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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 November 2023 (02.11.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/212626 A2

(51) International Patent Classification:
C12N 9/22 (2006.01) *C12N 15/85* (2006.01)

(21) International Application Number:
PCT/US2023/066276

(22) International Filing Date:
27 April 2023 (27.04.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/336,383 29 April 2022 (29.04.2022) US

(71) Applicant: **PIONEER HI-BRED INTERNATIONAL, INC.** [US/US]; 7100 NW 62nd Avenue, PO Box 1014, Johnston, Iowa 50131-1014 (US).

(72) Inventors: **VAN GINKLE, Elizabeth Sommers**; 7250 NW 62nd Avenue, PO Box 552, Johnston, Iowa 50131-1014 (US). **YOUNG, Joshua K**; 7250 NW 62nd Avenue, PO Box 552, Johnston, Iowa 50131-1014 (US).

(74) Agent: **RIVAS, Marcos P.** et al.; 9330 Zionsville Road, Indianapolis, Indiana 46268 (US).

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

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

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:
— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*
— *with sequence listing part of description (Rule 5.2(a))*

(54) **Title:** ENGINEERED CAS ENDONUCLEASE AND GUIDE RNA VARIANTS FOR IMPROVED GENOME EDITING

(57) **Abstract:** Compositions, methods, and systems are provided for genome modification of a target sequence in the genome of a cell, using novel engineered Cas endonucleases. These can include a guide polynucleotide/endonuclease system to modify or alter target sequences in the genome of a cell or organism. Also provided are novel effectors and endonuclease systems and elements comprising such systems. Compositions, methods, and systems are also provided that include a guide polynucleotide/endonuclease system comprising at least one endonuclease, optionally covalently or non-covalently linked to, or assembled with, at least one additional protein subunit or substrate.

WO 2023/212626 A2

ENGINEERED CAS ENDONUCLEASE VARIANTS FOR IMPROVED GENOME EDITING

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to United States Provisional Application No. 63/336,383, filed April 29, 2022, which is hereby incorporated herein by reference in its entirety.

REFERENCE TO SEQUENCE LISTING SUBMITTED ELECTRONICALLY

[0002] The official copy of the sequence listing is submitted concurrently with the specification as an xml formatted sequence listing with a file named 9208-WO-PCT.ST26 created on April 24, 2023, having a size of 58,600 bytes, which is part of the specification and is herein incorporated by reference in its entirety.

FIELD

[0003] The disclosure relates to the field of molecular biology, in particular to compositions of novel Cas endonuclease systems, and compositions and methods for editing or modifying the genome of a cell.

BACKGROUND

[0004] Recombinant DNA technology has made it possible to insert DNA sequences at targeted genomic locations and/or modify specific endogenous chromosomal sequences. Site-specific integration techniques, which employ site-specific recombination systems, as well as other types of recombination technologies, have been used to generate targeted insertions of genes of interest in a variety of organism. Genome-editing techniques such as designer zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), or homing meganucleases, are available for producing targeted genome perturbations, but these systems tend to have low specificity and employ designed nucleases that need to be redesigned for each target site, which renders them costly and time-consuming to prepare.

[0005] Newer technologies utilizing archaeal or bacterial adaptive immunity systems have been identified, called CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats), which comprise different domains of effector proteins that encompass a variety of activities (DNA recognition, binding, and optionally cleavage).

[0006] Despite the identification and characterization of some of these systems, there remains a need for engineering novel effectors and systems, as well as demonstrating activity in eukaryotes, particularly animals and plants, to effect editing of endogenous and previously introduced heterologous polynucleotides.

[0007] Herein are described novel engineered Cas polypeptides and endonucleases, and methods and compositions for use thereof.

SUMMARY

[0008] The compositions, methods, and systems disclosed herein are based, at least in part, on the discovery of Cas polypeptides that have been engineered (changed relative to naturally occurring Cas polypeptides) to make variants having surprisingly improved activity. The disclosed Cas variants provide improved activity in binding to and/or editing of a target site on a polynucleotide sequence. In particular examples, the disclosed Cas variants provide improved activity at lower temperatures, relative to the wildtype Cas polypeptide.

[0009] Accordingly, disclosed herein are compositions of novel engineered Cas polypeptides, systems comprising the engineered Cas polypeptides, and methods of use thereof. The disclosed engineered Cas polypeptides are capable of being guided by a guide polynucleotide to target double-stranded DNA in a PAM-dependent fashion. In some embodiments, the engineered Cas polypeptides are active endonucleases capable of introducing a break at the target site of the target double-stranded DNA. In other embodiments, the Cas polypeptide comprises one or more mutations that render it incapable of double-strand cutting, but permits single-strand cutting. In some embodiments, the Cas polypeptide comprises one or more mutations that render it incapable of cleaving either or both strands of a double-stranded polynucleotide, but it retains the ability to bind to a target polynucleotide sequence.

[0010] In one aspect, a novel engineered Cas polypeptide is provided that comprises a sequence having at least 90% amino acid sequence identity to SEQ ID NO:18 and comprises one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341. In some examples, the novel engineered Cas polypeptide comprises two, three, four, five, six, seven, eight, nine, ten or eleven of the following amino acid changes at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341. Thus, for example, in one example, the novel engineered Cas polypeptide comprises a Tyrosine at amino acid position 123, a Threonine at position 266, and a Proline at position 295 relative to an alignment with SEQ ID NO:18. The provided engineered Cas polypeptide is capable of site specifically binding to a target site of a polynucleotide.

[0011] In a second aspect, a novel engineered Cas polypeptide is provided that comprises at least one zinc-finger-like domain and a tri-split RuvC domain (comprising non-contiguous RuvC-I domain, RuvC-II domain, and RuvC-III domain) and which comprises one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341. In some examples, the novel engineered Cas polypeptide comprises two, three, four, five, six, seven, eight, nine, ten or eleven of the following amino acid changes at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341. For example, in one example, the novel engineered Cas polypeptide comprises a Tyrosine at amino acid position 123, a Threonine at position 266, and a Proline at position 295 relative to an alignment with SEQ ID NO:18. The provided engineered Cas polypeptide is capable of site specifically binding to a target site of a polynucleotide.

[0012] The provided novel engineered Cas polypeptide can have at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% amino acid sequence identity to any one of SEQ ID NOs:19 to 39, preferably wherein the Cas polypeptide includes one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341. For example, in one example, the novel engineered Cas polypeptide comprises a Tyrosine at amino acid position 123, a Threonine at position 266, and a Proline at position 295 relative to an alignment with SEQ ID NO:18. In preferred examples, the engineered Cas polypeptide is capable of site specifically binding to a target site of a polynucleotide.

[0013] The engineered Cas polypeptide disclosed herein can be an active endonuclease that cleaves double stranded DNA polynucleotides. Alternatively, the engineered Cas polypeptide disclosed herein can be inactivated, thereby reducing or eliminating its endonuclease activity. For example, amino acid residues that are essential for endonuclease activity are identified by bold and underlining in Fig.3. Thus, altering (e.g. by substitution, deletion, or insertion at the

site of) one or more essential amino acids for endonuclease can produce an inactive Cas polypeptide.

[0014] In particular examples of each of the foregoing disclosed aspects, the novel engineered Cas polypeptide can demonstrate greater endonuclease (DNA cleavage) activity, as compared to wildtype Cas-alpha 8 (SEQ ID NO:18) on the same DNA substrate. For example, the novel engineered Cas polypeptide can have at least ten times (10X), at least fifteen times (15X), at least twenty times (20X), at least twenty-five times (25X), at least fifty times (50X), at least seventy-five times (75X), at least eighty times (80X), at least ninety times (90X), at least 100 times (100X), at least 125 times (125X) greater endonuclease activity, as compared to wildtype Cas-alpha 8 (SEQ ID NO:18) on the same DNA substrate.

[0015] In some examples, the novel engineered Cas polypeptide provided herein demonstrates greater endonuclease activity at lower temperature ranges, as compared to wildtype Cas-alpha 8 (SEQ ID NO:18) on the same DNA substrate. For example, the novel engineered Cas polypeptide demonstrates much better endonuclease activity (relative to wildtype SEQ ID NO:18) at a temperature of about 37 degrees Celsius or less, about 35 degrees Celsius or less, about 30 degrees Celsius or less, about 25 degrees Celsius or less, or about 20 degrees or less.

[0016] In particular examples of the foregoing, the novel engineered Cas polypeptide comprises fewer than 500 amino acids in length, fewer than 475 amino acids in length, fewer than 450 amino acids in length, or fewer than 425 amino acids in length.

[0017] In some examples, the disclosed engineered Cas polypeptide is provided with a guide polynucleotide. The guide polynucleotide comprises a region of complementarity to the polynucleotide's target site. When combined, the disclosed engineered Cas polypeptide and guide polynucleotide can form a complex that binds the target site sequence on double stranded DNA. In particular examples, the complex of Cas polypeptide and guide polynucleotide can cleave the target site sequence on double stranded DNA (e.g., on genomic DNA).

[0018] Also provided herein is a synthetic composition that comprises the disclosed engineered Cas polypeptide, a target double-stranded DNA polynucleotide; and a guide polynucleotide comprising a variable targeting domain that comprises a region of complementarity to a target double-stranded DNA polynucleotide. The Cas polypeptide recognizes a PAM sequence on the target double-stranded DNA polynucleotide, and the guide polynucleotide and the Cas polypeptide form a complex that binds the target double-stranded DNA polynucleotide. The PAM sequence can comprise a thymine dinucleotide (TT).

[0019] In some examples, the engineered Cas polypeptide disclosed herein is part of a fusion protein. For example, the engineered Cas polypeptide can be joined, via linker, to a heterologous nuclease domain such as a deaminase.

[0020] In another aspect, provided is a in any of the compositions or methods, at least one component that has been optimized for expression in a eukaryotic cell, particularly a plant cell, a fungal cell, or an animal cell, is provided.

[0021] In one aspect, a synthetic composition is provided, comprising: a eukaryotic cell and a heterologous CRISPR-Cas effector; wherein said heterologous CRISPR-Cas effector protein is any example of the novel engineered Cas polypeptide disclosed herein. For example, the eukaryotic cell can be a human, non-human, animal, bacterial, fungal, insect, yeast, non-conventional yeast, or plant cell and the engineered Cas polypeptide can be (a) Cas polypeptide having at least 90% amino acid sequence identity to SEQ ID NO:18 and which comprises one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341 or (b) Cas polypeptide that comprises at least one zinc-finger-like domain and a tri-split RuvC domain (comprising non-contiguous RuvC-I domain, RuvC-II domain, and RuvC-III domain) and which comprises one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341 or (c) Cas polypeptide that has at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% amino acid sequence identity to any one of SEQ ID NOs:19 to 39, preferably wherein the Cas polypeptide includes one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341.

[0022] In another aspect, provided herein is a polynucleotide encoding any example of the novel engineered Cas polypeptide disclosed herein. Thus, provided herein is a polynucleotide that comprises a sequence encoding (a) Cas polypeptide having at least 90% amino acid

sequence identity to SEQ ID NO:18 and which comprises one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341 or (b) Cas polypeptide that comprises at least one zinc-finger-like domain and a tri-split RuvC domain (comprising non-contiguous RuvC-I domain, RuvC-II domain, and RuvC-III domain) and which comprises one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341 or (c) Cas polypeptide that has at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% amino acid sequence identity to any one of SEQ ID NOs:19 to 39, preferably wherein the Cas polypeptide includes one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341.

[0023] The polynucleotide encoding the novel engineered Cas polypeptide disclosed herein can further comprise a heterologous polynucleotide. The heterologous polynucleotide may be a noncoding regulatory expression element such as a promoter, intron, enhancer, or terminator; a donor polynucleotide; a polynucleotide modification template, optionally comprising at least one nucleotide modification as compared to the sequence of a polynucleotide in a cell; a transgene; a guide RNA; a guide DNA; a guide RNA-DNA hybrid; an endonuclease; a nuclear localization signal; and a cell transit peptide.

[0024] In a further aspect, methods are provided for using any of the compositions disclosed herein. In some methods, the disclosed Cas polypeptide or endonuclease binds to a target sequence of a polynucleotide, for example in the genome of a cell or *in vitro*. In some embodiments, the disclosed Cas polypeptide or endonuclease forms a complex with a guide polynucleotide, for example a guide RNA. In some methods, the complex recognizes, binds to, and optionally creates a nick (one strand) or a break (two strands) in the polynucleotide at or near the target sequence. In some examples of the method, the nick or break is repaired via Non-

Homologous End Joining (NHEJ). In additional examples, the nick or break is repaired via Homology-Directed Repair (HDR) or via Homologous Recombination (HR), with a polynucleotide modification template or a donor DNA molecule.

[0025] In any aspect, the engineered Cas polypeptide or endonuclease disclosed herein may be used in a synthetic composition (e.g., one that comprises a cell, guide polynucleotide, and/or target polynucleotide sequence), and incubated at a temperature of less than about 45 degrees Celsius, e.g., a temperature of about 40 degrees Celsius or less, about 37 degrees Celsius or less, about 35 degrees Celsius or less, about 30 degrees Celsius or less, about 28 degrees Celsius or less, or about 25 degrees Celsius or less. For example, the engineered Cas polypeptide or endonuclease used at the foregoing temperature can be (a) Cas polypeptide having at least 90% amino acid sequence identity to SEQ ID NO:18 and which comprises one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341 or (b) Cas polypeptide that comprises at least one zinc-finger-like domain and a tri-split RuvC domain (comprising non-contiguous RuvC-I domain, RuvC-II domain, and RuvC-III domain) and which comprises one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341 or (c) Cas polypeptide that has at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% amino acid sequence identity to any one of SEQ ID NOs:19 to 39, preferably wherein the Cas polypeptide includes one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341. Also provided is a method that includes contacting a polynucleotide with any engineered Cas endonuclease disclosed herein (including those specifically disclosed above) and creating a break in the polynucleotide at a temperature of less than about 45 degrees Celsius (e.g., a temperature of about 40 degrees Celsius or less, about 35 degrees Celsius or less, about 30 degrees Celsius or less, about 28 degrees Celsius or less, or

about 25 degrees Celsius or less). This break can be used to generate a targeted modification or altered target site (such as a base edit, deletion, or insertion) in the polynucleotide.

[0026] The novel engineered Cas endonucleases described herein are capable of creating a double-strand break in, or adjacent to, a target polynucleotide that comprises an appropriate PAM, and to which it is directed by a guide polynucleotide, in any prokaryotic or eukaryotic cell. In some cases, the cell is a plant cell or an animal cell or a fungal cell. In some cases, a plant cell is selected from the group consisting of maize, soybean, cotton, wheat, canola, oilseed rape, sorghum, rice, rye, barley, millet, oats, sugarcane, turfgrass, switchgrass, alfalfa, sunflower, tobacco, peanut, potato, tobacco, *Arabidopsis*, safflower, and tomato.

[0027] In another aspect, the engineered Cas polypeptide described herein comprises one or mutations that provide a nuclease inactivated or dead Cas polypeptide. For example, the engineered Cas polypeptide disclosed herein can be altered to include a substitution/deletion/insertion at one or more of amino acids at the position equivalent to position 225 or position 324, or position 401 in an alignment with SEQ ID NO:18. See e.g., Fig. 3. The disclosed inactivated engineered Cas polypeptide can be linked to an effector or effector protein, which can be a molecule that recognizes, binds to, and/or cleaves or nicks a polynucleotide target. The disclosed inactivated engineered Cas polypeptide can be linked to a base editing molecule, e.g., a deaminase, for targeted base editing. The disclosed inactivated engineered Cas polypeptide optionally linked to an effector or effector protein, can be used for targeted deliver of an effector molecule at a temperature of about 45 degrees Celsius or less, about 40 degrees Celsius or less, about 37 degrees Celsius or less, about 35 degrees Celsius or less, about 30 degrees Celsius or less, about 25 degrees Celsius or less, or about 20 degrees Celsius or less.

BRIEF DESCRIPTION OF THE DRAWINGS AND THE SEQUENCE LISTING

[0028] The disclosure can be more fully understood from the following detailed description and the accompanying drawings and Sequence Listing, which form a part of this application.

[0029] **FIG. 1** is a schematic drawing of an example of a Cas endonuclease variant yeast expression vector. Depicted numerals 1-16 refer to each of SEQ ID Nos:1-16, respectively. Thus, numeral 1 refers to SEQ ID NO:1 (ROX3 promoter); 2 refers to SEQ ID NO:2 (SV40 NLS coding sequence); 3 refers to SEQ ID NO:3 (yeast optimized Cas-alpha 8 gene); 4 refers to SEQ ID NO:4 (CYC1 terminator); 5 refers to SEQ ID NO:5 (SNR52 promoter) 6 refers to SEQ ID NO:6 (Cas endonuclease sgRNA or Cas-alpha recognition domain); 7 refers to SEQ ID NO:7 (Cas endonuclease sgRNA variable targeting domain 1); 8 refers to SEQ ID NO:8 (Cas endonuclease sgRNA variable targeting domain 2); 9 refers to SEQ ID NO:9 (Cas endonuclease

sgRNA variable targeting domain 3); 10 refers to SEQ ID NO:10 (Cas endonuclease sgRNA variable targeting domain 4); 11 refers to SEQ ID NO:11 (Cas endonuclease sgRNA 1) ; 12 refers to SEQ ID NO:12 (Cas endonuclease sgRNA 2); 13 refers to SEQ ID NO:13 (Cas endonuclease sgRNA 3); 14 refers to SEQ ID NO:14 (Cas endonuclease sgRNA 4); 15 refers to SEQ ID NO:15 (Hepatitis Delta Virus ribozyme); and 16 refers to SEQ ID NO:16 (SUP4 terminator). SEQ ID Nos. sequences are provided as examples.

[0030] **FIG. 2** is a schema for target cleavage detection in *S. cerevisiae*, wherein target cleavage and cellular repair results in the formation of a non-functional *ade2* gene that results in adenine auxotrophy and the switch from a white to a red/pink cellular phenotype. This color switch allows for identification of phenotypic differences and can be used to select cells expressing a Cas endonuclease variant and/or associated guide RNA with improved targeted DSB activity. Colonies with functional *ade2* gene are white. Colonies with non-functional *ade2* gene are red. Colony that are mottled red and white indicate multiple sectors containing a non-functional *ade2* gene. Colony with a single sector containing a non-functional *ade2* gene are white with a red section.

[0031] **FIG. 3** is the polypeptide sequence for a wildtype Cas12f (Cas-alpha 8) endonuclease (SEQ ID NO:18). Key catalytic residues required for endonuclease activity are shown in bold underlined font. Zinc finger domain indicated by a dashed underline.

[0032] The sequence descriptions and sequence listing attached hereto comply with the rules governing nucleotide and amino acid sequence disclosures in patent applications as set forth in 37 C.F.R. §§1.821 and 1.825. The sequence descriptions comprise the three letter codes for amino acids as defined in 37 C.F.R. §§ 1.821 and 1.825, which are incorporated herein by reference. Nucleic acid sequences listed in the accompanying sequence listing and referenced herein are shown using standard letter abbreviations for nucleotide bases. Only one strand of each nucleic acid sequence is shown, but the complementary strand is understood to be included by any reference to the displayed strand.

[0033] **SEQ ID NO:1** is a *Saccharomyces cerevisiae* ROX3 promoter sequence.

[0034] **SEQ ID NO:2** is artificial sequence encoding SV40 NLS sequence.

[0035] **SEQ ID NO:3** is a yeast-optimized sequence coding sequence for Cas-alpha 8 endonuclease.

[0036] **SEQ ID NO:4** is CYC1 terminator sequence (unknown origin).

[0037] **SEQ ID NO:5** is *Saccharomyces cerevisiae* DNA SNR52 promoter sequence, incorporates a guanine base pair at its 3' terminus.

[0038] **SEQ ID NO:6** is a coding sequence for Cas-alpha 8 endonuclease sgRNA (Cas endonuclease recognition domain).

[0039] **SEQ ID NO:7** is a coding sequence for Cas-alpha 8 endonuclease sgRNA variable targeting domain 1 sequence.

[0040] **SEQ ID NO:8** is a coding sequence for Cas-alpha 8 endonuclease sgRNA variable targeting domain 2 sequence.

[0041] **SEQ ID NO:9** is a coding sequence for Cas-alpha 8 endonuclease sgRNA variable targeting domain 3 sequence.

