYLIL VDLGFDGSSDRDYE MKTAELFTAELGFM GARLGDTRKPDVCV YHGANGLIIDNKAY GKGYSLPIKQADEIY RYIEENKERDARLNP NQWWKVFDESVTH FRFAFISGSFTGGFK DRIELISMRSGICGA AVNSVNLLLMAEEL KSGRLDYEEWFQYF DCNDEISF	103	CTGCCCGTGAAGAGCGAGGTGAGCGTGTTCAAGGACTACCTGAG GACCCACCTGACCCACGTGGACCACAGGTACCTGATCCTGGTGGA CCTGGGCTTCGACGGCAGCAGCGACAGGGACTACGAGATGAAGA CCGCCGAGCTGTTCACCGCCGAGCTGGGCTTCATGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACCACGGCGCC AACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATCTACAGGTACATCGAGG AGAACAAGGAGAGGGACGCCAGGCTGAACCCCAACCAGTGGTGG AAGGTGTTCGACGAGAGCGTGACCCACTTCAGGTTCGCCTTCATC AGCGGCAGCTTCACCGGCGGCTTCAAGGACAGGATCGAGCTGATC AGCATGAGGAGCGGCATCTGCGGCGCCGCCGTGAACAGCGTGAA CCTGCTGCTGATGGCCGAGGAGCTGAAGAGCGGCAGGCTGGACT ACGAGGAGTGGTTCCAGTACTTCGACTGCAACGACGAGATCAGCT TC
23	ISVKSDMAVVKDSV RERLAHVSHEYLILI DLGFDGTSDRDYEI QTAELFTRELDFLGG RLGDTRKPDVCIYY GKDGMIIDNKAYGK GYSLPIKQADEMYR YLEENKERNEKINPN RWWKVFDEGVTDY RFAFVSGSFTGGFKD RLENIHMRSGLCGG AIDSVTLLLLAEELK AGRMEYSEFFRLFD CNDEVTF	104	ATCAGCGTGAAGAGCGACATGGCCGTGGTGAAGGACAGCGTGAG GGAGAGGCTGGCCCACGTGAGCCACGAGTACCTGATCCTGATCGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCGAGCTGTTCACCAGGGAGCTGGACTTCCTGGGCGGCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCATCTACTACGGCAAG GACGGCATGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACCTGGAGG AGAACAAGGAGAGGAACGAGAAGATCAACCCCAACAGGTGGTGG AAGGTGTTCGACGAGGGCGTGACCGACTACAGGTTCGCCTTCGTG AGCGGCAGCTTCACCGGCGGCTTCAAGGACAGGCTGGAGAACAT CCACATGAGGAGCGGCCTGTGCGGCGGCGCCATCGACAGCGTGA CCCTGCTGCTGCTGGCCGAGGAGCTGAAGGCCGGCAGGATGGAG TACAGCGAGTTCTTCAGGCTGTTCGACTGCAACGACGAGGTGACC TTC
24	ELKDKAADAVKAK FLKLTGLSMKYIELL DIAYDSSRNRDFEIL TADLFKNVYGLDA MHLGGGRKPDAIAQ TSHFGIIIDTKAYGN GYSKSISQEDEMVR YIEDNQQRSITRNSV EWWKNFNSSIPSTAF YFLWVSSKFVGKFD DQLLATYNRTNTCG GALNVEQLLIGAYK VKAGLLGIGQIPSYF KNKEIAW	105	GAGCTGAAGGACAAGGCCGCCGACGCCGTGAAGGCCAAGTTCCT GAAGCTGACCGGCCTGAGCATGAAGTACATCGAGCTGCTGGACAT CGCCTACGACAGCAGCAGGAACAGGGACTTCGAGATCCTGACCG CCGACCTGTTCAAGAACGTGTACGGCCTGGACGCCATGCACCTGG GCGGCGGCAGGAAGCCCGACGCCATCGCCCAGACCAGCCACTTC GGCATCATCATCGACACCAAGGCCTACGGCAACGGCTACAGCAA GAGCATCAGCCAGGAGGACGAGATGGTGAGGTACATCGAGGACA ACCAGCAGAGGAGCATCACCAGGAACAGCGTGGAGTGGTGGAAG AACTTCAACAGCAGCATCCCCAGCACCGCCTTCTACTTCCTGTGG GTGAGCAGCAAGTTCGTGGGCAAGTTCGACGACCAGCTGCTGGCC ACCTACAACAGGACCAACACCTGCGGCGGCGCCCTGAACGTGGA GCAGCTGCTGATCGGCGCCTACAAGGTGAAGGCCGGCCTGCTGGG CATCGGCCAGATCCCCAGCTACTTCAAGAACAAGGAGATCGCCTG G

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
25	ISVKSDMAVVKDSV RERLAHVSHEYLLLI DLGFDGTSDRDYEI QTAELLTRELDFLG GRLGDTRKPDVCIY YGKDGMIIDNKAYG KGYSLPIKQADEMY RYLEENKERNEKINP NRWWKVFDEGVTD YRFAFVSGSFTGGFK DRLENIHMRSGLCG GAIDSVTLLLLAEEL KAGRMEYSEFFRLF DCNDEVTF	106	ATCAGCGTGAAGAGCGACATGGCCGTGGTGAAGGACAGCGTGAG GGAGAGGCTGGCCCACGTGAGCCACGAGTACCTGCTGCTGATCGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCGAGCTGCTGACCAGGGAGCTGGACTTCCTGGGCGGCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCATCTACTACGGCAAG GACGGCATGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACCTGGAGG AGAACAAGGAGAGGAACGAGAAGATCAACCCCAACAGGTGGTGG AAGGTGTTCGACGAGGGCGTGACCGACTACAGGTTCGCCTTCGTG AGCGGCAGCTTCACCGGCGGCTTCAAGGACAGGCTGGAGAACAT CCACATGAGGAGCGGCCTGTGCGGCGGCGCCATCGACAGCGTGA CCCTGCTGCTGCTGGCCGAGGAGCTGAAGGCCGGCAGGATGGAG TACAGCGAGTTCTTCAGGCTGTTCGACTGCAACGACGAGGTGACC TTC
26	ELKDEQAEKRKAKF LKETNLPMKYIELLD IAYDGKRNRDFEIVT MELFRNVYRLHSKL LGGGRKPDGLLYQD RFGVIVDTKAYGKG YSKSINQADEMIRYI EDNKRRDENRNPIK WWEAFPDTIPQEEF YFMWVSSKFIGKFQ EQLDYTSNETQIKG AALNVEQLLLGADL VLKGQLHISDLPSYF QNKEIEF	107	GAGCTGAAGGACGAGCAGGCCGAGAAGAGGAAGGCCAAGTTCCT GAAGGAGACCAACCTGCCCATGAAGTACATCGAGCTGCTGGACA TCGCCTACGACGGCAAGAGGAACAGGGACTTCGAGATCGTGACC ATGGAGCTGTTCAGGAACGTGTACAGGCTGCACAGCAAGCTGCTG GGCGGCGGCAGGAAGCCCGACGGCCTGCTGTACCAGGACAGGTT CGGCGTGATCGTGGACACCAAGGCCTACGGCAAGGGCTACAGCA AGAGCATCAACCAGGCCGACGAGATGATCAGGTACATCGAGGAC AACAAGAGGAGGGACGAGAACAGGAACCCCATCAAGTGGTGGGA GGCCTTCCCCGACACCATCCCCCAGGAGGAGTTCTACTTCATGTG GGTGAGCAGCAAGTTCATCGGCAAGTTCCAGGAGCAGCTGGACT ACACCAGCAACGAGACCCAGATCAAGGGCGCCGCCCTGAACGTG GAGCAGCTGCTGCTGGGCGCCGACCTGGTGCTGAAGGGCCAGCTG CACATCAGCGACCTGCCCAGCTACTTCCAGAACAAGGAGATCGAG TTC
27	RNLDNVERDNRKAE FLAKTSLPPRFIELLS IAYESKSNRDFEMIT AELFKDVYGLGAVH LGNAKKPDALAFND DFGIIIDTKAYSNGY SKNINQEDEMVRYIE DNQIRSPDRNNNEW WLSFPPSIPENDFHF LWVSSYFTGRFEEQ LQETSARTGGTTGG ALDVEQLLIGGSLIQ EGSLAPHEVPAYMQ NRVIHF	108	AGGAACCTGGACAACGTGGAGAGGGACAACAGGAAGGCCGAGTT CCTGGCCAAGACCAGCCTGCCCCCCAGGTTCATCGAGCTGCTGAG CATCGCCTACGAGAGCAAGAGCAACAGGGACTTCGAGATGATCA CCGCCGAGCTGTTCAAGGACGTGTACGGCCTGGGCGCCGTGCACC TGGGCAACGCCAAGAAGCCCGACGCCCTGGCCTTCAACGACGACT TCGGCATCATCATCGACACCAAGGCCTACAGCAACGGCTACAGCA AGAACATCAACCAGGAGGACGAGATGGTGAGGTACATCGAGGAC AACCAGATCAGGAGCCCCGACAGGAACAACAACGAGTGGTGGCT GAGCTTCCCCCCCAGCATCCCCGAGAACGACTTCCACTTCCTGTG GGTGAGCAGCTACTTCACCGGCAGGTTCGAGGAGCAGCTGCAGG AGACCAGCGCCAGGACCGGCGGCACCACCGGCGGCGCCCTGGAC GTGGAGCAGCTGCTGATCGGCGGCAGCCTGATCCAGGAGGGCAG CCTGGCCCCCCACGAGGTGCCCGCCTACATGCAGAACAGGGTGAT CCACTTC
28	SPVKSEVSVFKDYL RTHLTHVDHRYLIL VDLGFDGSSDRDYE MKTAELFTAELGFM GARLGDTRKPDVCV YHGAHGLIIDNKAY GKGYSLPIKQADEIY RYIEENKERAVRLNP NQWWKVFDESVAH FRFAFISGSFTGGFK DRIELISMRSGICGA AVNSVNLLLMAEEL KSGRLNYEEWFQYF	109	AGCCCCGTGAAGAGCGAGGTGAGCGTGTTCAAGGACTACCTGAG GACCCACCTGACCCACGTGGACCACAGGTACCTGATCCTGGTGGA CCTGGGCTTCGACGGCAGCAGCGACAGGGACTACGAGATGAAGA CCGCCGAGCTGTTCACCGCCGAGCTGGGCTTCATGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACCACGGCGCC CACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATCTACAGGTACATCGAGG AGAACAAGGAGAGGGCCGTGAGGCTGAACCCCAACCAGTGGTGG AAGGTGTTCGACGAGAGCGTGGCCCACTTCAGGTTCGCCTTCATC AGCGGCAGCTTCACCGGCGGCTTCAAGGACAGGATCGAGCTGATC AGCATGAGGAGCGGCATCTGCGGCGCCGCCGTGAACAGCGTGAA CCTGCTGCTGATGGCCGAGGAGCTGAAGAGCGGCAGGCTGAACT ACGAGGAGTGGTTCCAGTACTTCGACTGCAACGACGAGATCAGCC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
	DCNDEISL		TG
29	TLVDIEKERKKAYFL KETSLSPRYIELLEIA FDPKRNRDFEVITAE LLKAGYGLKAKVLG GGRRPDGIAYTKDY GLIVDTKAYSNGYG KNIGQADEMIRYIED NQKRDNKRNPIEW WREFEVQIPANSYY YLWVSGRFTGRFDE QLVYTSSQTNTRGG ALEVEQLLWGADA VMKGKLNVSDLPK YMNNSIIKL	110	ACCCTGGTGGACATCGAGAAGGAGAGGAAGAAGGCCTACTTCCT GAAGGAGACCAGCCTGAGCCCCAGGTACATCGAGCTGCTGGAGA TCGCCTTCGACCCCAAGAGGAACAGGGACTTCGAGGTGATCACCG CCGAGCTGCTGAAGGCCGGCTACGGCCTGAAGGCCAAGGTGCTG GGCGGCGGCAGGAGGCCCGACGGCATCGCCTACACCAAGGACTA CGGCCTGATCGTGGACACCAAGGCCTACAGCAACGGCTACGGCA AGAACATCGGCCAGGCCGACGAGATGATCAGGTACATCGAGGAC AACCAGAAGAGGGACAACAAGAGGAACCCCATCGAGTGGTGGAG GGAGTTCGAGGTGCAGATCCCCGCCAACAGCTACTACTACCTGTG GGTGAGCGGCAGGTTCACCGGCAGGTTCGACGAGCAGCTGGTGT ACACCAGCAGCCAGACCAACACCAGGGGCGGCGCCCTGGAGGTG GAGCAGCTGCTGTGGGGCGCCGACGCCGTGATGAAGGGCAAGCT GAACGTGAGCGACCTGCCCAAGTACATGAACAACAGCATCATCA AGCTG
30	ELRDKVIEEQKAIFL QKTKLPLSYIELLEIA RDGKRSRDFELITIE LFKNIYKINARILGG ARKPDGVLYMPEFG VIVDTKAYADGYSK SIAQADEMIRYIEDN KRRDPSRNSTKWWE HFPTSIPANNFYFLW VSSVFVNKFHEQLS YTAQETQTVGAALS VEQLLLGADSVLKG NLTTEKFIDSFKNQE IVF	111	GAGCTGAGGGACAAGGTGATCGAGGAGCAGAAGGCCATCTTCCT GCAGAAGACCAAGCTGCCCCTGAGCTACATCGAGCTGCTGGAGAT CGCCAGGGACGGCAAGAGGAGCAGGGACTTCGAGCTGATCACCA TCGAGCTGTTCAAGAACATCTACAAGATCAACGCCAGGATCCTGG GCGGCGCCAGGAAGCCCGACGGCGTGCTGTACATGCCCGAGTTCG GCGTGATCGTGGACACCAAGGCCTACGCCGACGGCTACAGCAAG AGCATCGCCCAGGCCGACGAGATGATCAGGTACATCGAGGACAA CAAGAGGAGGGACCCCAGCAGGAACAGCACCAAGTGGTGGGAGC ACTTCCCCACCAGCATCCCCGCCAACAACTTCTACTTCCTGTGGGT GAGCAGCGTGTTCGTGAACAAGTTCCACGAGCAGCTGAGCTACAC CGCCCAGGAGACCCAGACCGTGGGCGCCGCCCTGAGCGTGGAGC AGCTGCTGCTGGGCGCCGACAGCGTGCTGAAGGGCAACCTGACC ACCGAGAAGTTCATCGACAGCTTCAAGAACCAGGAGATCGTGTTC
31	GATKSDLSLLKDDIR KKLNHINHKYLVLI DLGFDGTADRDYEL QTADLLTSELAFKG ARLGDSRKPDVCVY HDKNGLIIDNKAYG SGYSLPIKQADEML RYIEENQKRDKALN PNEWWTIFDDAVSK FNFAFVSGEFTGGFK DRLENISRRSYTNGA AINSVNLLLLAEEIK SGRISYGDAFTKFEC NDEIII	112	GGCGCCACCAAGAGCGACCTGAGCCTGCTGAAGGACGACATCAG GAAGAAGCTGAACCACATCAACCACAAGTACCTGGTGCTGATCG ACCTGGGCTTCGACGGCACCGCCGACAGGGACTACGAGCTGCAG ACCGCCGACCTGCTGACCAGCGAGCTGGCCTTCAAGGGCGCCAGG CTGGGCGACAGCAGGAAGCCCGACGTGTGCGTGTACCACGACAA GAACGGCCTGATCATCGACAACAAGGCCTACGGCAGCGGCTACA GCCTGCCCATCAAGCAGGCCGACGAGATGCTGAGGTACATCGAG GAGAACCAGAAGAGGGACAAGGCCCTGAACCCCAACGAGTGGTG GACCATCTTCGACGACGCCGTGAGCAAGTTCAACTTCGCCTTCGT GAGCGGCGAGTTCACCGGCGGCTTCAAGGACAGGCTGGAGAACA TCAGCAGGAGGAGCTACACCAACGGCGCCGCCATCAACAGCGTG AACCTGCTGCTGCTGGCCGAGGAGATCAAGAGCGGCAGGATCAG CTACGGCGACGCCTTCACCAAGTTCGAGTGCAACGACGAGATCAT CATC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
32	ELRNAALDKQKVNF INKTGLPMKYIELLE IAFDGSRNRDFEMV TADLFKNVYGFNSIL LGGGRKPDGLIFTDR FGVIIDTKAYGNGYS KSIGQEDEMVRYIED NQLRDSNRNSVEW WKNFDEKIESENFYF MWISSKFIGQFSDQL QSTSDRTNTKGAAL NVEQLLLGAAAARD GKLDINSLPIYMNNK EILW	113	GAGCTGAGGAACGCCGCCCTGGACAAGCAGAAGGTGAACTTCAT CAACAAGACCGGCCTGCCCATGAAGTACATCGAGCTGCTGGAGAT CGCCTTCGACGGCAGCAGGAACAGGGACTTCGAGATGGTGACCG CCGACCTGTTCAAGAACGTGTACGGCTTCAACAGCATCCTGCTGG GCGGCGGCAGGAAGCCCGACGGCCTGATCTTCACCGACAGGTTCG GCGTGATCATCGACACCAAGGCCTACGGCAACGGCTACAGCAAG AGCATCGGCCAGGAGGACGAGATGGTGAGGTACATCGAGGACAA CCAGCTGAGGGACAGCAACAGGAACAGCGTGGAGTGGTGGAAGA ACTTCGACGAGAAGATCGAGAGCGAGAACTTCTACTTCATGTGGA TCAGCAGCAAGTTCATCGGCCAGTTCAGCGACCAGCTGCAGAGCA CCAGCGACAGGACCAACACCAAGGGCGCCGCCCTGAACGTGGAG CAGCTGCTGCTGGGCGCCGCCGCCGCCAGGGACGGCAAGCTGGA CATCAACAGCCTGCCCATCTACATGAACAACAAGGAGATCCTGTG G
33	ELKDEQSEKRKAYF LKETNLPLKYIELLD IAYDGKRNRDFEIVT MELFRNVYRLQSKL LGGVRKPDGLLYKH RFGIIVDTKAYGEGY SKSISQADEMIRYIE DNKRRDENRNSTK WWEHFPDCIPKQSF YFMWVSSKFVGKFQ EQLDYTANETKTNG AALNVEQLLWGAD LVAKGKLDISQLPSY FQNKEIEF	114	GAGCTGAAGGACGAGCAGAGCGAGAAGAGGAAGGCCTACTTCCT GAAGGAGACCAACCTGCCCCTGAAGTACATCGAGCTGCTGGACAT CGCCTACGACGGCAAGAGGAACAGGGACTTCGAGATCGTGACCA TGGAGCTGTTCAGGAACGTGTACAGGCTGCAGAGCAAGCTGCTGG GCGGCGTGAGGAAGCCCGACGGCCTGCTGTACAAGCACAGGTTC GGCATCATCGTGGACACCAAGGCCTACGGCGAGGGCTACAGCAA GAGCATCAGCCAGGCCGACGAGATGATCAGGTACATCGAGGACA ACAAGAGGAGGGACGAGAACAGGAACAGCACCAAGTGGTGGGA GCACTTCCCCGACTGCATCCCCAAGCAGAGCTTCTACTTCATGTG GGTGAGCAGCAAGTTCGTGGGCAAGTTCCAGGAGCAGCTGGACT ACACCGCCAACGAGACCAAGACCAACGGCGCCGCCCTGAACGTG GAGCAGCTGCTGTGGGGCGCCGACCTGGTGGCCAAGGGCAAGCT GGACATCAGCCAGCTGCCCAGCTACTTCCAGAACAAGGAGATCG AGTTC
34	HNNKFKNYLRENSE LSFKFIELIDIAYDGN RNRDMEIITAELLKE IYGLNVKLLGGGRK PDILAYTDDIGIIIDT KAYKDGYGKQINQ ADEMIRYIEDNQRR DLIRNPNEWWRYFP KSISKEKIYFMWISS YFKNNFYEQVQYTA QETKSIGAALNVRQ LLLCADAIQKEVLSL DTFLGSFRNEEINL	115	CACAACAACAAGTTCAAGAACTACCTGAGGGAGAACAGCGAGCT GAGCTTCAAGTTCATCGAGCTGATCGACATCGCCTACGACGGCAA CAGGAACAGGGACATGGAGATCATCACCGCCGAGCTGCTGAAGG AGATCTACGGCCTGAACGTGAAGCTGCTGGGCGGCGGCAGGAAG CCCGACATCCTGGCCTACACCGACGACATCGGCATCATCATCGAC ACCAAGGCCTACAAGGACGGCTACGGCAAGCAGATCAACCAGGC CGACGAGATGATCAGGTACATCGAGGACAACCAGAGGAGGGACC TGATCAGGAACCCCAACGAGTGGTGGAGGTACTTCCCCAAGAGC ATCAGCAAGGAGAAGATCTACTTCATGTGGATCAGCAGCTACTTC AAGAACAACTTCTACGAGCAGGTGCAGTACACCGCCCAGGAGAC CAAGAGCATCGGCGCCGCCCTGAACGTGAGGCAGCTGCTGCTGTG CGCCGACGCCATCCAGAAGGAGGTGCTGAGCCTGGACACCTTCCT GGGCAGCTTCAGGAACGAGGAGATCAACCTG
35	LPVKSEVSILKDYLR SHLTHIDHKYLILVD LGYDGTSDRDYEIQ TAQLLTAELSFLGGR LGDTRKPDVCIYYE DNGLIIDNKAYGKG YSLPMKQADEMYR YIEENKERSELLNPN CWWNIFDKDVKTFH FAFLSGEFTGGFRDR LNHISMRSGMRGAA VNSANLLIMAEKLK AGTMEYEEFFRLFD TNDEILF	116	CTGCCCGTGAAGAGCGAGGTGAGCATCCTGAAGGACTACCTGAG GAGCCACCTGACCCACATCGACCACAAGTACCTGATCCTGGTGGA CCTGGGCTACGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGCCGAGCTGAGCTTCCTGGGCGGCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCATCTACTACGAGGAC AACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATGAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAGCGAGCTGCTGAACCCCAACTGCTGGTGG AACATCTTCGACAAGGACGTGAAGACCTTCCACTTCGCCTTCCTG AGCGGCGAGTTCACCGGCGGCTTCAGGGACAGGCTGAACCACAT CAGCATGAGGAGCGGCATGAGGGGCGCCGCCGTGAACAGCGCCA ACCTGCTGATCATGGCCGAGAAGCTGAAGGCCGGCACCATGGAG TACGAGGAGTTCTTCAGGCTGTTCGACACCAACGACGAGATCCTG TTC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
36	LPVKSQVSILKDYLR SYLSHVDHKYLILLD LGFDGTSDRDYEIW TAQLLTAELSFLGGR LGDTRKPDVCIYYE DNGLIIDNKAYGKG YSLPIKQADEMYRYI EENKERSDLLNPNC WWNIFGEGVKTFRF AFLSGEFTGGFKDRL NHISMRSGIKGAAV NSANLLIMAEQLKS GTMSYEEFFQLFDY NDEIIF	117	CTGCCCGTGAAGAGCCAGGTGAGCATCCTGAAGGACTACCTGAG GAGCTACCTGAGCCACGTGGACCACAAGTACCTGATCCTGCTGGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCTGGA CCGCCCAGCTGCTGACCGCCGAGCTGAGCTTCCTGGGCGGCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCATCTACTACGAGGAC AACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAGCGACCTGCTGAACCCCAACTGCTGGTGG AACATCTTCGGCGAGGGCGTGAAGACCTTCAGGTTCGCCTTCCTG AGCGGCGAGTTCACCGGCGGCTTCAAGGACAGGCTGAACCACAT CAGCATGAGGAGCGGCATCAAGGGCGCCGCCGTGAACAGCGCCA ACCTGCTGATCATGGCCGAGCAGCTGAAGAGCGGCACCATGAGCT ACGAGGAGTTCTTCCAGCTGTTCGACTACAACGACGAGATCATCT TC
37	VSKTNILELKDNTRE KLVYLDHRYLSLFD LAYDDKASRDFEIQ TIDLLINELQFKGLR LGERRKPDGIISYGV NGVIIDNKAYSKGY NLPIRQADEMIRYIQ ENQSRDEKLNPNKW WENFEEETSKFNYL FISSKFISGFKKNLQY IADRTGVNGGAINV ENLLCFAEMLKSGK LEYNDFFNQYNNDE IIM	118	GTGAGCAAGACCAACATCCTGGAGCTGAAGGACAACACCAGGGA GAAGCTGGTGTACCTGGACCACAGGTACCTGAGCCTGTTCGACCT GGCCTACGACGACAAGGCCAGCAGGGACTTCGAGATCCAGACCA TCGACCTGCTGATCAACGAGCTGCAGTTCAAGGGCCTGAGGCTGG GCGAGAGGAGGAAGCCCGACGGCATCATCAGCTACGGCGTGAAC GGCGTGATCATCGACAACAAGGCCTACAGCAAGGGCTACAACCT GCCCATCAGGCAGGCCGACGAGATGATCAGGTACATCCAGGAGA ACCAGAGCAGGGACGAGAAGCTGAACCCCAACAAGTGGTGGGAG AACTTCGAGGAGGAGACCAGCAAGTTCAACTACCTGTTCATCAGC AGCAAGTTCATCAGCGGCTTCAAGAAGAACCTGCAGTACATCGCC GACAGGACCGGCGTGAACGGCGGCGCCATCAACGTGGAGAACCT GCTGTGCTTCGCCGAGATGCTGAAGAGCGGCAAGCTGGAGTACA ACGACTTCTTCAACCAGTACAACAACGACGAGATCATCATG
38	LPVKSQVSILKDYLR SCLSHVDHKYLILLD LGFDGTSDRDYEIQT AQLLTAELSFLGGRL GDTRKPDVCIYYED NGLIIDNKAYGKGY SLPIKQADEMYRYIE ENKERSELLNPNCW WNIFDEGVKTFRFA FLSGEFTGGFKDRLN HISMRSGIKGAAVNS ANLLIIAEQLKSGTM SYEEFFQLFDQNDEI TV	119	CTGCCCGTGAAGAGCCAGGTGAGCATCCTGAAGGACTACCTGAG GAGCTGCCTGAGCCACGTGGACCACAAGTACCTGATCCTGCTGGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGCCGAGCTGAGCTTCCTGGGCGGCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCATCTACTACGAGGAC AACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAGCGAGCTGCTGAACCCCAACTGCTGGTGG AACATCTTCGACGAGGGCGTGAAGACCTTCAGGTTCGCCTTCCTG AGCGGCGAGTTCACCGGCGGCTTCAAGGACAGGCTGAACCACAT CAGCATGAGGAGCGGCATCAAGGGCGCCGCCGTGAACAGCGCCA ACCTGCTGATCATCGCCGAGCAGCTGAAGAGCGGCACCATGAGCT ACGAGGAGTTCTTCCAGCTGTTCGACCAGAACGACGAGATCACCG TG
39	MSSKSEISVIKDNIR KRLNHINHKYLVLID LGFDGTADRDYELQ TADLLTSELSFKGAR LGDTRKPDVCVYHG TNGLIIDNKAYGKG YSLPIKQADEMLRYI EENQKRDKSLNPNE WWTIFDDAVSKFNF AFVSGEFTGGFKDR LENISRRSSVNGAAI NSVNLLLLAEEIKSG RMSYSDAFKNFDCN	120	ATGAGCAGCAAGAGCGAGATCAGCGTGATCAAGGACAACATCAG GAAGAGGCTGAACCACATCAACCACAAGTACCTGGTGCTGATCG ACCTGGGCTTCGACGGCACCGCCGACAGGGACTACGAGCTGCAG ACCGCCGACCTGCTGACCAGCGAGCTGAGCTTCAAGGGCGCCAG GCTGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACCACGGCA CCAACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTAC AGCCTGCCCATCAAGCAGGCCGACGAGATGCTGAGGTACATCGA GGAGAACCAGAAGAGGGACAAGAGCCTGAACCCCAACGAGTGGT GGACCATCTTCGACGACGCCGTGAGCAAGTTCAACTTCGCCTTCG TGAGCGGCGAGTTCACCGGCGGCTTCAAGGACAGGCTGGAGAAC ATCAGCAGGAGGAGCAGCGTGAACGGCGCCGCCATCAACAGCGT GAACCTGCTGCTGCTGGCCGAGGAGATCAAGAGCGGCAGGATGA GCTACAGCGACGCCTTCAAGAACTTCGACTGCAACAAGGAGATCA