[0042] **SEQ ID NO:10** is a coding sequence for Cas-alpha 8 endonuclease sgRNA variable targeting domain 4 sequence.

[0043] **SEQ ID NO:11** is an artificial sequence encoding Cas-alpha 8 sgRNA 1

[0044] **SEQ ID NO:12** is an artificial sequence encoding Cas-alpha 8 sgRNA 2

[0045] **SEQ ID NO:13** is an artificial sequence encoding Cas-alpha 8 sgRNA 3

[0046] **SEQ ID NO:14** is an artificial sequence encoding Cas-alpha 8 sgRNA 4

[0047] **SEQ ID NO:15** is a sequence encoding *Hepatitis Delta Virus* ribozyme sequence.

[0048] **SEQ ID NO:16** is a SUP4 terminator sequence

[0049] **SEQ ID NO:17** is an SV40 NLS amino acid sequence

[0050] **SEQ ID NO:18** is a wildtype Cas-alpha 8 endonuclease

[0051] **SEQ ID NO:19** is a Cas-alpha 8 endonuclease variant (I123Y)

[0052] **SEQ ID NO:20** is a Cas-alpha 8 endonuclease variant (L226Q)

[0053] **SEQ ID NO:21** is a Cas-alpha 8 endonuclease variant (A231E)

[0054] **SEQ ID NO:22** is a Cas-alpha 8 endonuclease variant (A231T)

[0055] **SEQ ID NO:23** is a Cas-alpha 8 endonuclease variant (A231Y)

[0056] **SEQ ID NO:24** is a Cas-alpha 8 endonuclease variant (R266T)

[0057] **SEQ ID NO:25** is a Cas-alpha 8 endonuclease variant (A295P)

[0058] **SEQ ID NO:26** is a Cas-alpha 8 endonuclease variant (T301)

[0059] **SEQ ID NO:27** is a Cas-alpha 8 endonuclease variant (Y305H)

[0060] **SEQ ID NO:28** is a Cas-alpha 8 endonuclease variant (R335D)

[0061] **SEQ ID NO:29** is a Cas-alpha 8 endonuclease variant (R335E)

[0062] **SEQ ID NO:30** is a Cas-alpha 8 endonuclease variant (R335P)

[0063] **SEQ ID NO:31** is a Cas-alpha 8 endonuclease variant (R335Q)

[0064] **SEQ ID NO:32** is a Cas-alpha 8 endonuclease variant (F336D)

[0065] **SEQ ID NO:33** is a Cas-alpha 8 endonuclease variant (F336E)

[0066] **SEQ ID NO:34** is a Cas-alpha 8 endonuclease variant (F336V)

[0067] **SEQ ID NO:35** is a Cas-alpha 8 endonuclease variant (F337I)

[0068] **SEQ ID NO:36** is a Cas-alpha 8 endonuclease variant (F337T)

[0069] **SEQ ID NO:37** is a Cas-alpha 8 endonuclease variant (F337V)

[0070] **SEQ ID NO:38** is a Cas-alpha 8 endonuclease variant (F341P)

[0071] **SEQ ID NO:39** is a Cas-alpha 8 endonuclease variant (I123Y, R266T, A295P)

DETAILED DESCRIPTION

[0072] The temperature optimum of the native Cas endonuclease is above the typical biological temperatures of some organisms, including plants and yeast. Because of this, Cas endonuclease would require a heat shock of approximately 45 degrees Celsius for optimal activity. For some applications, it may be beneficial to modify this property. Herein are presented methods and compositions for novel engineered CRISPR effectors, systems, and elements comprising such effectors, including, but not limiting to, novel endonucleases, novel guide polynucleotide/endonuclease complexes, guide polynucleotides, guide RNA elements, Cas proteins, and endonucleases, as well as proteins comprising an endonuclease functionality (domain). Compositions and methods are also provided for direct delivery of endonucleases, cleavage ready complexes, guide RNAs, and guide RNA/Cas endonuclease complexes. The present disclosure further includes compositions and methods for genome modification of a target sequence in the genome of a cell, for gene editing, and for inserting a polynucleotide of interest into the genome of a cell. The variants identified should improve genome editing outcomes in a variety of cell types including human and aid in the wide-spread adoption of this miniature RNA-guided Cas nuclease.

[0073] Terms used in the claims and specification are defined as set forth below unless otherwise specified. It must be noted that, as used in the specification and the appended claims, the singular forms "a," "an" and "the" include plural referents unless the context clearly dictates otherwise.

[0074] As used herein, “nucleic acid” means a polynucleotide and includes a single or a double-stranded polymer of deoxyribonucleotide or ribonucleotide bases. Nucleic acids may also include fragments and modified nucleotides. Thus, the terms “polynucleotide”, “nucleic acid sequence”, “nucleotide sequence” and “nucleic acid fragment” are used interchangeably to denote a polymer of RNA and/or DNA and/or RNA-DNA that is single- or double-stranded, optionally comprising synthetic, non-natural, or altered nucleotide bases. Nucleotides (usually found in their 5’-monophosphate form) are referred to by their single letter designation as follows: “A” for adenosine or deoxyadenosine (for RNA or DNA, respectively), “C” for cytosine or deoxycytosine, “G” for guanosine or deoxyguanosine, “U” for uridine, “T” for

deoxythymidine, “R” for purines (A or G), “Y” for pyrimidines (C or T), “K” for G or T, “H” for A or C or T, “I” for inosine, and “N” for any nucleotide.

[0075] The term “genome” as it applies to a prokaryotic and eukaryotic cell or organism cells encompasses not only chromosomal DNA found within the nucleus, but organelle DNA found within subcellular components (e.g., mitochondria, or plastid) of the cell.

[0076] “Open reading frame” is abbreviated ORF.

[0077] The term "selectively hybridizes" includes reference to hybridization, under stringent hybridization conditions, of a nucleic acid sequence to a specified nucleic acid target sequence to a detectably greater degree (e.g., at least 2-fold over background) than its hybridization to non-target nucleic acid sequences and to the substantial exclusion of non-target nucleic acids. Selectively hybridizing sequences typically have about at least 80% sequence identity, or 90% sequence identity, up to and including 100% sequence identity (i.e., fully complementary) with each other.

[0078] The term "stringent conditions" or “stringent hybridization conditions” includes reference to conditions under which a probe will selectively hybridize to its target sequence in an *in vitro* hybridization assay. Stringent conditions are sequence-dependent and will be different in different circumstances. By controlling the stringency of the hybridization and/or washing conditions, target sequences can be identified which are 100% complementary to the probe (homologous probing). Alternatively, stringency conditions can be adjusted to allow some mismatching in sequences so that lower degrees of similarity are detected (heterologous probing). Generally, a probe is less than about 1000 nucleotides in length, optionally less than 500 nucleotides in length. Typically, stringent conditions will be those in which the salt concentration is less than about 1.5 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salt(s)) at pH 7.0 to 8.3, and at least about 30°C for short probes (e.g., 10 to 50 nucleotides) and at least about 60°C for long probes (e.g., greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide. Exemplary low stringency conditions include hybridization with a buffer solution of 30 to 35% formamide, 1 M NaCl, 1% SDS (sodium dodecyl sulphate) at 37°C, and a wash in 1X to 2X SSC (20X SSC = 3.0 M NaCl/0.3 M trisodium citrate) at 50 to 55°C. Exemplary moderate stringency conditions include hybridization in 40 to 45% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 0.5X to 1X SSC at 55 to 60°C. Exemplary high stringency conditions include hybridization in 50% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 0.1X SSC at 60 to 65°C.

[0079] By "homology" is meant DNA sequences that are similar. For example, a "region of homology to a genomic region" that is found on the donor DNA is a region of DNA that has a similar sequence to a given "genomic region" in the cell or organism genome. A region of homology can be of any length that is sufficient to promote homologous recombination at the cleaved target site. For example, the region of homology can comprise at least 5-10, 5-15, 5-20, 5-25, 5-30, 5-35, 5-40, 5-45, 5- 50, 5-55, 5-60, 5-65, 5- 70, 5-75, 5-80, 5-85, 5-90, 5-95, 5-100, 5-200, 5-300, 5-400, 5-500, 5-600, 5-700, 5-800, 5-900, 5-1000, 5-1100, 5-1200, 5-1300, 5-1400, 5-1500, 5-1600, 5-1700, 5-1800, 5-1900, 5-2000, 5-2100, 5-2200, 5-2300, 5-2400, 5-2500, 5-2600, 5-2700, 5-2800, 5-2900, 5-3000, 5-3100 or more bases in length such that the region of homology has sufficient homology to undergo homologous recombination with the corresponding genomic region. "Sufficient homology" indicates that two polynucleotide sequences have sufficient structural similarity to act as substrates for a homologous recombination reaction. The structural similarity includes overall length of each polynucleotide fragment, as well as the sequence similarity of the polynucleotides. Sequence similarity can be described by the percent sequence identity over the whole length of the sequences, and/or by conserved regions comprising localized similarities such as contiguous nucleotides having 100% sequence identity, and percent sequence identity over a portion of the length of the sequences.

[0080] As used herein, a "genomic region" is a segment of a chromosome in the genome of a cell that is present on either side of the target site or, alternatively, also comprises a portion of the target site. The genomic region can comprise at least 5-10, 5-15, 5-20, 5-25, 5-30, 5-35, 5-40, 5-45, 5- 50, 5-55, 5-60, 5-65, 5- 70, 5-75, 5-80, 5-85, 5-90, 5-95, 5-100, 5-200, 5-300, 5-400, 5-500, 5-600, 5-700, 5-800, 5-900, 5-1000, 5-1100, 5-1200, 5-1300, 5-1400, 5-1500, 5-1600, 5-1700, 5-1800, 5-1900, 5-2000, 5-2100, 5-2200, 5-2300, 5-2400, 5-2500, 5-2600, 5-2700, 5-2800. 5-2900, 5-3000, 5-3100 or more bases such that the genomic region has sufficient homology to undergo homologous recombination with the corresponding region of homology.

[0081] As used herein, "homologous recombination" (HR) includes the exchange of DNA fragments between two DNA molecules at the sites of homology. The frequency of homologous recombination is influenced by a number of factors. Different organisms vary with respect to the amount of homologous recombination and the relative proportion of homologous to non-homologous recombination. Generally, the length of the region of homology affects the frequency of homologous recombination events: the longer the region of homology, the greater the frequency. The length of the homology region needed to observe homologous recombination is also species-variable. In many cases, at least 5 kb of homology has been utilized, but homologous recombination has been observed with as little as 25-50 bp of homology. See, for

example, Singer *et al.*, (1982) Cell 31:25-33; Shen and Huang, (1986) Genetics 112:441-57; Watt *et al.*, (1985) *Proc. Natl. Acad. Sci. USA* 82:4768-72, Sugawara and Haber, (1992) *Mol Cell Biol* 12:563-75, Rubnitz and Subramani, (1984) *Mol Cell Biol* 4:2253-8; Ayares *et al.*, (1986) *Proc. Natl. Acad. Sci. USA* 83:5199-203; Liskay *et al.*, (1987) *Genetics* 115:161-7.

[0082] "Sequence identity" or "identity" in the context of nucleic acid or polypeptide sequences refers to the nucleic acid bases or amino acid residues in two sequences that are the same when aligned for maximum correspondence over a specified comparison window.

[0083] The term "percentage of sequence identity" refers to the value determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide or polypeptide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the results by 100 to yield the percentage of sequence identity. Useful examples of percent sequence identities include, but are not limited to, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, or 95%, or any percentage from 50% to 100%. These identities can be determined using any of the programs described herein.

[0084] Sequence alignments and percent identity or similarity calculations may be determined using a variety of comparison methods designed to detect homologous sequences including, but not limited to, the MegAlign™ program of the LASERGENE bioinformatics computing suite (DNASTAR Inc., Madison, WI). Within the context of this application it will be understood that where sequence analysis software is used for analysis, that the results of the analysis will be based on the "default values" of the program referenced, unless otherwise specified. As used herein "default values" will mean any set of values or parameters that originally load with the software when first initialized.

[0085] The "Clustal V method of alignment" corresponds to the alignment method labeled Clustal V (described by Higgins and Sharp, (1989) *CABIOS* 5:151-153; Higgins *et al.*, (1992) *Comput Appl Biosci* 8:189-191) and found in the MegAlign™ program of the LASERGENE bioinformatics computing suite (DNASTAR Inc., Madison, WI). For multiple alignments, the default values correspond to GAP PENALTY=10 and GAP LENGTH PENALTY=10. Default parameters for pairwise alignments and calculation of percent identity of protein sequences using the Clustal method are KTUPLE=1, GAP PENALTY=3, WINDOW=5 and DIAGONALS

SAVED=5. For nucleic acids these parameters are KTUPLE=2, GAP PENALTY=5, WINDOW=4 and DIAGONALS SAVED=4. After alignment of the sequences using the Clustal V program, it is possible to obtain a "percent identity" by viewing the "sequence distances" Table in the same program. The "Clustal W method of alignment" corresponds to the alignment method labeled Clustal W (described by Higgins and Sharp, (1989) *CABIOS* 5:151-153; Higgins *et al.*, (1992) *Comput Appl Biosci* 8:189-191) and found in the MegAlign™ v6.1 program of the LASERGENE bioinformatics computing suite (DNASTAR Inc., Madison, WI). Default parameters for multiple alignment (GAP PENALTY=10, GAP LENGTH PENALTY=0.2, Delay Divergen Seqs (%)=30, DNA Transition Weight=0.5, Protein Weight Matrix=Gonnet Series, DNA Weight Matrix=IUB). After alignment of the sequences using the Clustal W program, it is possible to obtain a "percent identity" by viewing the "sequence distances" Table in the same program. Unless otherwise stated, sequence identity/similarity values provided herein refer to the value obtained using GAP Version 10 (GCG, Accelrys, San Diego, CA) using the following parameters:% identity and % similarity for a nucleotide sequence using a gap creation penalty weight of 50 and a gap length extension penalty weight of 3, and the nwsgapdna.cmp scoring matrix; % identity and % similarity for an amino acid sequence using a GAP creation penalty weight of 8 and a gap length extension penalty of 2, and the BLOSUM62 scoring matrix (Henikoff and Henikoff, (1989) *Proc. Natl. Acad. Sci. USA* 89:10915). GAP uses the algorithm of Needleman and Wunsch, (1970) *J Mol Biol* 48:443-53, to find an alignment of two complete sequences that maximizes the number of matches and minimizes the number of gaps. GAP considers all possible alignments and gap positions and creates the alignment with the largest number of matched bases and the fewest gaps, using a gap creation penalty and a gap extension penalty in units of matched bases. "BLAST" is a searching algorithm provided by the National Center for Biotechnology Information (NCBI) used to find regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches to identify sequences having sufficient similarity to a query sequence such that the similarity would not be predicted to have occurred randomly. BLAST reports the identified sequences and their local alignment to the query sequence. It is well understood by one skilled in the art that many levels of sequence identity are useful in identifying polypeptides from other species or modified naturally or synthetically wherein such polypeptides have the same or similar function or activity. Useful examples of percent identities include, but are not limited to, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90% or 95%, or any percentage from 50% to 100%. Indeed, any amino acid identity from 50% to 100% may be useful in describing the present disclosure, such as 51%, 52%, 53%,

54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99%.

[0086] Polynucleotide and polypeptide sequences, variants thereof, and the structural relationships of these sequences can be described by the terms “homology”, “homologous”, “substantially identical”, “substantially similar” and “corresponding substantially” which are used interchangeably herein. These refer to polypeptide or nucleic acid sequences wherein changes in one or more amino acids or nucleotide bases do not affect the function of the molecule, such as the ability to mediate gene expression or to produce a certain phenotype. These terms also refer to modification(s) of nucleic acid sequences that do not substantially alter the functional properties of the resulting nucleic acid relative to the initial, unmodified nucleic acid. These modifications include deletion, substitution, and/or insertion of one or more nucleotides in the nucleic acid fragment. Substantially similar nucleic acid sequences encompassed may be defined by their ability to hybridize (under moderately stringent conditions, e.g., 0.5X SSC, 0.1% SDS, 60°C) with the sequences exemplified herein, or to any portion of the nucleotide sequences disclosed herein and which are functionally equivalent to any of the nucleic acid sequences disclosed herein. Stringency conditions can be adjusted to screen for moderately similar fragments, such as homologous sequences from distantly related organisms, to highly similar fragments, such as genes that duplicate functional enzymes from closely related organisms. Post-hybridization washes determine stringency conditions.

[0087] A "centimorgan" (cM) or "map unit" is the distance between two polynucleotide sequences, linked genes, markers, target sites, loci, or any pair thereof, wherein 1% of the products of meiosis are recombinant. Thus, a centimorgan is equivalent to a distance equal to a 1% average recombination frequency between the two linked genes, markers, target sites, loci, or any pair thereof.

[0088] An "isolated" or "purified" nucleic acid molecule, polynucleotide, polypeptide, or protein, or biologically active portion thereof, is substantially or essentially free from components that normally accompany or interact with the polynucleotide or protein as found in its naturally occurring environment. Thus, an isolated or purified polynucleotide or polypeptide or protein is substantially free of other cellular material, or culture medium when produced by recombinant techniques, or substantially free of chemical precursors or other chemicals when chemically synthesized. Optimally, an "isolated" polynucleotide is free of sequences (optimally protein encoding sequences) that naturally flank the polynucleotide (i.e., sequences located at the 5' and 3' ends of the polynucleotide) in the genomic DNA of the organism from which the

polynucleotide is derived. For example, in various embodiments, the isolated polynucleotide can contain less than about 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb, or 0.1 kb of nucleotide sequence that naturally flank the polynucleotide in genomic DNA of the cell from which the polynucleotide is derived. Isolated polynucleotides may be purified from a cell in which they naturally occur. Conventional nucleic acid purification methods known to skilled artisans may be used to obtain isolated polynucleotides. The term also embraces recombinant polynucleotides and chemically synthesized polynucleotides.

[0089] The term "fragment" refers to a contiguous set of nucleotides or amino acids. In one embodiment, a fragment is 2, 3, 4, 5, 6, 7 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or greater than 20 contiguous nucleotides. In one embodiment, a fragment is 2, 3, 4, 5, 6, 7 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or greater than 20 contiguous amino acids. A fragment may or may not exhibit the function of a sequence sharing some percent identity over the length of said fragment.

[0090] The terms "fragment that is functionally equivalent" and "functionally equivalent fragment" are used interchangeably herein. These terms refer to a portion or subsequence of an isolated nucleic acid fragment or polypeptide that displays the same activity or function as the longer sequence from which it derives. In one example, the fragment retains the ability to alter gene expression or produce a certain phenotype whether or not the fragment encodes an active protein. For example, the fragment can be used in the design of genes to produce the desired phenotype in a modified plant. Genes can be designed for use in suppression by linking a nucleic acid fragment, whether or not it encodes an active enzyme, in the sense or antisense orientation relative to a plant promoter sequence.

[0091] "Gene" includes a nucleic acid fragment that expresses a functional molecule such as, but not limited to, a specific protein, including regulatory sequences preceding (5' non-coding sequences) and following (3' non-coding sequences) the coding sequence. "Native gene" refers to a gene as found in its natural endogenous location with its own regulatory sequences.

[0092] By the term "endogenous" it is meant a sequence or other molecule that naturally occurs in a cell or organism. In one aspect, an endogenous polynucleotide is normally found in the genome of a cell; that is, not heterologous.

[0093] An "allele" is one of several alternative forms of a gene occupying a given locus on a chromosome. When all the alleles present at a given locus on a chromosome are the same, that plant is homozygous at that locus. If the alleles present at a given locus on a chromosome differ, that plant is heterozygous at that locus.

[0094] "Coding sequence" refers to a polynucleotide sequence which codes for a specific amino acid sequence. "Regulatory sequences" refer to nucleotide sequences located upstream (5' non-coding sequences), within, or downstream (3' non-coding sequences) of a coding sequence, and which influence the transcription, RNA processing or stability, or translation of the associated coding sequence. Regulatory sequences include, but are not limited to, promoters, translation leader sequences, 5' untranslated sequences, 3' untranslated sequences, introns, polyadenylation target sequences, RNA processing sites, effector binding sites, and stem-loop structures.

[0095] A "mutated gene" is a gene that has been altered through human intervention. Such a "mutated gene" has a sequence that differs from the sequence of the corresponding non-mutated gene by at least one nucleotide addition, deletion, or substitution. In certain embodiments of the disclosure, the mutated gene comprises an alteration that results from a guide polynucleotide/Cas endonuclease system as disclosed herein. A mutated plant is a plant comprising a mutated gene.

[0096] As used herein, a "targeted mutation" is a mutation in a gene (referred to as the target gene), including a native gene, that was made by altering a target sequence within the target gene using any method known to one skilled in the art, including a method involving a guided Cas endonuclease system as disclosed herein.

[0097] The terms "knock-out", "gene knock-out" and "genetic knock-out" are used interchangeably herein. A knock-out represents a DNA sequence of a cell that has been rendered partially or completely inoperative by targeting with a Cas protein; for example, a DNA sequence prior to knock-out could have encoded an amino acid sequence, or could have had a regulatory function (e.g., promoter).

[0098] The terms "knock-in", "gene knock-in, "gene insertion" and "genetic knock-in" are used interchangeably herein. A knock-in represents the replacement or insertion of a DNA sequence at a specific DNA sequence in cell by targeting with a Cas protein (for example by homologous recombination (HR), wherein a suitable donor DNA polynucleotide is also used). examples of knock-ins are a specific insertion of a heterologous amino acid coding sequence in a coding region of a gene, or a specific insertion of a transcriptional regulatory element in a genetic locus.

[0099] By "domain" it is meant a contiguous stretch of nucleotides (that can be RNA, DNA, and/or RNA-DNA-combination sequence) or amino acids.

[0100] The term "conserved domain" or "motif" means a set of polynucleotides or amino acids conserved at specific positions along an aligned sequence of evolutionarily related

proteins. While amino acids at other positions can vary between homologous proteins, amino acids that are highly conserved at specific positions indicate amino acids that are essential to the structure, the stability, or the activity of a protein. Because they are identified by their high degree of conservation in aligned sequences of a family of protein homologues, they can be used as identifiers, or "signatures", to determine if a protein with a newly determined sequence belongs to a previously identified protein family.

[0101] A "codon-modified gene" or "codon-preferred gene" or "codon-optimized gene" is a gene having its frequency of codon usage designed to mimic the frequency of preferred codon usage of the host cell.

[0102] An "optimized" polynucleotide is a sequence that has been optimized for improved expression in a particular heterologous host cell.

[0103] A "plant-optimized nucleotide sequence" is a nucleotide sequence that has been optimized for expression in plants, particularly for increased expression in plants. A plant-optimized nucleotide sequence includes a codon-optimized gene. A plant-optimized nucleotide sequence can be synthesized by modifying a nucleotide sequence encoding a protein such as, for example, a Cas endonuclease as disclosed herein, using one or more plant-preferred codons for improved expression. *See*, for example, Campbell and Gowri (1990) *Plant Physiol.* 92:1-11 for a discussion of host-preferred codon usage.

[0104] A "promoter" is a region of DNA involved in recognition and binding of RNA polymerase and other proteins to initiate transcription. The promoter sequence consists of proximal and more distal upstream elements, the latter elements often referred to as enhancers. An "enhancer" is a DNA sequence that can stimulate promoter activity and may be an innate element of the promoter or a heterologous element inserted to enhance the level or tissue-specificity of a promoter. Promoters may be derived in their entirety from a native gene, or be composed of different elements derived from different promoters found in nature, and/or comprise synthetic DNA segments. It is understood by those skilled in the art that different promoters may direct the expression of a gene in different tissues or cell types, or at different stages of development, or in response to different environmental conditions. It is further recognized that since in most cases the exact boundaries of regulatory sequences have not been completely defined, DNA fragments of some variation may have identical promoter activity.

[0105] Promoters that cause a gene to be expressed in most cell types at most times are commonly referred to as "constitutive promoters". The term "inducible promoter" refers to a promoter that selectively express a coding sequence or functional RNA in response to the presence of an endogenous or exogenous stimulus, for example by chemical compounds

(chemical inducers) or in response to environmental, hormonal, chemical, and/or developmental signals. Inducible or regulated promoters include, for example, promoters induced or regulated by light, heat, stress, flooding or drought, salt stress, osmotic stress, phytohormones, wounding, or chemicals such as ethanol, abscisic acid (ABA), jasmonate, salicylic acid, or safeners.

[0106] "Translation leader sequence" refers to a polynucleotide sequence located between the promoter sequence of a gene and the coding sequence. The translation leader sequence is present in the mRNA upstream of the translation start sequence. The translation leader sequence may affect processing of the primary transcript to mRNA, mRNA stability or translation efficiency. Examples of translation leader sequences have been described (e.g., Turner and Foster, (1995) *Mol Biotechnol* 3:225-236).

[0107] "3' non-coding sequences", "transcription terminator" or "termination sequences" refer to DNA sequences located downstream of a coding sequence and include polyadenylation recognition sequences and other sequences encoding regulatory signals capable of affecting mRNA processing or gene expression. The polyadenylation signal is usually characterized by affecting the addition of polyadenylic acid tracts to the 3' end of the mRNA precursor. The use of different 3' non-coding sequences is exemplified by Ingelbrecht *et al.*, (1989) *Plant Cell* 1:671-680.

[0108] "RNA transcript" refers to the product resulting from RNA polymerase-catalyzed transcription of a DNA sequence. When the RNA transcript is a perfect complimentary copy of the DNA sequence, it is referred to as the primary transcript or pre-mRNA. An RNA transcript is referred to as the mature RNA or mRNA when it is an RNA sequence derived from post-transcriptional processing of the primary transcript pre-mRNA. "Messenger RNA" or "mRNA" refers to the RNA that is without introns and that can be translated into protein by the cell. "cDNA" refers to a DNA that is complementary to, and synthesized from, an mRNA template using the enzyme reverse transcriptase. The cDNA can be single-stranded or converted into double-stranded form using the Klenow fragment of DNA polymerase I. "Sense" RNA refers to RNA transcript that includes the mRNA and can be translated into protein within a cell or *in vitro*. "Antisense RNA" refers to an RNA transcript that is complementary to all or part of a target primary transcript or mRNA, and that blocks the expression of a target gene (see, e.g., U.S. Patent No. 5,107,065). The complementarity of an antisense RNA may be with any part of the specific gene transcript, i.e., at the 5' non-coding sequence, 3' non-coding sequence, introns, or the coding sequence. "Functional RNA" refers to antisense RNA, ribozyme RNA, or other RNA that may not be translated but yet has an effect on cellular processes. The terms

"complement" and "reverse complement" are used interchangeably herein with respect to mRNA transcripts and are meant to define the antisense RNA of the message.

[0109] The term "genome" refers to the entire complement of genetic material (genes and non-coding sequences) that is present in each cell of an organism, or virus or organelle; and/or a complete set of chromosomes inherited as a (haploid) unit from one parent.

[0110] The term "operably linked" refers to the association of nucleic acid sequences on a single nucleic acid fragment so that the function of one is regulated by the other. For example, a promoter is operably linked with a coding sequence when it is capable of regulating the expression of that coding sequence (i.e., the coding sequence is under the transcriptional control of the promoter). Coding sequences can be operably linked to regulatory sequences in a sense or antisense orientation. In another example, the complementary RNA regions can be operably linked, either directly or indirectly, 5' to the target mRNA, or 3' to the target mRNA, or within the target mRNA, or a first complementary region is 5' and its complement is 3' to the target mRNA.

[0111] Generally, "host" refers to an organism or cell into which a heterologous component (polynucleotide, polypeptide, other molecule, cell) has been introduced. As used herein, a "host cell" refers to an *in vivo* or *in vitro* eukaryotic cell, prokaryotic cell (e.g., bacterial or archaeal cell), or cell from a multicellular organism (e.g., a cell line) cultured as a unicellular entity, into which a heterologous polynucleotide or polypeptide has been introduced. In some embodiments, the cell is selected from the group consisting of: an archaeal cell, a bacterial cell, a eukaryotic cell, a eukaryotic single-cell organism, a somatic cell, a germ cell, a stem cell, a plant cell, an algal cell, an animal cell, in invertebrate cell, a vertebrate cell, a fish cell, a frog cell, a bird cell, an insect cell, a mammalian cell, a pig cell, a cow cell, a goat cell, a sheep cell, a rodent cell, a rat cell, a mouse cell, a non-human primate cell, and a human cell. In some cases, the cell is *in vitro*. In some cases, the cell is *in vivo*.

[0112] The term "recombinant" refers to an artificial combination of two otherwise separated segments of sequence, e.g., by chemical synthesis, or manipulation of isolated segments of nucleic acids by genetic engineering techniques.

[0113] The terms "plasmid", "vector" and "cassette" refer to a linear or circular extra chromosomal element often carrying genes that are not part of the central metabolism of the cell, and usually in the form of double-stranded DNA. Such elements may be autonomously replicating sequences, genome integrating sequences, phage, or nucleotide sequences, in linear or circular form, of a single- or double-stranded DNA or RNA, derived from any source, in which a number of nucleotide sequences have been joined or recombined into a unique

construction which is capable of introducing a polynucleotide of interest into a cell. "Transformation cassette" refers to a specific vector comprising a gene and having elements in addition to the gene that facilitates transformation of a particular host cell. "Expression cassette" refers to a specific vector comprising a gene and having elements in addition to the gene that allow for expression of that gene in a host.

[0114] The terms "recombinant DNA molecule", "recombinant DNA construct", "expression construct", "construct", and "recombinant construct" are used interchangeably herein. A recombinant DNA construct comprises an artificial combination of nucleic acid sequences, e.g., regulatory and coding sequences that are not all found together in nature. For example, a recombinant DNA construct may comprise regulatory sequences and coding sequences that are derived from different sources, or regulatory sequences and coding sequences derived from the same source, but arranged in a manner different than that found in nature. Such a construct may be used by itself or may be used in conjunction with a vector. If a vector is used, then the choice of vector is dependent upon the method that will be used to introduce the vector into the host cells as is well known to those skilled in the art. For example, a plasmid vector can be used. The skilled artisan is well aware of the genetic elements that must be present on the vector in order to successfully transform, select and propagate host cells. The skilled artisan will also recognize that different independent transformation events may result in different levels and patterns of expression (Jones *et al.*, (1985) *EMBO J* 4:2411-2418; De Almeida *et al.*, (1989) *Mol Gen Genetics* 218:78-86), and thus that multiple events are typically screened in order to obtain lines displaying the desired expression level and pattern. Such screening may be accomplished standard molecular biological, biochemical, and other assays including Southern analysis of DNA, Northern analysis of mRNA expression, PCR, real time quantitative PCR (qPCR), reverse transcription PCR (RT-PCR), immunoblotting analysis of protein expression, enzyme, or activity assays, and/or phenotypic analysis.

[0115] The term "heterologous" refers to the difference between the original environment, location, or composition of a particular polynucleotide or polypeptide sequence and its current environment, location, or composition. Non-limiting examples include differences in taxonomic derivation (*e.g.*, a polynucleotide sequence obtained from *Zea mays* would be heterologous if inserted into the genome of an *Oryza sativa* plant, or of a different variety or cultivar of *Zea mays*; or a polynucleotide obtained from a bacterium was introduced into a cell of a plant), or sequence (*e.g.*, a polynucleotide sequence obtained from *Zea mays*, isolated, modified, and re-introduced into a maize plant). As used herein, "heterologous" in reference to a sequence can refer to a sequence that originates from a different species, variety, foreign species, or, if from

the same species, is substantially modified from its native form in composition and/or genomic locus by deliberate human intervention. For example, a promoter operably linked to a heterologous polynucleotide is from a species different from the species from which the polynucleotide was derived, or, if from the same/analogous species, one or both are substantially modified from their original form and/or genomic locus, or the promoter is not the native promoter for the operably linked polynucleotide. Alternatively, one or more regulatory region(s) and/or a polynucleotide provided herein may be entirely synthetic. In another example, a target polynucleotide for cleavage by a Cas endonuclease may be of a different organism than that of the Cas endonuclease. In another example, a Cas endonuclease and guide RNA may be introduced to a target polynucleotide with an additional polynucleotide that acts as a template or donor for insertion into the target polynucleotide, wherein the additional polynucleotide is heterologous to the target polynucleotide and/or the Cas endonuclease.

[0116] The term "expression", as used herein, refers to the production of a functional end-product (e.g., an mRNA, guide RNA, or a protein) in either precursor or mature form.

[0117] A "mature" protein refers to a post-translationally processed polypeptide (i.e., one from which any pre- or propeptides present in the primary translation product have been removed).

[0118] "Precursor" protein refers to the primary product of translation of mRNA (i.e., with pre- and propeptides still present). Pre- and propeptides may be but are not limited to intracellular localization signals.

[0119] "CRISPR" (Clustered Regularly Interspaced Short Palindromic Repeats) loci refers to certain genetic loci encoding components of DNA cleavage systems, for example, used by bacterial and archaeal cells to destroy foreign DNA (Horvath and Barrangou, 2010, *Science* 327:167-170; WO2007025097, published 01 March 2007). A CRISPR locus can consist of a CRISPR array, comprising short direct repeats (CRISPR repeats) separated by short variable DNA sequences (called spacers), which can be flanked by diverse Cas (CRISPR-associated) genes.

[0120] As used herein, an "effector" or "effector protein" is a protein that encompasses an activity including recognizing, binding to, and/or cleaving or nicking a polynucleotide target. An effector, or effector protein, may also be an endonuclease. The "effector complex" of a CRISPR system includes Cas proteins involved in crRNA and target recognition and binding. Some of the component Cas proteins may additionally comprise domains involved in target polynucleotide cleavage.

[0121] The term "Cas protein" refers to a polypeptide encoded by a Cas (CRISPR-associated) gene. Cas proteins include proteins encoded by a gene in a *cas* locus and include adaptation molecules as well as interference molecules. An interference molecule of a bacterial adaptive immunity complex includes endonucleases. A Cas endonuclease described herein comprises one or more nuclease domains. A Cas endonuclease includes but is not limited to the novel engineered Cas endonuclease polypeptide disclosed herein, a Cas9 protein, a Cpf1 (Cas12) protein, a C2c1 protein, a C2c2 protein, a C2c3 protein, Cas3, Cas3-HD, Cas 5, Cas7, Cas8, Cas10, or combinations or complexes of these. A Cas protein may be a "Cas endonuclease" or "Cas effector protein", that when in complex with a suitable polynucleotide component, is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a specific polynucleotide target sequence. The Cas endonucleases of the disclosure include those having one or more RuvC nuclease domains. A Cas protein is further defined as a functional fragment or functional variant of a native Cas protein, or a protein that shares at least 50%, between 50% and 55%, at least 55%, between 55% and 60%, at least 60%, between 60% and 65%, at least 65%, between 65% and 70%, at least 70%, between 70% and 75%, at least 75%, between 75% and 80%, at least 80%, between 80% and 85%, at least 85%, between 85% and 90%, at least 90%, between 90% and 95%, at least 95%, between 95% and 96%, at least 96%, between 96% and 97%, at least 97%, between 97% and 98%, at least 98%, between 98% and 99%, at least 99%, between 99% and 100%, or 100% sequence identity with at least 50, between 50 and 100, at least 100, between 100 and 150, at least 150, between 150 and 200, at least 200, between 200 and 250, at least 250, between 250 and 300, at least 300, between 300 and 350, at least 350, between 350 and 400, at least 400, between 400 and 450, at least 500, or greater than 500 contiguous amino acids of a native Cas protein, and retains at least partial activity of the native sequence.

[0122] The terms "functional fragment", "fragment that is functionally equivalent" and "functionally equivalent fragment" of a Cas endonuclease are used interchangeably herein and refer to a portion or subsequence of the Cas endonuclease of the present disclosure in which the ability to recognize, bind to, and optionally unwind, nick, or cleave (introduce a single or double-strand break in) the target site is retained. The portion or subsequence of the Cas endonuclease can comprise a complete or partial (functional) peptide of any one of its domains such as for example, but not limiting to a complete of functional part of a Cas3 HD domain, a complete of functional part of a Cas3 Helicase domain, complete of functional part of a protein (such as but not limiting to a Cas5, Cas5d, Cas7 and Cas8b1).

[0123] The terms “functional variant”, “variant that is functionally equivalent” and “functionally equivalent variant” of a Cas endonuclease or Cas effector protein, including Cas endonuclease described herein, are used interchangeably herein, and refer to a variant of the Cas effector protein disclosed herein in which the ability to recognize, bind to, and optionally unwind, nick, or cleave all or part of a target sequence is retained.

[0124] A Cas endonuclease may also include a multifunctional Cas endonuclease. The term “multifunctional Cas endonuclease” and “multifunctional Cas endonuclease polypeptide” are used interchangeably herein and includes reference to a single polypeptide that has Cas endonuclease functionality (comprising at least one protein domain that can act as a Cas endonuclease) and at least one other functionality, such as but not limited to, the functionality to form a complex (comprises at least a second protein domain that can form a complex with other proteins). In one aspect, the multifunctional Cas endonuclease comprises at least one additional protein domain relative (either internally, upstream (5’), downstream (3’), or both internally 5’ and 3’, or any combination thereof) to those domains typical of a Cas endonuclease.

[0125] The terms “cascade” and “cascade complex” are used interchangeably herein and include reference to a multi-subunit protein complex that can assemble with a polynucleotide forming a polynucleotide-protein complex (PNP). Cascade is a PNP that relies on the polynucleotide for complex assembly and stability, and for the identification of target nucleic acid sequences. Cascade functions as a surveillance complex that finds and optionally binds target nucleic acids that are complementary to a variable targeting domain of the guide polynucleotide.

[0126] The terms ”cleavage-ready Cascade”, “crCascade”, ”cleavage-ready Cascade complex”, “crCascade complex”, ”cleavage-ready Cascade system”, “CRC” and “crCascade system”, are used interchangeably herein and include reference to a multi-subunit protein complex that can assemble with a polynucleotide forming a polynucleotide-protein complex (PNP), wherein one of the cascade proteins is a Cas endonuclease capable of recognizing, binding to, and optionally unwinding, nicking, or cleaving all or part of a target sequence.

[0127] The terms “5’-cap” and “7-methylguanylate (m7G) cap” are used interchangeably herein. A 7-methylguanylate residue is located on the 5′ terminus of messenger RNA (mRNA) in eukaryotes. RNA polymerase II (Pol II) transcribes mRNA in eukaryotes. Messenger RNA capping occurs generally as follows: the most terminal 5’ phosphate group of the mRNA transcript is removed by RNA terminal phosphatase, leaving two terminal phosphates. A guanosine monophosphate (GMP) is added to the terminal phosphate of the transcript by a

guanylyl transferase, leaving a 5′-5′ triphosphate-linked guanine at the transcript terminus. Finally, the 7-nitrogen of this terminal guanine is methylated by a methyl transferase.

[0128] The terminology “not having a 5’-cap” herein is used to refer to RNA having, for example, a 5’-hydroxyl group instead of a 5’-cap. Such RNA can be referred to as “uncapped RNA”, for example. Uncapped RNA can better accumulate in the nucleus following transcription, since 5’-capped RNA is subject to nuclear export. One or more RNA components herein are uncapped.

[0129] As used herein, the term “guide polynucleotide”, relates to a polynucleotide sequence that can form a complex with a Cas endonuclease, including the Cas endonuclease described herein, and enables the Cas endonuclease to recognize, optionally bind to, and optionally cleave a DNA target site. The guide polynucleotide sequence can be an RNA sequence, a DNA sequence, or a combination thereof (an RNA-DNA combination sequence).

[0130] The terms “functional fragment”, “fragment that is functionally equivalent” and “functionally equivalent fragment” of a guide RNA, crRNA or tracrRNA are used interchangeably herein, and refer to a portion or subsequence of the guide RNA, crRNA or tracrRNA, respectively, of the present disclosure in which the ability to function as a guide RNA, crRNA or tracrRNA, respectively, is retained.

[0131] The terms “functional variant “, “variant that is functionally equivalent” and “functionally equivalent variant” of a guide RNA, crRNA or tracrRNA (respectively) are used interchangeably herein, and refer to a variant of the guide RNA, crRNA or tracrRNA, respectively, of the present disclosure in which the ability to function as a guide RNA, crRNA or tracrRNA, respectively, is retained.