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
	KEITI		CCATC
40	RNLDKVERDSRKAE FLAKTSLPPRFIELLS IAYESKSNRDFEMIT AEFFKDVYGLGAVH LGNARKPDALAFTD NFGIVIDTKAYSNGY SKNINQEDEMVRYIE DNQIRSPERNKNEW WLSFPPSIPENNFHF LWVSSYFTGYFEEQ LQETSDRAGGMTGG ALDIEQLLIGGSLVQ EGKLAPHDIPEYMQ NRVIHF	121	AGGAACCTGGACAAGGTGGAGAGGGACAGCAGGAAGGCCGAGTT CCTGGCCAAGACCAGCCTGCCCCCCAGGTTCATCGAGCTGCTGAG CATCGCCTACGAGAGCAAGAGCAACAGGGACTTCGAGATGATCA CCGCCGAGTTCTTCAAGGACGTGTACGGCCTGGGCGCCGTGCACC TGGGCAACGCCAGGAAGCCCGACGCCCTGGCCTTCACCGACAACT TCGGCATCGTGATCGACACCAAGGCCTACAGCAACGGCTACAGCA AGAACATCAACCAGGAGGACGAGATGGTGAGGTACATCGAGGAC AACCAGATCAGGAGCCCCGAGAGGAACAAGAACGAGTGGTGGCT GAGCTTCCCCCCCAGCATCCCCGAGAACAACTTCCACTTCCTGTG GGTGAGCAGCTACTTCACCGGCTACTTCGAGGAGCAGCTGCAGGA GACCAGCGACAGGGCCGGCGGCATGACCGGCGGCGCCCTGGACA TCGAGCAGCTGCTGATCGGCGGCAGCCTGGTGCAGGAGGGCAAG CTGGCCCCCCACGACATCCCCGAGTACATGCAGAACAGGGTGATC CACTTC
41	APVKSEVSLCKDILR SHLTHVDHKYLILL DLGFDGTSDRDYEI QTAQLLTAELDFKG ARLGDTRKPDVCVY YGEDGLILDNKAYG KGYSLPIKQADEMY RYIEENKERNERLNP NKWWEIFDKDVVR YHFAFVSGTFTGGF KERLDNIRMRSGICG AAVNSMNLLLMAE ELKSGRLGYKECFA LFDCNDEIAF	122	GCCCCCGTGAAGAGCGAGGTGAGCCTGTGCAAGGACATCCTGAG GAGCCACCTGACCCACGTGGACCACAAGTACCTGATCCTGCTGGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGCCGAGCTGGACTTCAAGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACTACGGCGAG GACGGCCTGATCCTGGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAACGAGAGGCTGAACCCCAACAAGTGGTGG GAGATCTTCGACAAGGACGTGGTGAGGTACCACTTCGCCTTCGTG AGCGGCACCTTCACCGGCGGCTTCAAGGAGAGGCTGGACAACAT CAGGATGAGGAGCGGCATCTGCGGCGCCGCCGTGAACAGCATGA ACCTGCTGCTGATGGCCGAGGAGCTGAAGAGCGGCAGGCTGGGC TACAAGGAGTGCTTCGCCCTGTTCGACTGCAACGACGAGATCGCC TTC
42	SCVKDEVNDIVDRV RVKLKNIDHKYLILI SLAYSDETERTKKN SDARDFEIQTAELFT KELGFNGIRLGESNK PDVLISFGANGTIIDN KSYKDGFNIPRVTSD QMIRYINENNQRTT QLNPNEWWKNFDSS VSNYTFLFVTSFLKG SFKNQIEYISNATNG TRGAAINVESLLYIS EDIKSGKIKQSDFYS EFKNDEIVY	123	AGCTGCGTGAAGGACGAGGTGAACGACATCGTGGACAGGGTGAG GGTGAAGCTGAAGAACATCGACCACAAGTACCTGATCCTGATCAG CCTGGCCTACAGCGACGAGACCGAGAGGACCAAGAAGAACAGCG ACGCCAGGGACTTCGAGATCCAGACCGCCGAGCTGTTCACCAAGG AGCTGGGCTTCAACGGCATCAGGCTGGGCGAGAGCAACAAGCCC GACGTGCTGATCAGCTTCGGCGCCAACGGCACCATCATCGACAAC AAGAGCTACAAGGACGGCTTCAACATCCCCAGGGTGACCAGCGA CCAGATGATCAGGTACATCAACGAGAACAACCAGAGGACCACCC AGCTGAACCCCAACGAGTGGTGGAAGAACTTCGACAGCAGCGTG AGCAACTACACCTTCCTGTTCGTGACCAGCTTCCTGAAGGGCAGC TTCAAGAACCAGATCGAGTACATCAGCAACGCCACCAACGGCAC CAGGGGCGCCGCCATCAACGTGGAGAGCCTGCTGTACATCAGCG AGGACATCAAGAGCGGCAAGATCAAGCAGAGCGACTTCTACAGC GAGTTCAAGAACGACGAGATCGTGTAC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
43	SQGDKAREQLKAKF LAKTNLLPRYVELL DIAYDSKRNRDFEM VTAELFNFAYLLPA VHLGGVRKPDALVA TKKFGIIVDTKAYAN GYSRNANQADEMA RYITENQKRDPKTNP NRWWDNFDARIPPN AYYFLWVSSFFTGQ FDDQLSYTAHRTNT HGGALNVEQLLIGA NMIQTGQLDRNKLP EYMQDKEITF	124	AGCCAGGGCGACAAGGCCAGGGAGCAGCTGAAGGCCAAGTTCCT GGCCAAGACCAACCTGCTGCCCAGGTACGTGGAGCTGCTGGACAT CGCCTACGACAGCAAGAGGAACAGGGACTTCGAGATGGTGACCG CCGAGCTGTTCAACTTCGCCTACCTGCTGCCCGCCGTGCACCTGG GCGGCGTGAGGAAGCCCGACGCCCTGGTGGCCACCAAGAAGTTC GGCATCATCGTGGACACCAAGGCCTACGCCAACGGCTACAGCAG GAACGCCAACCAGGCCGACGAGATGGCCAGGTACATCACCGAGA ACCAGAAGAGGGACCCCAAGACCAACCCCAACAGGTGGTGGGAC AACTTCGACGCCAGGATCCCCCCCAACGCCTACTACTTCCTGTGG GTGAGCAGCTTCTTCACCGGCCAGTTCGACGACCAGCTGAGCTAC ACCGCCCACAGGACCAACACCCACGGCGGCGCCCTGAACGTGGA GCAGCTGCTGATCGGCGCCAACATGATCCAGACCGGCCAGCTGGA CAGGAACAAGCTGCCCGAGTACATGCAGGACAAGGAGATCACCT TC
44	KVQKSNILDVIEKCR EKINNIPHEYLALIP MSFDENESTMFEIKT IELLTEHCKFDGLHC GGASKPDGLIYSED YGVIIDTKSYKDGFN IQTPERDKMKRYIEE NQNRNPQHNKTRW WDEFPHNISNFLFLF VSGKFGGNFKEQLRI LSEQTNNTLGGALSS YVLLNIAEQIAINKID HCDFKTRISCLDEVA	125	AAGGTGCAGAAGAGCAACATCCTGGACGTGATCGAGAAGTGCAG GGAGAAGATCAACAACATCCCCCACGAGTACCTGGCCCTGATCCC CATGAGCTTCGACGAGAACGAGAGCACCATGTTCGAGATCAAGA CCATCGAGCTGCTGACCGAGCACTGCAAGTTCGACGGCCTGCACT GCGGCGGCGCCAGCAAGCCCGACGGCCTGATCTACAGCGAGGAC TACGGCGTGATCATCGACACCAAGAGCTACAAGGACGGCTTCAAC ATCCAGACCCCCGAGAGGGACAAGATGAAGAGGTACATCGAGGA GAACCAGAACAGGAACCCCCAGCACAACAAGACCAGGTGGTGGG ACGAGTTCCCCCACAACATCAGCAACTTCCTGTTCCTGTTCGTGAG CGGCAAGTTCGGCGGCAACTTCAAGGAGCAGCTGAGGATCCTGA GCGAGCAGACCAACAACACCCTGGGCGGCGCCCTGAGCAGCTAC GTGCTGCTGAACATCGCCGAGCAGATCGCCATCAACAAGATCGAC CACTGCGACTTCAAGACCAGGATCAGCTGCCTGGACGAGGTGGCC
45	VPVKSEVSLCKDYL RSYLTHVDHKYLILL DLGFDGTSDRDYEI QTAQLLTAELDFKG ARLGDTRKPDVCVY YGEDGLIIDNKAYG KGYSLPIKQADEIYR YIEENKKRDEKLNP NKWWEIFDKGVVR YHFAFVSGAFTGGF KERLDNIRMRSGICG AAINSMNLLLMAEE LKSGRLGYEECFALF DCNDEITF	126	GTGCCCGTGAAGAGCGAGGTGAGCCTGTGCAAGGACTACCTGAG GAGCTACCTGACCCACGTGGACCACAAGTACCTGATCCTGCTGGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGCCGAGCTGGACTTCAAGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACTACGGCGAG GACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATCTACAGGTACATCGAGG AGAACAAGAAGAGGGACGAGAAGCTGAACCCCAACAAGTGGTGG GAGATCTTCGACAAGGGCGTGGTGAGGTACCACTTCGCCTTCGTG AGCGGCGCCTTCACCGGCGGCTTCAAGGAGAGGCTGGACAACAT CAGGATGAGGAGCGGCATCTGCGGCGCCGCCATCAACAGCATGA ACCTGCTGCTGATGGCCGAGGAGCTGAAGAGCGGCAGGCTGGGC TACGAGGAGTGCTTCGCCCTGTTCGACTGCAACGACGAGATCACC TTC
46	VPVKSEVSLCKDYL RSHLNHVDHRYLIL LDLGFDGTSDRDYEI QTAQLLTGELNFKG ARLGDTRKPDVCVY YGEDGLIIDNKAYG KGYSLPIKQADEMY RYIEENKERNEKLNP NKWWEIFDKDVIHY HFAFVSGAFTGGFK ERLENIRMRSGIYGA AVNSMNLLLMAEEL KSGRLDYKECFKLF DCNDEIVL	127	GTGCCCGTGAAGAGCGAGGTGAGCCTGTGCAAGGACTACCTGAG GAGCCACCTGAACCACGTGGACCACAGGTACCTGATCCTGCTGGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGGCGAGCTGAACTTCAAGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACTACGGCGAG GACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAACGAGAAGCTGAACCCCAACAAGTGGTGG GAGATCTTCGACAAGGACGTGATCCACTACCACTTCGCCTTCGTG AGCGGCGCCTTCACCGGCGGCTTCAAGGAGAGGCTGGAGAACAT CAGGATGAGGAGCGGCATCTACGGCGCCGCCGTGAACAGCATGA ACCTGCTGCTGATGGCCGAGGAGCTGAAGAGCGGCAGGCTGGAC TACAAGGAGTGCTTCAAGCTGTTCGACTGCAACGACGAGATCGTG CTG

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
47	VPVKSEVSLLKDYL RSHLVHVDHKYLVL LDLGFDGTSDRDYEI QTAQLLTGELNFKG ARLGDTRKPDVCVY YGEDGLIIDNKAYG KGYSLPIKQADEMY RYIEENKERNEKLNP NKWWEIFGNDVIHY HFAFVSGAFTGGFK ERLDNIRMRSGIYGA AVNSMNLLLLAEEL KSGRLGYKECFKLF DCNDEIVL	128	GTGCCCGTGAAGAGCGAGGTGAGCCTGCTGAAGGACTACCTGAG GAGCCACCTGGTGCACGTGGACCACAAGTACCTGGTGCTGCTGGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGGCGAGCTGAACTTCAAGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACTACGGCGAG GACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAACGAGAAGCTGAACCCCAACAAGTGGTGG GAGATCTTCGGCAACGACGTGATCCACTACCACTTCGCCTTCGTG AGCGGCGCCTTCACCGGCGGCTTCAAGGAGAGGCTGGACAACAT CAGGATGAGGAGCGGCATCTACGGCGCCGCCGTGAACAGCATGA ACCTGCTGCTGCTGGCCGAGGAGCTGAAGAGCGGCAGGCTGGGC TACAAGGAGTGCTTCAAGCTGTTCGACTGCAACGACGAGATCGTG CTG
48	ECVKDNVVDIKDRV RNKLIHLDHKYLALI DLAYSDAASRAKKN ADAREFEIQTADLFT KELSFNGQRLGDSR KPDVIISYGLDGTIV DNKSYKDGFNISRT CADEMSRYINENNL RQKSLNPNEWWKN FDSTITAYTFLFITSY LKGQFEDQLEYVSN ANGGIKGAAIGVESL LYLSEGIKAGRISHA DFYSNFNNKEMIY	129	GAGTGCGTGAAGGACAACGTGGTGGACATCAAGGACAGGGTGAG GAACAAGCTGATCCACCTGGACCACAAGTACCTGGCCCTGATCGA CCTGGCCTACAGCGACGCCGCCAGCAGGGCCAAGAAGAACGCCG ACGCCAGGGAGTTCGAGATCCAGACCGCCGACCTGTTCACCAAGG AGCTGAGCTTCAACGGCCAGAGGCTGGGCGACAGCAGGAAGCCC GACGTGATCATCAGCTACGGCCTGGACGGCACCATCGTGGACAAC AAGAGCTACAAGGACGGCTTCAACATCAGCAGGACCTGCGCCGA CGAGATGAGCAGGTACATCAACGAGAACAACCTGAGGCAGAAGA GCCTGAACCCCAACGAGTGGTGGAAGAACTTCGACAGCACCATC ACCGCCTACACCTTCCTGTTCATCACCAGCTACCTGAAGGGCCAG TTCGAGGACCAGCTGGAGTACGTGAGCAACGCCAACGGCGGCAT CAAGGGCGCCGCCATCGGCGTGGAGAGCCTGCTGTACCTGAGCG AGGGCATCAAGGCCGGCAGGATCAGCCACGCCGACTTCTACAGC AACTTCAACAACAAGGAGATGATCTAC
49	IAKSDFSIIKDNIRRK LQYVNHKYLLLIDL GFDSDSNRDYEIQTA ELLTTELAFKGARL GDTRKPDVCVYYGE NGLIIDNKAYSKGYS LPMSQADEMVRYIE ENKARQSSINPNQW WKIFEDTVCNFNYA FVSGEFTGGFKDRL NNICERTRVSGGAIN TINLLLLAEELKSGR MSYPKCFSYFDTND EVHI	130	ATCGCCAAGAGCGACTTCAGCATCATCAAGGACAACATCAGGAG GAAGCTGCAGTACGTGAACCACAAGTACCTGCTGCTGATCGACCT GGGCTTCGACAGCGACAGCAACAGGGACTACGAGATCCAGACCG CCGAGCTGCTGACCACCGAGCTGGCCTTCAAGGGCGCCAGGCTGG GCGACACCAGGAAGCCCGACGTGTGCGTGTACTACGGCGAGAAC GGCCTGATCATCGACAACAAGGCCTACAGCAAGGGCTACAGCCT GCCCATGAGCCAGGCCGACGAGATGGTGAGGTACATCGAGGAGA ACAAGGCCAGGCAGAGCAGCATCAACCCCAACCAGTGGTGGAAG ATCTTCGAGGACACCGTGTGCAACTTCAACTACGCCTTCGTGAGC GGCGAGTTCACCGGCGGCTTCAAGGACAGGCTGAACAACATCTGC GAGAGGACCAGGGTGAGCGGCGGCGCCATCAACACCATCAACCT GCTGCTGCTGGCCGAGGAGCTGAAGAGCGGCAGGATGAGCTACC CCAAGTGCTTCAGCTACTTCGACACCAACGACGAGGTGCACATC
50	LKYLGIKKQNRAFEI ITAELFNTSYKLSAT HLGGGRRPDVLVYN DNFGIIVDTKAYKD GYGRNVNQEDEMV RYITENNIRKQDINK NDWWKYFSKSIPST SYYHLWISSQFVGM FSDQLRETSSRTGEN GGAMNVEQLLIGAN QVLNNVLDPNCLPK YMENKEIIF	131	CTGAAGTACCTGGGCATCAAGAAGCAGAACAGGGCCTTCGAGAT CATCACCGCCGAGCTGTTCAACACCAGCTACAAGCTGAGCGCCAC CCACCTGGGCGGCGGCAGGAGGCCCGACGTGCTGGTGTACAACG ACAACTTCGGCATCATCGTGGACACCAAGGCCTACAAGGACGGCT ACGGCAGGAACGTGAACCAGGAGGACGAGATGGTGAGGTACATC ACCGAGAACAACATCAGGAAGCAGGACATCAACAAGAACGACTG GTGGAAGTACTTCAGCAAGAGCATCCCCAGCACCAGCTACTACCA CCTGTGGATCAGCAGCCAGTTCGTGGGCATGTTCAGCGACCAGCT GAGGGAGACCAGCAGCAGGACCGGCGAGAACGGCGGCGCCATGA ACGTGGAGCAGCTGCTGATCGGCGCCAACCAGGTGCTGAACAAC GTGCTGGACCCCAACTGCCTGCCCAAGTACATGGAGAACAAGGA GATCATCTTC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
51	VPVKSEVSLCKDYL RSHLNHVDHKYLIL LDLGFDGTSDRDYEI QTAQLLTGELNFKG ARLGDTRKPDVCVY YGEDGLIIDNKAYG KGYSLPIKQADEMY RYIEENKERNEKLNP NKWWEIFDKDVIHY HFAFVSGAFTGGFR ERLENIRMRSGIYGA AVNSMNLLLMAEEL KSGRLGYKECFKLF DCNDEIVL	132	GTGCCCGTGAAGAGCGAGGTGAGCCTGTGCAAGGACTACCTGAG GAGCCACCTGAACCACGTGGACCACAAGTACCTGATCCTGCTGGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGGCGAGCTGAACTTCAAGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACTACGGCGAG GACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAACGAGAAGCTGAACCCCAACAAGTGGTGG GAGATCTTCGACAAGGACGTGATCCACTACCACTTCGCCTTCGTG AGCGGCGCCTTCACCGGCGGCTTCAGGGAGAGGCTGGAGAACAT CAGGATGAGGAGCGGCATCTACGGCGCCGCCGTGAACAGCATGA ACCTGCTGCTGATGGCCGAGGAGCTGAAGAGCGGCAGGCTGGGC TACAAGGAGTGCTTCAAGCTGTTCGACTGCAACGACGAGATCGTG CTG
52	VPVKSEVSLLKDYL RTHLLHVDHRYLILL DLGFDGTSDRDYEI QTAQLLTGELNFKG ARLGDTRKPDVCVY YGEDGLIIDNKAYG KGYSLPIKQADEMY RYIEENKERNEKLNP NKWWEIFDNDVIHY HFAFISGAFTGGFKE RLDNIRMRSGIYGA AVNSMNLLLMAEEL KSGRLGYKECFKLF DCNDEIVL	133	GTGCCCGTGAAGAGCGAGGTGAGCCTGCTGAAGGACTACCTGAG GACCCACCTGCTGCACGTGGACCACAGGTACCTGATCCTGCTGGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGGCGAGCTGAACTTCAAGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACTACGGCGAG GACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAACGAGAAGCTGAACCCCAACAAGTGGTGG GAGATCTTCGACAACGACGTGATCCACTACCACTTCGCCTTCATC AGCGGCGCCTTCACCGGCGGCTTCAAGGAGAGGCTGGACAACAT CAGGATGAGGAGCGGCATCTACGGCGCCGCCGTGAACAGCATGA ACCTGCTGCTGATGGCCGAGGAGCTGAAGAGCGGCAGGCTGGGC TACAAGGAGTGCTTCAAGCTGTTCGACTGCAACGACGAGATCGTG CTG
53	VPVKSEVSLCKDYL RSHLNHVDHKYLIL LDLGFDGTSDRDYEI QTAQLLTGELNFKG ARLGDTRKPDVCVY YGEDGLIIDNKAYG KGYSLPIKQADEMY RYIEENKERNEKLNP NKWWEIFDNDVIHY HFAFVSGAFTGGFR ERLENIRMRSGIYGA AVNSMNLLLMAEEL KSGRLGYKECFKLF DCNDEIVL	134	GTGCCCGTGAAGAGCGAGGTGAGCCTGTGCAAGGACTACCTGAG GAGCCACCTGAACCACGTGGACCACAAGTACCTGATCCTGCTGGA CCTGGGCTTCGACGGCACCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGGCGAGCTGAACTTCAAGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACTACGGCGAG GACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAACGAGAAGCTGAACCCCAACAAGTGGTGG GAGATCTTCGACAACGACGTGATCCACTACCACTTCGCCTTCGTG AGCGGCGCCTTCACCGGCGGCTTCAGGGAGAGGCTGGAGAACAT CAGGATGAGGAGCGGCATCTACGGCGCCGCCGTGAACAGCATGA ACCTGCTGCTGATGGCCGAGGAGCTGAAGAGCGGCAGGCTGGGC TACAAGGAGTGCTTCAAGCTGTTCGACTGCAACGACGAGATCGTG CTG
54	VPVKSEMSLLKDYL RTHLLHVDHRYLILL DLGFDGASDRDYEI QTAQLLTGELNFKG ARLGDTRKPDVCVY YGEDGLIIDNKAYG KGYSLPIKQADEMY RYIEENKERNEKLNP NKWWEIFDNDVIHY HFAFVSGAFTGGFK ERLDNIRMRSGIYGA AVNSMNLLLMAEEL KSGRLGYKECFKLF	135	GTGCCCGTGAAGAGCGAGATGAGCCTGCTGAAGGACTACCTGAG GACCCACCTGCTGCACGTGGACCACAGGTACCTGATCCTGCTGGA CCTGGGCTTCGACGGCGCCAGCGACAGGGACTACGAGATCCAGA CCGCCCAGCTGCTGACCGGCGAGCTGAACTTCAAGGGCGCCAGGC TGGGCGACACCAGGAAGCCCGACGTGTGCGTGTACTACGGCGAG GACGGCCTGATCATCGACAACAAGGCCTACGGCAAGGGCTACAG CCTGCCCATCAAGCAGGCCGACGAGATGTACAGGTACATCGAGG AGAACAAGGAGAGGAACGAGAAGCTGAACCCCAACAAGTGGTGG GAGATCTTCGACAACGACGTGATCCACTACCACTTCGCCTTCGTG AGCGGCGCCTTCACCGGCGGCTTCAAGGAGAGGCTGGACAACAT CAGGATGAGGAGCGGCATCTACGGCGCCGCCGTGAACAGCATGA ACCTGCTGCTGATGGCCGAGGAGCTGAAGAGCGGCAGGCTGGGC TACAAGGAGTGCTTCAAGCTGTTCGACTGCAACGACGAGATCGTG