[0132] The terms “single guide RNA" and “sgRNA” are used interchangeably herein and relate to a synthetic fusion of two RNA molecules, a crRNA (CRISPR RNA) comprising a variable targeting domain (linked to a tracr mate sequence that hybridizes to a tracrRNA), fused to a tracrRNA (trans-activating CRISPR RNA). The single guide RNA can comprise a crRNA or crRNA fragment and a tracrRNA or tracrRNA fragment of the type II CRISPR/Cas system that can form a complex with a type II Cas endonuclease, wherein said guide RNA/Cas endonuclease complex can direct the Cas endonuclease to a DNA target site, enabling the Cas endonuclease to recognize, optionally bind to, and optionally nick or cleave (introduce a single or double-strand break) the DNA target site.

[0133] The term “variable targeting domain” or “VT domain” is used interchangeably herein and includes a nucleotide sequence that can hybridize (is complementary) to one strand (nucleotide sequence) of a double strand DNA target site. The percent complementation between

the first nucleotide sequence domain (VT domain) and the target sequence can be at least 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 63%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or 100%. The variable targeting domain can be at least 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 or 30 nucleotides in length. In some embodiments, the variable targeting domain comprises a contiguous stretch of 12 to 30 nucleotides. The variable targeting domain can be composed of a DNA sequence, an RNA sequence, a modified DNA sequence, a modified RNA sequence, or any combination thereof.

[0134] The term "Cas endonuclease recognition domain" or "CER domain" (of a guide polynucleotide) is used interchangeably herein and includes a nucleotide sequence that interacts with a Cas endonuclease polypeptide. A CER domain comprises a (trans-acting) tracrNucleotide mate sequence followed by a tracrNucleotide sequence. The CER domain can be composed of a DNA sequence, an RNA sequence, a modified DNA sequence, a modified RNA sequence (see for example US20150059010A1, published 26 February 2015), or any combination thereof.

[0135] As used herein, the terms "guide polynucleotide/Cas endonuclease complex", "guide polynucleotide/Cas endonuclease system", " guide polynucleotide/Cas complex", "guide polynucleotide/Cas system" and "guided Cas system" "Polynucleotide-guided endonuclease" , "PGEN" are used interchangeably herein and refer to at least one guide polynucleotide and at least one Cas endonuclease, that are capable of forming a complex, wherein said guide polynucleotide/Cas endonuclease complex can direct the Cas endonuclease to a DNA target site, enabling the Cas endonuclease to recognize, bind to, and optionally nick or cleave (introduce a single or double-strand break) the DNA target site. A guide polynucleotide/Cas endonuclease complex herein can comprise Cas protein(s) and suitable polynucleotide component(s) of any of the known CRISPR systems (Horvath and Barrangou, 2010, *Science* 327:167-170; Makarova *et al.* 2015, *Nature Reviews Microbiology* Vol. 13:1-15; Zetsche *et al.*, 2015, Cell 163, 1-13; Shmakov *et al.*, 2015, *Molecular Cell* 60, 1-13).

[0136] The terms "guide RNA/Cas endonuclease complex", "guide RNA/Cas endonuclease system", " guide RNA/Cas complex", "guide RNA/Cas system", "gRNA/Cas complex", "gRNA/Cas system", "RNA-guided endonuclease" , "RGEN" are used interchangeably herein and refer to at least one RNA component and at least one Cas endonuclease that are capable of forming a complex , wherein said guide RNA/Cas endonuclease complex can direct the Cas endonuclease to a DNA target site, enabling the Cas endonuclease to recognize, bind to, and optionally nick or cleave (introduce a single or double-strand break) the DNA target site.

[0137] The terms “target site”, “target sequence”, “target site sequence, ”target DNA”, “target locus”, “genomic target site”, “genomic target sequence”, “genomic target locus” and “protospacer”, are used interchangeably herein and refer to a polynucleotide sequence such as, but not limited to, a nucleotide sequence on a chromosome, episome, a locus, or any other DNA molecule in the genome (including chromosomal, chloroplastic, mitochondrial DNA, plasmid DNA) of a cell, at which a guide polynucleotide/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave . The target site can be an endogenous site in the genome of a cell, or alternatively, the target site can be heterologous to the cell and thereby not be naturally occurring in the genome of the cell, or the target site can be found in a heterologous genomic location compared to where it occurs in nature. As used herein, terms “endogenous target sequence” and “native target sequence” are used interchangeable herein to refer to a target sequence that is endogenous or native to the genome of a cell and is at the endogenous or native position of that target sequence in the genome of the cell. An “artificial target site” or “artificial target sequence” are used interchangeably herein and refer to a target sequence that has been introduced into the genome of a cell. Such an artificial target sequence can be identical in sequence to an endogenous or native target sequence in the genome of a cell but be located in a different position (*i.e.*, a non-endogenous or non-native position) in the genome of a cell.

[0138] A “protospacer adjacent motif” (PAM) herein refers to a short nucleotide sequence adjacent to a target sequence (protospacer) that is recognized (targeted) by a guide polynucleotide/Cas endonuclease system described herein. The Cas endonuclease may not successfully recognize a target DNA sequence if the target DNA sequence is not followed by a PAM sequence. The sequence and length of a PAM herein can differ depending on the Cas protein or Cas protein complex used. The PAM sequence can be of any length but is typically 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 nucleotides long.

[0139] An “altered target site”, “altered target sequence”, “modified target site”, “modified target sequence” are used interchangeably herein and refer to a target sequence as disclosed herein that comprises at least one alteration when compared to non-altered target sequence. Such “alterations” include, for example: (i) replacement of at least one nucleotide, (ii) a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, (iv) a chemical alteration of at least one nucleotide, or (v) any combination of (i) – (iv).

[0140] A “modified nucleotide” or “edited nucleotide” refers to a nucleotide sequence of interest that comprises at least one alteration when compared to its non-modified nucleotide sequence. Such “alterations” include, for example: (i) replacement of at least one nucleotide, (ii)

a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, (iv) a chemical alteration of at least one nucleotide, or (v) any combination of (i) – (iv).

[0141] Methods for "modifying a target site" and "altering a target site" are used interchangeably herein and refer to methods for producing an altered target site.

[0142] As used herein, "donor DNA" is a DNA construct that comprises a polynucleotide of interest to be inserted into the target site of a Cas endonuclease.

[0143] The term "polynucleotide modification template" includes a polynucleotide that comprises at least one nucleotide modification when compared to the nucleotide sequence to be edited. A nucleotide modification can be at least one nucleotide substitution, addition, or deletion. Optionally, the polynucleotide modification template can further comprise homologous nucleotide sequences flanking the at least one nucleotide modification, wherein the flanking homologous nucleotide sequences provide sufficient homology to the desired nucleotide sequence to be edited.

[0144] The term "plant-optimized Cas endonuclease" herein refers to a Cas protein, including a multifunctional Cas protein, encoded by a nucleotide sequence that has been optimized for expression in a plant cell or plant.

[0145] A "plant-optimized nucleotide sequence encoding a Cas endonuclease", "plant-optimized construct encoding a Cas endonuclease" and a "plant-optimized polynucleotide encoding a Cas endonuclease" are used interchangeably herein and refer to a nucleotide sequence encoding a Cas protein, or a variant or functional fragment thereof, that has been optimized for expression in a plant cell or plant. A plant comprising a plant-optimized Cas endonuclease includes a plant comprising the nucleotide sequence encoding for the Cas sequence and/or a plant comprising the Cas endonuclease protein. In one aspect, the plant-optimized Cas endonuclease nucleotide sequence is a maize-optimized, rice-optimized, wheat-optimized, soybean-optimized, cotton-optimized, or canola-optimized Cas endonuclease.

[0146] The term "plant" generically includes whole plants, plant organs, plant tissues, seeds, plant cells, seeds, and progeny of the same. The plant is a monocot or dicot. Plant cells include, without limitation, cells from seeds, suspension cultures, embryos, meristematic regions, callus tissue, leaves, roots, shoots, gametophytes, sporophytes, pollen, and microspores. A "plant element" is intended to reference either a whole plant or a plant component, which may comprise differentiated and/or undifferentiated tissues, for example but not limited to plant tissues, parts, and cell types. In one embodiment, a plant element is one of the following: whole plant, seedling, meristematic tissue, ground tissue, vascular tissue, dermal tissue, seed, leaf, root, shoot, stem, flower, fruit, stolon, bulb, tuber, corm, keiki, shoot, bud, tumor tissue, and various forms of cells

and culture (e.g., single cells, protoplasts, embryos, callus tissue). It should be noted that a protoplast is not technically an “intact” plant cell (as naturally found with all components), as protoplasts lack a cell wall. The term "plant organ" refers to plant tissue or a group of tissues that constitute a morphologically and functionally distinct part of a plant. As used herein, a "plant element" is synonymous to a "portion" of a plant, and refers to any part of the plant, and can include distinct tissues and/or organs, and may be used interchangeably with the term "tissue" throughout. Similarly, a "plant reproductive element" is intended to generically reference any part of a plant that is able to initiate other plants via either sexual or asexual reproduction of that plant, for example but not limited to seed, seedling, root, shoot, cutting, scion, graft, stolon, bulb, tuber, corm, keiki, or bud. The plant element may be in plant or in a plant organ, tissue culture, or cell culture.

[0147] “Progeny” comprises any subsequent generation of a plant.

[0148] As used herein, the term “plant part” refers to plant cells, plant protoplasts, plant cell tissue cultures from which plants can be regenerated, plant calli, plant clumps, and plant cells that are intact in plants or parts of plants such as embryos, pollen, ovules, seeds, leaves, flowers, branches, fruit, kernels, ears, cobs, husks, stalks, roots, root tips, anthers, and the like, as well as the parts themselves. Grain is intended to mean the mature seed produced by commercial growers for purposes other than growing or reproducing the species. Progeny, variants, and mutants of the regenerated plants are also included within the scope of the disclosure, provided that these parts comprise the introduced polynucleotides.

[0149] The term “monocotyledonous” or “monocot” refers to the subclass of angiosperm plants also known as “monocotyledoneae”, whose seeds typically comprise only one embryonic leaf, or cotyledon. The term includes references to whole plants, plant elements, plant organs (e.g., leaves, stems, roots, etc.), seeds, plant cells, and progeny of the same.

[0150] The term “dicotyledonous” or “dicot” refers to the subclass of angiosperm plants also knows as “dicotyledoneae”, whose seeds typically comprise two embryonic leaves, or cotyledons. The term includes references to whole plants, plant elements, plant organs (e.g., leaves, stems, roots, etc.), seeds, plant cells, and progeny of the same.

[0151] As used herein, a "male sterile plant" is a plant that does not produce male gametes that are viable or otherwise capable of fertilization. As used herein, a "female sterile plant" is a plant that does not produce female gametes that are viable or otherwise capable of fertilization. It is recognized that male-sterile and female-sterile plants can be female-fertile and male- fertile, respectively. It is further recognized that a male fertile (but female sterile) plant can produce

viable progeny when crossed with a female fertile plant and that a female fertile (but male sterile) plant can produce viable progeny when crossed with a male fertile plant.

[0152] The term "non-conventional yeast" herein refers to any yeast that is not a *Saccharomyces* (e.g., *S. cerevisiae*) or *Schizosaccharomyces* yeast species. (see "Non-Conventional Yeasts in Genetics, Biochemistry and Biotechnology: Practical Protocols", K. Wolf, K.D. Breunig, G. Barth, Eds., Springer-Verlag, Berlin, Germany, 2003).

[0153] The term "crossed" or "cross" or "crossing" in the context of this disclosure means the fusion of gametes via pollination to produce progeny (i.e., cells, seeds, or plants). The term encompasses both sexual crosses (the pollination of one plant by another) and selfing (self-pollination, i.e., when the pollen and ovule (or microspores and megaspores) are from the same plant or genetically identical plants).

[0154] The term "introgression" refers to the transmission of a desired allele of a genetic locus from one genetic background to another. For example, introgression of a desired allele at a specified locus can be transmitted to at least one progeny plant via a sexual cross between two parent plants, where at least one of the parent plants has the desired allele within its genome. Alternatively, for example, transmission of an allele can occur by recombination between two donor genomes, e.g., in a fused protoplast, where at least one of the donor protoplasts has the desired allele in its genome. The desired allele can be, e.g., a transgene, a modified (mutated or edited) native allele, or a selected allele of a marker or QTL.

[0155] The term "isoline" is a comparative term, and references organisms that are genetically identical, but differ in treatment. In one example, two genetically identical maize plant embryos may be separated into two different groups, one receiving a treatment (such as the introduction of a CRISPR-Cas effector endonuclease) and one control that does not receive such treatment. Any phenotypic differences between the two groups may thus be attributed solely to the treatment and not to any inherency of the plant's endogenous genetic makeup.

[0156] "Introducing" is intended to mean presenting to a target, such as a cell or organism, a polynucleotide or polypeptide or polynucleotide-protein complex, in such a manner that the component(s) gains access to the interior of a cell of the organism or to the cell itself.

[0157] A "polynucleotide of interest" includes any nucleotide sequence encoding a protein or polypeptide that improves desirability of crops, i.e. a trait of agronomic interest. Polynucleotides of interest include, but are not limited to: polynucleotides encoding important traits for agronomics, herbicide-resistance, insecticidal resistance, disease resistance, nematode resistance, herbicide resistance, microbial resistance, fungal resistance, viral resistance, fertility or sterility, grain characteristics, commercial products, phenotypic marker, or any other trait of

agronomic or commercial importance. A polynucleotide of interest may additionally be utilized in either the sense or anti-sense orientation. Further, more than one polynucleotide of interest may be utilized together, or "stacked", to provide additional benefit.

[0158] A "complex trait locus" includes a genomic locus that has multiple transgenes genetically linked to each other.

[0159] The compositions and methods herein may provide for an improved "agronomic trait" or "trait of agronomic importance" or "trait of agronomic interest" to a plant, which may include, but not be limited to, the following: disease resistance, drought tolerance, heat tolerance, cold tolerance, salinity tolerance, metal tolerance, herbicide tolerance, improved water use efficiency, improved nitrogen utilization, improved nitrogen fixation, pest resistance, herbivore resistance, pathogen resistance, yield improvement, health enhancement, vigor improvement, growth improvement, photosynthetic capability improvement, nutrition enhancement, altered protein content, altered oil content, increased biomass, increased shoot length, increased root length, improved root architecture, modulation of a metabolite, modulation of the proteome, increased seed weight, altered seed carbohydrate composition, altered seed oil composition, altered seed protein composition, altered seed nutrient composition, as compared to an isoline plant not comprising a modification derived from the methods or compositions herein.

[0160] "Agronomic trait potential" is intended to mean a capability of a plant element for exhibiting a phenotype, preferably an improved agronomic trait, at some point during its life cycle, or conveying said phenotype to another plant element with which it is associated in the same plant.

[0161] The terms "decreased," "fewer," "slower" and "increased" "faster" "enhanced" "greater" as used herein refers to a decrease or increase in a characteristic of the modified plant element or resulting plant compared to an unmodified plant element or resulting plant. For example, a decrease in a characteristic may be at least 1%, at least 2%, at least 3%, at least 4%, at least 5%, between 5% and 10%, at least 10%, between 10% and 20%, at least 15%, at least 20%, between 20% and 30%, at least 25%, at least 30%, between 30% and 40%, at least 35%, at least 40%, between 40% and 50%, at least 45%, at least 50%, between 50% and 60%, at least about 60%, between 60% and 70%, between 70% and 80%, at least 75%, at least about 80%, between 80% and 90%, at least about 90%, between 90% and 100%, at least 100%, between 100% and 200%, at least 200%, at least about 300%, at least about 400%) or more lower than the untreated control and an increase may be at least 1%, at least 2%, at least 3%, at least 4%, at least 5%, between 5% and 10%, at least 10%, between 10% and 20%, at least 15%, at least 20%, between 20% and 30%, at least 25%, at least 30%, between 30% and 40%, at least 35%, at least

40%, between 40% and 50%, at least 45%, at least 50%, between 50% and 60%, at least about 60%, between 60% and 70%, between 70% and 80%, at least 75%, at least about 80%, between 80% and 90%, at least about 90%, between 90% and 100%, at least 100%, between 100% and 200%, at least 200%, at least about 300%, at least about 400% or more higher than the untreated control.

[0162] As used herein, the term "before", in reference to a sequence position, refers to an occurrence of one sequence upstream, or 5', to another sequence.

[0163] The meaning of abbreviations is as follows: "sec" means second(s), "min" means minute(s), "h" means hour(s), "d" means day(s), "µL" means microliter(s), "mL" means milliliter(s), "L" means liter(s), "µM" means micromolar, "mM" means millimolar, "M" means molar, "mmol" means millimole(s), "µmole" or "umole" mean micromole(s), "g" means gram(s), "µg" or "ug" means microgram(s), "ng" means nanogram(s), "U" means unit(s), "bp" means base pair(s) and "kb" means kilobase(s).

Classification of CRISPR-Cas Systems

[0164] CRISPR-Cas systems have been classified according to sequence and structural analysis of components. Multiple CRISPR/Cas systems have been described including Class 1 systems, with multisubunit effector complexes (comprising type I, type III, and type IV), and Class 2 systems, with single protein effectors (comprising type II, type V, and type VI) (Makarova *et al.* 2015, *Nature Reviews Microbiology* Vol. 13:1-15; Zetsche *et al.*, 2015, *Cell* 163, 1-13; Shmakov *et al.*, 2015, *Molecular Cell* 60, 1-13; Haft *et al.*, 2005, *Computational Biology, PLoS Comput Biol* 1(6):e60; and Koonin *et al.* 2017, *Curr Opinion Microbiology* 37:67-78).

[0165] A CRISPR-Cas system comprises, at a minimum, a CRISPR RNA (crRNA) molecule and at least one CRISPR-associated (Cas) protein to form crRNA ribonucleoprotein (crRNP) effector complexes. CRISPR-Cas loci comprise an array of identical repeats interspersed with DNA-targeting spacers that encode the crRNA components and an operon-like unit of *cas* genes encoding the Cas protein components. The resulting ribonucleoprotein complex recognizes a polynucleotide in a sequence-specific manner (Jore *et al.*, *Nature Structural & Molecular Biology* 18, 529–536 (2011)). The crRNA serves as a guide RNA for sequence specific binding of the effector (protein or complex) to double strand DNA sequences, by forming base pairs with the complementary DNA strand while displacing the noncomplementary strand to form a so-called R-loop. (Jore *et al.*, 2011. *Nature Structural & Molecular Biology* 18, 529–536).

[0166] RNA transcripts of CRISPR loci (pre-crRNA) are cleaved specifically in the repeat sequences by CRISPR associated (Cas) endoribonucleases in type I and type III systems or by

RNase III in type II systems. The number of CRISPR-associated genes at a given CRISPR locus can vary between species.

[0167] Different *cas* genes that encode proteins with different domains are present in different CRISPR systems. The *cas* operon comprises genes that encode for one or more effector endonucleases, as well as other Cas proteins. Protein subunits include those described in Makarova *et al.* 2011, *Nat Rev Microbiol.* 2011 9(6):467–477; Makarova *et al.* 2015, *Nature Reviews Microbiology* Vol. 13:1-15; and Koonin *et al.* 2017, *Current Opinion Microbiology* 37:67-78). The types of domains include those involved in Expression (pre-crRNA processing, for example Cas 6 or RNaseIII), Interference (including an effector module for crRNA and target binding, as well as domain(s) for target cleavage), Adaptation (spacer insertion, for example Cas1 or Cas2), and Ancillary (regulation or helper or unknown function). Some domains may serve more than one purpose, for example Cas9 comprises domains for endonuclease functionality as well as for target cleavage, among others.

[0168] The Cas endonuclease is guided by a single CRISPR RNA (crRNA) through direct RNA-DNA base-pairing to recognize a DNA target site that is in close vicinity to a protospacer adjacent motif (PAM) (Jore, M.M. *et al.*, 2011, *Nat. Struct. Mol. Biol.* 18:529-536, Westra, E.R. *et al.*, 2012, *Molecular Cell* 46:595-605, and Sinkunas, T. *et al.*, 2013, *EMBO J.* 32:385-394).

Class I CRISPR-Cas Systems

[0169] Class I CRISPR-Cas systems comprise Types I, III, and IV. A characteristic feature of Class I systems is the presence of an effector endonuclease complex instead of a single protein. A Cascade complex comprises an RNA recognition motif (RRM) and a nucleic acid-binding domain that is the core fold of the diverse RAMP (Repeat-Associated Mysterious Proteins) protein superfamily (Makarova *et al.* 2013, *Biochem Soc Trans* 41, 1392-1400; Makarova *et al.* 2015, *Nature Reviews Microbiology* Vol. 13:1-15). RAMP protein subunits include Cas5 and Cas7 (which comprise the skeleton of the crRNA–effector complex), wherein the Cas5 subunit binds the 5' handle of the crRNA and interacts with the large subunit, and often includes Cas6 which is loosely associated with the effector complex and typically functions as the repeat-specific RNase in the pre-crRNA processing (Charpentier *et al.*, *FEMS Microbiol Rev* 2015, 39:428-441; Niewoehner *et al.*, *RNA* 2016, 22:318-329).

[0170] Type I CRISPR-Cas systems comprise a complex of effector proteins, termed Cascade (CRISPR-associated complex for antiviral defense) comprising at a minimum Cas5 and Cas7. The effector complex functions together with a single CRISPR RNA (crRNA) and Cas3 to defend against invading viral DNA (Brouns, S.J.J. *et al. Science* 321:960-964; Makarova *et al.* 2015, *Nature Reviews Microbiology* Vol. 13:1-15). Type I CRISPR-Cas loci comprise the

signature gene *cas3* (or a variant cas3' or *cas3''*), which encodes a metal-dependent nuclease that possesses a single-stranded DNA (ssDNA)-stimulated superfamily 2 helicase with a demonstrated capacity to unwind double stranded DNA (dsDNA) and RNA–DNA duplexes (Makarova *et al.* 2015, Nature Reviews; Microbiology Vol. 13:1-15). Following target recognition, the Cas3 endonuclease is recruited to the Cascade-crRNA-target DNA complex to cleave and degrade the DNA target (Westra, E.R. *et al.* (2012) *Molecular Cell* 46:595-605, Sinkunas, T. *et al.* (2011) *EMBO J.* 30:1335-1342, and Sinkunas, T. *et al.* (2013) *EMBO J.* 32:385-394). In some type I systems, Cas6 can be the active endonuclease that is responsible for crRNA processing, and Cas5 and Cas7 function as non-catalytic RNA-binding proteins; although in type I-C systems, crRNA processing can be catalyzed by Cas5 (Makarova *et al.* 2015, *Nature Reviews Microbiology* Vol. 13:1-15). Type I systems are divided into seven subtypes (Makarova *et al.* 2011, *Nat Rev Microbiol.* 2011 9(6):467–477; Koonin *et al.* 2017, *Curr Opinion Microbiology* 37:67-78). A modified type I CRISPR-associated complex for adaptive antiviral defense (Cascade) comprising at least the protein subunits Cas7, Cas5 and Cas6, wherein one of these subunits is synthetically fused to a Cas3 endonuclease or a modified restriction endonuclease, FokI, have been described (WO2013098244 published 04 July 4, 2013).

[0171] Type III CRISPR-Cas systems, comprising a plurality of *cas7* genes, target either ssRNA or ssDNA, and function as either an RNase as well as a target RNA-activated DNA nuclease (Tamulaitis *et al.*, *Trends in Microbiology* 25(10)49-61, 2017). Csm (Type III-A) and Cmr (Type III-B) complexes function as RNA-activated single-stranded (ss) DNases that couple the target RNA binding/cleavage with ssDNA degradation. Upon foreign DNA infection, the CRISPR RNA (crRNA)-guided binding of the Csm or Cmr complex to the emerging transcript recruits Cas10 DNase to the actively transcribed phage DNA, resulting in degradation of both the transcript and phage DNA, but not the host DNA. The Cas10 HD-domain is responsible for the ssDNase activity, and Csm3/Cmr4 subunits are responsible for the endoribonuclease activity of the Csm/Cmr complex. The 3′-flanking sequence of the target RNA is critical for the ssDNase activity of Csm/Cmr: the base pairing with the 5′-handle of crRNA protects host DNA from degradation.

[0172] Type IV systems, although comprising typical type I *cas5* and *cas7* domains in addition to a *cas8*-like domain, may lack the CRISPR array that is characteristic of most other CRISPR-Cas systems.

[0173] Class II CRISPR-Cas systems comprise Types II, V, and VI. A characteristic feature of Class II systems is the presence of a single Cas effector protein instead of an effector complex.

Types II and V Cas proteins comprise an RuvC endonuclease domain that adopts the RNase H fold.

[0174] Type II CRISPR/Cas systems employ a crRNA and tracrRNA (trans-activating CRISPR RNA) to guide the Cas endonuclease to its DNA target. The crRNA comprises a spacer region complementary to one strand of the double strand DNA target and a region that base pairs with the tracrRNA (trans-activating CRISPR RNA) forming an RNA duplex that directs the Cas endonuclease to cleave the DNA target, leaving a blunt end. Spacers are acquired through a not fully understood process involving Cas1 and Cas2 proteins. Type II CRISPR/Cas loci typically comprise *cas1* and *cas2* genes in addition to the *cas9* gene (Chylinski *et al.*, 2013, *RNA Biology* 10:726-737; Makarova *et al.* 2015, *Nature Reviews Microbiology* Vol. 13:1-15). Type II CRISR-Cas loci can encode a tracrRNA, which is partially complementary to the repeats within the respective CRISPR array and can comprise other proteins such as Csn1 and Csn2. The presence of *cas9* in the vicinity of *cas1* and *cas2* genes is the hallmark of type II loci (Makarova *et al.* 2015, *Nature Reviews Microbiology* Vol. 13:1-15).

[0175] Type V CRISPR/Cas systems comprise a single Cas endonuclease, including Cpf1 (Cas12) (Koonin *et al.*, *Curr Opinion Microbiology* 37:67-78, 2017), that is an active RNA-guided endonuclease that does not necessarily require the additional trans-activating CRISPR (tracr) RNA for target cleavage, unlike Cas9.

[0176] Type VI CRISPR-Cas systems comprise a *cas13* gene that encodes a nuclease with two HEPN (Higher Eukaryotes and Prokaryotes Nucleotide-binding) domains but no HNH or RuvC domains and are not dependent upon tracrRNA activity. The majority of HEPN domains comprise conserved motifs that constitute a metal-independent endoRNase active site (Anantharam *et al.*, *Biol Direct* 8:15, 2013). Because of this feature, it is thought that type VI systems act on RNA targets instead of the DNA targets that are common to other CRISPR-Cas systems.

[0177] Novel Cas endonuclease CRISPR-Cas Systems

[0178] Disclosed herein is a novel CRISPR-Cas system, components thereof, and methods of using said components. The system comprises a novel Cas effector protein, Cas endonuclease.

[0179] The novel CRISPR-Cas system components described herein may comprise one or more subunits from different Cas systems, subunits derived or modified from more than one different bacterial or archaeal prokaryote, and/or synthetic or engineered components.

[0180] Described herein is a newly engineered Cas polypeptides and polynucleotides encoding the newly engineered Cas polypeptides.

[0181] **CRISPR-Cas System Components**

[0182] Cas Proteins

[0183] A number of proteins may be encoded in the CRISPR *cas* operon, including those involved in adaptation (spacer insertion), interference (effector module target binding, target nicking or cleavage – *e.g.* endonuclease activity), expression (pre-crRNA processing), regulation, or other.

[0184] Two proteins, Cas1 and Cas2, are conserved among many CRISPR systems (for example, as described in Koonin *et al.*, *Curr Opinion Microbiology* 37:67-78, 2017). Cas1 is a metal-dependent DNA-specific endonuclease that produces double-stranded DNA fragments. In some systems Cas1 forms a stable complex with Cas2, which is essential to spacer acquisition and insertion for CRISPR systems (Nuñez *et al.*, *Nature Str Mol Biol* 21:528-534, 2014).

[0185] A number of other proteins have been identified across different systems, including Cas4 (which may have similarity to a RecB nuclease) and is thought to play a role in the capture of new viral DNA sequences for incorporation into the CRISPR array (Zhang et al., PLOS One 7(10):e47232, 2012).

[0186] Some proteins may encompass a plurality of functions. For example, Cas9, the signature protein of Class 2 type II systems, has been demonstrated to be involved in pre-crRNA processing, target binding, as well as target cleavage.

[0187] In some native systems, such as the Cas endonuclease CRISPR system from *Syntrophomonas palmitatica*, no genes encoding Cas1, Cas2, or Cas4 proteins were detect near the endonuclease gene.

Cas Endonucleases and Effectors

[0188] Endonucleases are enzymes that cleave the phosphodiester bond within a polynucleotide chain and include restriction endonucleases that cleave DNA at specific sites without damaging the bases. Examples of endonucleases include restriction endonucleases, meganucleases, TAL effector nucleases (TALENs), zinc finger nucleases, and Cas (CRISPR-associated) effector endonucleases.

[0189] Cas endonucleases, either as single effector proteins or in an effector complex with other components, unwind the DNA duplex at the target sequence and optionally cleave at least one DNA strand, as mediated by recognition of the target sequence by a polynucleotide (such as, but not limited to, a crRNA or guide RNA) that is in complex with the Cas effector protein. Such recognition and cutting of a target sequence by a Cas endonuclease typically occur if the correct protospacer-adjacent motif (PAM) is located at or adjacent to the 3' end of the DNA target sequence. Alternatively, a Cas endonuclease herein may lack DNA cleavage or nicking activity but can still specifically bind to a DNA target sequence when complexed with a suitable

RNA component. (See also U.S. Patent Application US20150082478 published 19 March 2015 and US20150059010 published 26 February 2015).

[0190] Cas endonucleases may occur as individual effectors (Class 2 CRISPR systems) or as part of larger effector complexes (Class I CRISPR systems).

[0191] Cas endonucleases that have been described include, but are not limited to, for example, Cas3 (a feature of Class 1 type I systems), Cas9 (a feature of Class 2 type II systems) and Cas12 (Cpf1) (a feature of Class 2 type V systems).

[0192] Cas3 (and its variants Cas3' and Cas3'') functions as a single-stranded DNA nuclease (HD domain) and an ATP-dependent helicase. A variant of the Cas3 endonuclease can be obtained by disabling the functional activity of one or both domains of the Cas3 endonuclease poly peptide. Disabling the ATPase dependent helicase activity (by deletion, knockout of the Cas3-helicase domain, or through mutagenesis of critical residues or by assembling the reaction in the absence of ATP as described previously (Sinkunas, T. *et al.*, 2013, *EMBO J.* 32:385-394) can convert the cleavage ready Cascade comprising the modified Cas3 endonuclease into a nickase (as the HD domain is still functional). Disabling the HD endonuclease activity can be accomplished by any method known in the art, such as but not limited to, mutagenesis of critical residues of the HD domain, can convert the cleavage ready Cascade comprising the modified Cas3 endonuclease into a helicase. Disabling the both the Cas helicase and Cas3 HD endonuclease activity can be accomplished by any method known in the art, such as but not limited to, mutagenesis of critical residues of both the helicase and HD domains, can convert the cleavage ready Cascade comprising the modified Cas3 endonuclease into a binder protein that binds to a target sequence.

[0193] Cas9 (formerly referred to as Cas5, Csn1, or Csx12) is a Cas endonuclease that forms a complex with a crNucleotide and a tracrNucleotide, or with a single guide polynucleotide, for specifically recognizing and cleaving all or part of a DNA target sequence. Cas9 recognizes a 3' GC-rich PAM sequence on the target dsDNA. A Cas9 protein comprises a RuvC nuclease with an HNH (H-N-H) nuclease adjacent to the RuvC-II domain. The RuvC nuclease and HNH nuclease each can cleave a single DNA strand at a target sequence (the concerted action of both domains leads to DNA double-strand cleavage, whereas activity of one domain leads to a nick). In general, the RuvC domain comprises subdomains I, II and III, where domain I is located near the N-terminus of Cas9 and subdomains II and III are located in the middle of the protein, flanking the HNH domain (Hsu *et al.*, 2013, *Cell* 157:1262-1278). Cas9 endonucleases are typically derived from a type II CRISPR system, which includes a DNA cleavage system utilizing a Cas9 endonuclease in complex with at least one polynucleotide component. For

example, a Cas9 can be in complex with a CRISPR RNA (crRNA) and a trans-activating CRISPR RNA (tracrRNA). In another example, a Cas9 can be in complex with a single guide RNA (Makarova *et al.* 2015, *Nature Reviews Microbiology* Vol. 13:1-15).

[0194] Cas12 (formerly referred to as Cpf1, and variants c2c1, c2c3, CasX, and CasY) comprise an RuvC nuclease domain and produced staggered, 5' overhangs on the dsDNA target. Some variants do not require a tracrRNA, unlike the functionality of Cas9. Cas12 and its variants recognize a 5' AT-rich PAM sequence on the target dsDNA. An insert domain, called Nuc, of the Cas12a protein has been demonstrated to be responsible for target strand cleavage (Yamano *et al.*, *Cell* 2016, 165:949-962). Additional mutation studies in other Cas12 proteins demonstrated the Nuc domain contributes to guide and target binding, with the RuvC domain responsible for cleavage (Swarts *et al.*, *Mol Cell* 2017, 66:221-233 e224).

[0195] Cas endonucleases and effector proteins can be used for targeted genome editing (via simplex and multiplex double-strand breaks and nicks) and targeted genome regulation (via tethering of epigenetic effector domains to either the Cas protein or sgRNA. A Cas endonuclease can also be engineered to function as an RNA-guided recombinase, and via RNA tethers could serve as a scaffold for the assembly of multiprotein and nucleic acid complexes (Mali *et al.*, 2013, *Nature Methods Vol.* 10:957-963).

Cas endonucleases and Variants Thereof

[0196] A Cas endonuclease, or a functional variant thereof, is defined as a functional RNA-guided, PAM-dependent dsDNA cleavage protein of fewer than about 500 amino acids, comprising: a C-terminal RuvC catalytic domain split into three subdomains and further comprising bridge-helix and one or more Zinc finger motif(s); and an N-terminal Rec subunit with a helical bundle, WED wedge-like (or "Oligonucleotide Binding Domain", OBD) domain, and, optionally, a Zinc finger motif.

[0197] The novel Cas endonuclease variant proteins disclosed herein include effector proteins (endonucleases). The wildtype (WT) Cas-alpha 8 protein recognizes a relatively simple PAM sequence of TT at or near the target site of the target double-stranded polydeoxyribonucleotide.

[0198] Functional variants of Cas endonuclease are capable of double-strand break or single-strand nick activity, which activity may be less than the activity of the WT Cas endonuclease, approximately the same, or of even greater activity. Different levels of activity may have different uses according to the practitioner's desires. An engineered Cas endonuclease disclosed herein, or functional variant thereof, comprises, when aligned to SEQ ID NO:18, relative to the amino acid position numbers of SEQ ID NO:18, one or more substitutions, insertions, or

deletions at relative amino acid positions 123, 226, 231, 266, 295, 301, 305, 335, 336, 337, and 341. In particular examples, the provided novel engineered Cas polypeptide is an endonuclease that comprises at least 50%, between 50% and 55%, at least 55%, between 55% and 60%, at least 60%, between 60% and 65%, at least 65%, between 65% and 70%, at least 70%, between 70% and 75%, at least 75%, between 75% and 80%, at least 80%, between 80% and 85%, at least 85%, between 85% and 90%, at least 90%, between 90% and 95%, at least 95%, between 95% and 96%, at least 96%, between 96% and 97%, at least 97%, between 97% and 98%, at least 98%, between 98% and 99%, at least 99%, between 99% and 100%, or 100% sequence identity with at least 50, between 50 and 100, at least 100, between 100 and 150, at least 150, between 150 and 200, at least 200, between 200 and 250, at least 250, between 250 and 300, at least 300, between 300 and 350, at least 350, between 350 and 400, at least 400, or between 400 and 422, or 422 contiguous amino acids of any of SEQ ID NO:18. Preferably, the engineered Cas endonuclease retains the essential amino acids at positions, when aligned to SEQ ID NO:18, correspond to the Aspartate at position 225, the Glutamate at position 324, and the Aspartate at position 401 relative to an alignment with SEQ ID NO:18. See Fig. 3 herein. And in particular examples, the engineered Cas endonuclease comprises, when aligned to SEQ ID NO:18, Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341 relative to the amino acid position numbers of SEQ ID NO:18.

[0199] An engineered Cas endonuclease disclosed herein, or functional variant thereof, comprises at least 50%, between 50% and 55%, at least 55%, between 55% and 60%, at least 60%, between 60% and 65%, at least 65%, between 65% and 70%, at least 70%, between 70% and 75%, at least 75%, between 75% and 80%, at least 80%, between 80% and 85%, at least 85%, between 85% and 90%, at least 90%, between 90% and 95%, at least 95%, between 95% and 96%, at least 96%, between 96% and 97%, at least 97%, between 97% and 98%, at least 98%, between 98% and 99%, at least 99%, between 99% and 100%, or 100% sequence identity with at least 50, between 50 and 100, at least 100, between 100 and 150, at least 150, between 150 and 200, at least 200, between 200 and 250, at least 250, between 250 and 300, at least 300, between 300 and 350, at least 350, between 350 and 400, at least 400, or between 400 and 422, or 422 contiguous amino acids of any one of SEQ ID NOs:19-39. Preferably, the engineered Cas endonuclease retains the essential amino acids at positions, when aligned to SEQ ID NO:18, correspond to the Aspartate at position 225, the Glutamate at position 324, and the Aspartate at position 401 relative to an alignment with SEQ ID NO:18. See Fig. 3.

[0200] "functional fragment" of a Cas endonuclease variant endonuclease refers to a polynucleotide of fewer than 422 amino acids that comprises at least 50%, between 50% and 55%, at least 55%, between 55% and 60%, at least 60%, between 60% and 65%, at least 65%, between 65% and 70%, at least 70%, between 70% and 75%, at least 75%, between 75% and 80%, at least 80%, between 80% and 85%, at least 85%, between 85% and 90%, at least 90%, between 90% and 95%, at least 95%, between 95% and 96%, at least 96%, between 96% and 97%, at least 97%, between 97% and 98%, at least 98%, between 98% and 99%, at least 99%, between 99% and 100%, or 100% sequence identity with at least 50, between 50 and 100, at least 100, between 100 and 150, at least 150, between 150 and 200, at least 200, between 200 and 250, at least 250, between 250 and 300, at least 300, between 300 and 350, at least 350, between 350 and 400, at least 400, or between 400 and 422, or 422 contiguous amino acids of any one of SEQ ID NOs:19-39. Preferably, the engineered Cas endonuclease retains the essential amino acids at positions, when aligned to SEQ ID NO:18, correspond to the Aspartate at position 225, the Glutamate at position 324, and the Aspartate at position 401 relative to an alignment with SEQ ID NO:18. See Fig. 3.

[0201] RuvC domains have been demonstrated in the literature to encompass endonuclease functionality. A Cas endonuclease may be isolated or identified from a locus that comprises a *Cas endonuclease* gene encoding an effector protein, and an array comprising a plurality repeats.

[0202] Zinc finger motifs are domains that coordinate one or more zinc ions, usually through Cysteine and Histidine sidechains, to stabilize their fold. Zinc fingers are named for the pattern of Cysteine and Histidine residues that coordinate the zinc ion (*e.g.*, C4 means a zinc ion is coordinated by four Cysteine residues; C3H means a zinc ion is coordinated by three Cysteine residues and one Histidine residue).

[0203] Cas endonuclease proteins comprise one or more Zinc Finger (ZFN) coordination motif(s) that may form a Zinc binding domain. Zinc Finger-like motifs can aid in target and non-target strand separation and loading of the guide RNA into the DNA target. Cas endonuclease proteins comprising one or more Zinc Finger motifs may provide additional stability to the ribonucleoprotein complex on the target polynucleotide. Cas endonuclease proteins comprise C4 or C3H zinc binding domains.

[0204] As used herein, a "domain" is synonymous with "motif". For example, a zinc finger domain and zinc finger motif are used synonymously. Similarly, a zinc binding domain and zinc binding motif are used synonymously.

[0205] Cas endonucleases are RNA-guided endonucleases capable of binding to, and cleaving, a double-strand DNA target that comprises: (1) a sequence sharing homology with a nucleotide sequence of the guide RNA, and (2) a PAM sequence.

[0206] A Cas endonuclease is functional as a double-strand-break-inducing agent, and may also be a nickase, or a single-strand-break inducing agent. In some aspects, a catalytically inactive Cas endonuclease may be used to target or recruit to a target DNA sequence but not induce cleavage. In some aspects, a catalytically inactive Cas endonuclease protein may be used with a functional endonuclease, to cleave a target sequence. In some aspects, a catalytically inactive Cas endonuclease protein may be combined with a base editing molecule, such as a deaminase. In some aspects, a deaminase may be a cytidine deaminase. In some aspects, a deaminase may be an adenine deaminase. In some aspects, a deaminase may be ADAR-2.