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
	DCNDEIVL		CTG
55	ILVDKEREMRKAKF LKETVLDSKFISLLD LAADATKSRDFEIVT AELFKEAYNLNSVL LGGSNKPDGLVFTD DFGILLDTKAYKNG FSIYAKDRDQMIRY VDDNNKRDKIRNPN EWWKSFSPLIPNDKF YYLWVSNFFKGQFK NQIEYVNRETNTYG AVLNVEQLLYGADA VIKGIINPNKLHEYFS NDEIKF	136	ATCCTGGTGGACAAGGAGAGGGAGATGAGGAAGGCCAAGTTCCT GAAGGAGACCGTGCTGGACAGCAAGTTCATCAGCCTGCTGGACCT GGCCGCCGACGCCACCAAGAGCAGGGACTTCGAGATCGTGACCG CCGAGCTGTTCAAGGAGGCCTACAACCTGAACAGCGTGCTGCTGG GCGGCAGCAACAAGCCCGACGGCCTGGTGTTCACCGACGACTTCG GCATCCTGCTGGACACCAAGGCCTACAAGAACGGCTTCAGCATCT ACGCCAAGGACAGGGACCAGATGATCAGGTACGTGGACGACAAC AACAAGAGGGACAAGATCAGGAACCCCAACGAGTGGTGGAAGAG CTTCAGCCCCCTGATCCCCAACGACAAGTTCTACTACCTGTGGGT GAGCAACTTCTTCAAGGGCCAGTTCAAGAACCAGATCGAGTACGT GAACAGGGAGACCAACACCTACGGCGCCGTGCTGAACGTGGAGC AGCTGCTGTACGGCGCCGACGCCGTGATCAAGGGCATCATCAACC CCAACAAGCTGCACGAGTACTTCAGCAACGACGAGATCAAGTTC
56	TVDEKERLELKEYFI SNTRIPSKYITLLDLA YDGNANRDFEIVTA ELFKDIFKLQSKHM GGTRKPDILIWTDKF GVIADTKAYSKGYK KNISEADKMVRYVN ENTNRNKVDNTNE WWNSFDSRIPKDAY YFLWISSEFVGKFDE QLTETSSRTGRNGAS INVYQLLRGADLVQ KSKFNIHDLPNLMQ NNEIKF	137	ACCGTGGACGAGAAGGAGAGGCTGGAGCTGAAGGAGTACTTCAT CAGCAACACCAGGATCCCCAGCAAGTACATCACCCTGCTGGACCT GGCCTACGACGGCAACGCCAACAGGGACTTCGAGATCGTGACCG CCGAGCTGTTCAAGGACATCTTCAAGCTGCAGAGCAAGCACATGG GCGGCACCAGGAAGCCCGACATCCTGATCTGGACCGACAAGTTCG GCGTGATCGCCGACACCAAGGCCTACAGCAAGGGCTACAAGAAG AACATCAGCGAGGCCGACAAGATGGTGAGGTACGTGAACGAGAA CACCAACAGGAACAAGGTGGACAACACCAACGAGTGGTGGAACA GCTTCGACAGCAGGATCCCCAAGGACGCCTACTACTTCCTGTGGA TCAGCAGCGAGTTCGTGGGCAAGTTCGACGAGCAGCTGACCGAG ACCAGCAGCAGGACCGGCAGGAACGGCGCCAGCATCAACGTGTA CCAGCTGCTGAGGGGCGCCGACCTGGTGCAGAAGAGCAAGTTCA ACATCCACGACCTGCCCAACCTGATGCAGAACAACGAGATCAAGT TC
57	TLQKSDIEKFKNQLR TELTNIDHSYLKGIDI ASKKTTTNVENTEF EAISTKVFTDELGFF GEHLGGSNKPDGLI WDNDCAIILDSKAY SEGFPLTASHTDAM GRYLRQFKERKEEIK PTWWDIAPDNLANT YFAYVSGSFSGNYK AQLQKFRQDTNHM GGALEFVKLLLLAN NYKAHKMSINEVKE SILDYNISY	138	ACCCTGCAGAAGAGCGACATCGAGAAGTTCAAGAACCAGCTGAG GACCGAGCTGACCAACATCGACCACAGCTACCTGAAGGGCATCG ACATCGCCAGCAAGAAGACCACCACCAACGTGGAGAACACCGAG TTCGAGGCCATCAGCACCAAGGTGTTCACCGACGAGCTGGGCTTC TTCGGCGAGCACCTGGGCGGCAGCAACAAGCCCGACGGCCTGAT CTGGGACAACGACTGCGCCATCATCCTGGACAGCAAGGCCTACAG CGAGGGCTTCCCCCTGACCGCCAGCCACACCGACGCCATGGGCAG GTACCTGAGGCAGTTCAAGGAGAGGAAGGAGGAGATCAAGCCCA CCTGGTGGGACATCGCCCCCGACAACCTGGCCAACACCTACTTCG CCTACGTGAGCGGCAGCTTCAGCGGCAACTACAAGGCCCAGCTGC AGAAGTTCAGGCAGGACACCAACCACATGGGCGGCGCCCTGGAG TTCGTGAAGCTGCTGCTGCTGGCCAACAACTACAAGGCCCACAAG ATGAGCATCAACGAGGTGAAGGAGAGCATCCTGGACTACAACAT CAGCTAC
58	VKEKTDAALVKERV RLQLHNINHKYLALI DYAFSGKNNSRDFE VYTIDLLVNELTFGG LHLGGTRKPDGIFY HGSNGIIIDNKAYAK GFVITRNMADEMIR YVQENNDRNPERNP NCWWKGFPHDVTR YNYVFISSMFKGEV EHMLDNIRQSTGIDG CVLTIENLLYYADAI	139	GTGAAGGAGAAGACCGACGCCGCCCTGGTGAAGGAGAGGGTGAG GCTGCAGCTGCACAACATCAACCACAAGTACCTGGCCCTGATCGA CTACGCCTTCAGCGGCAAGAACAACAGCAGGGACTTCGAGGTGT ACACCATCGACCTGCTGGTGAACGAGCTGACCTTCGGCGGCCTGC ACCTGGGCGGCACCAGGAAGCCCGACGGCATCTTCTACCACGGCA GCAACGGCATCATCATCGACAACAAGGCCTACGCCAAGGGCTTCG TGATCACCAGGAACATGGCCGACGAGATGATCAGGTACGTGCAG GAGAACAACGACAGGAACCCCGAGAGGAACCCCAACTGCTGGTG GAAGGGCTTCCCCCACGACGTGACCAGGTACAACTACGTGTTCAT CAGCAGCATGTTCAAGGGCGAGGTGGAGCACATGCTGGACAACA TCAGGCAGAGCACCGGCATCGACGGCTGCGTGCTGACCATCGAG AACCTGCTGTACTACGCCGACGCCATCAAGGGCGGCACCCTGAGC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
	KGGTLSKATFINGFN ANKEMVF		AAGGCCACCTTCATCAACGGCTTCAACGCCAACAAGGAGATGGTG TTC
59	VKETTDSVIIKDRVR LKLHHVNHKYLTLI DYAFSGKNNCMDFE VYTIDLLVNELAFN GVHLGGTRKPDGIF YHNRNGIIIDNKAYS HGFTLSRAMADEMI RYIQENNDRNPERN PNKWWENFDKGVN QFNFVFISSLFKGEIE HMLTNIKQSTDGVE GCVLSAENLLYFAE AMKSGVMPKTEFIS YFGAGKEIQF	140	GTGAAGGAGACCACCGACAGCGTGATCATCAAGGACAGGGTGAG GCTGAAGCTGCACCACGTGAACCACAAGTACCTGACCCTGATCGA CTACGCCTTCAGCGGCAAGAACAACTGCATGGACTTCGAGGTGTA CACCATCGACCTGCTGGTGAACGAGCTGGCCTTCAACGGCGTGCA CCTGGGCGGCACCAGGAAGCCCGACGGCATCTTCTACCACAACAG GAACGGCATCATCATCGACAACAAGGCCTACAGCCACGGCTTCAC CCTGAGCAGGGCCATGGCCGACGAGATGATCAGGTACATCCAGG AGAACAACGACAGGAACCCCGAGAGGAACCCCAACAAGTGGTGG GAGAACTTCGACAAGGGCGTGAACCAGTTCAACTTCGTGTTCATC AGCAGCCTGTTCAAGGGCGAGATCGAGCACATGCTGACCAACATC AAGCAGAGCACCGACGGCGTGGAGGGCTGCGTGCTGAGCGCCGA GAACCTGCTGTACTTCGCCGAGGCCATGAAGAGCGGCGTGATGCC CAAGACCGAGTTCATCAGCTACTTCGGCGCCGGCAAGGAGATCCA GTTC
60	SACKADITELKDKIR KSLKVLDHKYLVLV DLAYSDASTKSKKN SDAREFEIQTADLFT KELKFDGMRLGDSN RPDVIISHDNFGTIID NKSYKDGFNIDKKC ADEMSRYINENQRRI PELPKNEWWKNFD VNVDIFTFLFITSYLK GNFKDQLEYISKSQS DIKGAAISVEHLLYI SEKVKNGSMDKADF FKLFNNDEIRV	141	AGCGCCTGCAAGGCCGACATCACCGAGCTGAAGGACAAGATCAG GAAGAGCCTGAAGGTGCTGGACCACAAGTACCTGGTGCTGGTGG ACCTGGCCTACAGCGACGCCAGCACCAAGAGCAAGAAGAACAGC GACGCCAGGGAGTTCGAGATCCAGACCGCCGACCTGTTCACCAAG GAGCTGAAGTTCGACGGCATGAGGCTGGGCGACAGCAACAGGCC CGACGTGATCATCAGCCACGACAACTTCGGCACCATCATCGACAA CAAGAGCTACAAGGACGGCTTCAACATCGACAAGAAGTGCGCCG ACGAGATGAGCAGGTACATCAACGAGAACCAGAGGAGGATCCCC GAGCTGCCCAAGAACGAGTGGTGGAAGAACTTCGACGTGAACGT GGACATCTTCACCTTCCTGTTCATCACCAGCTACCTGAAGGGCAA CTTCAAGGACCAGCTGGAGTACATCAGCAAGAGCCAGAGCGACA TCAAGGGCGCCGCCATCAGCGTGGAGCACCTGCTGTACATCAGCG AGAAGGTGAAGAACGGCAGCATGGACAAGGCCGACTTCTTCAAG CTGTTCAACAACGACGAGATCAGGGTG
61	VLKDKHLEKIKEKF LENTSLDPRFISLIEIS RDKKQNRAFEIITAE LFNTSYNLSAIHLGG GRRPDVLAYNDNFG IIVDTKAYKNGYGR NVNQEDEMVRYITE NKIRKQDISKNNWW KYFSKSIPSTSYYHL WISSEFVGMFSDQL RETSSRTGENGGAM NVEQLLIGANQVLN NVLDPNRLPEYMEN KEIIF	142	GTGCTGAAGGACAAGCACCTGGAGAAGATCAAGGAGAAGTTCCT GGAGAACACCAGCCTGGACCCCAGGTTCATCAGCCTGATCGAGAT CAGCAGGGACAAGAAGCAGAACAGGGCCTTCGAGATCATCACCG CCGAGCTGTTCAACACCAGCTACAACCTGAGCGCCATCCACCTGG GCGGCGGCAGGAGGCCCGACGTGCTGGCCTACAACGACAACTTC GGCATCATCGTGGACACCAAGGCCTACAAGAACGGCTACGGCAG GAACGTGAACCAGGAGGACGAGATGGTGAGGTACATCACCGAGA ACAAGATCAGGAAGCAGGACATCAGCAAGAACAACTGGTGGAAG TACTTCAGCAAGAGCATCCCCAGCACCAGCTACTACCACCTGTGG ATCAGCAGCGAGTTCGTGGGCATGTTCAGCGACCAGCTGAGGGA GACCAGCAGCAGGACCGGCGAGAACGGCGGCGCCATGAACGTGG AGCAGCTGCTGATCGGCGCCAACCAGGTGCTGAACAACGTGCTGG ACCCCAACAGGCTGCCCGAGTACATGGAGAACAAGGAGATCATC TTC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
62	ALKDKHLEKIKEKF LENTSLDPRFISLIEIS RDKKQNRAFEIITAE LFNTSYKLSATHLG GGRRPDVLVYNDNF GIIVDTKAYKDGYG RNVNQEDEMVRYIT ENNIRKQDINKNDW WKYFSKSIPSTSYYH LWISSQFVGMFSDQ LRETSSRTGENGGA MNVEQLLIGANQVL NNVLDPNCLPKYME NKEIIF	143	GCCCTGAAGGACAAGCACCTGGAGAAGATCAAGGAGAAGTTCCT GGAGAACACCAGCCTGGACCCCAGGTTCATCAGCCTGATCGAGAT CAGCAGGGACAAGAAGCAGAACAGGGCCTTCGAGATCATCACCG CCGAGCTGTTCAACACCAGCTACAAGCTGAGCGCCACCCACCTGG GCGGCGGCAGGAGGCCCGACGTGCTGGTGTACAACGACAACTTC GGCATCATCGTGGACACCAAGGCCTACAAGGACGGCTACGGCAG GAACGTGAACCAGGAGGACGAGATGGTGAGGTACATCACCGAGA ACAACATCAGGAAGCAGGACATCAACAAGAACGACTGGTGGAAG TACTTCAGCAAGAGCATCCCCAGCACCAGCTACTACCACCTGTGG ATCAGCAGCCAGTTCGTGGGCATGTTCAGCGACCAGCTGAGGGAG ACCAGCAGCAGGACCGGCGAGAACGGCGGCGCCATGAACGTGGA GCAGCTGCTGATCGGCGCCAACCAGGTGCTGAACAACGTGCTGGA CCCCAACTGCCTGCCCAAGTACATGGAGAACAAGGAGATCATCTT C
63	VLEKSDIEKFKNQLR TELTNIDHSYLKGIDI ASKKKTSNVENTEF EAISTKIFTDELGFSG KHLGGSNKPDGLLW DDDCAIILDSKAYSE GFPLTASHTDAMGR YLRQFTERKEEIKPT WWDIAPEHLDNTYF AYVSGSFSGNYKEQ LQKFRQDTNHLGGA LEFVKLLLLANNYK TQKMSKKEVKKSIL DYNISY	144	GTGCTGGAGAAGAGCGACATCGAGAAGTTCAAGAACCAGCTGAG GACCGAGCTGACCAACATCGACCACAGCTACCTGAAGGGCATCG ACATCGCCAGCAAGAAGAAGACCAGCAACGTGGAGAACACCGAG TTCGAGGCCATCAGCACCAAGATCTTCACCGACGAGCTGGGCTTC AGCGGCAAGCACCTGGGCGGCAGCAACAAGCCCGACGGCCTGCT GTGGGACGACGACTGCGCCATCATCCTGGACAGCAAGGCCTACA GCGAGGGCTTCCCCCTGACCGCCAGCCACACCGACGCCATGGGCA GGTACCTGAGGCAGTTCACCGAGAGGAAGGAGGAGATCAAGCCC ACCTGGTGGGACATCGCCCCCGAGCACCTGGACAACACCTACTTC GCCTACGTGAGCGGCAGCTTCAGCGGCAACTACAAGGAGCAGCT GCAGAAGTTCAGGCAGGACACCAACCACCTGGGCGGCGCCCTGG AGTTCGTGAAGCTGCTGCTGCTGGCCAACAACTACAAGACCCAGA AGATGAGCAAGAAGGAGGTGAAGAAGAGCATCCTGGACTACAAC ATCAGCTAC
64	AEADVTSEKIKNHF RRVTELPERYLELLD IAFDHKRNRDFEMV TAGLFKDVYGLESV HLGGANKPDGVVY NDNFGIILDTKAYEN GYGKHISQIDEMVR YIDDNRLRDTTRNP NKWWENFDADIPSD QFYYLWVSGKFLPN FAEQLKQTNYRSHA NGGGLEVQQLLLGA DAVKRRKLDVNTIP NYMKNEVITL	145	GCCGAGGCCGACGTGACCAGCGAGAAGATCAAGAACCACTTCAG GAGGGTGACCGAGCTGCCCGAGAGGTACCTGGAGCTGCTGGACA TCGCCTTCGACCACAAGAGGAACAGGGACTTCGAGATGGTGACC GCCGGCCTGTTCAAGGACGTGTACGGCCTGGAGAGCGTGCACCTG GGCGGCGCCAACAAGCCCGACGGCGTGGTGTACAACGACAACTT CGGCATCATCCTGGACACCAAGGCCTACGAGAACGGCTACGGCA AGCACATCAGCCAGATCGACGAGATGGTGAGGTACATCGACGAC AACAGGCTGAGGGACACCACCAGGAACCCCAACAAGTGGTGGGA GAACTTCGACGCCGACATCCCCAGCGACCAGTTCTACTACCTGTG GGTGAGCGGCAAGTTCCTGCCCAACTTCGCCGAGCAGCTGAAGCA GACCAACTACAGGAGCCACGCCAACGGCGGCGGCCTGGAGGTGC AGCAGCTGCTGCTGGGCGCCGACGCCGTGAAGAGGAGGAAGCTG GACGTGAACACCATCCCCAACTACATGAAGAACGAGGTGATCAC CCTG
65	AEADLNSEKIKNHY RKITNLPEKYIELLDI AFDHRRHQDFEIVT AGLFKDCYGLSSIHL GGQNKPDGVVFNN KFGIILDTKAYEKGY GMHIGQIDEMCRYI DDNKKRDIVRQPNE WWKNFGDNIPKDQF YYLWISGKFLPRFNE QLKQTHYRTSINGG GLEVSQLLLGANAA MKGKLDVNTLPKH	146	GCCGAGGCCGACCTGAACAGCGAGAAGATCAAGAACCACTACAG GAAGATCACCAACCTGCCCGAGAAGTACATCGAGCTGCTGGACAT CGCCTTCGACCACAGGAGGCACCAGGACTTCGAGATCGTGACCGC CGGCCTGTTCAAGGACTGCTACGGCCTGAGCAGCATCCACCTGGG CGGCCAGAACAAGCCCGACGGCGTGGTGTTCAACAACAAGTTCG GCATCATCCTGGACACCAAGGCCTACGAGAAGGGCTACGGCATG CACATCGGCCAGATCGACGAGATGTGCAGGTACATCGACGACAA CAAGAAGAGGGACATCGTGAGGCAGCCCAACGAGTGGTGGAAGA ACTTCGGCGACAACATCCCCAAGGACCAGTTCTACTACCTGTGGA TCAGCGGCAAGTTCCTGCCCAGGTTCAACGAGCAGCTGAAGCAGA CCCACTACAGGACCAGCATCAACGGCGGCGGCCTGGAGGTGAGC CAGCTGCTGCTGGGCGCCAACGCCGCCATGAAGGGCAAGCTGGA CGTGAACACCCTGCCCAAGCACATGAACAACCAGGTGATCAAGCT

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
	MNNQVIKL		G
66	VLKDAALQKTKNTL LNELTEIDPADIEVIE MSWKKATTRSQNTL EATLFEVKVVEIFKK YFELNGEHLGGQNR PDGAVYYNSTYGIIL DTKAYSNGYNIPVD QQREMVDYITDVID KNQNVTPNRWWEA FPATLLKNNIYYLW VAGGFTGKYLDQLT RTHNQTNMDGGAM TTEVLLRLANKVSS GNLKTTDIPKLMTN KLILS	147	GTGCTGAAGGACGCCGCCCTGCAGAAGACCAAGAACACCCTGCT GAACGAGCTGACCGAGATCGACCCCGCCGACATCGAGGTGATCG AGATGAGCTGGAAGAAGGCCACCACCAGGAGCCAGAACACCCTG GAGGCCACCCTGTTCGAGGTGAAGGTGGTGGAGATCTTCAAGAA GTACTTCGAGCTGAACGGCGAGCACCTGGGCGGCCAGAACAGGC CCGACGGCGCCGTGTACTACAACAGCACCTACGGCATCATCCTGG ACACCAAGGCCTACAGCAACGGCTACAACATCCCCGTGGACCAG CAGAGGGAGATGGTGGACTACATCACCGACGTGATCGACAAGAA CCAGAACGTGACCCCCAACAGGTGGTGGGAGGCCTTCCCCGCCAC CCTGCTGAAGAACAACATCTACTACCTGTGGGTGGCCGGCGGCTT CACCGGCAAGTACCTGGACCAGCTGACCAGGACCCACAACCAGA CCAACATGGACGGCGGCGCCATGACCACCGAGGTGCTGCTGAGG CTGGCCAACAAGGTGAGCAGCGGCAACCTGAAGACCACCGACAT CCCCAAGCTGATGACCAACAAGCTGATCCTGAGC
67	AEADLDSERIKNHY RKITNLPEKYIELLDI AFDHHRHQDFEIITA GLFKDCYGLSSIHLG GQNKPDGVVFNGKF GIILDTKAYEKGYG MHINQIDEMCRYIED NKQRDKIRQPNEW WNNFGDNIPENKFY YLWVSGKFLPKFNE QLKQTHYRTGINGG GLEVSQLLLGADAV MKGALNVNILPTYM HNNVIQ	148	GCCGAGGCCGACCTGGACAGCGAGAGGATCAAGAACCACTACAG GAAGATCACCAACCTGCCCGAGAAGTACATCGAGCTGCTGGACAT CGCCTTCGACCACCACAGGCACCAGGACTTCGAGATCATCACCGC CGGCCTGTTCAAGGACTGCTACGGCCTGAGCAGCATCCACCTGGG CGGCCAGAACAAGCCCGACGGCGTGGTGTTCAACGGCAAGTTCG GCATCATCCTGGACACCAAGGCCTACGAGAAGGGCTACGGCATG CACATCAACCAGATCGACGAGATGTGCAGGTACATCGAGGACAA CAAGCAGAGGGACAAGATCAGGCAGCCCAACGAGTGGTGGAACA ACTTCGGCGACAACATCCCCGAGAACAAGTTCTACTACCTGTGGG TGAGCGGCAAGTTCCTGCCCAAGTTCAACGAGCAGCTGAAGCAG ACCCACTACAGGACCGGCATCAACGGCGGCGGCCTGGAGGTGAG CCAGCTGCTGCTGGGCGCCGACGCCGTGATGAAGGGCGCCCTGAA CGTGAACATCCTGCCCACCTACATGCACAACAACGTGATCCAG
68	EISDIALQKEKAYFY KNTALSKRHISILEIA FDGSKNRDLEILSAE VFKDYYQLESIHLG GGLKPDGIAFNQNF GIIVDTKAYKGVYS RSRAEADKMFRYIE DNKKRDPKRNQSL WWRSFNEHIPANNF YFLWISGKFQRNFD TQINQLNYETGYRG GALSARQFLIGADAI QKGKIDINDLPSYFN NSVISF	149	GAGATCAGCGACATCGCCCTGCAGAAGGAGAAGGCCTACTTCTAC AAGAACACCGCCCTGAGCAAGAGGCACATCAGCATCCTGGAGAT CGCCTTCGACGGCAGCAAGAACAGGGACCTGGAGATCCTGAGCG CCGAGGTGTTCAAGGACTACTACCAGCTGGAGAGCATCCACCTGG GCGGCGGCCTGAAGCCCGACGGCATCGCCTTCAACCAGAACTTCG GCATCATCGTGGACACCAAGGCCTACAAGGGCGTGTACAGCAGG AGCAGGGCCGAGGCCGACAAGATGTTCAGGTACATCGAGGACAA CAAGAAGAGGGACCCCAAGAGGAACCAGAGCCTGTGGTGGAGGA GCTTCAACGAGCACATCCCCGCCAACAACTTCTACTTCCTGTGGA TCAGCGGCAAGTTCCAGAGGAACTTCGACACCCAGATCAACCAGC TGAACTACGAGACCGGCTACAGGGGCGGCGCCCTGAGCGCCAGG CAGTTCCTGATCGGCGCCGACGCCATCCAGAAGGGCAAGATCGAC ATCAACGACCTGCCCAGCTACTTCAACAACAGCGTGATCAGCTTC
69	TSREKSRLNLKEYFV SNTNLPNKFITLLDL AYDGKANRDFELIT SELFREIYKLNTRHL GGTRKPDILIWNENF GIIADTKAYSKGYK KNISEEDKMVRYIDE NIKRSKDYNPNEWW KVFDNEISSNNYFYL WISSEFIGKFEEQLQ ETAQRTNVKGASIN	150	ACCAGCAGGGAGAAGAGCAGGCTGAACCTGAAGGAGTACTTCGT GAGCAACACCAACCTGCCCAACAAGTTCATCACCCTGCTGGACCT GGCCTACGACGGCAAGGCCAACAGGGACTTCGAGCTGATCACCA GCGAGCTGTTCAGGGAGATCTACAAGCTGAACACCAGGCACCTG GGCGGCACCAGGAAGCCCGACATCCTGATCTGGAACGAGAACTT CGGCATCATCGCCGACACCAAGGCCTACAGCAAGGGCTACAAGA AGAACATCAGCGAGGAGGACAAGATGGTGAGGTACATCGACGAG AACATCAAGAGGAGCAAGGACTACAACCCCAACGAGTGGTGGAA GGTGTTCGACAACGAGATCAGCAGCAACAACTACTTCTACCTGTG GATCAGCAGCGAGTTCATCGGCAAGTTCGAGGAGCAGCTGCAGG AGACCGCCCAGAGGACCAACGTGAAGGGCGCCAGCATCAACGTG

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
	VYQLLMGAHKVQT KELNVNSIPKYMNN TEIKF		TACCAGCTGCTGATGGGCGCCCACAAGGTGCAGACCAAGGAGCT GAACGTGAACAGCATCCCCAAGTACATGAACAACACCGAGATCA AGTTC
70	NCIKDSIIDIKDRVRT KLVHLDHKYLALID LAFSDADTRTKKNS DAREFEIQTADLFTK ELSFNGQRLGDSRK PDIIISFDKIGTIIDNK SYKDGFNISRPCADE MIRYINENNLRKKSL NANEWWNKFDPTIT AYSFLFITSYLKGQF QEQLEYISNANGGIK GAAIGIENLLYLSEA LKSGKISHKDFYQNF NNKEITY	151	AACTGCATCAAGGACAGCATCATCGACATCAAGGACAGGGTGAG GACCAAGCTGGTGCACCTGGACCACAAGTACCTGGCCCTGATCGA CCTGGCCTTCAGCGACGCCGACACCAGGACCAAGAAGAACAGCG ACGCCAGGGAGTTCGAGATCCAGACCGCCGACCTGTTCACCAAGG AGCTGAGCTTCAACGGCCAGAGGCTGGGCGACAGCAGGAAGCCC GACATCATCATCAGCTTCGACAAGATCGGCACCATCATCGACAAC AAGAGCTACAAGGACGGCTTCAACATCAGCAGGCCCTGCGCCGA CGAGATGATCAGGTACATCAACGAGAACAACCTGAGGAAGAAGA GCCTGAACGCCAACGAGTGGTGGAACAAGTTCGACCCCACCATCA CCGCCTACAGCTTCCTGTTCATCACCAGCTACCTGAAGGGCCAGT TCCAGGAGCAGCTGGAGTACATCAGCAACGCCAACGGCGGCATC AAGGGCGCCGCCATCGGCATCGAGAACCTGCTGTACCTGAGCGA GGCCCTGAAGAGCGGCAAGATCAGCCACAAGGACTTCTACCAGA ACTTCAACAACAAGGAGATCACCTAC
71	LPQKDQVQQQQDEL RPMLKNVDHRYLQL VELALDSDQNSEYS QFEQLTMELVLKHL DFDGKPLGGSNKPD GIAWDNDGNFIIFDT KAYNKGYSLAGNT DKVKRYIDDVRDRD TSRTSTWWQLVPKS IDVHNLLRFVYVSG NFTGNYMKLLDSLR SWSNAQGGLASVEK LLLTSELYLRNMYS HQELIDSWTDNNVK H	152	CTGCCCCAGAAGGACCAGGTGCAGCAGCAGCAGGACGAGCTGAG GCCCATGCTGAAGAACGTGGACCACAGGTACCTGCAGCTGGTGG AGCTGGCCCTGGACAGCGACCAGAACAGCGAGTACAGCCAGTTC GAGCAGCTGACCATGGAGCTGGTGCTGAAGCACCTGGACTTCGAC GGCAAGCCCCTGGGCGGCAGCAACAAGCCCGACGGCATCGCCTG GGACAACGACGGCAACTTCATCATCTTCGACACCAAGGCCTACAA CAAGGGCTACAGCCTGGCCGGCAACACCGACAAGGTGAAGAGGT ACATCGACGACGTGAGGGACAGGGACACCAGCAGGACCAGCACC TGGTGGCAGCTGGTGCCCAAGAGCATCGACGTGCACAACCTGCTG AGGTTCGTGTACGTGAGCGGCAACTTCACCGGCAACTACATGAAG CTGCTGGACAGCCTGAGGAGCTGGAGCAACGCCCAGGGCGGCCT GGCCAGCGTGGAGAAGCTGCTGCTGACCAGCGAGCTGTACCTGA GGAACATGTACAGCCACCAGGAGCTGATCGACAGCTGGACCGAC AACAACGTGAAGCAC
72	TTDAVVVKDRARV RLHNINHKYLTLIDY AFSGKNNCTEFEIYT IDLLVNELAFNGIHL GGTRKPDGIFDYNQ QGIIIDNKAYSKGFTI TRSMADEMVRYVQ ENNDRNPERNKTQ WWLNFGDNVNHFN FVFISSMFKGEVRH MLNNIKQSTGVDGC VLTAENLLYFADAI KGGTVKRTDFINLF GKNDEL	153	ACCACCGACGCCGTGGTGGTGAAGGACAGGGCCAGGGTGAGGCT GCACAACATCAACCACAAGTACCTGACCCTGATCGACTACGCCTT CAGCGGCAAGAACAACTGCACCGAGTTCGAGATCTACACCATCG ACCTGCTGGTGAACGAGCTGGCCTTCAACGGCATCCACCTGGGCG GCACCAGGAAGCCCGACGGCATCTTCGACTACAACCAGCAGGGC ATCATCATCGACAACAAGGCCTACAGCAAGGGCTTCACCATCACC AGGAGCATGGCCGACGAGATGGTGAGGTACGTGCAGGAGAACAA CGACAGGAACCCCGAGAGGAACAAGACCCAGTGGTGGCTGAACT TCGGCGACAACGTGAACCACTTCAACTTCGTGTTCATCAGCAGCA TGTTCAAGGGCGAGGTGAGGCACATGCTGAACAACATCAAGCAG AGCACCGGCGTGGACGGCTGCGTGCTGACCGCCGAGAACCTGCTG TACTTCGCCGACGCCATCAAGGGCGGCACCGTGAAGAGGACCGA CTTCATCAACCTGTTCGGCAAGAACGACGAGCTG