[0207] A Cas endonuclease, effector protein, or functional fragment thereof, for use in the disclosed methods, can be isolated from a native source, or from, a recombinant source where the genetically modified host cell is modified to express the nucleic acid sequence encoding the protein. Alternatively, the Cas protein can be produced using cell free protein expression systems or be synthetically produced. Effector Cas nucleases may be isolated and introduced into a heterologous cell or may be modified from its native form to exhibit a different type or magnitude of activity than what it would exhibit in its native source. Such modifications include but are not limited to fragments, variants, substitutions, deletions, and insertions. Cas endonuclease WT compositions are described in WO2020123887 published on 16 July 2020.

[0208] Fragments and variants of Cas endonucleases and Cas endonuclease effector proteins can be obtained via methods such as site-directed mutagenesis and synthetic construction. Methods for measuring endonuclease activity are well known in the art such as, but not limiting to, WO2013166113 published 07 November 2013, WO2016186953 published 24 November 2016, and WO2016186946 published 24 November 2016.

[0209] The Cas endonuclease can comprise a modified form of the Cas polypeptide. The modified form of the Cas polypeptide can include an amino acid change (e.g., deletion, insertion, or substitution) that reduces the naturally occurring nuclease activity of the Cas protein. For example, in some instances, the modified form of the Cas protein has less than 50%, less than 40%, less than 30%, less than 20%, less than 10%, less than 5%, or less than 1% of the nuclease activity of the corresponding wild-type Cas polypeptide (US20140068797 published 06 March 2014). In some cases, the modified form of the Cas polypeptide has no substantial nuclease activity and is referred to as catalytically “inactivated Cas” or “deactivated Cas (dCas).” An inactivated Cas/deactivated Cas includes a deactivated Cas endonuclease (dCas). A catalytically

inactive Cas effector protein can be fused to a heterologous sequence to induce or modify activity.

[0210] A Cas endonuclease can be part of a fusion protein comprising one or more heterologous protein domains (e.g., 1, 2, 3, or more domains in addition to the Cas protein). Such a fusion protein may comprise any additional protein sequence, and optionally a linker sequence between any two domains, such as between Cas and a first heterologous domain. Examples of protein domains that may be fused to a Cas protein herein include, without limitation, epitope tags (e.g., histidine [His], V5, FLAG, influenza hemagglutinin [HA], myc, VSV-G, thioredoxin [Trx]), reporters (e.g., glutathione-5-transferase [GST], horseradish peroxidase [HRP], chloramphenicol acetyltransferase [CAT], beta-galactosidase, beta-glucuronidase [GUS], luciferase, green fluorescent protein [GFP], HcRed, DsRed, cyan fluorescent protein [CFP], yellow fluorescent protein [YFP], blue fluorescent protein [BFP]), and domains having one or more of the following activities: methylase activity, demethylase activity, transcription activation activity (e.g., VP16 or VP64), transcription repression activity, transcription release factor activity, histone modification activity, RNA cleavage activity and nucleic acid binding activity. A Cas protein can also be in fusion with a protein that binds DNA molecules or other molecules, such as maltose binding protein (MBP), S-tag, Lex A DNA binding domain (DBD), GAL4A DNA binding domain, and herpes simplex virus (HSV) VP16.

[0211] A catalytically active and/or inactive Cas endonuclease can be fused to a heterologous sequence (US20140068797 published 06 March 2014). Suitable fusion partners include, but are not limited to, a polypeptide that provides an activity that indirectly increases transcription by acting directly on the target DNA or on a polypeptide (e.g., a histone or other DNA-binding protein) associated with the target DNA. Additional suitable fusion partners include, but are not limited to, a polypeptide that provides for methyltransferase activity, demethylase activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity, or demyristoylation activity. Further suitable fusion partners include, but are not limited to, a polypeptide that directly provides for increased transcription of the target nucleic acid (e.g., a transcription activator or a fragment thereof, a protein or fragment thereof that recruits a transcription activator, a small molecule/drug-responsive transcription regulator, etc.). A partially active or catalytically inactive Cas endonuclease can also be fused to another protein or domain, for example Clo51 or FokI

nuclease, to generate double-strand breaks (Guilinger *et al. Nature Biotechnology*, volume 32, number 6, June 2014).

[0212] A catalytically active or inactive Cas protein, such as the Cas endonuclease protein described herein, can also be in fusion with a molecule that directs editing of single or multiple bases in a polynucleotide sequence, for example a site-specific deaminase that can change the identity of a nucleotide, for example from C•G to T•A or an A•T to G•C (Gaudelli et al., Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage." Nature (2017); Nishida et al. "Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems." Science 353 (6305) (2016); Komor et al. "Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage." Nature 533 (7603) (2016):420-4. A base editing fusion protein may comprise, for example, an active (double strand break creating), partially active (nickase) or deactivated (catalytically inactive) Cas endonuclease and a deaminase (such as, but not limited to, a cytidine deaminase, an adenine deaminase, APOBEC1, APOBEC3A, BE2, BE3, BE4, ABEs, or the like). Base edit repair inhibitors and glycosylase inhibitors (*e.g.*, uracil glycosylase inhibitor (to prevent uracil removal)) are contemplated as other components of a base editing system, in some embodiments.

[0213] The Cas endonucleases described herein can be expressed and purified by methods known in the art, for example as described in WO/2016/186953 published 24 November 2016.

[0214] Many Cas endonucleases have been described to date that can recognize specific PAM sequences (WO2016186953 published 24 November 2016, WO2016186946 published 24 November 2016, and Zetsche B *et al.* 2015. *Cell* 163, 1013) and cleave the target DNA at a specific position. It is understood that based on the methods and embodiments described herein utilizing a novel guided Cas system one skilled in the art can now tailor these methods such that they can utilize any guided endonuclease system.

[0215] A Cas effector protein can comprise a heterologous nuclear localization sequence (NLS). A heterologous NLS amino acid sequence herein may be of sufficient strength to drive accumulation of a Cas protein in a detectable amount in the nucleus of a yeast cell herein, for example. An NLS may comprise one (monopartite) or more (e.g., bipartite) short sequences (e.g., 2 to 20 residues) of basic, positively charged residues (e.g., lysine and/or arginine), and can be located anywhere in a Cas amino acid sequence but such that it is exposed on the protein surface. An NLS may be operably linked to the N-terminus or C-terminus of a Cas protein herein, for example. Two or more NLS sequences can be linked to a Cas protein, for example, such as on both the N- and C-termini of a Cas protein. The Cas endonuclease gene can be operably linked to a SV40 nuclear targeting signal upstream of the Cas codon region and a bipartite VirD2

nuclear localization signal (Tinland *et al.* (1992) *Proc. Natl. Acad. Sci. USA* 89:7442-6) downstream of the Cas codon region. Non-limiting examples of suitable NLS sequences herein include those disclosed in U.S. Patent Nos. 6,660,830 and 7,309,576.

[0216] Guide Polynucleotides

[0217] The guide polynucleotide enables target recognition, binding, and optionally cleavage by the Cas endonuclease, and can be a single molecule or a double molecule. The guide polynucleotide sequence can be an RNA sequence, a DNA sequence, or a combination thereof (an RNA-DNA combination sequence). Optionally, the guide polynucleotide can comprise at least one nucleotide, phosphodiester bond or linkage modification such as, but not limited, to Locked Nucleic Acid (LNA), 5-methyl dC, 2,6-Diaminopurine, 2'-Fluoro A, 2'-Fluoro U, 2'-O-Methyl RNA, phosphorothioate bond, linkage to a cholesterol molecule, linkage to a polyethylene glycol molecule, linkage to a spacer 18 (hexaethylene glycol chain) molecule, or 5' to 3' covalent linkage resulting in circularization. A guide polynucleotide that solely comprises ribonucleic acids is also referred to as a "guide RNA" or "gRNA" (US20150082478 published 19 March 2015 and US20150059010 published 26 February 2015). A guide polynucleotide may be engineered or synthetic.

[0218] The guide polynucleotide includes a chimeric non-naturally occurring guide RNA comprising regions that are not found together in nature (i.e., they are heterologous with each other). For example, a chimeric non-naturally occurring guide RNA comprising a first nucleotide sequence domain (referred to as Variable Targeting domain or VT domain) that can hybridize to a nucleotide sequence in a target DNA, linked to a second nucleotide sequence that can recognize the Cas endonuclease, such that the first and second nucleotide sequence are not found linked together in nature.

[0219] The guide polynucleotide can be a double molecule (also referred to as duplex guide polynucleotide) comprising a crNucleotide sequence (such as a crRNA) and a tracrNucleotide (such as a tracrRNA) sequence. In some cases, there is a linker polynucleotide that connects the crRNA and tracrRNA to form a single guide, for example an sgRNA.

[0220] The crNucleotide includes a first nucleotide sequence domain (referred to as Variable Targeting domain or VT domain) that can hybridize to a nucleotide sequence in a target DNA and a second nucleotide sequence (also referred to as a tracr mate sequence) that is part of a Cas endonuclease recognition (CER) domain. The tracr mate sequence can hybridized to a tracrNucleotide along a region of complementarity and together form the Cas endonuclease recognition domain or CER domain. The CER domain is capable of interacting with a Cas endonuclease polypeptide. The crNucleotide and the tracrNucleotide of the duplex guide

polynucleotide can be RNA, DNA, and/or RNA-DNA- combination sequences. In some embodiments, the crNucleotide molecule of the duplex guide polynucleotide is referred to as "crDNA" (when composed of a contiguous stretch of DNA nucleotides) or "crRNA" (when composed of a contiguous stretch of RNA nucleotides), or "crDNA-RNA" (when composed of a combination of DNA and RNA nucleotides). The crNucleotide can comprise a fragment of the crRNA naturally occurring in Bacteria and Archaea. The size of the fragment of the crRNA naturally occurring in Bacteria and Archaea that can be present in a crNucleotide disclosed herein can range from, but is not limited to, 2, 3, 4, 5, 6, 7, 8, 9,10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more nucleotides.

[0221] In some embodiments the tracrNucleotide is referred to as "tracrRNA" (when composed of a contiguous stretch of RNA nucleotides) or "tracrDNA" (when composed of a contiguous stretch of DNA nucleotides) or "tracrDNA-RNA" (when composed of a combination of DNA and RNA nucleotides. In one embodiment, the RNA that guides the RNA/ Cas9 endonuclease complex is a duplexed RNA comprising a duplex crRNA-tracrRNA. The tracrRNA (trans-activating CRISPR RNA) comprises, in the 5'-to-3' direction, (i) a sequence that anneals with the repeat region of CRISPR type II crRNA and (ii) a stem loop-comprising portion (Deltcheva *et al.*, *Nature* 471:602-607). The duplex guide polynucleotide can form a complex with a Cas endonuclease, wherein said guide polynucleotide/Cas endonuclease complex (also referred to as a guide polynucleotide/Cas endonuclease system) can direct the Cas endonuclease to a genomic target site, enabling the Cas endonuclease to recognize, bind to, and optionally nick or cleave (introduce a single or double-strand break) into the target site. (US20150082478 published 19 March 2015 and US20150059010 published 26 February 2015).

[0222] In one aspect, the guide polynucleotide is a guide polynucleotide capable of forming a PGEN as described herein, wherein said guide polynucleotide comprises a first nucleotide sequence domain that is complementary to a nucleotide sequence in a target DNA, and a second nucleotide sequence domain that interacts with said Cas endonuclease polypeptide disclosed herein.

[0223] In one aspect, the guide polynucleotide is a guide polynucleotide described herein, wherein the first nucleotide sequence and the second nucleotide sequence domain is selected from the group consisting of a DNA sequence, an RNA sequence, and a combination thereof.

[0224] In one aspect, the guide polynucleotide is a guide polynucleotide described herein, wherein the first nucleotide sequence and the second nucleotide sequence domain is selected from the group consisting of RNA backbone modifications that enhance stability, DNA backbone modifications that enhance stability, and a combination thereof (see Kanasty *et al.*,

2013, Common RNA-backbone modifications, *Nature Materials* 12:976-977; US20150082478 published 19 March 2015 and US20150059010 published 26 February 2015)

[0225] The guide RNA includes a dual molecule comprising a chimeric non-naturally occurring crRNA linked to at least one tracrRNA. A chimeric non-naturally occurring crRNA includes a crRNA that comprises regions that are not found together in nature (i.e., they are heterologous with each other. For example, a crRNA comprising a first nucleotide sequence domain (referred to as Variable Targeting domain or VT domain) that can hybridize to a nucleotide sequence in a target DNA, linked to a second nucleotide sequence (also referred to as a tracr mate sequence) such that the first and second sequence are not found linked together in nature.

[0226] The guide polynucleotide can also be a single molecule (also referred to as single guide polynucleotide) comprising a crNucleotide sequence linked to a tracrNucleotide sequence. The single guide polynucleotide comprises a first nucleotide sequence domain (referred to as Variable Targeting domain or VT domain) that can hybridize to a nucleotide sequence in a target DNA and a Cas endonuclease recognition domain (CER domain), that interacts with a Cas endonuclease polypeptide.

[0227] The VT domain and /or the CER domain of a single guide polynucleotide can comprise a RNA sequence, a DNA sequence, or an RNA-DNA-combination sequence. The single guide polynucleotide being comprised of sequences from the crNucleotide and the tracrNucleotide may be referred to as "single guide RNA" (when composed of a contiguous stretch of RNA nucleotides) or "single guide DNA" (when composed of a contiguous stretch of DNA nucleotides) or "single guide RNA-DNA" (when composed of a combination of RNA and DNA nucleotides). The single guide polynucleotide can form a complex with a Cas endonuclease, wherein said guide polynucleotide/Cas endonuclease complex (also referred to as a guide polynucleotide/Cas endonuclease system) can direct the Cas endonuclease to a genomic target site, enabling the Cas endonuclease to recognize, bind to, and optionally nick or cleave (introduce a single or double-strand break) the target site. (US20150082478 published 19 March 2015 and US20150059010 published 26 February 2015).

[0228] A chimeric non-naturally occurring single guide RNA (sgRNA) includes a sgRNA that comprises regions that are not found together in nature (i.e., they are heterologous with each other. For example, a sgRNA comprising a first nucleotide sequence domain (referred to as Variable Targeting domain or VT domain) that can hybridize to a nucleotide sequence in a target DNA linked to a second nucleotide sequence (also referred to as a tracr mate sequence) that are not found linked together in nature.

[0229] The nucleotide sequence linking the crNucleotide and the tracrNucleotide of a single guide polynucleotide can comprise a RNA sequence, a DNA sequence, or an RNA-DNA combination sequence. In one embodiment, the nucleotide sequence linking the crNucleotide and the tracrNucleotide of a single guide polynucleotide (also referred to as "loop") can be at least 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, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 or 100 nucleotides in length. In another embodiment, the nucleotide sequence linking the crNucleotide and the tracrNucleotide of a single guide polynucleotide can comprise a tetraloop sequence, such as, but not limiting to a GAAA tetraloop sequence.

[0230] The guide polynucleotide can be produced by any method known in the art, including chemically synthesizing guide polynucleotides (such as but not limiting to Hendel *et al.* 2015, *Nature Biotechnology* 33, 985–989), *in vitro* generated guide polynucleotides, and/or self-splicing guide RNAs (such as but not limited to Xie *et al.* 2015, *PNAS* 112:3570-3575).

[0231] Protospacer Adjacent Motif (PAM)

[0232] A "protospacer adjacent motif" (PAM) herein refers to a short nucleotide sequence adjacent to a target sequence (protospacer) that can be recognized (targeted) by a guide polynucleotide/Cas endonuclease system. The Cas endonuclease may not successfully recognize a target DNA sequence if the target DNA sequence is not followed by a PAM sequence. The sequence and length of a PAM herein can differ depending on the Cas protein or Cas protein complex used. The PAM sequence can be of any length but is typically 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 nucleotides long.

[0233] A "randomized PAM" and "randomized protospacer adjacent motif" are used interchangeably herein and refer to a random DNA sequence adjacent to a target sequence (protospacer) that is recognized (targeted) by a guide polynucleotide/Cas endonuclease system. The randomized PAM sequence can be of any length but is typically 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 nucleotides long. A randomized nucleotide includes anyone of the nucleotides A, C, G or T.

[0234] Guide Polynucleotide/Cas Endonuclease Complexes

[0235] A guide polynucleotide/Cas endonuclease complex described herein is capable of recognizing, binding to, and optionally nicking, unwinding, or cleaving all or part of a target sequence.

[0236] A guide polynucleotide/Cas endonuclease complex that can cleave both strands of a DNA target sequence typically comprises a Cas protein that has all of its endonuclease domains in a functional state (e.g., wild type endonuclease domains or variants thereof retaining some or all activity in each endonuclease domain). Thus, a wildtype Cas protein (e.g., a Cas protein disclosed herein), or a variant thereof retaining some or all activity in each endonuclease domain of the Cas protein, is a suitable example of a Cas endonuclease that can cleave both strands of a DNA target sequence.

[0237] A guide polynucleotide/Cas endonuclease complex that can cleave one strand of a DNA target sequence can be characterized herein as having nickase activity (e.g., partial cleaving capability). A Cas nickase typically comprises one functional endonuclease domain that allows the Cas to cleave only one strand (i.e., make a nick) of a DNA target sequence. For example, a Cas9 nickase may comprise (i) a mutant, dysfunctional RuvC domain and (ii) a functional HNH domain (e.g., wild type HNH domain). As another example, a Cas9 nickase may comprise (i) a functional RuvC domain (e.g., wild type RuvC domain) and (ii) a mutant, dysfunctional HNH domain. Non-limiting examples of Cas9 nickases suitable for use herein are disclosed in US20140189896 published on 03 July 2014. A pair of Cas nickases can be used to increase the specificity of DNA targeting. In general, this can be done by providing two Cas nickases that, by virtue of being associated with RNA components with different guide sequences, target and nick nearby DNA sequences on opposite strands in the region for desired targeting. Such nearby cleavage of each DNA strand creates a double-strand break (i.e., a DSB with single-stranded overhangs), which is then recognized as a substrate for non-homologous-end-joining, NHEJ (prone to imperfect repair leading to mutations) or homologous recombination, HR. Each nick in these embodiments can be at least about 5, between 5 and 10, at least 10, between 10 and 15, at least15, between 15 and 20, at least 20, between 20 and 30, at least 30, between 30 and 40, at least 40, between 40 and 50, at least 50, between 50 and 60, at least 60, between 60 and 70, at least 70, between 70 and 80, at least 80, between 80 and 90, at least 90, between 90 and 100, or 100 or greater (or any integer between 5 and 100) bases apart from each other, for example. One or two Cas nickase proteins herein can be used in a Cas nickase pair. For example, a Cas9 nickase with a mutant RuvC domain, but functioning HNH domain (i.e., Cas9 HNH+/RuvC-), can be used (e.g., *Streptococcus pyogenes* Cas9 HNH+/RuvC-). Each Cas9 nickase (e.g., Cas9 HNH+/RuvC-) can be directed to specific DNA sites nearby each other (up to 100 base pairs apart) by using suitable RNA components herein with guide RNA sequences targeting each nickase to each specific DNA site.

[0238] A guide polynucleotide/Cas endonuclease complex in certain embodiments can bind to a DNA target site sequence but does not cleave any strand at the target site sequence. Such a complex may comprise a Cas protein in which all of its nuclease domains are mutant, dysfunctional. For example, a Cas9 protein that can bind to a DNA target site sequence but does not cleave any strand at the target site sequence, may comprise both a mutant, dysfunctional RuvC domain and a mutant, dysfunctional HNH domain. A Cas protein herein that binds, but does not cleave, a target DNA sequence can be used to modulate gene expression, for example, in which case the Cas protein could be fused with a transcription factor (or portion thereof) (e.g., a repressor or activator, such as any of those disclosed herein).

[0239] In one aspect, the guide polynucleotide/Cas endonuclease complex (PGEN) described herein is a PGEN, wherein said Cas endonuclease is a novel engineered Cas polypeptide provided herein, which is optionally covalently or non-covalently linked, or assembled to at least one protein subunit, or functional fragment thereof.