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
73	LPKKDNVQRQQDEL RPLLKHVDHRYLQL VELALDSSQNSEYS MLESMTMELLLTHL DFDGASLGGASKPD GIAWDKDGNFLIVD TKAYDNGYSLAGNT DKVARYIDDVRAKD PNRASTWWTQVPES LNVDDNLSFMYVSG SFTGNYQRLLKDLR ARTNARGGLTTVEK LLLTSEAYLAKSGY GHTQLLNDWTDDNI DH	154	CTGCCCAAGAAGGACAACGTGCAGAGGCAGCAGGACGAGCTGAG GCCCCTGCTGAAGCACGTGGACCACAGGTACCTGCAGCTGGTGGA GCTGGCCCTGGACAGCAGCCAGAACAGCGAGTACAGCATGCTGG AGAGCATGACCATGGAGCTGCTGCTGACCCACCTGGACTTCGACG GCGCCAGCCTGGGCGGCGCCAGCAAGCCCGACGGCATCGCCTGG GACAAGGACGGCAACTTCCTGATCGTGGACACCAAGGCCTACGA CAACGGCTACAGCCTGGCCGGCAACACCGACAAGGTGGCCAGGT ACATCGACGACGTGAGGGCCAAGGACCCCAACAGGGCCAGCACC TGGTGGACCCAGGTGCCCGAGAGCCTGAACGTGGACGACAACCT GAGCTTCATGTACGTGAGCGGCAGCTTCACCGGCAACTACCAGAG GCTGCTGAAGGACCTGAGGGCCAGGACCAACGCCAGGGGCGGCC TGACCACCGTGGAGAAGCTGCTGCTGACCAGCGAGGCCTACCTGG CCAAGAGCGGCTACGGCCACACCCAGCTGCTGAACGACTGGACC GACGACAACATCGACCAC
74	QIKDKYLEDLKLEL YKKTNLPNKYYEM VDIAYDGKRNREFEI YTSDLMQEIYGFKT TLLGGTRKPDVVSY SDAHGYIIDTKAYA NGYRKEIKQEDEMV RYIEDNQLKDVLRN PNKWWECFDDAEH KKEYYFLWISSKFV GEFSSQLQDTSRRTG IKGGAVNIVQLLLG AHLVYSGEISKDQF AAYMNNTEINF	155	CAGATCAAGGACAAGTACCTGGAGGACCTGAAGCTGGAGCTGTA CAAGAAGACCAACCTGCCCAACAAGTACTACGAGATGGTGGACA TCGCCTACGACGGCAAGAGGAACAGGGAGTTCGAGATCTACACC AGCGACCTGATGCAGGAGATCTACGGCTTCAAGACCACCCTGCTG GGCGGCACCAGGAAGCCCGACGTGGTGAGCTACAGCGACGCCCA CGGCTACATCATCGACACCAAGGCCTACGCCAACGGCTACAGGA AGGAGATCAAGCAGGAGGACGAGATGGTGAGGTACATCGAGGAC AACCAGCTGAAGGACGTGCTGAGGAACCCCAACAAGTGGTGGGA GTGCTTCGACGACGCCGAGCACAAGAAGGAGTACTACTTCCTGTG GATCAGCAGCAAGTTCGTGGGCGAGTTCAGCAGCCAGCTGCAGG ACACCAGCAGGAGGACCGGCATCAAGGGCGGCGCCGTGAACATC GTGCAGCTGCTGCTGGGCGCCCACCTGGTGTACAGCGGCGAGATC AGCAAGGACCAGTTCGCCGCCTACATGAACAACACCGAGATCAA CTTC
75	MNPRNEIVIAKHLSG GNRPEIVCYHPEDKP DHGLILDSKAYKSG FTIPSGERDKMVRYI EEYITKNQLQNPNE WWKNLKGAEYPGI VGFGFISNSFLGHYR KQLDYIMRRTKIKG SSITTEHLLKTVEDV LSEKGNVIDFFKYFL E	156	ATGAACCCCAGGAACGAGATCGTGATCGCCAAGCACCTGAGCGG CGGCAACAGGCCCGAGATCGTGTGCTACCACCCCGAGGACAAGC CCGACCACGGCCTGATCCTGGACAGCAAGGCCTACAAGAGCGGC TTCACCATCCCCAGCGGCGAGAGGGACAAGATGGTGAGGTACAT CGAGGAGTACATCACCAAGAACCAGCTGCAGAACCCCAACGAGT GGTGGAAGAACCTGAAGGGCGCCGAGTACCCCGGCATCGTGGGC TTCGGCTTCATCAGCAACAGCTTCCTGGGCCACTACAGGAAGCAG CTGGACTACATCATGAGGAGGACCAAGATCAAGGGCAGCAGCAT CACCACCGAGCACCTGCTGAAGACCGTGGAGGACGTGCTGAGCG AGAAGGGCAACGTGATCGACTTCTTCAAGTACTTCCTGGAG
76	EIKNQEIEELKQIAL NKYTALPSEWVELIE ISRDKDQSTIFEMKV AELFKTCYRIKSLHL GGASKPDCLLWDDS FSVIVDAKAYKDGF PFQASEKDKMVRYL RECERKDKAENATE WWNNFPPELNSNQL FFMFASSFFSSTAEK HLESVSIASKFSGCA WDVDNLLSGANFFL QNPQATLQYHLIRV FSNKVVD	157	GAGATCAAGAACCAGGAGATCGAGGAGCTGAAGCAGATCGCCCT GAACAAGTACACCGCCCTGCCCAGCGAGTGGGTGGAGCTGATCG AGATCAGCAGGGACAAGGACCAGAGCACCATCTTCGAGATGAAG GTGGCCGAGCTGTTCAAGACCTGCTACAGGATCAAGAGCCTGCAC CTGGGCGGCGCCAGCAAGCCCGACTGCCTGCTGTGGGACGACAG CTTCAGCGTGATCGTGGACGCCAAGGCCTACAAGGACGGCTTCCC CTTCCAGGCCAGCGAGAAGGACAAGATGGTGAGGTACCTGAGGG AGTGCGAGAGGAAGGACAAGGCCGAGAACGCCACCGAGTGGTGG AACAACTTCCCCCCCGAGCTGAACAGCAACCAGCTGTTCTTCATG TTCGCCAGCAGCTTCTTCAGCAGCACCGCCGAGAAGCACCTGGAG AGCGTGAGCATCGCCAGCAAGTTCAGCGGCTGCGCCTGGGACGTG GACAACCTGCTGAGCGGCGCCAACTTCTTCCTGCAGAACCCCCAG GCCACCCTGCAGTACCACCTGATCAGGGTGTTCAGCAACAAGGTG GTGGAC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
77	LPHKDNVIKQQDEL RPMLKHVNHKYLQ LVELAFESSRNSEYS QFETLTMELVLKYL DFSGKSLGGANKPD GIAWDPLGNFLIFDT KAYKHGYTLSNNTD RVARYINDVRDKDI QRISRWWQSIPTYID VKNKLQFVYISGSFT GHYLRLLNDLRSRT RAKGGLVTVEKLLL TTERYLAEADYTHK ELFDDWMDDNIEH	158	CTGCCCCACAAGGACAACGTGATCAAGCAGCAGGACGAGCTGAG GCCCATGCTGAAGCACGTGAACCACAAGTACCTGCAGCTGGTGGA GCTGGCCTTCGAGAGCAGCAGGAACAGCGAGTACAGCCAGTTCG AGACCCTGACCATGGAGCTGGTGCTGAAGTACCTGGACTTCAGCG GCAAGAGCCTGGGCGGCGCCAACAAGCCCGACGGCATCGCCTGG GACCCCCTGGGCAACTTCCTGATCTTCGACACCAAGGCCTACAAG CACGGCTACACCCTGAGCAACAACACCGACAGGGTGGCCAGGTA CATCAACGACGTGAGGGACAAGGACATCCAGAGGATCAGCAGGT GGTGGCAGAGCATCCCCACCTACATCGACGTGAAGAACAAGCTG CAGTTCGTGTACATCAGCGGCAGCTTCACCGGCCACTACCTGAGG CTGCTGAACGACCTGAGGAGCAGGACCAGGGCCAAGGGCGGCCT GGTGACCGTGGAGAAGCTGCTGCTGACCACCGAGAGGTACCTGG CCGAGGCCGACTACACCCACAAGGAGCTGTTCGACGACTGGATG GACGACAACATCGAGCAC
78	RISPSNLEQTKQQLR EELINLDHQYLDILD FSIAGNVGARQFEV RIVELLNEIIIAKHLS GGNRPEIIGFNPKEN PEDCIIMDSKAYKEG FNIPANERDKMIRYV EEYNAKDNTLNNNK WWKNFESPNYPTNQ VKFSFVSSSFIGQFT NQLTYINNRTNVNG SAITAETLLRKVENV MNVNTEYNLNNFFE ELGSNTLVA	159	AGGATCAGCCCCAGCAACCTGGAGCAGACCAAGCAGCAGCTGAG GGAGGAGCTGATCAACCTGGACCACCAGTACCTGGACATCCTGGA CTTCAGCATCGCCGGCAACGTGGGCGCCAGGCAGTTCGAGGTGAG GATCGTGGAGCTGCTGAACGAGATCATCATCGCCAAGCACCTGAG CGGCGGCAACAGGCCCGAGATCATCGGCTTCAACCCCAAGGAGA ACCCCGAGGACTGCATCATCATGGACAGCAAGGCCTACAAGGAG GGCTTCAACATCCCCGCCAACGAGAGGGACAAGATGATCAGGTA CGTGGAGGAGTACAACGCCAAGGACAACACCCTGAACAACAACA AGTGGTGGAAGAACTTCGAGAGCCCCAACTACCCCACCAACCAG GTGAAGTTCAGCTTCGTGAGCAGCAGCTTCATCGGCCAGTTCACC AACCAGCTGACCTACATCAACAACAGGACCAACGTGAACGGCAG CGCCATCACCGCCGAGACCCTGCTGAGGAAGGTGGAGAACGTGA TGAACGTGAACACCGAGTACAACCTGAACAACTTCTTCGAGGAGC TGGGCAGCAACACCCTGGTGGCC
79	TFDSTVADNLKNLIL PKLKELDHKYLQAI DIAYKRSNTTNHEN TLLEVLSADLFTKE MDYHGKHLGGANK PDGFVYDEETGWIL DSKAYRDGFAVTAH TTDAMGRYIDQYRD RDDKSTWWEDFPK DLPQTYFAYVSGFYI GKYQEQLQDFENRK HMKGGLIEVAKLILL AEKYKENKITHDQIT LQILNDHISQ	160	ACCTTCGACAGCACCGTGGCCGACAACCTGAAGAACCTGATCCTG CCCAAGCTGAAGGAGCTGGACCACAAGTACCTGCAGGCCATCGA CATCGCCTACAAGAGGAGCAACACCACCAACCACGAGAACACCC TGCTGGAGGTGCTGAGCGCCGACCTGTTCACCAAGGAGATGGACT ACCACGGCAAGCACCTGGGCGGCGCCAACAAGCCCGACGGCTTC GTGTACGACGAGGAGACCGGCTGGATCCTGGACAGCAAGGCCTA CAGGGACGGCTTCGCCGTGACCGCCCACACCACCGACGCCATGGG CAGGTACATCGACCAGTACAGGGACAGGGACGACAAGAGCACCT GGTGGGAGGACTTCCCCAAGGACCTGCCCCAGACCTACTTCGCCT ACGTGAGCGGCTTCTACATCGGCAAGTACCAGGAGCAGCTGCAG GACTTCGAGAACAGGAAGCACATGAAGGGCGGCCTGATCGAGGT GGCCAAGCTGATCCTGCTGGCCGAGAAGTACAAGGAGAACAAGA TCACCCACGACCAGATCACCCTGCAGATCCTGAACGACCACATCA GCCAG
80	PLDVVEQMKAELRP LLNHVNHRLLAIIDF SYNMSRGDDKRLED YTAQIYKLISHDTHL LAGPSRPDVVSVIND LGIIIDSKAYKQGFNI PQAEEDKMVRYLDE SIRRDPAINPTKWWE YLGASTEYVFQFVSS SFSSGASAKLRQIHR RSSIEGSIITAKNLLL LAENFLCTNTINIDL FRQNNEI	161	CCCCTGGACGTGGTGGAGCAGATGAAGGCCGAGCTGAGGCCCCT GCTGAACCACGTGAACCACAGGCTGCTGGCCATCATCGACTTCAG CTACAACATGAGCAGGGGCGACGACAAGAGGCTGGAGGACTACA CCGCCCAGATCTACAAGCTGATCAGCCACGACACCCACCTGCTGG CCGGCCCCAGCAGGCCCGACGTGGTGAGCGTGATCAACGACCTG GGCATCATCATCGACAGCAAGGCCTACAAGCAGGGCTTCAACATC CCCCAGGCCGAGGAGGACAAGATGGTGAGGTACCTGGACGAGAG CATCAGGAGGGACCCCGCCATCAACCCCACCAAGTGGTGGGAGT ACCTGGGCGCCAGCACCGAGTACGTGTTCCAGTTCGTGAGCAGCA GCTTCAGCAGCGGCGCCAGCGCCAAGCTGAGGCAGATCCACAGG AGGAGCAGCATCGAGGGCAGCATCATCACCGCCAAGAACCTGCT GCTGCTGGCCGAGAACTTCCTGTGCACCAACACCATCAACATCGA CCTGTTCAGGCAGAACAACGAGATC

SEQ ID NO	Amino Acid Sequence	SEQ ID NO	Back Translated Nucleic Acid Sequences
81	QLVPSYITQTKLRLS GLINYIDHSYFDLID LGFDGRQNRLYELRI VELLNLINSLKALHL SGGNRPEIIAYSPDV NPINGVIMDSKSYRG GFNIPNSERDKMIRY INEYNQKNPTLNSN RWWENFRAPDYPQS PLKYSFVSGNFIGQF LNQIQYILTQTGING GAITSEKLIEKVNAV LNPNISYTINNFFND LGCNRLVQ	162	CAGCTGGTGCCCAGCTACATCACCCAGACCAAGCTGAGGCTGAGC GGCCTGATCAACTACATCGACCACAGCTACTTCGACCTGATCGAC CTGGGCTTCGACGGCAGGCAGAACAGGCTGTACGAGCTGAGGAT CGTGGAGCTGCTGAACCTGATCAACAGCCTGAAGGCCCTGCACCT GAGCGGCGGCAACAGGCCCGAGATCATCGCCTACAGCCCCGACG TGAACCCCATCAACGGCGTGATCATGGACAGCAAGAGCTACAGG GGCGGCTTCAACATCCCCAACAGCGAGAGGGACAAGATGATCAG GTACATCAACGAGTACAACCAGAAGAACCCCACCCTGAACAGCA ACAGGTGGTGGGAGAACTTCAGGGCCCCCGACTACCCCCAGAGC CCCCTGAAGTACAGCTTCGTGAGCGGCAACTTCATCGGCCAGTTC CTGAACCAGATCCAGTACATCCTGACCCAGACCGGCATCAACGGC GGCGCCATCACCAGCGAGAAGCTGATCGAGAAGGTGAACGCCGT GCTGAACCCCAACATCAGCTACACCATCAACAACTTCTTCAACGA CCTGGGCTGCAACAGGCTGGTGCAG

[0095] In some embodiments, an endonuclease of the present disclosure can have a sequence of $X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10}X_{11}X_{12}X_{13}X_{14}X_{15}X_{16}X_{17}X_{18}X_{19}X_{20}X_{21}X_{22}X_{23}X_{24}X_{25}X_{26}X_{27}X_{28}X_{29}X_{30}X_{31}$ $X_{32}X_{33}X_{34}X_{35}X_{36}X_{37}X_{38}X_{39}X_{40}X_{41}X_{42}X_{43}KX_{44}X_{45}X_{46}X_{47}X_{48}X_{49}X_{50}X_{51}X_{52}X_{53}X_{54}X_{55}GX_{56}HLGGX_{57}$ $RX_{58}PDGX_{59}X_{60}X_{61}X_{62}X_{63}X_{64}X_{65}X_{66}X_{67}X_{68}X_{69}X_{70}X_{71}X_{72}X_{73}X_{74}GX_{75}IX_{76}DTKX_{77}YX_{78}X_{79}GYX_{80}L$ $PIX_{81}QX_{82}DEMX_{83}RYX_{84}X_{85}ENX_{86}X_{87}RX_{88}X_{89}X_{90}X_{91}NX_{92}NX_{93}WWX_{94}X_{95}X_{96}X_{97}X_{98}X_{99}X_{100}X_{101}$ $X_{102}X_{103}X_{104}X_{105}X_{106}FX_{107}X_{108}X_{109}X_{110}FX_{111}GX_{112}X_{113}X_{114}X_{115}X_{116}X_{117}X_{118}RX_{119}X_{120}X_{121}X_{122}X_{123}$ $X_{124}X_{125}X_{126}GX_{127}X_{128}X_{129}X_{130}X_{131}X_{132}X_{133}LLX_{134}X_{135}X_{136}X_{137}X_{138}X_{139}X_{140}X_{141}X_{142}X_{143}X_{144}X_{145}$ $X_{146}X_{147}X_{148}X_{149}X_{150}X_{151}X_{152}X_{153}FX_{154}X_{155}X_{156}X_{157}X_{158}X_{159}X_{160}$ (SEQ ID NO: 316), wherein X_1 is F, Q, N, D, or absent, X_2 is L, I, T, S, N, or absent, X_3 is V, I, G, A, E, T, or absent, X_4 is K, C, or absent, X_5 is G, S, or absent, X_6 is A, S, E, D, N, or absent, X_7 is M, I, V, Q, F, L, or absent, X_8 is E, S, T, N, or absent, X_9 is I, M, E, T, Q, or absent, X_{10} is K, S, L, I, T, E, or absent, X_{11} is K or absent, X_{12} is S, A, E, D, or absent, X_{13} is E, N, Q, K, or absent, X_{14} is L, M, V, or absent, X_{15} is R or absent, X_{16} is H, D, T, G, E, N, or absent, X_{17} is K, N, Q, E, A, or absent, X_{18} is L or absent, X_{19} is R, Q, N, T, D, or absent, X_{20} is H, M, V, N, T, or absent, X_{21} is V, L, I, or absent, X_{22} is P, S, or absent, X_{23} is H or absent, X_{24} is E, D, or absent, X_{25} is Y or absent, X_{26} is I, L, or absent, X_{27} is E, Q, G, S, A, Y, or absent, X_{28} is L or absent, X_{29} is I, V, L, or absent, X_{30} is E, D, or absent, X_{31} is I, L, or absent, X_{32} is A, S, or absent, X_{33} is Q, Y, F, or absent, X_{34} is D or absent, X_{35} is S, P, or absent, X_{36} is K, Y, Q, T, or absent, X_{37} is Q or absent, X_{38} is N or absent, X_{39} is R, K, or absent, X_{40} is L, I, or absent, X_{41} is L, F, or absent, X_{42} is E or absent, X_{43} is F, M, L, or absent, X_{44} is V, T, or I, X_{45} is V, M, L, or I, X_{46} is E, D, or Q, X_{47} is F or L, X_{48} is F or L, X_{49} is K, I, T, or V, X_{50} is K, N, or E, X_{51} is I or E, X_{52} is Y, F, or C, X_{53} is G, or N, X_{54} is Y, or F, X_{55} is R, S, N, E, K, or Q, X_{56} is K, S, L, V, or T, X_{57} is S, A, or V, X_{58} is K or R, X_{59} is A, I, or V, X_{60} is L, M, V, I, or C, X_{61} is F or Y, X_{62} is T, A, or S, X_{63} is K, E, or absent, X_{64} is D, E, or absent, X_{65} is E, A, or absent, X_{66} is N, K, or

absent, X_{67} is E, S, or absent, X_{68} is D, E, Q, A, or absent, X_{69} is G, V, K, N, or absent, X_{70} is L, G, E, S, or absent, X_{71} is V, S, K, T, E, or absent, X_{72} is L, H, K, E, Y, D, or A, X_{73} is N, G, or D, X_{74} is H, F, or Y, X_{75} is I, or V, X_{76} is L, V, or I, X_{77} is A or S, X_{78} is K or S, X_{79} is D, G, K, S, or N, X_{80} is R, N, S, or G, X_{81} is S, A, or G, X_{82} is A, I, or V, X_{83} is Q, E, I, or V, X_{84} is V or I, X_{85} is D, R, G, I, or E, X_{86} is N, I, or Q, X_{87} is K, D, T, E, or K, X_{88} is S, N, D, or E, X_{89} is Q, E, I, K, or A, X_{90} is V, H, R, K, L, or E, X_{91} is I, V, or R, X_{92} is P, S, T, or R, X_{93} is E, R, C, Q, or K, X_{94} is E, N, or K, X_{95} is I, V, N, E, or A, X_{96} is Y or F, X_{97} is P, G, or E, X_{98} is T, E, S, D, K, or N, X_{99} is S, D, K, G, N, or T, X_{100} is I, T, V, or L, X_{101} is T, N, G, or D, X_{102} is D, E, T, K, or I, X_{103} is F or Y, X_{104} is K or Y, X_{105} is F or Y, X_{106} is L, S, or M, X_{107} is V or I, X_{108} is S or A, X_{109} is G or A, X_{110} is F, Y, H, E, or K, X_{111} is Q, K, T, N, or I, X_{112} is D, N, or K, X_{113} is Y, F, I, or V, X_{114} is R, E, K, Q, or F, X_{115} is K, E, A, or N, X_{116} is Q or K, X_{117} is L or I, X_{118} is E, D, N, or Q, X_{119} is V, I, or L, X_{120} is S, N, F, T, or Q, X_{121} is H, I, C, or R, X_{122} is L, D, N, S, or F, X_{123} is T or K, X_{124} is K, G, or N, X_{125} is C, V, or I, X_{126} is Q, L, K, or Y, X_{127} is A, G, or N, X_{128} is V or A, X_{129} is M, L, I, V, or A, X_{130} is S, T, or D, X_{131} is V or I, X_{132} is E, Q, K, S, or I, X_{133} is Q, H, or T, X_{134} is L, R, or Y, X_{135} is G, I, L, or T, X_{136} is G, A, or V, X_{137} is E, N, or D, X_{138} is K, Y, D, E, A, or R, X_{139} is I, F, Y, or C, X_{140} is K or R, X_{141} is E, R, A, G, or T, X_{142} is G or N, X_{143} is S, I, K, R, or E, X_{144} is L, I, or M, X_{145} is T, S, D, or K, X_{146} is L, H, Y, R, T, or F, X_{147} is E, Y, I, M, A, or L, X_{148} is E, D, R, or G, X_{149} is V, F, M, L, or I, X_{150} is G, K, R, L, V, or E, X_{151} is K, N, D, L, H, or S, X_{152} is K, L, C, or absent, X_{153} is K, S, I, Y, M, or F, X_{154} is K, L, C, H, D, Q, or N, X_{155} is N or Y, X_{156} is D, K, T, E, C, or absent, X_{157} is E, V, R, or absent, X_{158} is I, F, L, or absent, X_{159} is V, Q, E, L, or absent, and X_{160} is F or absent.

[0096] In some embodiments, an endonuclease of the present disclosure can have a sequence of $X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10}X_{11}X_{12}X_{13}X_{14}X_{15}X_{16}X_{17}X_{18}X_{19}X_{20}X_{21}X_{22}X_{23}X_{24}X_{25}X_{26}X_{27}X_{28}X_{29}X_{30}X_{31}X_{32}X_{33}X_{34}X_{35}X_{36}X_{37}X_{38}X_{39}X_{40}X_{41}X_{42}X_{43}KX_{44}X_{45}X_{46}X_{47}X_{48}X_{49}X_{50}X_{51}X_{52}X_{53}X_{54}X_{55}GX_{56}HLGGX_{57}RX_{58}PDGX_{59}X_{60}X_{61}X_{62}X_{63}X_{64}X_{65}X_{66}X_{67}X_{68}X_{69}X_{70}X_{71}X_{72}X_{73}X_{74}GX_{75}IX_{76}DTKX_{77}YX_{78}X_{79}GYX_{80}LPIX_{81}QX_{82}DEMX_{83}RYX_{84}X_{85}ENX_{86}X_{87}RX_{88}X_{89}X_{90}X_{91}NX_{92}NX_{93}WWX_{94}X_{95}X_{96}X_{97}X_{98}X_{99}X_{100}X_{101}X_{102}X_{103}X_{104}X_{105}X_{106}FX_{107}X_{108}X_{109}X_{110}FX_{111}GX_{112}X_{113}X_{114}X_{115}X_{116}X_{117}X_{118}RX_{119}X_{120}X_{121}X_{122}X_{123}X_{124}X_{125}X_{126}GX_{127}X_{128}X_{129}X_{130}X_{131}X_{132}X_{133}LLX_{134}X_{135}X_{136}X_{137}X_{138}X_{139}X_{140}X_{141}X_{142}X_{143}X_{144}X_{145}X_{146}X_{147}X_{148}X_{149}X_{150}X_{151}X_{152}X_{153}FX_{154}X_{155}X_{156}X_{157}X_{158}X_{159}X_{160}$ (SEQ ID NO: 317), wherein X_1 is F, Q, N, or absent, X_2 is L, I, T, S, or absent, X_3 is V, I, G, A, E, T, or absent, X_4 is K, C, or absent, X_5 is G, S, or absent, X_6 is A, S, E, D, or absent, X_7 is M, I, V, Q, F, L, or absent, X_8 is E, S, T, or absent, X_9 is I, M, E, T, Q, or absent, X_{10} is K, S, L, I, T, E, or absent, X_{11} is K or absent, X_{12} is S, A, E, D, or absent, X_{13} is E, N, Q, K, or absent, X_{14} is L, M, V, or absent, X_{15} is R or absent, X_{16} is H, D, T, G, E, N, or absent, X_{17} is K, N, Q, E, A, or absent, X_{18} is L or absent, X_{19} is R, Q, N, T, D, or

absent, X_{20} is H, M, V, N, T, or absent, X_{21} is V, L, I, or absent, X_{22} is P, S, or absent, X_{23} is H or absent, X_{24} is E, D, or absent, X_{25} is Y or absent, X_{26} is I, L, or absent, X_{27} is E, Q, G, S, A, or absent, X_{28} is L or absent, X_{29} is I, V, L, or absent, X_{30} is E, D, or absent, X_{31} is I, L, or absent, X_{32} is A, S, or absent, X_{33} is Q, Y, F, or absent, X_{34} is D or absent, X_{35} is S, P, or absent, X_{36} is K, Y, Q, T, or absent, X_{37} is Q or absent, X_{38} is N or absent, X_{39} is R or absent, X_{40} is L, I, or absent, X_{41} is L, F, or absent, X_{42} is E or absent, X_{43} is F, M, L, or absent, X_{44} is V, T, or I, X_{45} is V, M, L, or I, X_{46} is E, D, or Q, X_{47} is F or L, X_{48} is F or L, X_{49} is K, I, T, or V, X_{50} is K, N, or E, X_{51} is I or E, X_{52} is Y, F, or C, X_{53} is G, or N, X_{54} is Y, or F, X_{55} is R, S, N, E, K, or Q, X_{56} is K, S, L, V, or T, X_{57} is S or A, X_{58} is K or R, X_{59} is A, I, or V, X_{60} is L, M, V, I, or C, X_{61} is F or Y, X_{62} is T, A, or S, X_{63} is K, E, or absent, X_{64} is D, E, or absent, X_{65} is E, A, or absent, X_{66} is N, K, or absent, X_{67} is E, S, or absent, X_{68} is D, E, Q, A, or absent, X_{69} is G, V, K, N, or absent, X_{70} is L, G, E, S, or absent, X_{71} is V, S, K, T, E, or absent, X_{72} is L, H, K, E, Y, D, or A, X_{73} is N, G, or D, X_{74} is H, F, or Y, X_{75} is I, or V, X_{76} is L, V, or I, X_{77} is A or S, X_{78} is K or S, X_{79} is D, G, K, S, or N, X_{80} is R, N, S, or G, X_{81} is S, A, or G, X_{82} is A, I, or V, X_{83} is Q, E, I, or V, X_{84} is V or I, X_{85} is D, R, G, I, or E, X_{86} is N, I, or Q, X_{87} is K, D, T, E, or K, X_{88} is S, N, D, or E, X_{89} is Q, E, I, K, or A, X_{90} is V, H, R, K, L, or E, X_{91} is I, V, or R, X_{92} is P, S, T, or R, X_{93} is E, R, C, Q, or K, X_{94} is E, N, or K, X_{95} is I, V, N, E, or A, X_{96} is Y or F, X_{97} is P, G, or E, X_{98} is T, E, S, D, K, or N, X_{99} is S, D, K, G, N, or T, X_{100} is I, T, V, or L, X_{101} is T, N, G, or D, X_{102} is D, E, T, K, or I, X_{103} is F or Y, X_{104} is K or Y, X_{105} is F or Y, X_{106} is L, S, or M, X_{107} is V or I, X_{108} is S or A, X_{109} is G or A, X_{110} is F, Y, H, E, or K, X_{111} is Q, K, T, N, or I, X_{112} is D, N, or K, X_{113} is Y, F, I, or V, X_{114} is R, E, K, Q, or F, X_{115} is K, E, A, or N, X_{116} is Q or K, X_{117} is L or I, X_{118} is E, D, N, or Q, X_{119} is V, I, or L, X_{120} is S, N, F, T, or Q, X_{121} is H, I, C, or R, X_{122} is L, D, N, S, or F, X_{123} is T or K, X_{124} is K, G, or N, X_{125} is C, V, or I, X_{126} is Q, L, K, or Y, X_{127} is A, G, or N, X_{128} is V or A, X_{129} is M, L, I, V, or A, X_{130} is S, T, or D, X_{131} is V or I, X_{132} is E, Q, K, S, or I, X_{133} is Q, H, or T, X_{134} is L, R, or Y, X_{135} is G, I, L, or T, X_{136} is G, A, or V, X_{137} is E, N, or D, X_{138} is K, Y, D, E, A, or R, X_{139} is I, F, Y, or C, X_{140} is K or R, X_{141} is E, R, A, G, or T, X_{142} is G or N, X_{143} is S, I, K, R, or E, X_{144} is L, I, or M, X_{145} is T, S, D, or K, X_{146} is L, H, Y, R, or T, X_{147} is E, Y, I, M, or A, X_{148} is E, D, R, or G, X_{149} is V, F, M, L, or I, X_{150} is G, K, R, L, V, or E, X_{151} is K, N, D, L, H, or S, X_{152} is K, L, C, or absent, X_{153} is K, S, I, Y, M, or F, X_{154} is K, L, C, H, D, Q, or N, X_{155} is N or Y, X_{156} is D, K, T, E, C, or absent, X_{157} is E, V, R, or absent, X_{158} is I, F, L, or absent, X_{159} is V, Q, E, L, or absent, and X_{160} is F or absent.

[0097] In some embodiments, an endonuclease of the present disclosure can have a sequence of X_1LVKSSX_2EEX_3KEELREKLX_4HLSHEYLX_5LX_6DLAYDSKQNRLFEMKVX_7ELLINECGYX_8G LHLGGSRKPDGIX_9YTEGLKX_{10}NYGIIIDTKAYSDGYNLPISQADEMERYIRENNTRN$X_{11}X_{12}$V

NPNEWWENFPX$_{13}$NINEFYFLFVSGHFKGNX$_{14}$EEQLERISIX$_{15}$TX$_{16}$IKGAAMSVX$_{17}$TLLLLAN EIKAGRLX$_{18}$LEEVX$_{19}$KYFDNKEIX$_{20}$F (SEQ ID NO: 318), wherein X$_1$ is F, Q, N, D, or absent, X$_2$ is M, I, V, Q, F, L, or absent, X$_3$ is K, S, L, I, T, E, or absent, X$_4$ is R, Q, N, T, D, or absent, X$_5$ is E, Q, G, S, A, Y, or absent, X$_6$ is I, V, L, or absent, X$_7$ is V, M, L, or I, X$_8$ is R, S, N, E, K, or Q, X$_9$ is L, M, V, I, or C, X$_{10}$ is L, H, K, E, Y, D, or A, X$_{11}$ is Q, E, I, K, or A, X$_{12}$ is V, H, R, K, L, or E, X$_{13}$ is T, E, S, D, K, or N, X$_{14}$ is Y, F, I, or V, X$_{15}$ is L, D, N, S, or F, X$_{16}$ is K, G, or N, X$_{17}$ is E, Q, K, S, or I, X$_{18}$ is T, S, D, or K, X$_{19}$ is G, K, R, L, V, or E, and X$_{20}$ is V, Q, E, L, or absent.

[0098] In some embodiments, an endonuclease of the present disclosure can have a sequence of X$_1$LVKSSX$_2$EEX$_3$KEELREKLX$_4$HLSHEYLX$_5$LX$_6$DLAYDSKQNRLFEMKVX$_7$ELLINECGYX$_8$G LHLGGSRKPDGIX$_9$YTEGLKX$_{10}$NYGIIIDTKAYSDGYNLPISQADEMERYIRENNTRNX$_{11}$X$_{12}$V NPNEWWENFPX$_{13}$NINEFYFLFVSGHFKGNX$_{14}$EEQLERISIX$_{15}$TX$_{16}$IKGAAMSVX$_{17}$TLLLLAN EIKAGRLX$_{18}$LEEVX$_{19}$KYFDNKEIX$_{20}$F (SEQ ID NO: 319), wherein X$_1$ is F, Q, N, or absent, X$_2$ is M, I, V, Q, F, L, or absent, X$_3$ is K, S, L, I, T, E, or absent, X$_4$ is R, Q, N, T, D, or absent, X$_5$ is E, Q, G, S, A, or absent, X$_6$ is I, V, L, or absent, X$_7$ is V, M, L, or I, X$_8$ is R, S, N, E, K, or Q, X$_9$ is L, M, V, I, or C, X$_{10}$ is L, H, K, E, Y, D, or A, X$_{11}$ is Q, E, I, K, or A, X$_{12}$ is V, H, R, K, L, or E, X$_{13}$ is T, E, S, D, K, or N, X$_{14}$ is Y, F, I, or V, X$_{15}$ is L, D, N, S, or F, X$_{16}$ is K, G, or N, X$_{17}$ is E, Q, K, S, or I, X$_{18}$ is T, S, D, or K, X$_{19}$ is G, K, R, L, V, or E, and X$_{20}$ is V, Q, E, L, or absent. In some aspects, the cleavage domain comprises a sequence selected from SEQ ID NO: 316 – SEQ ID NO: 319.

[0099] In some embodiments, an endonuclease of the present disclosure can have conserved amino acid residues at position 76 (D or E), position 98 (D), and position 100 (K), which together preserve catalytic function. In some embodiments, an endonuclease of the present disclosure can have conserved amino acid residues at position 114 (D) and position 118 (R), which together preserve dimerization of two cleavage domains.

[0100] In some embodiments, endonucleases disclosed herein (e.g., SEQ ID NO: 1 – SEQ ID NO: 81 (nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162)) can have at least 33.3% divergence from SEQ ID NO: 163 (FokI) and, is immunologically orthogonal to SEQ ID NO: 163 (FokI). In some embodiments, an immunologically orthogonal endonuclease (e.g., SEQ ID NO: 1 – SEQ ID NO: 81 (nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162)) can be administered to a patient that has already received, and is thus can have an adverse immune reaction to, FokI. In some embodiments, endonucleases disclosed herein (e.g., SEQ ID NO: 1 – SEQ ID NO: 81 (nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162)) can have at least 35%, at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, or at least 75% divergence from SEQ ID NO: 163 (FokI).

[0101] In some embodiments, an endonuclease disclosed herein (e.g., SEQ ID NO: 1 – SEQ ID NO: 81 (nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162)) can be fused to any nucleic acid binding domain disclosed herein to form a non-naturally occurring fusion protein. This fusion protein can have one or more of the following characteristics: (a) induces greater than 1% indels (insertions/deletions) at a target site; (b) the cleavage domain comprises a molecular weight of less than 23 kDa; (c) the cleavage domain comprises less than 196 amino acids; and (d) capable of cleaving across a spacer region greater than 24 base pairs. In some embodiments, the non-naturally occurring fusion protein can induce greater than 5%, greater than 10%, greater than 20%, greater than 30%, greater than 40%, greater than 50%, greater than 60%, greater than 70%, greater than 80%, or greater than 90% indels at the target site. In some embodiments, indels are generated via the non-homologous end joining (NHEJ) pathway upon administration of a genome editing complex disclosed herein to a subject. Indels can be measured using deep sequencing,

DNA Binding Domains Fused to SEQ ID NO: 1 – SEQ ID NO: 81 (Nucleic Acid Sequences of SEQ ID NO: 82 – SEQ ID NO: 162)

[0102] The present disclosure provides for novel compositions of endonucleases with modular nucleic acid binding domains (e.g., TALEs, RNBDs, or MAP-NBDs) described herein. In some instances the novel endonucleases can be fused to a DNA binding domain from *Xanthomonas spp.* (TALE), Ralstonia (RNBD), or an animal pathogen (MAP-NBD) resulting in genome editing complexes. A TALEN, RNBD-nuclease, or MAP-NBD-nuclease can include multiple components including the DNA binding domain, an optional linker, and a repressor domain. The genome editing complexes described herein can be used to selectively bind and cleave to a target gene sequence for genome editing purposes. For example, a DNA binding domain from *Xanthomonas, Ralstonia,* or an animal pathogen of the present disclosure can be used to direct the binding of a genome editing complex to a desired genomic sequence.

[0103] The genome editing complexes described herein, comprising a DNA binding domain fused to an endonuclease, can be used to edit genomic loci of interest by binding to a target nucleic acid sequence via the DNA binding domain and cleaving phosphodiester bonds of target double stranded DNA via the endonuclease.

[0104] In some aspects, DNA binding domains fused to nucleases can create a site-specific double-stranded DNA break when fused to a nuclease. Such breaks can then be subsequently repaired by cellular machinery, through either homology-dependent repair or non-homologous end joining (NHEJ). Genome editing, using DNA binding domains fused to nucleases described herein,, can thus

be used to delete a sequence of interest (e.g., an aberrantly expressed or mutated gene) or to introduce a nucleic acid sequence of interest (e.g., a functional gene). DNA binding domains of the present disclosure can be programmed to delivery virtually any nuclease, including those disclosed herein, to any target site for therapeutic purposes, including ex vivo engineered cell therapies obtained using the compositions disclosed herein or gene therapy by direct in vivo administration of the compositions disclosed herein. In addition, the DNA binding domain can bind to specific DNA sequences and in some cases they can activate the expression of host genes. In some instances, the disclosure provides for enzymes, *e.g.*, SEQ ID NO: 1 – SEQ ID NO: 81 (or any one of nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162) that can be fused to the DNA binding domains of TALEs, RNBDs, and MAP-NBDs. In some instances, enzymes of the disclosure, including SEQ ID NO: 1 (nucleic acid sequence of SEQ ID NO: 82), SEQ ID NO: 4 (nucleic acid sequence of SEQ ID NO: 85), and SEQ ID NO: 8 (nucleic acid sequence of SEQ ID NO: 89), can achieve greater than 30% indels via the NHEJ pathway on a target gene when fused to a DNA binding domain of a TALE, RNBD, and MAP-NBD.

[0105] A non-naturally occurring fusion protein of the disclosure, *e.g.*, any one of SEQ ID NO: 1 – SEQ ID NO: 81 (or any one of nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162) fused to a DNA binding domain, can comprise a repeat unit. A repeat unit can be from a wild-type DNA-binding domain (*Ralstonia solanacearum, Xanthomonas spp.*, or an animal pathogen) or a modified repeat unit enhanced for specific recognition of a particular nucleic acid base. A modified repeat unit can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25 or more mutations that can enhance the repeat module for specific recognition of a particular nucleic acid base. In some embodiments, a modified repeat unit is modified at amino acid position 2, 3, 4, 11, 12, 13, 21, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, or 35. In some embodiments, a modified repeat unit is modified at amino acid positions 12 or 13.

[0106] As described in further detail below, a non-naturally occurring fusion protein of the disclosure, *e.g.*, anyone of SEQ ID NO: 1 – SEQ ID NO: 81 (or any one of nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162) fused to a plurality of repeat units (e.g., derived from *Ralstonia solanacearum, Xanthomonas spp.*, or an animal pathogen), can further comprise a C-terminal truncation, which can served as a linker between the DNA binding domain and the nuclease.

[0107] A non-naturally occurring fusion protein of the disclosure, *e.g.*, anyone of SEQ ID NO: 1 – SEQ ID NO: 81 (or any one of nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162) fused to a DNA binding domain, can further comprise an N-terminal cap as described in further detail below. An N-terminal cap can be a polypeptide portion flanking the DNA-binding repeat unit. An N-

terminal cap can be any length and can comprise from 0 to 136 amino acid residues in length. An N-terminal cap can be 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 110, 120, or 130 amino acid residues in length. In some embodiments, an N-terminal cap can modulate structural stability of the DNA-binding repeat units. In some embodiments, an N-terminal cap can modulate nonspecific interactions. In some cases, an N-terminal cap can decrease nonspecific interaction. In some cases, an N-terminal cap can reduce off-target effect. As used here, off-target effect refers to the interaction of a genome editing complex with a sequence that is not the target binding site of interest. An N-terminal cap can further comprise a wild-type N-terminal cap sequence of a protein from *Ralstonia solanacearum, Xanthomonas spp.*, or an animal pathogen or can comprise a modified N-terminal cap sequence.

[0108] In some embodiments, a DNA binding domain comprises at least one repeat unit having a repeat variable diresidue (RVD), which contacts a target nucleic acid base. In some embodiments, a DNA binding domain comprises more than one repeat unit, each having an RVD, which contacts a target nucleic acid base. In some embodiments, the DNA binding domain comprises 1 to 50 RVDs. In some embodiments, the DNA binding domain components of the fusion proteins can be at least 14 RVDs, at least 15 RVDs, at least 16 RVDs, at least 17 RVDs, at least 18 RVDs, at least 19 RVDs, at least 20 RVDs in length, or at least 21 RVDs in length. In some embodiments, the DNA binding domains can be 16 to 21 RVDs in length.

[0109] In some embodiments, any one of the DNA binding domains described herein can bind to a region of interest of any gene. For example, the DNA binding domains described herein can bind upstream of the promoter region, upstream of the gene transcription start site, or downstream of the transcription start site. In certain embodiments, the DNA binding domain binding region is no farther than 50 base pairs downstream of the transcription start site. In some embodiments, the DNA binding domain is designed to bind in proximity to the transcription start site (TSS). In other embodiments, the TALE can be designed to bind in the 5' UTR region.

[0110] A DNA binding domain described herein can comprise between 1 to 50 repeat units. A DNA binding domain described herein can comprise between 5 and 45, between 8 to 45, between 10 to 40, between 12 to 35, between 15 to 30, between 20 to 30, between 8 to 40, between 8 to 35, between 8 to 30, between 10 to 35, between 10 to 30, between 10 to 25, between 10 to 20, or between 15 to 25 repeat units.

[0111] A DNA binding domain described herein can comprise at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 45, 50, or more repeat units. A DNA binding domain described herein can comprise 1, 2, 3,

4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 45, or 50 repeat units. A DNA binding domain described herein can comprise 5 repeat units. A DNA binding domain described herein can comprise 10 repeat units. A DNA binding domain described herein can comprise 11 repeat units. A DNA binding domain described herein can comprise 12 repeat units, or another suitable number.

[0112] A repeat unit of a DNA binding domain can be 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36 37, 38, 39 or 40 residues in length.

[0113] In some embodiments, the effector can be a protein secreted from Xanthomonas or Ralstonia bacteria upon plant infection. In some embodiments, the effector can be a protein that is a mutated form of, or otherwise derived from, a protein secreted from Xanthomonas or Ralstonia bacteria. The effector can further comprise a DNA-binding module which includes a variable number of about 33-35 amino acid residue repeat units. Each amino acid repeat unit recognizes one base pair through two adjacent amino acids (*e.g.*, at amino acid positions 12 and 13 of the repeat unit). As such, amino acid positions 12 and 13 of the repeat unit can also be referred to as repeat variable diresidue (RVD).

Linkers

[0114] A nuclease, *e.g.*, anyone of SEQ ID NO: 1 – SEQ ID NO: 81 (or any one of nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162) fused to a DNA binding domain (e.g., an RNBD, a MAP-NBD, a TALE), can further include a linker connecting SEQ ID NO: 1 – SEQ ID NO: 81 (or any one of nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162) to the DNA binding domain. A linker used herein can be a short flexible linker comprising 0 base pairs, 3 to 6 base pairs, 6 to 12 base pairs, 12 to 15 base pairs, 15 to 21 base pairs, 21 to 24 base pairs, 24 to 30 base pairs, 30 to 36 base pairs, 36 to 42 base pairs, 42 to 48 base pairs, or 1-48 base pairs. The nucleic acid sequence of the linker can encode for an amino acid sequence comprising 0 residues, 1-3 residues, 4-7 residues, 8-10 residues, 10-12 residues, 12-15 residues, or 1-15 residues. Linkers can include, but are not limited to, residues such as glycine, methionine, aspartic acid, alanine, lysine, serine, leucine, threonine, tryptophan, or any combination thereof.

[0115] When linking a repressor domain to an RNBD, MAP-NBD, or TALE, the linker can have a nucleic acid sequence of GGCGGTGGCGGAGGGATGGATGCTAAGTCACTAACTGCCTGGTCC (SEQ ID NO: 165) and an amino acid sequence of GGGGGMDAKSLTAWS (SEQ ID NO: 166).

[0116] A nuclease, *e.g.*, anyone of SEQ ID NO: 1 – SEQ ID NO: 81 (or any one of nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162) can be connected to a DNA binding domain via a linker, a linker can be between 1 to 70 amino acid residues in length. A linker can be from 5 to 45, from 5 to 40, from 5 to 35, from 5 to 30, from 5 to 25, from 5 to 20, from 5 to 15, from 10 to 40, from 10 to 35, from 10 to 30, from 10 to 25, from 10 to 20, from 12 to 40, from 12 to 35, from 12 to 30, from 12 to 25, from 12 to 20, from 14 to 40, from 14 to 35, from 14 to 30, from 14 to 25, from 14 to 20, from 14 to 16, from 15 to 40, from 15 to 35, from 15 to 30, from 15 to 25, from 15 to 20, from 15 to 18, from 18 to 40, from 18 to 35, from 18 to 30, from 18 to 25, from 18 to 24, from 20 to 40, from 20 to 35, from 20 to 30, from 25 to 30, from 25 to 70, from 30 to 70, from 5 to 70, from 35 to 70, from 40 to 70, from 45 to 70, from 50 to 70, from 55 to 70, from 60 to 70, or from 65 to 70 amino acid residues in length.

[0117] A linker for linking a nuclease, *e.g.*, anyone of SEQ ID NO: 1 – SEQ ID NO: 81 (or any one of nucleic acid sequences of SEQ ID NO: 82 – SEQ ID NO: 162) to a DNA binding domain can be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 55, 60, 65, or 70 amino acid residues in length.

[0118] In some embodiments, the linker can be the N-terminus of a naturally occurring *Ralstonia solanacearum*-derived protein, *Xanthomonas spp.*-derived protein, or *Legionella quateirensis*-derived protein, wherein any functional domain disclosed herein is fused to the N-terminus of the engineered DNA binding domain. In some embodiments, the linker comprising the N-terminus can comprise the full length naturally occurring N-terminus of a naturally occurring *Ralstonia solanacearum*-derived protein, *Xanthomonas spp.*-derived protein, or *Legionella quateirensis*-derived protein, or a truncation of the naturally occurring N-terminus, such as amino acid residues at positions 1 to 137 of the naturally occurring *Ralstonia solanacearum*-derived protein N-terminus (e.g., SEQ ID NO: 264), positions 1 (H) to 115 (S) of the naturally occurring *Ralstonia solanacearum*-derived protein N-terminus (SEQ ID NO: 320), positions 1 (N) to 115 (S) of the naturally occurring *Xanthomonas spp.*-derived protein N-terminus (SEQ ID NO: 321), or positions 1 (G) to 115 (K) of the naturally occurring *Legionella quateirensis*-derived protein N-terminus (SEQ ID NO: 322). In some embodiments, the linker can comprise amino acid residues at positions 1 to 120 of the naturally occurring *Ralstonia solanacearum*-derived protein (SEQ ID NO: 303), *Xanthomonas spp.*-derived protein (SEQ ID NO: 301), or *Legionella quateirensis*-derived protein (SEQ ID N): 304). In some embodiments, the linker can comprise the naturally occurring N-terminus of *Ralstonia solanacearum* truncated to any length. For example, the naturally occurring N-terminus of *Ralstonia solanacearum* can be truncated to amino acid residues at positions 1 to 120, 1

to 115, 1 to 50, 1 to 70, 1 to 100, 1 to 120, 1 to 130, 10 to 40, 60 to 100, or 100 to 120 and used at the N-terminus of the engineered DNA binding domain as a linker to a nuclease or a repressor.

[0119] In other embodiments, the linker can be the C-terminus of a naturally occurring *Ralstonia solanacearum*-derived protein, *Xanthomonas spp.*-derived protein, or animal pathogen-derived protein, wherein any functional domain disclosed herein is fused to the C-terminus of the engineered DNA binding domain. In some embodiments, the linker comprising the C-terminus can comprise the full length naturally occurring C-terminus of a naturally occurring *Ralstonia solanacearum*-derived protein, *Xanthomonas spp.*-derived protein, or animal pathogen-derived protein, or a truncation of the naturally occurring C-terminus, such as positions 1 to 63 of the naturally occurring *Ralstonia solanacearum*-derived protein (SEQ ID NO: 266), *Xanthomonas spp.*-derived protein (SEQ ID NO: 298), or *Legionella quateirensis*-derived protein (SEQ ID NO: 306). In some embodiments, the naturally occurring C-terminus of *Ralstonia solanacearum*-derived protein, *Xanthomonas spp.*-derived protein, or *Legionella quateirensis*-derived protein can be truncated to any length and used at the C-terminus of the engineered DNA binding domain and used as a linker to a nuclease or repressor. For example, the naturally occurring C-terminus of *Ralstonia solanacearum*-derived protein, *Xanthomonas spp.*-derived protein, or *Legionella quateirensis*-derived protein can be truncated to amino acid residues at positions 1 to 63, 1 to 50, 1 to 70, 1 to 100, 1 to 120, 1 to 130, 10 to 40, 60 to 100, or 100 to 120 and used at the C-terminus of the engineered DNA binding domain.

Functional Domains

[0120] An RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be linked to a functional domain. The functional domain can provide different types activity, such as genome editing, gene regulation (*e.g.*, activation or repression), or visualization of a genomic locus via imaging.