[0240] In one embodiment of the disclosure, the guide polynucleotide/Cas endonuclease complex is a guide polynucleotide/Cas endonuclease complex (PGEN) comprising at least one guide polynucleotide and at least one Cas endonuclease polypeptide, wherein said Cas endonuclease polypeptide comprises at least one protein subunit, or a functional fragment thereof, wherein said guide polynucleotide is a chimeric non-naturally occurring guide polynucleotide, wherein said guide polynucleotide/Cas endonuclease complex is capable of recognizing, binding to, and optionally nicking, unwinding, or cleaving all or part of a target sequence.

[0241] The Cas effector protein can be a Cas endonuclease effector protein as disclosed herein.

[0242] In one embodiment of the disclosure, the guide polynucleotide/Cas effector complex is a guide polynucleotide/Cas effector protein complex (PGEN) comprising at least one guide polynucleotide and a Cas endonuclease effector protein disclosed herein, wherein said guide polynucleotide/Cas effector protein complex is capable of recognizing, binding to, and optionally nicking, unwinding, or cleaving all or part of a target sequence.

[0243] The PGEN can be a guide polynucleotide/Cas effector protein complex, wherein said Cas effector protein further comprises one copy or multiple copies of at least one protein subunit, or a functional fragment thereof. In some embodiments, said protein subunit is selected from the group consisting of a Cas1 protein subunit, a Cas2 protein subunit, a Cas4 protein subunit, and any combination thereof. The PGEN can be a guide polynucleotide/Cas effector protein

complex, wherein said Cas effector protein further comprises at least two different protein subunits of selected from the group consisting of a Cas1, Cas2, and Cas4.

[0244] The PGEN can be a guide polynucleotide/Cas effector protein complex, wherein said Cas effector protein further comprises at least three different protein subunits, or functional fragments thereof, selected from the group consisting of Cas1, Cas2, and one additional Cas protein, optionally comprising Cas4.

[0245] In one aspect, the guide polynucleotide/Cas effector protein complex (PGEN) described herein is a PGEN, wherein said Cas effector protein is covalently or non-covalently linked to at least one protein subunit, or functional fragment thereof. The PGEN can be a guide polynucleotide/Cas effector protein complex, wherein said Cas effector protein polypeptide is covalently or non-covalently linked or assembled to one copy or multiple copies of at least one protein subunit, or a functional fragment thereof, selected from the group consisting of a Cas1 protein subunit, a Cas2 protein subunit, a one additional Cas protein optionally comprising Cas4 protein subunit, and any combination thereof. The PGEN can be a guide polynucleotide/Cas effector protein complex, wherein said Cas effector protein is covalently or non-covalently linked or assembled to at least two different protein subunits selected from the group consisting of a Cas1, a Cas2, and one additional Cas protein, optionally comprising Cas4. The PGEN can be a guide polynucleotide/Cas effector protein complex, wherein said Cas effector protein is covalently or non-covalently linked to at least three different protein subunits, or functional fragments thereof, selected from the group consisting of a Cas1, a Cas2, and one additional Cas protein, optionally comprising Cas4, and any combination thereof.

[0246] Any component of the guide polynucleotide/Cas effector protein complex, the guide polynucleotide/Cas effector protein complex itself, as well as the polynucleotide modification template(s) and/or donor DNA(s), can be introduced into a heterologous cell or organism by any method known in the art.

[0247] **Recombinant Constructs for Transformation of Cells**

[0248] The disclosed guide polynucleotides, Cas endonucleases, polynucleotide modification templates, donor DNAs, guide polynucleotide/Cas endonuclease systems disclosed herein, and any one combination thereof, optionally further comprising one or more polynucleotide(s) of interest, can be introduced into a cell. Cells include, but are not limited to, human, non-human, animal, bacterial, fungal, insect, yeast, non-conventional yeast, and plant cells as well as plants and seeds produced by the methods described herein.

[0249] In one aspect, disclosed herein is a method that includes introducing into a cell a novel engineered Cas polypeptide disclosed herein into a cell such as a human, non-human,

animal, bacterial, fungal, insect, yeast, non-conventional yeast, and plant cell. The method includes transformation and/or expression of a recombinant construct encoding the engineered Cas polypeptide, which can have at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% amino acid sequence identity to any one of SEQ ID NOs:19 to 39, preferably wherein the Cas polypeptide includes one or more of the following amino acids at the indicated positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341. For example, in one example, the novel engineered Cas polypeptide comprises a Tyrosine at amino acid position 123, a Threonine at position 266, and a Proline at position 295 relative to an alignment with SEQ ID NO:18. Also provided is a recombinant organism comprising the foregoing human, non-human, animal, bacterial, fungal, insect, yeast, non-conventional yeast, or plant cells comprising the engineered Cas polypeptide.

[0250] Standard recombinant DNA and molecular cloning techniques used herein are well known in the art and are described more fully in Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual; Cold Spring Harbor Laboratory*: Cold Spring Harbor, NY (1989). Transformation methods are well known to those skilled in the art and are described *infra*.

[0251] Vectors and constructs include circular plasmids, and linear polynucleotides, comprising a polynucleotide of interest and optionally other components including linkers, adapters, regulatory or analysis. In some examples a recognition site and/or target site can be comprised within an intron, coding sequence, 5' UTRs, 3' UTRs, and/or regulatory regions.

[0252] Components for Expression and Utilization of Novel CRISPR-Cas Systems in Prokaryotic and Eukaryotic cells

[0253] This disclosure further provides expression constructs for expressing in a prokaryotic or eukaryotic cell/organism a guide RNA/Cas system that is capable of recognizing, binding to, and optionally nicking, unwinding, or cleaving all or part of a target sequence.

[0254] In one embodiment, the expression constructs of the disclosure comprise a promoter operably linked to a nucleotide sequence encoding a Cas gene (or plant optimized, including a Cas endonuclease gene described herein) and a promoter operably linked to a guide RNA of the present disclosure. The promoter is capable of driving expression of an operably linked nucleotide sequence in a prokaryotic or eukaryotic cell/organism.

[0255] Nucleotide sequence modification of the guide polynucleotide, VT domain and/or CER domain can be selected from, but not limited to , the group consisting of a 5' cap, a 3' polyadenylated tail, a riboswitch sequence, a stability control sequence, a sequence that forms a dsRNA duplex, a modification or sequence that targets the guide poly nucleotide to a subcellular location, a modification or sequence that provides for tracking , a modification or sequence that provides a binding site for proteins , a Locked Nucleic Acid (LNA), a 5-methyl dC nucleotide, a 2,6-Diaminopurine nucleotide, a 2'-Fluoro A nucleotide, a 2'-Fluoro U nucleotide; a 2'-O-Methyl RNA nucleotide, a phosphorothioate bond, linkage to a cholesterol molecule, linkage to a polyethylene glycol molecule, linkage to a spacer 18 molecule, a 5' to 3' covalent linkage, or any combination thereof. These modifications can result in at least one additional beneficial feature, wherein the additional beneficial feature is selected from the group of a modified or regulated stability, a subcellular targeting, tracking, a fluorescent label, a binding site for a protein or protein complex, modified binding affinity to complementary target sequence, modified resistance to cellular degradation, and increased cellular permeability.

[0256] A method of expressing RNA components such as gRNA in eukaryotic cells for performing Cas9-mediated DNA targeting has been to use RNA polymerase III (Pol III) promoters, which allow for transcription of RNA with precisely defined, unmodified, 5'- and 3'-ends (DiCarlo *et al.*, *Nucleic Acids Res.* 41:4336-4343; Ma *et al.*, *Mol. Ther. Nucleic Acids* 3:e161). This strategy has been successfully applied in cells of several different species including maize and soybean (US20150082478 published 19 March 2015). Methods for expressing RNA components that do not have a 5' cap have been described (WO2016/025131 published 18 February 2016).

[0257] Various methods and compositions can be employed to obtain a cell or organism having a polynucleotide of interest inserted in a target site for a Cas endonuclease. Such methods can employ homologous recombination (HR) to provide integration of the polynucleotide of interest at the target site. In one method described herein, a polynucleotide of interest is introduced into the organism cell via a donor DNA construct.

[0258] The donor DNA construct further comprises a first and a second region of homology that flank the polynucleotide of interest. The first and second regions of homology of the donor DNA share homology to a first and a second genomic region, respectively, present in or flanking the target site of the cell or organism genome.

[0259] The donor DNA can be tethered to the guide polynucleotide. Tethered donor DNAs can allow for co-localizing target and donor DNA, useful in genome editing, gene insertion, and targeted genome regulation, and can also be useful in targeting post-mitotic cells where function

of endogenous HR machinery is expected to be highly diminished (Mali *et al.*, 2013, *Nature Methods* Vol. 10:957-963).

[0260] The amount of homology or sequence identity shared by a target and a donor polynucleotide can vary and includes total lengths and/or regions having unit integral values in the ranges of about 1-20 bp, 20-50 bp, 50-100 bp, 75-150 bp, 100-250 bp, 150-300 bp, 200-400 bp, 250-500 bp, 300-600 bp, 350-750 bp, 400-800 bp, 450-900 bp, 500-1000 bp, 600-1250 bp, 700-1500 bp, 800-1750 bp, 900-2000 bp, 1-2.5 kb, 1.5–3 kb, 2-4 kb, 2.5-5 kb, 3-6 kb, 3.5-7 kb, 4-8 kb, 5-10 kb, or up to and including the total length of the target site. These ranges include every integer within the range, for example, the range of 1-20 bp includes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20 bps. The amount of homology can also be described by percent sequence identity over the full aligned length of the two polynucleotides which includes percent sequence identity at least of about 50%, 55%, 60%, 65%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, between 98% and 99%, 99%, between 99% and 100%, or 100%. Sufficient homology includes any combination of polynucleotide length, global percent sequence identity, and optionally conserved regions of contiguous nucleotides or local percent sequence identity, for example sufficient homology can be described as a region of 75-150 bp having at least 80% sequence identity to a region of the target locus. Sufficient homology can also be described by the predicted ability of two polynucleotides to specifically hybridize under high stringency conditions, see, for example, Sambrook *et al.*, (1989) *Molecular Cloning: A Laboratory Manual*, (Cold Spring Harbor Laboratory Press, NY); *Current Protocols in Molecular Biology*, Ausubel *et al.*, Eds (1994) *Current Protocols*, (Greene Publishing Associates, Inc. and John Wiley & Sons, Inc.); and, Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes*, (Elsevier, New York).

[0261] The structural similarity between a given genomic region and the corresponding region of homology found on the donor DNA can be any degree of sequence identity that allows for homologous recombination to occur. For example, the amount of homology or sequence identity shared by the "region of homology" of the donor DNA and the "genomic region" of the organism genome can be at least 50%, 55%, 60%, 65%, 70%, 75%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or 100% sequence identity, such that the sequences undergo homologous recombination

[0262] The region of homology on the donor DNA can have homology to any sequence flanking the target site. While in some instances the regions of homology share significant

sequence homology to the genomic sequence immediately flanking the target site, it is recognized that the regions of homology can be designed to have sufficient homology to regions that may be further 5' or 3' to the target site. The regions of homology can also have homology with a fragment of the target site along with downstream genomic regions

[0263] In one embodiment, the first region of homology further comprises a first fragment of the target site and the second region of homology comprises a second fragment of the target site, wherein the first and second fragments are dissimilar.

[0264] Polynucleotides of Interest

[0265] Polynucleotides of interest are further described herein and include polynucleotides reflective of the commercial markets and interests of those involved in the development of the crop. Crops and markets of interest change, and as developing nations open up world markets, new crops and technologies will emerge also. In addition, as our understanding of agronomic traits and characteristics such as yield and heterosis increase, the choice of genes for genetic engineering will change accordingly.

[0266] General categories of polynucleotides of interest include, for example, genes of interest involved in information, such as zinc fingers, those involved in communication, such as kinases, and those involved in housekeeping, such as heat shock proteins. More specific polynucleotides of interest include, but are not limited to, genes involved in traits of agronomic interest such as but not limited to: crop yield, grain quality, crop nutrient content, starch and carbohydrate quality and quantity as well as those affecting kernel size, sucrose loading, protein quality and quantity, nitrogen fixation and/or utilization, fatty acid and oil composition, genes encoding proteins conferring resistance to abiotic stress (such as drought, nitrogen, temperature, salinity, toxic metals or trace elements, or those conferring resistance to toxins such as pesticides and herbicides), genes encoding proteins conferring resistance to biotic stress (such as attacks by fungi, viruses, bacteria, insects, and nematodes, and development of diseases associated with these organisms).

[0267] Agronomically important traits such as oil, starch, and protein content can be genetically altered in addition to using traditional breeding methods. Modifications include increasing content of oleic acid, saturated and unsaturated oils, increasing levels of lysine and sulfur, providing essential amino acids, and also modification of starch. Hordothionin protein modifications are described in U.S. Patent Nos. 5,703,049, 5,885,801, 5,885,802, and 5,990,389.

[0268] Polynucleotide sequences of interest may encode proteins involved in providing disease or pest resistance. By "disease resistance" or "pest resistance" is intended that the plants avoid the harmful symptoms that are the outcome of the plant-pathogen interactions. Pest

resistance genes may encode resistance to pests that have great yield drag such as rootworm, cutworm, European Corn Borer, and the like. Disease resistance and insect resistance genes such as lysozymes or cecropins for antibacterial protection, or proteins such as defensins, glucanases or chitinases for antifungal protection, or *Bacillus thuringiensis* endotoxins, protease inhibitors, collagenases, lectins, or glycosidases for controlling nematodes or insects are all examples of useful gene products. Genes encoding disease resistance traits include detoxification genes, such as against fumonisin (U.S. Patent No. 5,792,931); avirulence (avr) and disease resistance (R) genes (Jones *et al.* (1994) *Science* 266:789; Martin *et al.* (1993) *Science* 262:1432; and Mindrinos *et al.* (1994) *Cell* 78:1089); and the like. Insect resistance genes may encode resistance to pests that have great yield drag such as rootworm, cutworm, European Corn Borer, and the like. Such genes include, for example, *Bacillus thuringiensis* toxic protein genes (U.S. Patent Nos. 5,366,892; 5,747,450; 5,736,514; 5,723,756; 5,593,881; and Geiser *et al.* (1986) *Gene* 48:109); and the like.

[0269] An "herbicide resistance protein" or a protein resulting from expression of an "herbicide resistance-encoding nucleic acid molecule" includes proteins that confer upon a cell the ability to tolerate a higher concentration of an herbicide than cells that do not express the protein, or to tolerate a certain concentration of an herbicide for a longer period of time than cells that do not express the protein. Herbicide resistance traits may be introduced into plants by genes coding for resistance to herbicides that act to inhibit the action of acetolactate synthase (ALS, also referred to as acetohydroxyacid synthase, AHAS), in particular the sulfonylurea (UK:sulphonylurea) type herbicides, genes coding for resistance to herbicides that act to inhibit the action of glutamine synthase, such as phosphinothricin or basta (e.g., the bar gene), glyphosate (e.g., the EPSP synthase gene and the GAT gene), HPPD inhibitors (e.g., the HPPD gene) or other such genes known in the art. See, for example, US Patent Nos. 7,626,077, 5,310,667, 5,866,775, 6,225,114, 6,248,876, 7,169,970, 6,867,293, and 9,187,762. The *bar* gene encodes resistance to the herbicide basta, the *nptII* gene encodes resistance to the antibiotics kanamycin and geneticin, and the ALS-gene mutants encode resistance to the herbicide chlorsulfuron.

[0270] Furthermore, it is recognized that the polynucleotide of interest may also comprise antisense sequences complementary to at least a portion of the messenger RNA (mRNA) for a targeted gene sequence of interest. Antisense nucleotides are constructed to hybridize with the corresponding mRNA. Modifications of the antisense sequences may be made as long as the sequences hybridize to and interfere with expression of the corresponding mRNA. In this manner, antisense constructions having 70%, 80%, or 85% sequence identity to the

corresponding antisense sequences may be used. Furthermore, portions of the antisense nucleotides may be used to disrupt the expression of the target gene. Generally, sequences of at least 50 nucleotides, 100 nucleotides, 200 nucleotides, or greater may be used.

[0271] In addition, the polynucleotide of interest may also be used in the sense orientation to suppress the expression of endogenous genes in plants. Methods for suppressing gene expression in plants using polynucleotides in the sense orientation are known in the art. The methods generally involve transforming plants with a DNA construct comprising a promoter that drives expression in a plant operably linked to at least a portion of a nucleotide sequence that corresponds to the transcript of the endogenous gene. Typically, such a nucleotide sequence has substantial sequence identity to the sequence of the transcript of the endogenous gene, generally greater than about 65% sequence identity, about 85% sequence identity, or greater than about 95% sequence identity. See U.S. Patent Nos. 5,283,184 and 5,034,323.

[0272] The polynucleotide of interest can also be a phenotypic marker. A phenotypic marker is screenable or a selectable marker that includes visual markers and selectable markers whether it is a positive or negative selectable marker. Any phenotypic marker can be used. Specifically, a selectable or screenable marker comprises a DNA segment that allows one to identify or select for or against a molecule or a cell that comprises it, often under particular conditions. These markers can encode an activity, such as, but not limited to, production of RNA, peptide, or protein, or can provide a binding site for RNA, peptides, proteins, inorganic and organic compounds or compositions and the like.

[0273] Examples of selectable markers include, but are not limited to, DNA segments that comprise restriction enzyme sites; DNA segments that encode products which provide resistance against otherwise toxic compounds including antibiotics, such as, spectinomycin, ampicillin, kanamycin, tetracycline, Basta, neomycin phosphotransferase II (NEO) and hygromycin phosphotransferase (HPT)); DNA segments that encode products which are otherwise lacking in the recipient cell (e.g., tRNA genes, auxotrophic markers); DNA segments that encode products which can be readily identified (e.g., phenotypic markers such as β-galactosidase, GUS; fluorescent proteins such as green fluorescent protein (GFP), cyan (CFP), yellow (YFP), red (RFP), and cell surface proteins); the generation of new primer sites for PCR (e.g., the juxtaposition of two DNA sequence not previously juxtaposed), the inclusion of DNA sequences not acted upon or acted upon by a restriction endonuclease or other DNA modifying enzyme, chemical, etc.; and, the inclusion of a DNA sequences required for a specific modification (e.g., methylation) that allows its identification.

[0274] Additional selectable markers include genes that confer resistance to herbicidal compounds, such as sulphonylureas, glufosinate ammonium, bromoxynil, imidazolinones, and 2,4-dichlorophenoxyacetate (2,4-D). See for example, Acetolactase synthase (ALS) for resistance to sulfonylureas, imidazolinones, triazolopyrimidine sulfonamides, pyrimidinylsalicylates and sulphonylaminocarbonyl-triazolinones (Shaner and Singh, 1997, Herbicide Activity:*Toxicol Biochem Mol Biol* 69-110); glyphosate resistant 5-enolpyruvylshikimate-3-phosphate (EPSPS) (Saroha *et al.* 1998, *J. Plant Biochemistry & Biotechnology* Vol 7:65-72);

[0275] Polynucleotides of interest includes genes that can be stacked or used in combination with other traits, such as but not limited to herbicide resistance or any other trait described herein. Polynucleotides of interest and/or traits can be stacked together in a complex trait locus as described in US20130263324 published 03 Oct 2013 and in WO/2013/112686, published 01 August 2013.

[0276] A polypeptide of interest includes any protein or polypeptide that is encoded by a polynucleotide of interest described herein.

[0277] Further provided are methods for identifying at least one plant cell, comprising in its genome, a polynucleotide of interest integrated at the target site. A variety of methods are available for identifying those plant cells with insertion into the genome at or near to the target site. Such methods can be viewed as directly analyzing a target sequence to detect any change in the target sequence, including but not limited to PCR methods, sequencing methods, nuclease digestion, Southern blots, and any combination thereof. See, for example, US20090133152 published 21 May 2009. The method also comprises recovering a plant from the plant cell comprising a polynucleotide of interest integrated into its genome. The plant may be sterile or fertile. It is recognized that any polynucleotide of interest can be provided, integrated into the plant genome at the target site, and expressed in a plant.

[0278] <u>Optimization of Sequences for Expression in Plants</u>

[0279] Methods are available in the art for synthesizing plant-preferred genes. See, for example, U.S. Patent Nos. 5,380,831, and 5,436,391, and Murray *et al.* (1989) *Nucleic Acids Res.* 17:477-498. Additional sequence modifications are known to enhance gene expression in a plant host. These include, for example, elimination of one or more sequences encoding spurious polyadenylation signals, one or more exon-intron splice site signals, one or more transposon-like repeats, and other such well-characterized sequences that may be deleterious to gene expression. The G-C content of the sequence may be adjusted to levels average for a given plant host, as calculated by reference to known genes expressed in the host plant cell. When possible,

the sequence is modified to avoid one or more predicted hairpin secondary mRNA structures. Thus, "a plant-optimized nucleotide sequence" of the present disclosure comprises one or more of such sequence modifications.