A. Genome Editing Domains

[0121] For example, an RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be linked to a nuclease, wherein the RNBD provides specificity and targeting and the nuclease provides genome editing functionality. In some embodiments, the nuclease can be a cleavage domain, which dimerizes with another copy of the same cleavage domain to form an active full domain capable of cleaving DNA. In other embodiments, the nuclease can be a cleavage domain, which is capable of cleaving DNA without needing to dimerize. For example, a nuclease comprising a cleavage domain can be an endonuclease, such as FokI or Bfil. In some embodiments, two cleavage domains (e.g., FokI or Bfil) can be fused together to form a fully

functional single cleavage domain. When cleavage domains are used as the nuclease, two RNBDs can be engineered, the first RNBD binding to a top strand of a target nucleic acid sequence and comprising a first FokI cleavage domain and a second RNBD binding to a bottom strand of a target nucleic acid sequence and comprising a second FokI cleavage domain.

[0122] In some embodiments, a fully functional cleavage domain, capable of cleaving DNA without needing to dimerize include meganucleases, also referred to as homing endonucleases. For example, a meganuclease can include I-AniI or I-OnuI. In some embodiments, the nuclease can be a type IIS restriction enzyme, such as FokI or BfiI.

[0123] A nuclease domain fused to an RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be an endonuclease or an exonuclease. An endonuclease can include restriction endonucleases and homing endonucleases. An endonuclease can also include S1 Nuclease, mung bean nuclease, pancreatic DNase I, micrococcal nuclease, or yeast HO endonuclease. An exonuclease can include a 3'-5' exonuclease or a 5'-3' exonuclease. An exonuclease can also include a DNA exonuclease or an RNA exonuclease. Examples of exonuclease includes exonucleases I, II, III, IV, V, and VIII; DNA polymerase I, RNA exonuclease 2, and the like.

[0124] A nuclease domain fused to an RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be a restriction endonuclease (or restriction enzyme). In some instances, a restriction enzyme cleaves DNA at a site removed from the recognition site and has a separate binding and cleavage domains. In some instances, such restriction enzyme is a Type IIS restriction enzyme.

[0125] A nuclease domain fused to an RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be a Type IIS nuclease. A Type IIS nuclease can be FokI or BfiI. In some cases, a nuclease domain fused to an RNBD (e.g., *Ralstonia solanacearum*-derived) is FokI. In other cases, a nuclease domain fused to an RNBD (e.g., *Ralstonia solanacearum*-derived) is BfiI.

[0126] FokI can be a wild-type FokI or can comprise one or more mutations. In some cases, FokI can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more mutations. A mutation can enhance cleavage efficiency. A mutation can abolish cleavage activity. In some cases, a mutation can modulate homodimerization. For example, FokI can have a mutation at one or more amino acid residue positions 446, 447, 479, 483, 484, 486, 487, 490, 491, 496, 498, 499, 500, 531, 534, 537, and 538 to modulate homodimerization.

[0127] In some instances, a FokI cleavage domain is, for example, as described in Kim et al. "Hybrid restriction enzymes: Zinc finger fusions to Fok I cleavage domain," PNAS 93: 1156-1160 (1996), which is incorporated herein by reference in its entirety. In some cases, a FokI cleavage domain described herein has a sequence as follows: QLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLG GSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWW KVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEE VRRKFNNGEINF (SEQ ID NO: 163). In other instances, a FokI cleavage domain described herein is a FokI, for example, as described in U.S. Patent No. 8,586,526, which is incorporated herein by reference in its entirety.

[0128] An RNBD (e.g., *Ralstonia solanacearum*-derived) can be linked to a functional group that modifies DNA nucleotides, for example an adenosine deaminase.

[0129] In some embodiments, an RNBD (e.g., *Ralstonia solanacearum*-derived) can be linked to any nuclease as set forth in **TABLE 7** showing exemplary amino acid sequences (SEQ ID NO: 1 – SEQ ID NO: 81) of endonucleases for genome editing and the corresponding back-translated nucleic acid sequences (SEQ ID NO: 82 – SEQ ID NO: 162) of the endonucleases.

[0130] For purposes of gene editing, a first DNA binding domain (e.g., of a TALE, RNBD, or MAP-NBD) linked to a cleavage domain and a second DNA binding domain (e.g., of a TALE, RNBD, or MAP-NBD) linked to a cleavage domain can be provided. The first DNA binding domain (e.g., of a TALE, RNBD, or MAP-NBD) linked to a cleavage domain can recognize a top strand of double stranded DNA and bind to said region of double stranded DNA. The second DNA binding domain (e.g., of a TALE, RNBD, or MAP-NBD) linked to a cleavage domain can recognize a separate, non-overlapping bottom strand of double stranded DNA and bind to said region of double stranded DNA. The target nucleic acid sequence on the bottom strand can have its complementary nucleic acid sequence in the top strand positioned 10 to 20 nucleotides towards the 3' end from the first region. In some embodiments this stretch of 10 to 20 nucleotides can be referred to as the spacer region. In some embodiments, this first DNA binding domain (e.g., of a TALE, RNBD, or MAP-NBD) linked to a cleavage domain and the second DNA binding domain (e.g., of a TALE, RNBD, or MAP-NBD) linked to a cleavage domain both bind at a target site, allowing for dimerization of the two cleavage domains in the spacer region and allowing for catalytic activity and cleaving of the target DNA.

a. Potency and Specificity of Genome Editing

[0131] In some embodiments, the efficiency of genome editing with a genome editing complex of the present disclosure (e.g., any one of an RNBD, MAP-NBD, or TALE fused to any nuclease disclosed herein) can be determined. Specifically, the potency and specificity of the genome editing complex can indicate whether a particular modular nucleic acid binding domain fused to a nuclease provides efficient editing. Potency can be defined as the percent indels (insertions/deletions) that are generated via the non-homologous end joining (NHEJ) pathway at a target site after administering a modular nucleic acid binding domain fused to a nuclease to a subject. A modular nucleic acid binding domain can have a potency of greater than 50%, greater than 55%, greater than 60%, greater than 65%, greater than 70%, greater than 75%, greater than 80%, greater than 85%, greater than 90%, greater than 95%, greater than 92%, greater than 95%, greater than 97%, or greater than 99%. A modular nucleic acid binding domain can have a potency of from 50% to 100%, 50% to 60%, 60% to 70%, 70% to 80%, 80% to 90%, or 90% to 100%.

[0132] Specificity can be defined as a specificity ratio, wherein the ratio is the percent indels at a target site of interest over the percent indels at the top-ranked off-target site for a particular genome editing complex (e.g., any DNA binding domain linked to a nuclease described herein) of interest. A high specificity ratio would indicate that a modular nucleic acid binding domain fused to a nuclease edits primarily at the desired target site and exhibits fewer instances of undesirable, off-target editing. A low specificity ratio would indicate that a modular nucleic acid binding domain fused to a nuclease does not edit efficiently at the desired target site and/or can indicate that the modular nucleic acid binding domain fused to a nuclease exhibits high off-target activity. A modular nucleic acid binding domain can have a specificity ratio for the target site of at least 50:1, 55:1, 60:1, 65:1, 70:1, 75:1, 80:1, 85:1, 90:1, 92:1, 95:1, 97:1, 99:1, 50:2, 55:2, 60:2, 65:2, 70:2, 75:2, 80:2, 85:2, 90:2, 92:2, 95:2, 97:2, 99:2, 50:3, 55:3, 60:3, 65:3, 70:3, 75:3, 80:3, 85:3, 90:3, 92:3, 95:3, 97:3, 99:3, 50:4, 55:4, 60:4, 65:4, 70:4, 75:4, 80:4, 85:4, 90:4, 92:4, 95:4, 97:4, 99:4, 50:5, 55:5, 60:5, 65:5, 70:5, 75:5, 80:5, 85:5, 90:5, 92:5, 95:5, 97:5, or 99:5. Percent indels generated via non-homologous end joining (NHEJ) can be measured via deep sequencing techniques.

[0133] In some embodiments, the composition further comprises a cleavage domain linked to the modular nucleic acid binding domain to form a non-naturally occurring fusion protein. In some aspects, the modular nucleic acid binding domain comprises a potency for a target site greater than 65% and a specificity ratio for the target site of 50:1; and a functional domain; wherein the modular nucleic acid binding domain comprises a plurality of repeat units, wherein at least one repeat unit of the plurality comprises a binding region configured to bind to a target nucleic acid base in the target

site, wherein the potency comprises indel percentage at the target site, and wherein the specificity ratio comprises indel percentage at the target site over indel percentage at a top-ranked off-target site of the non-naturally occurring fusion protein.

[0134] In some embodiments, the repeat unit comprises a sequence of $A_{1-11}X_1X_2B_{14-35}$ (SEQ ID NO: 448), wherein each amino acid residue of A_{1-11} comprises any amino acid residue; wherein X_1X_2 comprises the binding region; wherein each amino acid residue of B_{14-35} comprises any amino acid; and wherein a first repeat unit of the plurality of repeat units comprises at least one residue in A_{1-11}, B_{14-35}, or a combination thereof that differs from a corresponding residue in a second repeat unit of the plurality of repeat units.

[0135] In some embodiments, the binding region comprises an amino acid residue at position 13 or an amino acid residue at position 12 and the amino acid residue at position 13. In further aspects, the amino acid residue at position 13 binds to the target nucleic acid base. In still further aspects, the amino acid residue at position 12 stabilizes the configuration of the binding region. In some aspects, the indel percentage is measured by deep sequencing. In some aspects, the modular nucleic acid binding domain further comprises one or more properties selected from the following: (a) binds the target site, wherein the target site comprises a 5' guanine; (b) comprises from 7 repeat units to 25 repeat units; and (c) upon binding to the target site, the modular nucleic acid binding domain is separated from a second modular nucleic acid binding domain bound to a second target site by from 2 to 50 base pairs.

[0136] The top-ranked off-target site for a composition (e.g., a modular nucleic acid binding domain linked to a cleavage domain) can be determined using the predicted report of genome-wide nuclease off-target sites (PROGNOS) ranking algorithms as described in Fine et al. (Nucleic Acids Res. 2014 Apr;42(6):e42. doi: 10.1093/nar/gkt1326. Epub 2013 Dec 30.). As described in Fine et al, the PROGNOS algorithm TALEN v2.0 can use the DNA target sequence as input; prior construction and experimental characterization of the specific nucleases are not necessary. Based on the differences between the sequence of a potential off-target site in the genome and the intended target sequence, the algorithm can generate a score that is used to rank potential off-target sites. If two (or more) potential off-target sites have equal scores, they can be further ranked by the type of genomic region annotated for each site with the following order: Exon > Promoter > Intron > Intergenic. A final ranking by chromosomal location can be employed as a tie-breaker to ensure consistency in the ranking order. Thus, a score can be generated for each potential off-target site.

B. Regulatory Domains

[0137] As another example, an RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be linked to a gene regulating domain. A gene regulation domain can be an activator or a repressor. For example, an RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be linked to an activation domain, such as VP16, VP64, p65, p300 catalytic domain, TET1 catalytic domain, TDG, Ldb1 self-associated domain, SAM activator (VP64, p65, HSF1), or VPR (VP64, p65, Rta). Alternatively, an RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be linked to a repressor, such as KRAB, Sin3a, LSD1, SUV39H1, G9A (EHMT2), DNMT1, DNMT3A-DNMT3L, DNMT3B, KOX, TGF-beta-inducible early gene (TIEG), v-erbA, SID, MBD2, MBD3, Rb, or MeCP2.

[0138] In some embodiments, an RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be linked to a DNA modifying protein, such as DNMT3a. An RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be linked to a chromatin-modifying protein, such as lysine-specific histone demethylase 1 (LSD1). An RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be linked to a protein that is capable of recruiting other proteins, such as KRAB. The DNA modifying protein (e.g., DNMT3a) and proteins capable of recruiting other proteins (e.g., KRAB) can serve as repressors of transcription. Thus, RNBDs (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), linked to a DNA modifying protein (e.g., DNMT3a) or a domain capable of recruiting other proteins (e.g., KRAB, a domain found in transcriptional repressors, such as Kox1) can provide gene repression functionality, can serve as transcription factors, wherein the RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), provides specificity and targeting and the DNA modifying protein and the protein capable of recruiting other proteins provides gene repression functionality, which can be referred to as a TALE-transcription factor (TALE-TF), RNBD-transcription factor (RNBD-TF), or MAP-NBD-transcription factor (MAP-NBD-TF).

[0139] In some embodiments, expression of the target gene can be reduced by at least 5%, at least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 92%, at least 95%, at least 97%, or at least 99% by using a DNA binding domain fused to a repression domain (*e.g.*, an RNBD-TF, a MAP-NBD-TF, or TALE-TF) of the present disclosure as compared to non-treated cells. In some embodiments, expression of the

target gene can be reduced by 5% to 10%, 10% to 15%, 15% to 20%, 20%, to 25%, 25% to 30%, 30% to 35%, 35% to 40%, 40% to 45%, 45% to 50%, 50% to 55%, 55% to 60%, 60% to 65%, 65% to 70%, 70% to 75%, 75% to 80%, 80% to 85%, 85% to 90%, 90% to 95%, or 95% to 99% by using an RNBD-TF, a MAP-NBD-TF, or TALE-TF of the present disclosure as compared to non-treated cells. In some embodiments, expression of the checkpoint gene can be reduced by over 90% by using an RNBD-TF, a MAP-NBD-TF, or TALE-TF of the present disclosure as compared to non-treated cells.

[0140] In some embodiments, repression of the target gene with a DNA binding domain fused to a repression domain (*e.g.,* an RNBD-TF, a MAP-NBD-TF, or TALE-TF) of the present disclosure and subsequent reduced expression of the target gene can last for at least 1 day, at least 2 days, at least 3 days, at least 4 days, at least 5 days, at least 6 days, at least 7 days, at least 8 days, at least 9 days, at least 10 days, at least 11 days, at least 12 days, at least 13 days, at least 14 days, at least 15 days, at least 16 days, at least 17 days, at least 18 days, at least 19 days, at least 20 days, at least 21 days, at least 22 days, at least 23 days, at least 24 days, at least 25 days, at least 26 days, at least 27 days, or at least 28 days. In some embodiments, repression of the target gene with an RNBD-TF, a MAP-NBD-TF, or TALE-TF of the present disclosure and subsequent reduced expression of the target gene can last for 1 days to 3 days, 3 days to 5 days, 5 days to 7 days, 7 days to 9 days, 9 days to 11 days, 11 days to 13 days, 13 days to 15 days, 15 days to 17 days, 17 days to 19 days, 19 days to 21 days, 21 days to 23 days, 23 days to 25 days, or 25 days to 28 days.

[0141] In various aspects, the present disclosure provides a method of identifying a target binding site in a target gene of a cell, the method comprising: (a) contacting a cell with an engineered genomic regulatory complex comprising a DNA binding domain, a repressor domain, and a linker; (b) measuring expression of the target gene; and (c) determining expression of the target gene is repressed by at least 50%, at least 60%, at least 70%, at least 80%, at least 85%, at least 90%, at least 92%, at least 95%, at least 97%, or at least 99% for at least 3 days, wherein the target gene is selected from: a checkpoint gene and a T cell surface receptor.

[0142] In some aspects, expression of the target gene is repressed in at least 75%, at least 80%, at least 85%, at least 90%, at least 95%, or at least 99% of a plurality of the cells. In some aspects, the engineered genomic regulatory complex is undetectable after at least 3 days. In some aspects, determining the engineered genomic regulatory complex is undetectable is measured by qPCR, imaging of a FLAG-tag, or a combination thereof. In some aspects, the measuring expression of the target gene comprises flow cytometry quantification of expression of the target gene.

[0143] In some embodiments, repression of the target gene with a DNA binding domain fused to a repression domain (*e.g.*, an RNBD-TF, a MAP-NBD-TF, or TALE-TF) of the present disclosure can last even after the DNA binding domain-gene regulator becomes undetectable. The DNA binding domain fused to a repression domain (*e.g.*, an RNBD-TF, a MAP-NBD-TF, or TALE-TF) can become undetectable after at least 3 days. In some embodiments, the DNA binding domain fused to a repression domain (*e.g.*, an RNBD-TF, a MAP-NBD-TF, or TALE-TF) can become undetectable after at least 1 day, at least 2 days, at least 3 days, at least 4 days, at least 5 days, at least 6 days, at least 1 week, at least 2 weeks, at least 3 weeks, or at least 4 weeks. In some embodiments, qPCR or imaging via the FLAG-tag can be used to confirm that the DNA binding domain fused to a repression domain (*e.g.*, an RNBD-TF, a MAP-NBD-TF, or TALE-TF) is no longer detectable.

C. Imaging Moieties

[0144] An RNBD (e.g., *Ralstonia solanacearum*-derived), or another binding domain (*e.g.,* MAP-NBD or TALE), can be linked to a fluorophore, such as Hydroxycoumarin, methoxycoumarin, Alexa fluor, aminocoumarin, Cy2, FAM, Alexa fluor 488, Fluorescein FITC, Alexa fluor 430, Alexa fluor 532, HEX, Cy3, TRITC, Alexa fluor 546, Alexa fluor 555, R-phycoerythrin (PE), Rhodamine Red-X, Tamara, Cy3.5, Rox, Alexa fluor 568, Red 613, Texas Red, Alexa fluor 594, Alexa fluor 633, Allophycocyanin, Alexa fluor 633, Cy5, Alexa fluor 660, Cy5.5, TruRed, Alexa fluor 680, Cy7, GFP, or mCHERRY. An RNBD (e.g., *Ralstonia solanacearum*-derived) can be linked to a biotinylation reagent.

Genes and Indications of Interest

[0145] In some embodiments, genome editing can be performed by fusing a nuclease of the present disclosure with a DNA binding domain for a particular genomic locus of interest. Genetic modification can involve introducing a functional gene for therapeutic purposes, knocking out a gene for therapeutic gene, or engineering a cell ex vivo (*e.g.*, HSCs or CAR T cells) to be administered back into a subject in need thereof. For example, the genome editing complex can have a target site within PDCD1, CTLA4, LAG3, TET2, BTLA, HAVCR2, CCR5, CXCR4, TRA, TRB, B2M, albumin, HBB, HBA1, TTR, NR3C1, CD52, erythroid specific enhancer of the BCL11A gene, CBLB, TGFBR1, SERPINA1, HBV genomic DNA in infected cells, CEP290, DMD, CFTR, IL2RG, CS-1, or any combination thereof. In some embodiments, a genome editing complex can cleave double stranded DNA at a target site in order to insert a chimeric antigen receptor (CAR), alpha-L iduronidase (IDUA), iduronate-2-sulfatase (IDS), or Factor 9 (F9). Cells, such as hematopoietic stem cells (HSCs) and T cells, can be engineered ex vivo with the genome editing

complex. Alternatively, genome editing complexes can be directly administered to a subject in need thereof.

[0146] The subject receiving treatment can be suffering from a disease such as transthyretin amyloidosis (ATTR), HIV, glioblastoma multiforme, cancer, acute lymphoblastic leukemia, acute myeloid leukemia, beta-thalassemia, sickle cell disease, MPSI, MPSII, Hemophilia B, multiple myeloma, melanoma, sarcoma, Leber congenital amaurosis (LCA10), CD19 malignancies, BCMA-related malignancies, duchenne muscular dystrophy (DMD), cystic fibrosis, alpha-1 antitrypsin deficiency, X-linked severe combined immunodeficiency (X-SCID), or Hepatitis B.

Samples for Analysis

[0147] In some aspects, described herein include methods of modifying the genetic material of a target cell utilizing an RNBD described herein. A sample described herein may be a fresh sample. The sample may be a live sample.

[0148] The sample may be a cell sample. The cell sample may be obtained from the cells or tissue of an animal. The animal cell may comprise a cell from an invertebrate, fish, amphibian, reptile, or mammal. The mammalian cell may be obtained from a primate, ape, equine, bovine, porcine, canine, feline, or rodent. The mammal may be a primate, ape, dog, cat, rabbit, ferret, or the like. The rodent may be a mouse, rat, hamster, gerbil, hamster, chinchilla, or guinea pig. The bird cell may be from a canary, parakeet, or parrot. The reptile cell may be from a turtle, lizard, or snake. The fish cell may be from a tropical fish. For example, the fish cell may be from a zebrafish (such as Danio rerio). The amphibian cell may be from a frog. An invertebrate cell may be from an insect, arthropod, marine invertebrate, or worm. The worm cell may be from a nematode (such as Caenorhabditis elegans). The arthropod cell may be from a tarantula or hermit crab.

[0149] The cell sample may be obtained from a mammalian cell. For example, the mammalian cell may be an epithelial cell, connective tissue cell, hormone secreting cell, a nerve cell, a skeletal muscle cell, a blood cell, an immune system cell, or a stem cell. A cell may be a fresh cell, live cell, fixed cell, intact cell, or cell lysate. Cell samples can be any primary cell, such as a hematopoetic stem cell (HSCs) or naïve or stimulated T cells (*e.g.*, CD4+ T cells).

[0150] Cell samples may be cells derived from a cell line, such as an immortalized cell line. Exemplary cell lines include, but are not limited to, 293A cell line, 293FT cell line, 293F cell line, 293 H cell line, HEK 293 cell line, CHO DG44 cell line, CHO-S cell line, CHO-K1 cell line, Expi293F™ cell line, Flp-In™ T-REx™ 293 cell line, Flp-In™-293 cell line, Flp-In™-3T3 cell line, Flp-In™-BHK cell line, Flp-In™-CHO cell line, Flp-In™-CV-1 cell line, Flp-In™-Jurkat cell line,

FreeStyle™ 293-F cell line, FreeStyle™ CHO-S cell line, GripTite™ 293 MSR cell line, GS-CHO cell line, HepaRG™ cell line, T-REx™ Jurkat cell line, Per.C6 cell line, T-REx™-293 cell line, T-REx™-CHO cell line, T-REx™-HeLa cell line, NC-HIMT cell line, PC12 cell line, A549 cells, and K562 cells.

[0151] In some embodiments, an RNBD of the present disclosure can be used to modify a target cell. The target cell can itself be unmodified or modified. For example, an unmodified cell can be edited with an RNBD of the present disclosure to introduce an insertion, deletion, or mutation in its genome. In some embodiments, a modified cell already having a mutation can be repaired with an RNBD of the present disclosure.

[0152] In some instances, a target cell is a cell comprising one or more single nucleotide polymorphism (SNP). In some instances, an RNBD-nuclease described herein is designed to target and edit a target cell comprising a SNP.

[0153] In some cases, a target cell is a cell that does not contain a modification. For example, a target cell can comprise a genome without genetic defect (e.g., without genetic mutation) and an RNBD-nuclease described herein can be used to introduce a modification (e.g., a mutation) within the genome.

[0154] The cell sample may be obtained from cells of a primate. The primate may be a human, or a non-human primate. The cell sample may be obtained from a human. For example, the cell sample may comprise cells obtained from blood, urine, stool, saliva, lymph fluid, cerebrospinal fluid, synovial fluid, cystic fluid, ascites, pleural effusion, amniotic fluid, chorionic villus sample, vaginal fluid, interstitial fluid, buccal swab sample, sputum, bronchial lavage, Pap smear sample, or ocular fluid. The cell sample may comprise cells obtained from a blood sample, an aspirate sample, or a smear sample.

[0155] The cell sample may be a circulating tumor cell sample. A circulating tumor cell sample may comprise lymphoma cells, fetal cells, apoptotic cells, epithelia cells, endothelial cells, stem cells, progenitor cells, mesenchymal cells, osteoblast cells, osteocytes, hematopoietic stem cells (HSC) (e.g., a CD34+ HSC), foam cells, adipose cells, transcervical cells, circulating cardiocytes, circulating fibrocytes, circulating cancer stem cells, circulating myocytes, circulating cells from a kidney, circulating cells from a gastrointestinal tract, circulating cells from a lung, circulating cells from reproductive organs, circulating cells from a central nervous system, circulating hepatic cells, circulating cells from a spleen, circulating cells from a thymus, circulating cells from a thyroid, circulating cells from an endocrine gland, circulating cells from a parathyroid, circulating cells from a pituitary, circulating cells from an adrenal gland, circulating cells from islets of Langerhans,

circulating cells from a pancreas, circulating cells from a hypothalamus, circulating cells from prostate tissues, circulating cells from breast tissues, circulating cells from circulating retinal cells, circulating ophthalmic cells, circulating auditory cells, circulating epidermal cells, circulating cells from the urinary tract, or combinations thereof.

[0156] The cell can be a T cell. For example, in some embodiments, the T cell can be an engineered T cell transduced to express a chimeric antigen receptor (CAR). The CAR T cell can be engineered to bind to BCMA, CD19, CD22, WT1, L1CAM, MUC16, ROR1, or LeY.

[0157] A cell sample may be a peripheral blood mononuclear cell sample.

[0158] A cell sample may comprise cancerous cells. The cancerous cells may form a cancer which may be a solid tumor or a hematologic malignancy. The cancerous cell sample may comprise cells obtained from a solid tumor. The solid tumor may include a sarcoma or a carcinoma. Exemplary sarcoma cell sample may include, but are not limited to, cell sample obtained from alveolar rhabdomyosarcoma, alveolar soft part sarcoma, ameloblastoma, angiosarcoma, chondrosarcoma, chordoma, clear cell sarcoma of soft tissue, dedifferentiated liposarcoma, desmoid, desmoplastic small round cell tumor, embryonal rhabdomyosarcoma, epithelioid fibrosarcoma, epithelioid hemangioendothelioma, epithelioid sarcoma, esthesioneuroblastoma, Ewing sarcoma, extrarenal rhabdoid tumor, extraskeletal myxoid chondrosarcoma, extraskeletal osteosarcoma, fibrosarcoma, giant cell tumor, hemangiopericytoma, infantile fibrosarcoma, inflammatory myofibroblastic tumor, Kaposi sarcoma, leiomyosarcoma of bone, liposarcoma, liposarcoma of bone, malignant fibrous histiocytoma (MFH), malignant fibrous histiocytoma (MFH) of bone, malignant mesenchymoma, malignant peripheral nerve sheath tumor, mesenchymal chondrosarcoma, myxofibrosarcoma, myxoid liposarcoma, myxoinflammatory fibroblastic sarcoma, neoplasms with perivascular epitheioid cell differentiation, osteosarcoma, parosteal osteosarcoma, neoplasm with perivascular epitheioid cell differentiation, periosteal osteosarcoma, pleomorphic liposarcoma, pleomorphic rhabdomyosarcoma, PNET/extraskeletal Ewing tumor, rhabdomyosarcoma, round cell liposarcoma, small cell osteosarcoma, solitary fibrous tumor, synovial sarcoma, or telangiectatic osteosarcoma.

[0159] Exemplary carcinoma cell samples may include, but are not limited to, cell samples obtained from an anal cancer, appendix cancer, bile duct cancer (i.e., cholangiocarcinoma), bladder cancer, brain tumor, breast cancer, cervical cancer, colon cancer, cancer of Unknown Primary (CUP), esophageal cancer, eye cancer, fallopian tube cancer, gastroenterological cancer, kidney cancer, liver cancer, lung cancer, medulloblastoma, melanoma, oral cancer, ovarian cancer, pancreatic cancer, parathyroid disease, penile cancer, pituitary tumor, prostate cancer, rectal cancer, skin cancer,

stomach cancer, testicular cancer, throat cancer, thyroid cancer, uterine cancer, vaginal cancer, or vulvar cancer.

[0160] The cancerous cell sample may comprise cells obtained from a hematologic malignancy. Hematologic malignancy may comprise a leukemia, a lymphoma, a myeloma, a non-Hodgkin's lymphoma, or a Hodgkin's lymphoma. The hematologic malignancy may be a T-cell based hematologic malignancy. The hematologic malignancy may be a B-cell based hematologic malignancy. Exemplary B-cell based hematologic malignancy may include, but are not limited to, chronic lymphocytic leukemia (CLL), small lymphocytic lymphoma (SLL), high risk CLL, a non-CLL/SLL lymphoma, prolymphocytic leukemia (PLL), follicular lymphoma (FL), diffuse large B-cell lymphoma (DLBCL), mantle cell lymphoma (MCL), Waldenström's macroglobulinemia, multiple myeloma, extranodal marginal zone B cell lymphoma, nodal marginal zone B cell lymphoma, Burkitt's lymphoma, non-Burkitt high grade B cell lymphoma, primary mediastinal B-cell lymphoma (PMBL), immunoblastic large cell lymphoma, precursor B-lymphoblastic lymphoma, B cell prolymphocytic leukemia, lymphoplasmacytic lymphoma, splenic marginal zone lymphoma, plasma cell myeloma, plasmacytoma, mediastinal (thymic) large B cell lymphoma, intravascular large B cell lymphoma, primary effusion lymphoma, or lymphomatoid granulomatosis. Exemplary T-cell based hematologic malignancy may include, but are not limited to, peripheral T-cell lymphoma not otherwise specified (PTCL-NOS), anaplastic large cell lymphoma, angioimmunoblastic lymphoma, cutaneous T-cell lymphoma, adult T-cell leukemia/lymphoma (ATLL), blastic NK-cell lymphoma, enteropathy-type T-cell lymphoma, hematosplenic gamma-delta T-cell lymphoma, lymphoblastic lymphoma, nasal NK/T-cell lymphomas, or treatment-related T-cell lymphomas.

[0161] A cell sample described herein may comprise a tumor cell line sample. Exemplary tumor cell line sample may include, but are not limited to, cell samples from tumor cell lines such as 600MPE, AU565, BT-20, BT-474, BT-483, BT-549, Evsa-T, Hs578T, MCF-7, MDA-MB-231, SkBr3, T-47D, HeLa, DU145, PC3, LNCaP, A549, H1299, NCI-H460, A2780, SKOV-3/Luc, Neuro2a, RKO, RKO-AS45-1, HT-29, SW1417, SW948, DLD-1, SW480, Capan-1, MC/9, B72.3, B25.2, B6.2, B38.1, DMS 153, SU.86.86, SNU-182, SNU-423, SNU-449, SNU-475, SNU-387, Hs 817.T, LMH, LMH/2A, SNU-398, PLHC-1, HepG2/SF, OCI-Ly1, OCI-Ly2, OCI-Ly3, OCI-Ly4, OCI-Ly6, OCI-Ly7, OCI-Ly10, OCI-Ly18, OCI-Ly19, U2932, DB, HBL-1, RIVA, SUDHL2, TMD8, MEC1, MEC2, 8E5, CCRF-CEM, MOLT-3, TALL-104, AML-193, THP-1, BDCM, HL-60, Jurkat, RPMI 8226, MOLT-4, RS4, K-562, KASUMI-1, Daudi, GA-10, Raji, JeKo-1, NK-92, and Mino.

[0162] A cell sample may comprise cells obtained from a biopsy sample, necropsy sample, or autopsy sample.

[0163] The cell samples (such as a biopsy sample) may be obtained from an individual by any suitable means of obtaining the sample using well-known and routine clinical methods. Procedures for obtaining tissue samples from an individual are well known. For example, procedures for drawing and processing tissue sample such as from a needle aspiration biopsy are well-known and may be employed to obtain a sample for use in the methods provided. Typically, for collection of such a tissue sample, a thin hollow needle is inserted into a mass such as a tumor mass for sampling of cells that, after being stained, will be examined under a microscope.

[0164] A cell may be a live cell. A cell may be a eukaryotic cell. A cell may be a yeast cell. A cell may be a plant cell. A cell may be obtained from an agricultural plant.

EXAMPLES

[0165] These examples are provided for illustrative purposes only and not to limit the scope of the claims provided herein.

EXAMPLE 1

Genome Editing Complexes and Gene Regulators with Expanded Repeat Units

[0166] This example describes genome editing complexes and gene regulators with expanded repeat units. DNA binding domains (e.g., RNBD, MAP-NBD, TALE) are engineered from a plurality of repeat units and fused to a nuclease disclosed herein (e.g., FokI or SEQ ID NO: 1 – SEQ ID NO: 81), an activation domain (VP16, VP64, p65, p300 catalytic domain, TET1 catalytic domain, TDG, Ldb1 self-associated domain, SAM activator (VP64, p65, HSF1), or VPR (VP64, p65, Rta), or a repression domain (e.g., KRAB, Sin3a, LSD1, SUV39H1, G9A (EHMT2), DNMT1, DNMT3A-DNMT3L, DNMT3B, KOX, TGF-beta-inducible early gene (TIEG), v-erbA, SID, MBD2, MBD3, Rb, or MeCP2). At least one repeat unit of the DNA binding domain has greater than 39 amino acid residues and binds to a target nucleotide. The expanded repeat unit has altered affinity for its target nucleotide. The DNA binding domain with expanded repeat units exhibits altered binding to a target gene.

EXAMPLE 2

Genome Editing Complexes and Gene Regulators with Contracted Repeat Units

[0167] This example describes genome editing complexes and gene regulators with contracted repeat units. DNA binding domains (e.g., RNBD, MAP-NBD, TALE) are engineered from a

plurality of repeat units and fused to a nuclease disclosed herein (e.g., FokI or SEQ ID NO: 1 – SEQ ID NO: 81), an activation domain (VP16, VP64, p65, p300 catalytic domain, TET1 catalytic domain, TDG, Ldb1 self-associated domain, SAM activator (VP64, p65, HSF1), or VPR (VP64, p65, Rta), or a repression domain (e.g., KRAB, Sin3a, LSD1, SUV39H1, G9A (EHMT2), DNMT1, DNMT3A-DNMT3L, DNMT3B, KOX, TGF-beta-inducible early gene (TIEG), v-erbA, SID, MBD2, MBD3, Rb, or MeCP2). At least one repeat unit of the DNA binding domain has less than 32 amino acid residues and binds to a target nucleotide. The contracted repeat unit has altered affinity for its target nucleotide. The DNA binding domain with contracted repeat units exhibits altered binding to a target gene. (e.g., RNBD, MAP-NBD, TALE) are engineered from a plurality of repeat units

EXAMPLE 3

Genome Editing Complexes and Gene Regulators with Gapped Repeat Units Having Recognition Sites

[0168] This example describes genome editing complexes and gene regulators with gapped repeat units having recognition sites. DNA binding domains (e.g., RNBD, MAP-NBD, TALE) are engineered from a plurality of repeat units and fused via a linker to a nuclease disclosed herein(e.g., FokI or SEQ ID NO: 1 – SEQ ID NO: 81), an activation domain (VP16, VP64, p65, p300 catalytic domain, TET1 catalytic domain, TDG, Ldb1 self-associated domain, SAM activator (VP64, p65, HSF1), or VPR (VP64, p65, Rta), or a repression domain (e.g., KRAB, Sin3a, LSD1, SUV39H1, G9A (EHMT2), DNMT1, DNMT3A-DNMT3L, DNMT3B, KOX, TGF-beta-inducible early gene (TIEG), v-erbA, SID, MBD2, MBD3, Rb, or MeCP2). Said linker has a recognition site for a small molecule, a protease, or a kinase or serves as a localization signal. Said linker having a recognition site separating each repeat unit from a neighboring repeat unit within the DNA binding domain and are, thus, gapped. Engineered DNA binding domains with gapped repeat units exhibit genome editing or gene regulation activity along with secondary activity.

EXAMPLE 4

Genome Editing with DNA Binding Domain comprising Expanded Repeat Units and fused to a Nuclease

[0169] This example illustrates genome editing with a DNA binding domain comprising expanded repeat units and fused to a nuclease. A DNA binding domain (e.g., RNBD, MAP-NBD, TALE) in which at least one repeat unit has greater than 39 amino acid residues is fused to a cleavage domain, such as an endonuclease to form a genome editing complex. The DNA binding domain is fused to

the nuclease optionally, via a naturally occurring linker, a variant or truncation of a naturally occurring linker, or a synthetic linker.

Direct Administration to Introduce a Gene

[0170] The genome editing complex is administered directly to a subject in need thereof and is taken up by a cell. The subject has a disease. The DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to introduce a gene. The introduced gene is a mutated gene or a functional gene.

[0171] Factor IX. The genome editing complex with a cleavage domain introduces a double strand break into the albumin gene locus (e.g., into intron 1) concomitant with delivery to the cell of an ectopic nucleic acid bearing a cDNA of the factor IX gene. The double strand break leads to the integration of the ectopic nucleic acid into intron 1 of the albumin gene; the factor IX protein is secreted by the cell into the circulation. The target cell is a hepatocyte and the subject in need thereof has Hemophilia B.

***Ex Vivo* Engineering of a Cell to Introduce a Gene**

[0172] The genome editing complex is transfected into cells *ex vivo* along with an ectopic nucleic acid bearing a gene. Upon transfection of cells *ex vivo*, the DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to introduce an ectopically provided gene (also provided to the cell) in the region cleaved by the genome editing complex. The resulting engineered cells with modified DNA are administered to a subject in need thereof. The subject has a disease.

[0173] CAR. The genome editing complex with a cleavage domain introduces a chimeric antigen receptor (CAR) by editing the DNA of a target cell. The target cell is a T cell and the subject has cancer, such as a blood cancer. Upon administration of the engineered cells to a subject, the engineered CAR T cells effectively eliminate cancer in the subject.

Direct Administration to Partially or Completely Knock Out a Gene

[0174] The genome editing complex is administered directly to a subject in need thereof and is taken up by a cell. The subject has a disease. The DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to partially or completely knock out a gene.

[0175] TTR. The genome editing complex with a cleavage domain partially or completely knocks out the transthyretin (TTR) gene by editing the DNA of a target cell. The target cell is a liver cell and the subject in need thereof has transthyretin amyloidosis (ATTR).

[0176] SERPINA1. The genome editing complex with a cleavage domain partially or completely knocks out the SERPINA1 gene by editing the DNA of a target cell. The target cell is a liver cell and the subject in need thereof has alpha-1 antitrypsin deficiency (dA1AT def).

***Ex Vivo* Engineering of a Cell to Partially or Completely Knock Out a Gene or a Gene Regulatory Region**

[0177] The genome editing complex is transfected in cells *ex vivo*. Upon transfection of cells *ex vivo*, the DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to partially or completely knock out a gene or a gene regulatory region. The subject has a disease.

[0178] BCL11A Enhancer. The genome editing complex with a cleavage domain partially or completely knocks out the BCL11A erythroid enhancer by editing the DNA of a target cell. The target cell is an HPSC and the subject in need thereof has b-thalassemia or sickle cell disease.

[0179] CCR5. The genome editing complex with a cleavage domain partially or completely knocks the CCR5 gene by editing the DNA of a target cell, thereby allowing for introduction of a mutated version of CCR5. Target cells, in which mutated versions of CCR5 are introduced via the action of the genome editing complex, are not infected by HIV via the modified CCR5 receptor. The target cell is a T cell or a hematopoietic stem cell (HPSC) and the subject has HIV.

[0180] Upon administration of the genome editing complex directly to a subject or upon administration of an engineered cell with DNA that has been modified with the genome editing complex, the disease symptoms are eliminated or reduced.

EXAMPLE 5

Genome Editing with a DNA Binding Domain comprising Contracted Repeat Units and fused to a Nuclease

[0181] This example illustrates genome editing with a DNA binding domain comprising contracted repeat units and fused to a nuclease. A DNA binding domain (e.g., RNBD, MAP-NBD, TALE) in which at least one repeat unit has less than 32 amino acid residues is fused to a cleavage domain, such as an endonuclease to form a genome editing complex. The DNA binding domain is fused to the nuclease optionally, via a naturally occurring linker, a variant or truncation of a naturally occurring linker, or a synthetic linker.

Direct Administration to Introduce a Gene

[0182] The genome editing complex is administered directly to a subject in need thereof and is taken up by a cell. The subject has a disease. The DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to introduce a gene. The introduced gene is a mutated gene or a functional gene.

[0183] Factor IX. The genome editing complex with a cleavage domain introduces a double strand break into the albumin gene locus (e.g., into intron 1) concomitant with delivery to the cell of an ectopic nucleic acid bearing a cDNA of the factor IX gene. The double strand break leads to the integration of the ectopic nucleic acid into intron 1 of the albumin gene; the factor IX protein is secreted by the cell into the circulation. The target cell is a hepatocyte and the subject in need thereof has Hemophilia B.

***Ex Vivo* Engineering of a Cell to Introduce a Gene**

[0184] The genome editing complex is transfected into cells *ex vivo* along with an ectopic nucleic acid bearing a gene. Upon transfection of cells *ex vivo*, the DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to introduce an ectopically provided gene (also provided to the cell) in the region cleaved by the genome editing complex. The resulting engineered cells with modified DNA are administered to a subject in need thereof. The subject has a disease.

[0185] CAR. The genome editing complex with a cleavage domain introduces a chimeric antigen receptor (CAR) by editing the DNA of a target cell. The target cell is a T cell and the subject has cancer, such as a blood cancer. Upon administration of the engineered cells to a subject, the engineered CAR T cells effectively eliminate cancer in the subject.

Direct Administration to Partially or Completely Knock Out a Gene

[0186] The genome editing complex is administered directly to a subject in need thereof and is taken up by a cell. The subject has a disease. The DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to partially or completely knock out a gene.

[0187] TTR. The genome editing complex with a cleavage domain partially or completely knocks out the transthyretin (TTR) gene by editing the DNA of a target cell. The target cell is a liver cell and the subject in need thereof has transthyretin amyloidosis (ATTR).

[0188] SERPINA1. The genome editing complex with a cleavage domain partially or completely knocks out the SERPINA1 gene by editing the DNA of a target cell. The target cell is a liver cell and the subject in need thereof has alpha-1 antitrypsin deficiency (dA1AT def).

***Ex Vivo* Engineering of a Cell to Partially or Completely Knock Out a Gene or a Gene Regulatory Region**

[0189] The genome editing complex is transfected in cells *ex vivo*. Upon transfection of cells *ex vivo*, the DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to partially or completely knock out a gene or a gene regulatory region. The subject has a disease.

[0190] BCL11A Enhancer. The genome editing complex with a cleavage domain partially or completely knocks out the BCL11A erythroid enhancer by editing the DNA of a target cell. The target cell is an HPSC and the subject in need thereof has b-thalassemia or sickle cell disease.

[0191] CCR5. The genome editing complex with a cleavage domain partially or completely knocks the CCR5 gene by editing the DNA of a target cell, thereby allowing for introduction of a mutated version of CCR5. Target cells, in which mutated versions of CCR5 are introduced via the action of the genome editing complex, are not infected by HIV via the modified CCR5 receptor. The target cell is a T cell or a hematopoietic stem cell (HPSC) and the subject has HIV.

[0192] Upon administration of the genome editing complex directly to a subject or upon administration of an engineered cell with DNA that has been modified with the genome editing complex, the disease symptoms are eliminated or reduced.

EXAMPLE 6

Genome Editing with DNA Binding Domain Having Gapped Repeat Units and fused to a Nuclease

[0193] This example illustrates genome editing DNA binding domains fused to a nuclease, wherein the DNA binding domains have gapped repeat units. A DNA binding domain (e.g., RNBD, MAP-NBD, TALE) in which all repeat units are separated from neighboring repeat units with a linker comprising a recognition site is fused to a cleavage domain, such as an endonuclease to form a genome editing complex. Said linker has a recognition site for a small molecule, a protease, or a kinase or serves as a localization signal. The DNA binding domain is fused to the nuclease optionally, via a naturally occurring linker, a variant or truncation of a naturally occurring linker, or a synthetic linker.

Direct Administration to Introduce a Gene

[0194] The genome editing complex is administered directly to a subject in need thereof and is taken up by a cell. The subject has a disease. The DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to introduce a gene. The introduced gene is a mutated gene or a functional gene.

[0195] Factor IX. The genome editing complex with a cleavage domain introduces a double strand break into the albumin gene locus (e.g., into intron 1) concomitant with delivery to the cell of an ectopic nucleic acid bearing a cDNA of the factor IX gene. The double strand break leads to the integration of the ectopic nucleic acid into intron 1 of the albumin gene; the factor IX protein is secreted by the cell into the circulation. The target cell is a hepatocyte and the subject in need thereof has Hemophilia B.

***Ex Vivo* Engineering of a Cell to Introduce a Gene**

[0196] The genome editing complex is transfected into cells *ex vivo* along with an ectopic nucleic acid bearing a gene. Upon transfection of cells *ex vivo*, the DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to introduce an ectopically provided gene (also provided to the cell) in the region cleaved by the genome editing complex. The resulting engineered cells with modified DNA are administered to a subject in need thereof. The subject has a disease.

[0197] CAR. The genome editing complex with a cleavage domain introduces a chimeric antigen receptor (CAR) by editing the DNA of a target cell. The target cell is a T cell and the subject has cancer, such as a blood cancer. Upon administration of the engineered cells to a subject, the engineered CAR T cells effectively eliminate cancer in the subject.

Direct Administration to Partially or Completely Knock Out a Gene

[0198] The genome editing complex is administered directly to a subject in need thereof and is taken up by a cell. The subject has a disease. The DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to partially or completely knock out a gene.

[0199] TTR. The genome editing complex with a cleavage domain partially or completely knocks out the transthyretin (TTR) gene by editing the DNA of a target cell. The target cell is a liver cell and the subject in need thereof has transthyretin amyloidosis (ATTR).

[0200] SERPINA1. The genome editing complex with a cleavage domain partially or completely knocks out the SERPINA1 gene by editing the DNA of a target cell. The target cell is a liver cell and the subject in need thereof has alpha-1 antitrypsin deficiency (dA1AT def).

***Ex Vivo* Engineering of a Cell to Partially or Completely Knock Out a Gene or a Gene Regulatory Region**

[0201] The genome editing complex is transfected in cells *ex vivo*. Upon transfection of cells *ex vivo*, the DNA binding domain of the genome editing complex binds a region of DNA in a target cell and the cleavage domain induces a double strand break in the DNA of the target cell to partially or completely knock out a gene or a gene regulatory region. The subject has a disease.

[0202] BCL11A Enhancer. The genome editing complex with a cleavage domain partially or completely knocks out the BCL11A erythroid enhancer by editing the DNA of a target cell. The target cell is an HPSC and the subject in need thereof has b-thalassemia or sickle cell disease.

[0203] CCR5. The genome editing complex with a cleavage domain partially or completely knocks the CCR5 gene by editing the DNA of a target cell, thereby allowing for introduction of a mutated version of CCR5. Target cells, in which mutated versions of CCR5 are introduced via the action of the genome editing complex, are not infected by HIV via the modified CCR5 receptor. The target cell is a T cell or a hematopoietic stem cell (HPSC) and the subject has HIV.

[0204] Upon administration of the genome editing complex directly to a subject or upon administration of an engineered cell with DNA that has been modified with the genome editing complex, the disease symptoms are eliminated or reduced.

EXAMPLE 7

TALE Protein with N-terminus Fragment

[0205] A DNA binding protein engineered to have a shortened N-terminus derived from a TALE protein was generated. U.S. Patent No. 8,586,526 shows that while the N-terminus region (referred to as N-cap) from a TALE protein can be shortened by deleting amino acids at the N-terminus, deleting amino acids beyond amino acid position N+134 decreased DNA binding affinity, with the decrease in DNA binding apparent even with deletion of amino acids beyond amino acid position N+137. U.S. Patent No. 8,586,526 concluded that amino acid sequence from N+1 through N+137 are required for binding to DNA while the first 152 amino acids of the N-cap sequence are dispensable.

[0206] However, it has been discovered that further deleting amino acids till position N+116 surprising leads to recovery of DNA binding. Even shorter N-terminus regions such as a fragment

having deletion till position N+111 also retains DNA binding activity. Deleting amino acids till position N+106 significantly decreases DNA binding. Further deletion of the N-terminus region, such as, deleting amino acids till position N+101 does not lead to recovery of DNA binding. See Fig. 2.

[0207] TALEN monomers recognizing 5′-TTTCTGTCACCAATCCT-3′ (SEQ ID NO: 449) and 5′-TCCCCTCCACCCCACAGT-3′ (SEQ ID NO: 450) in the human *AAVS1* locus were engineered to harbor N-terminus regions that included deletions encompassing residues N137-116, N137-111, N137-106 and N137-101. While these residues are numbered with reference to the N+137 construct in U.S. Patent No. 8,586,526, N137-116 refers to deletion of amino acids starting at the N-terminus of the N-cap sequence (N+228) and extending through amino acid residue 116 such that the resulting fragment retains amino acids residues from position N+115 to position N+1, and so on. The amino acid sequence of the N-terminal truncation del_N137-116 is set forth in SEQ ID NO:321. The amino acid sequence of the N-terminal truncation del_N137-111 is set forth in SEQ ID NO:447.

[0208] NK562 cells were transfected with 2 μg plasmid DNA for each TALEN monomer using an AMAXA™ Nucleofector™ 96-well Shuttle™ system as per the manufacturer's recommendations. Full length TALEN monomers were included ("AAVS1 control"), together with N137-116/full length and full length/N137-116 heterodimers. Cells were cold shocked at 30°C and genomic DNA was harvested at 72 h using QuickExtract™ (Lucigen). Indel rates were determined by amplicon sequencing. The TALE repeats present in the TALE monomers have the sequence LTPDQVVAIAS(RVD)GGKQALETVQRLLPVLCQDHG (SEQ ID NO: 451), with a RVD selected based on the target sequence.

[0209] Fig. 2 represents DNA binding activity assayed by measuring nuclease activity of Fok1 fused to C-terminus of the polypeptides. AAVS1 control data set correspond to TALENS using the standard full-length N-terminus (N+288 to N+1). N-terminal truncation del_N137-116 (N-terminus extending from N+115 to N+1) showed higher activity than standard full-length N-terminus (N+288 to N+1). N-terminal truncation del_N137-111 (N-terminus extending from N+110 to N+1) was also active. Further truncation del_ N137-106 (N-terminus extending from N+105 to N+1) significantly decreased DNA binding. Further deletion of the N-terminus region del_N137-101 (N-terminus extending from N+100 to N+1) did not lead to recovery of DNA binding. Thus, a fragment of the N-terminus of a TALE protein extending from N+115 to N+1 shows full activity. Mock/GFP is a negative control. The AAVS1/del_N137-116 data shows that an N1-115 TALEN monomer can be combined with a monomer comprising full-length N-terminus region of a TALE protein.

[0210] While preferred embodiments of the present invention have been shown and , it will be apparent to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now 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 may be employed in practicing the invention. 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. A polypeptide comprising a modular nucleic acid binding domain comprising a plurality of repeat units, wherein a repeat unit of the plurality of repeat units recognizes a target nucleic acid base and wherein the plurality of repeat units has one or more of the following characteristics:
   (a) at least one repeat unit comprising greater than 39 amino acid residues;
   (b) at least one repeat unit comprising greater than 35 amino acid residues derived from the genus of *Ralstonia*;
   (c) at least one repeat unit comprising less than 32 amino acid residues; and
   (d) each repeat unit of the plurality of repeat units is separated from a neighboring repeat unit by a linker comprising a recognition site.
2. The composition of claim 1, wherein the at least one repeat unit comprises an amino acid selected from glycine, alanine, threonine or histidine at a position after an amino acid residue at position 35.
3. The polypeptide any one of claims 1-2, wherein the at least one repeat unit comprises an amino acid selected from glycine, alanine, threonine or histidine at a position after an amino acid residue at position 39.
4. The polypeptide of any one of claims 1-3, wherein the recognition site is for a small molecule, a protease, or a kinase.
5. The polypeptide of any one of claims 1-4, wherein the recognition site serves as a localization signal.
6. The polypeptide of any one of claims 1-5, wherein the polypeptide further comprises a cleavage domain linked to the modular nucleic acid binding domain to form a non-naturally occurring fusion protein.
7. The polypeptide of claim 6, wherein the modular nucleic acid binding domain comprises a potency for a target site greater than 65% and a specificity ratio for the target site of 50:1; and a functional domain; wherein the modular nucleic acid binding domain comprises a plurality of repeat units, wherein at least one repeat unit of the plurality comprises a binding region configured to bind to a target nucleic acid base in the target site, wherein the potency comprises indel percentage at the target site, and wherein the specificity ratio comprises indel percentage at the target site over indel percentage at a top-ranked off-target site of the non-naturally occurring fusion protein.

8. The polypeptide of any one of claims 1-7, wherein the repeat unit comprises a sequence of $A_{1-11}X_1X_2B_{14-35}$ (SEQ ID NO: 448),
   wherein each amino acid residue of A_{1-11} comprises any amino acid residue;
   wherein X_1X_2 comprises the binding region;
   wherein each amino acid residue of B_{14-35} comprises any amino acid; and
wherein a first repeat unit of the plurality of repeat units comprises at least one residue in A_{1-11}, B_{14-35}, or a combination thereof that differs from a corresponding residue in a second repeat unit of the plurality of repeat units.
9. The polypeptide of any one of claims 7-8, wherein the binding region comprises an amino acid residue at position 13 or an amino acid residue at position 12 and the amino acid residue at position 13.
10. The composition of claim 9, wherein the amino acid residue at position 13 binds to the target nucleic acid base.
11. The polypeptide of any one of claims 9-10, wherein the amino acid residue at position 12 stabilizes the configuration of the binding region.
12. The polypeptide of any one of claims 7-11, wherein the indel percentage is measured by deep sequencing.
13. The polypeptide of any one of claims 7-12, wherein the modular nucleic acid binding domain further comprises one or more properties selected from the following:
    (a) binds the target site, wherein the target site comprises a 5' guanine;
    (b) comprises from 7 repeat units to 25 repeat units; and
    (c) upon binding to the target site, the modular nucleic acid binding domain is separated from a second modular nucleic acid binding domain bound to a second target site by from 2 to 50 base pairs.

14. The polypeptide of any one of claims 1-13, wherein the plurality of repeat units comprises a *Ralstonia* repeat unit, a *Xanthomonas* repeat unit, a *Legionella* repeat unit, or any combination thereof.

15. The polypeptide of claim 14, wherein the *Ralstonia* repeat unit is a *Ralstonia solanacearum* repeat unit, the *Xanthomonas* repeat unit is a *Xanthomonas spp.* repeat unit, and the Legionella repeat unit is a *Legionella quateirensis* repeat unit.

16. The polypeptide of any one of claims 8-15, wherein the B_{14-35} of at least one repeat unit of the plurality of repeat units has at least 92% sequence identity to GGKQALEAVRAQLLDLRAAPYG (SEQ ID NO: 280).

17. The polypeptide of any one of claims 7-16, wherein the binding region comprises HD binding to cytosine, NG binding to thymidine, NK binding to guanine, SI binding to adenosine, RS binding to adenosine, HN binding to guanine, or NT binds to adenosine.

18. The polypeptide of any one of claims 1-17, wherein the at least one repeat unit comprises any one of SEQ ID NO: 267 – SEQ ID NO: 279.

19. The polypeptide of any one of claims 1-18, wherein the at least one repeat unit comprises at least 80% sequence identity with any one of SEQ ID NO: 168 – SEQ ID NO: 263.

20. The polypeptide of any one of claims 1-19, wherein the at least one repeat unit comprises at least 80% sequence identity with SEQ ID NO: 209, SEQ ID NO: 197, SEQ ID NO: 233, SEQ ID NO: 253, SEQ ID NO: 203, or SEQ ID NO: 218.

21. The polypeptide of any one of claims 1-20, wherein the at least one repeat unit comprises any one of SEQ ID NO: 168 – SEQ ID NO: 263.

22. The polypeptide of any one of claims 1-21, wherein the at least one repeat unit comprises SEQ ID NO: 209, SEQ ID NO: 197, SEQ ID NO: 233, SEQ ID NO: 253, SEQ ID NO: 203, or SEQ ID NO: 218.

23. The polypeptide of any one of claims 1-22, wherein the target nucleic acid base is cytosine, guanine, thymidine, adenosine, uracil, or a combination thereof.

24. The polypeptide of any one of claims 1-23, wherein the modular nucleic acid binding domain comprises an N-terminus amino acid sequence, a C-terminus amino acid sequence, or a combination thereof.

25. The polypeptide of claim 24, wherein the N-terminus amino acid sequence is from *Xanthomonas spp.*, *Legionella quateirensis*, or *Ralstonia solanacearum*.

26. The polypeptide of any one of claims 24-25, wherein the N-terminus amino acid sequence comprises at least 80% sequence identity to SEQ ID NO: 264, SEQ ID NO: 300, SEQ ID

NO: 335, SEQ ID NO: 303, SEQ ID NO: 301, SEQ ID NO: 304, or SEQ ID NO: 320, SEQ ID NO: 321, or SEQ ID NO: 322.

27. The polypeptide of any one of claims 24-26, wherein the N-terminus amino acid sequence comprises SEQ ID NO: 264, SEQ ID NO: 300, SEQ ID NO: 335, SEQ ID NO: 303, SEQ ID NO: 301, SEQ ID NO: 304, or SEQ ID NO: 320, SEQ ID NO: 321, or SEQ ID NO: 322.

28. The polypeptide of any one of claims 24-27, wherein the C-terminus amino acid sequence is from *Xanthomonas spp.*, *Legionella quateirensis*, or *Ralstonia solanacearum*.

29. The polypeptide of any one of claims 24-28, wherein the C-terminus amino acid sequence comprises at least 80% sequence identity to SEQ ID NO: 266, SEQ ID NO: 298, or SEQ ID NO: 306.

30. The polypeptide of any one of claims 24-29, wherein the C-terminus amino acid sequence comprises SEQ ID NO: 266, SEQ ID NO: 298, or SEQ ID NO: 306.

31. The polypeptide of any one of claims 24-30, wherein the C-terminus amino acid sequence serves as a linker between the modular nucleic acid binding domain and a functional domain.

32. The polypeptide of any one of claims 1-31, wherein the modular nucleic acid binding domain comprises a half repeat.

33. The polypeptide of claim 32, wherein the half repeat comprises at least 80% sequence identity to SEQ ID NO: 265, SEQ ID NO: 327 – SEQ ID NO: 334, or SEQ ID NO: 290.

34. The polypeptide of any one of claims 31-33, wherein the functional domain is a cleavage domain or a repression domain.

35. The polypeptide of claim 34, wherein the cleavage domain comprises at least 33.3% divergence from SEQ ID NO: 163 and is immunologically orthogonal to SEQ ID NO: 163.

36. The polypeptide of any one of claims 34-35, comprising one or more of the following characteristics:

(a) induces greater than 1% indels at the target site;

(b) the cleavage domain comprises a molecular weight of less than 23 kDa;

(c) the cleavage domain comprises less than 196 amino acids; and

(d) capable of cleaving across a spacer region greater than 24 base pairs.

37. The polypeptide of any one of claims 34-36, wherein the polypeptide induces greater than 5%, greater than 10%, greater than 20%, greater than 30%, greater than 40%, greater than 50%, greater than 60%, greater than 70%, greater than 80%, or greater than 90% indels at the target site.
38. The polypeptide of any one of claims 34-37, wherein the cleavage domain comprises at least 35%, at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, or at least 75% divergence from SEQ ID NO: 163.
39. The polypeptide of any one of claims 34-38, wherein the cleavage domain comprises a sequence selected from SEQ ID NO: 316 – SEQ ID NO: 319.
40. The polypeptide of any one of claims 34-39, wherein the cleavage domain comprises a nucleic acid sequence encoding for a sequence having at least 80% sequence identity with SEQ ID NO: 1 – SEQ ID NO: 81.
41. The polypeptide of any one of claims 34-39, wherein the cleavage domain comprises a nucleic acid sequence encoding for a sequence selected from SEQ ID NO: 1 – SEQ ID NO: 81.
42. The polypeptide of any one of claims 34-41, wherein the nucleic acid sequence comprises at least 80% sequence identity with SEQ ID NO: 82 – SEQ ID NO: 162.
43. The polypeptide of any one of claims 34-42, wherein the nucleotide sequence encoding for the sequence comprises any one of SEQ ID NO: 82 – SEQ ID NO: 162.
44. The polypeptide of claim 34, wherein the repression domain comprises KRAB, Sin3a, LSD1, SUV39H1, G9A (EHMT2), DNMT1, DNMT3A-DNMT3L, DNMT3B, KOX, TGF-beta-inducible early gene (TIEG), v-erbA, SID, MBD2, MBD3, Rb, or MeCP2.
45. The polypeptide of any one of claims 1-44, wherein the plurality of repeat units comprises 3 to 60 repeat units.
46. The polypeptide of any one of claims 7-45, wherein the target site is a nucleic acid sequence within a PDCD1 gene, a CTLA4 gene, a LAG3 gene, a TET2 gene, a BTLA gene, a HAVCR2 gene, a CCR5 gene, a CXCR4 gene, a TRA gene, a TRB gene, a B2M gene, an albumin gene, a HBB gene, a HBA1 gene, a TTR gene, a NR3C1 gene, a CD52 gene, an erythroid specific enhancer of the BCL11A gene, a CBLB gene, a TGFBR1 gene, a SERPINA1 gene, a HBV genomic DNA in infected cells, a CEP290 gene, a DMD gene, a CFTR gene, or an IL2RG gene.

47. The polypeptide of any one of claims 7-46, wherein a nucleic acid sequence encoding a chimeric antigen receptor (CAR), alpha-L iduronidase (IDUA), iduronate-2-sulfatase (IDS), or Factor 9 (F9), is inserted at the target site.
48. A method of genome editing, the method comprising:

administering the polypeptide of any one of claims 1-43 or 45-47 and inducing a double stranded break.
49. A method of gene repression, the method comprising administering the polypeptide of any one of claims 1-34 or 44-47 and repressing gene expression.
50. A non-naturally occurring DNA-binding polypeptide comprising from N-terminus to C-terminus:

an N-terminus region comprises at least residues N+110 to N+1 of a TALE protein, wherein the N-terminus region does not include residues N+288 to N+116 of the TALE protein;

a plurality of TALE-repeat units, the TALE repeat units comprising a repeat variable di-residue (RVD); and

a C-terminus region of a TALE protein.
51. The DNA binding polypeptide of claim 50, wherein the N-terminus region comprises residues N+1 up to N+115 of the TALE protein.
52. The DNA binding polypeptide of claim 50, wherein the N-terminus region comprises residues N+1 up to N+110 of the TALE protein.
53. The DNA binding polypeptide of any one of claims 50-52, wherein the C-terminus region comprises residues C+1 to C+63 of the TALE protein.
54. The DNA binding polypeptide of any one of claims 50-53, wherein the N-terminus region consists of residues N+1 to N+115 of the TALE protein.
55. The DNA binding polypeptide of any one of claims 50-54, wherein the TALE repeat units are ordered from N-terminus to C-terminus to specifically bind to a target nucleic acid in genomic DNA.
56. The DNA binding polypeptide of any one of claims 50-55, wherein a heterologous functional domain is conjugated to the N-terminus and/or C-terminus.
57. The DNA binding polypeptide of claim 56, wherein the functional domain comprises an enzyme, a transcriptional activator, a transcriptional repressor, or a DNA nucleotide modifier.
58. The DNA binding polypeptide of claim 57, wherein the enzyme is a nuclease, a DNA modifying protein, or a chromatin modifying protein.

59. The DNA binding polypeptide of claim 58, wherein the nuclease is a cleavage domain or a half-cleavage domain.
60. The DNA binding polypeptide of claim 59, wherein the cleavage domain or half-cleavage domain comprises a type IIS restriction enzyme.
61. The DNA binding polypeptide of claim 60, wherein the type IIS restriction enzyme comprises FokI or Bfil.
62. The DNA binding polypeptide of claim 58, wherein the chromatin modifying protein is lysine-specific histone demethylase 1 (LSD1).
63. The DNA binding polypeptide of claim 57, wherein the transcriptional activator comprises VP16, VP64, p65, p300 catalytic domain, TET1 catalytic domain, TDG, Ldb1 self-associated domain, SAM activator (VP64, p65, HSF1), or VPR (VP64, p65, Rta).
64. The DNA binding polypeptide of claim 57, wherein the transcriptional repressor comprises KRAB, Sin3a, LSD1, SUV39H1, G9A (EHMT2), DNMT1, DNMT3A-DNMT3L, DNMT3B, KOX, TGF-beta-inducible early gene (TIEG), v-erbA, SID, MBD2, MBD3, Rb, or MeCP2.
65. The DNA binding polypeptide claim 57, wherein the DNA nucleotide modifier is adenosine deaminase.
66. The DNA binding polypeptide of any of claims 55-65, wherein the target nucleic acid is within a PDCD 1 gene, a CTLA4 gene, a LAG3 gene, a TET2 gene, a ETLA gene, a HA VCR2 gene, a CCR5 gene, a CXCR4 gene, a TRA gene, a TRE gene, a E2M gene, an albumin gene, a HEE gene, a HEAl gene, a TTR gene, a NR3Cl gene, a CD52 gene, an erythroid specific enhancer of the ECLllA gene, a CELE gene, a TGFERl gene, a SERPINAl gene, a HEV genomic DNA in infected cells, a CEP290 gene, a DMD gene, a CFTR gene, or an IL2RG gene.

67. The DNA binding polypeptide of any of claims 55-76, wherein the heterologous functional domain comprises a fluorophore or a detectable tag.
68. A nucleic acid encoding the polypeptide of any of claims 50-68.
69. The nucleic acid of claim 68, wherein the nucleic acid is operably linked to a promoter sequence that confers expression of the polypeptide.
70. The nucleic acid of claim 68 or 69, wherein the sequence of the nucleic acid is codon optimized for expression of the polypeptide in a human cell.
71. The nucleic acid of any one of claims 68-70, wherein the nucleic acid is a deoxyribonucleic acid (DNA).

72. The nucleic acid of any one of claims 68-70, wherein the nucleic acid is a ribonucleic acid (RNA).
73. A vector comprising the nucleic acid of any of claims 68-71.
74. The vector of claim 78, wherein the vector is a viral vector.
75. A host cell comprising the nucleic acid of any of claims 68-72 or the vector of claim 73 or 74.
76. A host cell that expresses the polypeptide of any of claims 50-67.
77. A pharmaceutical composition comprising the polypeptide of any of claims 50-67 and a pharmaceutically acceptable excipient.
78. A pharmaceutical composition comprising the nucleic acid of any of claims 68-72 or the vector of claim 73 or 74 and a pharmaceutically acceptable excipient.
79. A method of modulating expression of an endogenous gene in a cell, the method comprising:

introducing into the cell the polypeptide of claim 56,

wherein the DNA binding polypeptide binds to a target nucleic acid sequence present in the endogenous gene and the heterologous functional domain modulates expression of the endogenous gene.
80. The method of claim 79, wherein the polypeptide is introduced as a nucleic acid encoding the polypeptide.
81. The method of claim 80, wherein the nucleic acid is a deoxyribonucleic acid (DNA).
82. The method of claim 80, wherein the nucleic acid is a ribonucleic acid (RNA).
83. The method of any of claims 80-82, wherein the sequence of the nucleic acid is codon optimized for expression in a human cell.
84. The method of any of claims 79-83, wherein the functional domain is a transcriptional activator and the target nucleic acid sequence is present in an expression control region of the gene, wherein the polypeptide increases expression of the gene.
85. The method of claim 84, wherein the transcriptional activator comprises VP16, VP64, p65, p300 catalytic domain, TET1 catalytic domain, TDG, Ldb1 self-associated domain, SAM activator (VP64, p65, HSF1), or VPR (VP64, p65, Rta).
86. The method of any of claims 79-83, wherein the functional domain is a transcriptional repressor and the target nucleic acid sequence is present in an expression control region of the gene, wherein the polypeptide decreases expression of the gene.
87. The method of claim 76, wherein the transcriptional repressor comprises KRAB, Sin3a, LSD1, SUV39H1, G9A (EHMT2), DNMT1, DNMT3A-DNMT3L, DNMT3B, KOX, TGF-beta-inducible early gene (TIEG), v-erbA, SID, MBD2, MBD3, Rb, or MeCP2.

88. The method of any of claims 79-87, wherein the gene is a PDCD 1 gene, a CTLA4 gene, a LAG3 gene, a TET2 gene, a ETLA gene, a HA VCR2 gene, a CCR5 gene, a CXCR4 gene, a TRA gene, a TRE gene, a E2M gene, an albumin gene, a HEE gene, a HEAl gene, a TTR gene, a NR3Cl gene, a CD52 gene, an erythroid specific enhancer of the ECLllA gene, a CELE gene, a TGFERl gene, a SERPINAl gene, a HEV genomic DNA in infected cells, a CEP290 gene, a DMD gene, a CFTR gene, or an IL2RG gene.
89. The method of any of claims 86-88, wherein the expression control region of the gene comprises a promoter region of the gene.
90. The method of any of claims 79-83, wherein the functional domain is a nuclease comprising a cleavage domain or a half-cleavage domain and the endogenous gene is inactivated by cleavage.
91. The method of claim 90, wherein inactivation occurs via non-homologous end joining (NHEJ).
92. The method of claims 90 or 91, wherein the DNA binding polypeptide is a first polypeptide that binds to a first target nucleic acid sequence in the gene and comprises a half-cleavage domain and the method comprises introducing a second DNA binding polypeptide that binds to a second target nucleic acid sequence in the gene and comprises a half-cleavage domain.
93. The method of claim 92, wherein the first target nucleic acid sequence and the second target sequence are spaced apart in the gene and the two half-cleavage domains mediate a cleavage of the gene sequence at a location in between the first and second target nucleic acid sequences, wherein the second DNA binding polypeptide comprises from N-terminus to C-terminus:

    an N-terminus region comprising residues N+1 to up to N+115 of a TALE protein or a full-length N-terminus region of a TALE protein;
    a plurality of TALE-repeat units, the TALE repeat units comprising a repeat variable di-residue (RVD); and
    a C-terminus region of a TALE protein.

94. The method of claim 93, wherein the C-terminus region is a full length region of the TALE protein.
95. The method of claim 93, wherein the C-terminus region is a fragment of the C-terminus region of the TALE protein.

96. The method of claim 93, wherein the C-terminus region extends from C+1 to C+63 of the TALE protein.
97. The method of any of claims 90-96, wherein the cleavage domain or the cleavage half domain comprises FokI or Bfil.
98. The method of any of claims 90-96, wherein the cleavage domain comprises a meganuclease.
99. The method of any of claims 90-98, wherein the gene is a PDCD 1 gene, a CTLA4 gene, a LAG3 gene, a TET2 gene, a ETLA gene, a HA VCR2 gene, a CCR5 gene, a CXCR4 gene, a TRA gene, a TRE gene, a E2M gene, an albumin gene, a HEE gene, a HEAl gene, a TTR gene, a NR3Cl gene, a CD52 gene, an erythroid specific enhancer of the ECLllA gene, a CELE gene, a TGFERl gene, a SERPINAl gene, a HEV genomic DNA in infected cells, a CEP290 gene, a DMD gene, a CFTR gene, or an IL2RG gene.

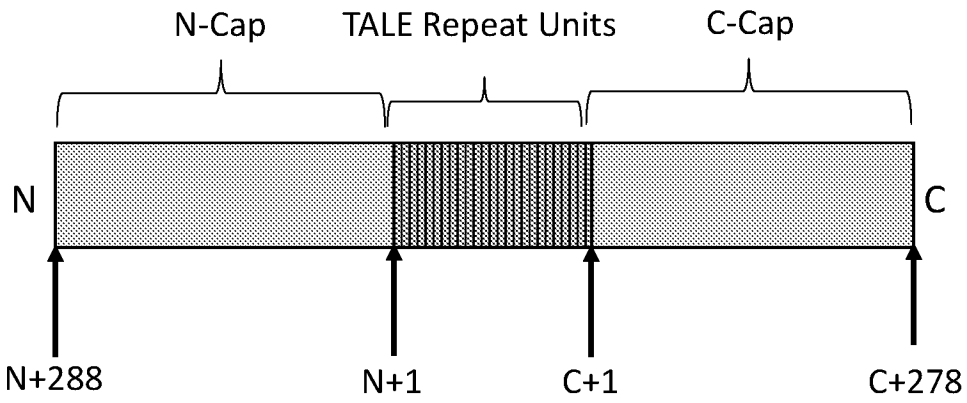

Fig. 1A

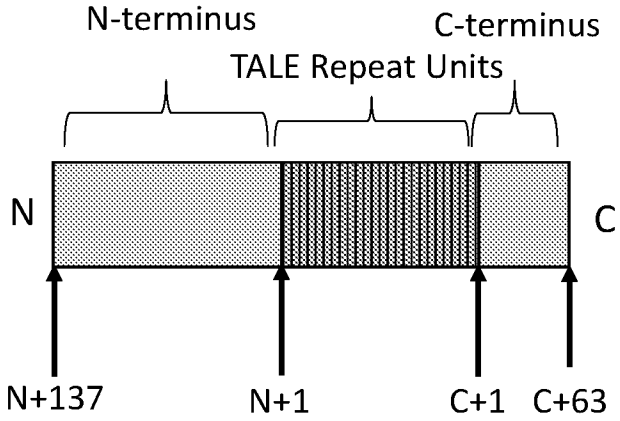

Fig. 1B

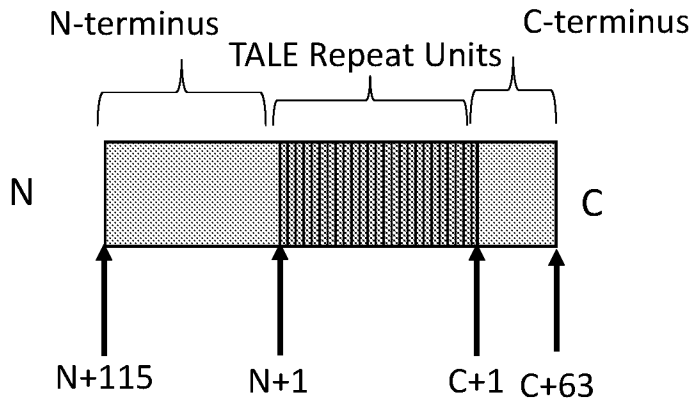

Fig. 1C

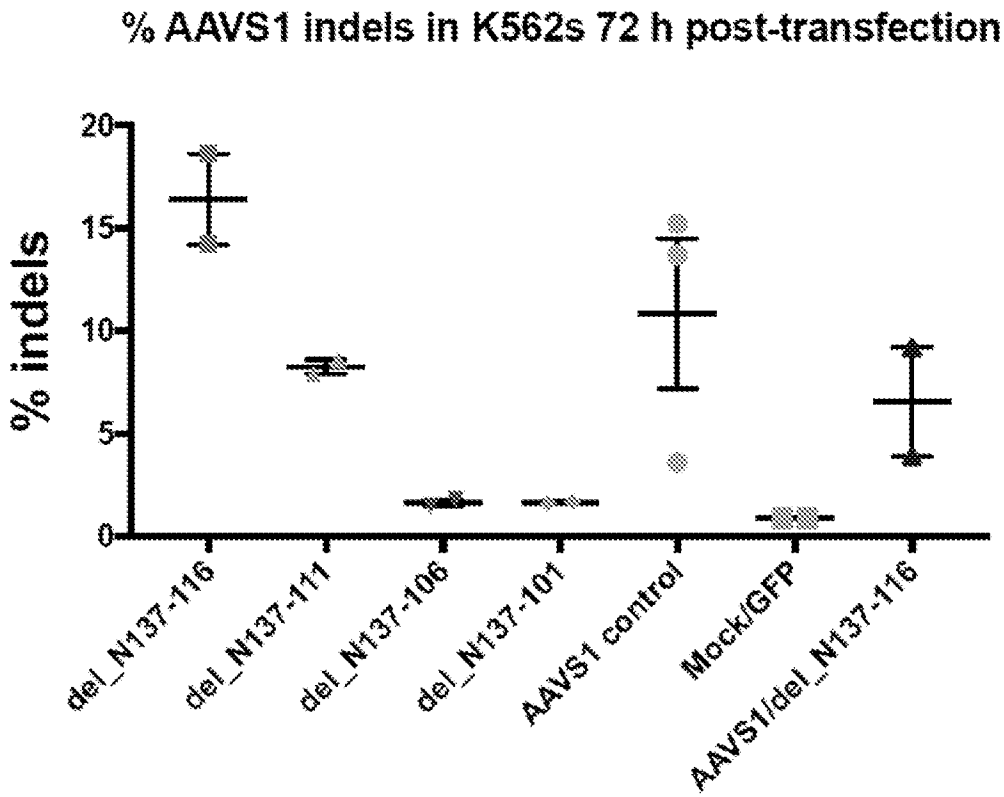
% AAVS1 indels in K562s 72 h post-transfection
% indels
20
15
10
5
0
del_N137-116
del_N137-111
del_N137-106
del_N137-101
AAVS1 control
Mock/GFP
AAVS1/del_N137-116

Fig. 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US19/39326

A. CLASSIFICATION OF SUBJECT MATTER

IPC - C12Q 1/68; C12N 9/22; C07K 14/195 (2019.01)

CPC - C12Q 1/68; C12N 9/22; C07K 14/195

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	(RICHTER, A et al.) A TAL Effector Repeat Architecture for Frameshift Binding. Nature Communications. 11 March 2014; Vol. 5; pages 1-9; abstract; page 2, column 2, paragraph 3; page 3, column 2, paragraph 1; page 9, column 1, paragraph 6; supplementary table 1; DOI: 10.1038/ncomms4447	1-2, 3/1-2
A	US 2014/0134741 A1 (SANGAMO BIOSCIENCES, INC.) 15 May 2014; whole document	1-2, 3/1-2
A	US 2013/0210151 A1 (UNIVERSITY OF WESTERN ONTARIO) 15 August 2013; whole document	1-2, 3/1-2
P, X	WO 2018/140654 A1 (ALTIUS INSTITUTE FOR BIOMEDICAL SCIENCES) 02 August 2018; whole document	1-2, 3/1-2

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

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

Date of the actual completion of the international search	Date of mailing of the international search report
04 November 2019 (04.11.2019)	19 NOV 2019
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Shane Thomas Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US19/39326

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

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

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

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

3. ☒ Claims Nos.: 4-49, 54-99
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

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

This International Searching Authority found multiple inventions in this international application, as follows:

-***- Please see within the next Supplemental Page-***-

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

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

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

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

Group 1 - Claims 1-3

Remark on Protest

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

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

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US19/39326

-***-Continued from Box III Observations where unity of invention is lacking -***-

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

Group I, Claims 1-3 are directed toward DNA binding domains comprising a plurality of repeat units, wherein each repeat unit is expanded or contracted in length and wherein each repeat unit is separated from a neighboring repeat unit by a linker.

Group II, Claims 50-53 are directed toward DNA binding proteins that include a fragment of the N-cap sequence of a TALE protein.

The inventions listed as Groups I and II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: the special technical features of Group I include DNA binding domains comprising a plurality of repeat units of expanded or contracted, wherein at least one repeat unit is derived from the genus Ralstonia, not present in Group II; the special technical features of Group II include a TALE protein, not present in Group I.

Groups I and II share the technical features including: a polypeptide comprising a modular nucleic acid binding domain comprising a plurality of repeat units, wherein a repeat unit of the plurality of repeat units recognizes a target nucleic acid base and wherein the plurality of repeat units comprises (d) each repeat unit of the plurality of repeat units is separated from a neighboring repeat unit by a linker comprising a recognition site.

However, these shared technical features are previously disclosed by US 2014/0134741 A1 to Sangamo Biosciences, Inc. (hereinafter 'Sangamo').

Sangamo discloses a polypeptide (claim 1) comprising a nucleic acid binding domain (claim 1) comprising a plurality of repeat units (a plurality of TALE (repeat units); paragraph [0075]; claim 1), wherein a repeat unit of the plurality of repeat units recognizes a target nucleic acid base (wherein a repeat unit of the plurality of repeat units recognizes a target nucleic acid base; paragraph [0075]) and wherein the plurality of repeat units comprises (d) each repeat unit of the plurality of repeat units is separated from a neighboring repeat unit by a linker comprising a recognition site each repeat unit of the plurality of repeat units is separated from a neighboring repeat unit by a linker comprising a recognition site ((d) each repeat unit of the plurality of repeat units is separated from a neighboring repeat unit by RVDs (a linker comprising a recognition site); paragraphs [0021]; claim 1).

Since none of the special technical features of the Groups I and II inventions are found in more than one of the inventions, and since all of the shared technical features are previously disclosed by the Sangamo reference, unity of invention is lacking.