[0280] Expression Elements

[0281] Any polynucleotide encoding a Cas protein or other CRISPR system component disclosed herein may be functionally linked to a heterologous expression element, to facilitate transcription or regulation in a host cell. Such expression elements include but are not limited to a promoter, leader, intron, and terminator. Expression elements may be “minimal” – meaning a shorter sequence derived from a native source, that still functions as an expression regulator or modifier. Alternatively, an expression element may be “optimized” – meaning that its polynucleotide sequence has been altered from its native state in order to function with a more desirable characteristic in a particular host cell (for example, but not limited to, a bacterial promoter may be “maize-optimized” to improve its expression in corn plants). Alternatively, an expression element may be “synthetic” – meaning that it is designed in silico and synthesized for use in a host cell. Synthetic expression elements may be entirely synthetic, or partially synthetic (comprising a fragment of a naturally occurring polynucleotide sequence).

[0282] It has been shown that certain promoters are able to direct RNA synthesis at a higher rate than others. These are called “strong promoters”. Certain other promoters have been shown to direct RNA synthesis at higher levels only in particular types of cells or tissues and are often referred to as “tissue specific promoters”, or “tissue-preferred promoters” if the promoters direct RNA synthesis preferably in certain tissues but also in other tissues at reduced levels.

[0283] A plant promoter includes a promoter capable of initiating transcription in a plant cell. For a review of plant promoters, see, Potenza *et al.*, 2004, *In vitro Cell Dev Biol* 40:1-22; Porto *et al.*, 2014, Molecular Biotechnology (2014), 56(1), 38-49.

[0284] Constitutive promoters include, for example, the core CaMV 35S promoter (Odell *et al.*, (1985) *Nature* 313:810-2); rice actin (McElroy *et al.*, (1990) *Plant Cell* 2:163-71); ubiquitin (Christensen *et al.*, (1989) *Plant Mol Biol* 12:619-32; ALS promoter (U.S. Patent No. 5,659,026) and the like.

[0285] Tissue-preferred promoters can be utilized to target enhanced expression within a particular plant tissue. Tissue-preferred promoters include, for example, WO2013103367 published 11 July 2013, Kawamata *et al.*, (1997) *Plant Cell Physiol* 38:792-803; Hansen *et al.*, (1997) *Mol Gen Genet* 254:337-43; Russell *et al.*, (1997) *Transgenic Res* 6:157-68; Rinehart *et al.*, (1996) *Plant Physiol* 112:1331-41; Van Camp *et al.*, (1996) *Plant Physiol* 112:525-35; Canevascini *et al.*, (1996) *Plant Physiol* 112:513-524; Lam, (1994) *Results Probl Cell Differ*

20:181-96; and Guevara-Garcia *et al.*, (1993) *Plant J* 4:495-505. Leaf-preferred promoters include, for example, Yamamoto *et al.*, (1997) *Plant J* 12:255-65; Kwon *et al.*, (1994) *Plant Physiol* 105:357-67; Yamamoto *et al.*, (1994) *Plant Cell Physiol* 35:773-8; Gotor *et al.*, (1993) *Plant J* 3:509-18; Orozco *et al.*, (1993) *Plant Mol Biol* 23:1129-38; Matsuoka *et al.*, (1993) *Proc. Natl. Acad. Sci. USA* 90:9586-90; Simpson *et al.*, (1958) *EMBO J* 4:2723-9; Timko *et al.*, (1988) *Nature* 318:57-8. Root-preferred promoters include, for example, Hire *et al.*, (1992) *Plant Mol Biol* 20:207-18 (soybean root-specific glutamine synthase gene); Miao *et al.*, (1991) *Plant Cell* 3:11-22 (cytosolic glutamine synthase (GS)); Keller and Baumgartner, (1991) *Plant Cell* 3:1051-61 (root-specific control element in the GRP 1.8 gene of French bean); Sanger *et al.*, (1990) *Plant Mol Biol* 14:433-43 (root-specific promoter of *A. tumefaciens* mannopine synthase (MAS)); Bogusz *et al.*, (1990) *Plant Cell* 2:633-41 (root-specific promoters isolated from *Parasponia andersonii* and *Trema tomentosa*); Leach and Aoyagi, (1991) *Plant Sci* 79:69-76 (*A. rhizogenes* rolC and rolD root-inducing genes); Teeri *et al.*, (1989) *EMBO J* 8:343-50 (*Agrobacterium* wound-induced TR1' and TR2' genes); VfENOD-GRP3 gene promoter (Kuster *et al.*, (1995) *Plant Mol Biol* 29:759-72); and rolB promoter (Capana *et al.*, (1994) *Plant Mol Biol* 25:681-91; phaseolin gene (Murai *et al.*, (1983) *Science* 23:476-82; Sengopta-Gopalen *et al.*, (1988) *Proc. Natl. Acad. Sci. USA* 82:3320-4). See also, U.S. Patent Nos. 5,837,876; 5,750,386; 5,633,363; 5,459,252; 5,401,836; 5,110,732 and 5,023,179.

[0286] Seed-preferred promoters include both seed-specific promoters active during seed development, as well as seed-germinating promoters active during seed germination. See, Thompson *et al.*, (1989) *BioEssays* 10:108. Seed-preferred promoters include, but are not limited to, Cim1 (cytokinin-induced message); cZ19B1 (maize 19 kDa zein); and milps (myo-inositol-1-phosphate synthase); and for example those disclosed in WO2000011177 published 02 March 2000 and U.S. Patent 6,225,529. For dicots, seed-preferred promoters include, but are not limited to, bean β-phaseolin, napin, β-conglycinin, soybean lectin, cruciferin, and the like. For monocots, seed-preferred promoters include, but are not limited to, maize 15 kDa zein, 22 kDa zein, 27 kDa gamma zein, *waxy*, shrunken 1, shrunken 2, globulin 1, oleosin, and nuc1. See also, WO2000012733 published 09 March 2000, where seed-preferred promoters from *END1* and *END2* genes are disclosed.

[0287] Chemical inducible (regulated) promoters can be used to modulate the expression of a gene in a prokaryotic and eukaryotic cell or organism through the application of an exogenous chemical regulator. The promoter may be a chemical-inducible promoter, where application of the chemical induces gene expression, or a chemical-repressible promoter, where application of the chemical represses gene expression. Chemical-inducible promoters include, but are not

limited to, the maize In2-2 promoter, activated by benzene sulfonamide herbicide safeners (De Veylder *et al.*, (1997) *Plant Cell Physiol* 38:568-77), the maize GST promoter (GST-II-27, WO1993001294 published 21 January 1993), activated by hydrophobic electrophilic compounds used as pre-emergent herbicides, and the tobacco PR-1a promoter (Ono *et al.*, (2004) *Biosci Biotechnol Biochem* 68:803-7) activated by salicylic acid. Other chemical-regulated promoters include steroid-responsive promoters (see, for example, the glucocorticoid-inducible promoter (Schena *et al.*, (1991) *Proc. Natl. Acad. Sci. USA* 88:10421-5; McNellis *et al.*, (1998) *Plant J* 14:247-257); tetracycline-inducible and tetracycline-repressible promoters (Gatz *et al.*, (1991) *Mol Gen Genet* 227:229-37; U.S. Patent Nos. 5,814,618 and 5,789,156).

[0288] Pathogen inducible promoters induced following infection by a pathogen include, but are not limited to those regulating expression of PR proteins, SAR proteins, beta-1,3-glucanase, chitinase, *etc.*

[0289] A stress-inducible promoter includes the RD29A promoter (Kasuga *et al.* (1999) Nature Biotechnol. 17:287-91). One of ordinary skill in the art is familiar with protocols for simulating stress conditions such as drought, osmotic stress, salt stress and temperature stress and for evaluating stress tolerance of plants that have been subjected to simulated or naturally occurring stress conditions.

[0290] Another example of an inducible promoter useful in plant cells, is the ZmCAS1 promoter, described in US20130312137 published 21 November 2013.

[0291] New promoters of various types useful in plant cells are constantly being discovered; numerous examples may be found in the compilation by Okamuro and Goldberg, (1989) In *The Biochemistry of Plants*, Vol. 115, Stumpf and Conn, eds (New York, NY Academic Press), pp. 1-82.

[0292] **Modification of Genomes with Novel CRISPR-Cas System Components**

[0293] As described herein, a guided Cas endonuclease can recognize, bind to a DNA target sequence, and introduce a single strand (nick) or double-strand break. Once a single or double-strand break is induced in the DNA, the cell's DNA repair mechanism is activated to repair the break. Error-prone DNA repair mechanisms can produce mutations at double-strand break sites. The most common repair mechanism to bring the broken ends together is the nonhomologous end-joining (NHEJ) pathway (Bleuyard *et al.*, (2006) *DNA Repair* 5:1-12). The structural integrity of chromosomes is typically preserved by the repair, but deletions, insertions, or other rearrangements (such as chromosomal translocations) are possible (Siebert and Puchta, 2002, *Plant Cell* 14:1121-31; Pacher *et al.*, 2007, *Genetics* 175:21-9).

[0294] DNA double-strand breaks appear to be an effective factor to stimulate homologous recombination pathways (Puchta *et al.*, (1995) *Plant Mol Biol* 28:281-92; Tzfira and White, (2005) *Trends Biotechnol* 23:567-9; Puchta, (2005) *J Exp Bot* 56:1-14). Using DNA-breaking agents, a two- to nine-fold increase of homologous recombination was observed between artificially constructed homologous DNA repeats in plants (Puchta *et al.*, (1995) *Plant Mol Biol* 28:281-92). In maize protoplasts, experiments with linear DNA molecules demonstrated enhanced homologous recombination between plasmids (Lyznik *et al.*, (1991) *Mol Gen Genet* 230:209-18).

[0295] Homology-directed repair (HDR) is a mechanism in cells to repair double-stranded and single stranded DNA breaks. Homology-directed repair includes homologous recombination (HR) and single-strand annealing (SSA) (Lieber. 2010 *Annu. Rev. Biochem.* 79:181-211). The most common form of HDR is called homologous recombination (HR), which has the longest sequence homology requirements between the donor and acceptor DNA. Other forms of HDR include single-stranded annealing (SSA) and breakage-induced replication, and these require shorter sequence homology relative to HR. Homology-directed repair at nicks (single-stranded breaks) can occur via a mechanism distinct from HDR at double-strand breaks (Davis and Maizels. *PNAS* (0027-8424), 111 (10), p. E924-E932).

[0296] Alteration of the genome of a prokaryotic and eukaryotic cell or organism cell, for example, through homologous recombination (HR), is a powerful tool for genetic engineering. Homologous recombination has been demonstrated in plants (Halfter *et al.*, (1992) *Mol Gen Genet* 231:186-93) and insects (Dray and Gloor, 1997, *Genetics* 147:689-99). Homologous recombination has also been accomplished in other organisms. For example, at least 150-200 bp of homology was required for homologous recombination in the parasitic protozoan *Leishmania* (Papadopoulou and Dumas, (1997) *Nucleic Acids Res* 25:4278-86). In the filamentous fungus *Aspergillus nidulans*, gene replacement has been accomplished with as little as 50 bp flanking homology (Chaveroche *et al.*, (2000) *Nucleic Acids Res* 28:e97). Targeted gene replacement has also been demonstrated in the ciliate *Tetrahymena thermophila* (Gaertig *et al.*, (1994) *Nucleic Acids Res* 22:5391-8). In mammals, homologous recombination has been most successful in the mouse using pluripotent embryonic stem cell lines (ES) that can be grown in culture, transformed, selected, and introduced into a mouse embryo (Watson *et al.*, 1992, Recombinant DNA, 2nd Ed., Scientific American Books distributed by WH Freeman & Co.).

[0297] Gene Targeting

[0298] The guide polynucleotide/Cas systems described herein can be used for gene targeting.

[0299] In general, DNA targeting can be performed by cleaving one or both strands at a specific polynucleotide sequence in a cell with a Cas protein associated with a suitable polynucleotide component. Once a single or double-strand break is induced in the DNA, the cell's DNA repair mechanism is activated to repair the break via nonhomologous end-joining (NHEJ) or Homology-Directed Repair (HDR) processes which can lead to modifications at the target site.

[0300] The length of the DNA sequence at the target site can vary, and includes, for example, target sites that are at least 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, or more than 30 nucleotides in length. It is further possible that the target site can be palindromic, that is, the sequence on one strand reads the same in the opposite direction on the complementary strand. The nick/cleavage site can be within the target sequence or the nick/cleavage site could be outside of the target sequence. In another variation, the cleavage could occur at nucleotide positions immediately opposite each other to produce a blunt end cut or, in other cases, the incisions could be staggered to produce single-stranded overhangs, also called "sticky ends", which can be either 5' overhangs, or 3' overhangs. Active variants of genomic target sites can also be used. Such active variants can comprise at least 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or more sequence identity to the given target site, wherein the active variants retain biological activity and hence are capable of being recognized and cleaved by a Cas endonuclease.

[0301] Assays to measure the single or double-strand break of a target site by an endonuclease are known in the art and generally measure the overall activity and specificity of the agent on DNA substrates comprising recognition sites.

[0302] A targeting method herein can be performed in such a way that two or more DNA target sites are targeted in the method, for example. Such a method can optionally be characterized as a multiplex method. Two, three, four, five, six, seven, eight, nine, ten, or more target sites can be targeted at the same time in certain embodiments. A multiplex method is typically performed by a targeting method herein in which multiple different RNA components are provided, each designed to guide a guide polynucleotide/Cas endonuclease complex to a unique DNA target site.

[0303] Gene Editing

[0304] The process for editing a genomic sequence combining DSB and modification templates generally comprises: introducing into a host cell a DSB-inducing agent, or a nucleic acid encoding a DSB-inducing agent, that recognizes a target sequence in the chromosomal sequence and is able to induce a DSB in the genomic sequence, and at least one polynucleotide

modification template comprising at least one nucleotide alteration when compared to the nucleotide sequence to be edited. The polynucleotide modification template can further comprise nucleotide sequences flanking the at least one nucleotide alteration, in which the flanking sequences are substantially homologous to the chromosomal region flanking the DSB. Genome editing using DSB-inducing agents, such as Cas-gRNA complexes, has been described, for example in US20150082478 published on 19 March 2015, WO2015026886 published on 26 February 2015, WO2016007347 published 14 January 2016, and WO/2016/025131 published on 18 February 2016.

[0305] Some uses for guide RNA/Cas endonuclease systems have been described (see for example:US20150082478 A1 published 19 March 2015, WO2015026886 published 26 February 2015, and US20150059010 published 26 February 2015) and include but are not limited to modifying or replacing nucleotide sequences of interest (such as a regulatory elements), insertion of polynucleotides of interest, gene knock-out, gene-knock in, modification of splicing sites and/or introducing alternate splicing sites, modifications of nucleotide sequences encoding a protein of interest, amino acid and/or protein fusions, and gene silencing by expressing an inverted repeat into a gene of interest.

[0306] Proteins may be altered in various ways including amino acid substitutions, deletions, truncations, and insertions. Methods for such manipulations are generally known. For example, amino acid sequence variants of the protein(s) can be prepared by mutations in the DNA. Methods for mutagenesis and nucleotide sequence alterations include, for example, Kunkel, (1985) *Proc. Natl. Acad. Sci. USA* 82:488-92; Kunkel *et al.*, (1987) *Meth Enzymol* 154:367-82; U.S. Patent No. 4,873,192; Walker and Gaastra, eds. (1983) *Techniques in Molecular Biology* (MacMillan Publishing Company, New York) and the references cited therein. Guidance regarding amino acid substitutions not likely to affect biological activity of the protein is found, for example, in the model of Dayhoff *et al.*, (1978) *Atlas of Protein Sequence and Structure* (Natl Biomed Res Found, Washington, D.C.). Conservative substitutions, such as exchanging one amino acid with another having similar properties, may be preferable. Conservative deletions, insertions, and amino acid substitutions are not expected to produce radical changes in the characteristics of the protein, and the effect of any substitution, deletion, insertion, or combination thereof can be evaluated by routine screening assays. Assays for double-strand-break-inducing activity are known and generally measure the overall activity and specificity of the agent on DNA substrates comprising target sites.

[0307] Described herein are methods for genome editing with a Cas endonuclease and complexes with a Cas endonuclease and a guide polynucleotide. Following characterization of

the guide RNA and PAM sequence, components of the endonuclease and associated CRISPR RNA (crRNA) may be utilized to modify chromosomal DNA in other organisms including plants. To facilitate optimal expression and nuclear localization (for eukaryotic cells), the genes comprising the complex may be optimized as described in WO2016186953 published 24 November 2016, and then delivered into cells as DNA expression cassettes by methods known in the art. The components necessary to comprise an active complex may also be delivered as RNA with or without modifications that protect the RNA from degradation or as mRNA capped or uncapped (Zhang, Y. *et al.*, 2016, *Nat. Commun.* 7:12617) or Cas protein guide polynucleotide complexes (WO2017070032 published 27 April 2017), or any combination thereof. Additionally, a part or part(s) of the complex and crRNA may be expressed from a DNA construct while other components are delivered as RNA with or without modifications that protect the RNA from degradation or as mRNA capped or uncapped (Zhang et al. 2016 *Nat. Commun.* 7:12617) or Cas protein guide polynucleotide complexes (WO2017070032 published 27 April 2017) or any combination thereof. To produce crRNAs *in-vivo*, tRNA derived elements may also be used to recruit endogenous RNAses to cleave crRNA transcripts into mature forms capable of guiding the complex to its DNA target site, as described, for example, in WO2017105991 published 22 June 2017. Nickase complexes may be utilized separately or concertedly to generate a single or multiple DNA nicks on one or both DNA strands. Furthermore, the cleavage activity of the Cas endonuclease may be deactivated by altering key catalytic residues in its cleavage domain (Sinkunas, T. *et al.*, 2013, *EMBO J.* 32:385-394) resulting in an RNA guided helicase that may be used to enhance homology directed repair, induce transcriptional activation, or remodel local DNA structures. Moreover, the activity of the Cas cleavage and helicase domains may both be knocked-out and used in combination with other DNA cutting, DNA nicking, DNA binding, transcriptional activation, transcriptional repression, DNA remodeling, DNA deamination, DNA unwinding, DNA recombination enhancing, DNA integration, DNA inversion, and DNA repair agents.

[0308] The transcriptional direction of the tracrRNA for the CRISPR-Cas system (if present) and other components of the CRISPR-Cas system (such as variable targeting domain, crRNA repeat, loop, anti-repeat) can be deduced as described in WO2016186946 published 24 November 2016, and WO2016186953 published 24 November 2016.

[0309] As described herein, once the appropriate guide RNA requirement is established, the PAM preferences for each new system disclosed herein may be examined. If the cleavage complex results in degradation of the randomized PAM library, the complex can be converted into a nickase by disabling the ATPase dependent helicase activity either through mutagenesis

of critical residues or by assembling the reaction in the absence of ATP as described previously (Sinkunas, T. *et al.*, 2013, *EMBO J.* 32:385-394). Two regions of PAM randomization separated by two protospacer targets may be utilized to generate a double-stranded DNA break which may be captured and sequenced to examine the PAM sequences that support cleavage by the respective complex.

[0310] Provided herein is a method for modifying a target site in the genome of a cell, the method comprising introducing into a cell at least one PGEN described herein, and identifying at least one cell that has a modification at said target, wherein the modification at said target site is selected from the group consisting of (i) a replacement of at least one nucleotide, (ii) a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, the chemical alteration of at least one nucleotide, and (v) any combination of (i) – (iv).

[0311] The nucleotide to be edited can be located within or outside a target site recognized and cleaved by a Cas endonuclease. In one example, the at least one nucleotide modification is not a modification at a target site recognized and cleaved by a Cas endonuclease. In another example, there are 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, 30, 40, 50, 100, 200, 300, 400, 500, 600, 700, 900 or 1000 nucleotides between the at least one nucleotide to be edited and the genomic target site.

[0312] A knock-out may be produced by an indel (insertion or deletion of nucleotide bases in a target DNA sequence through NHEJ), or by specific removal of sequence that reduces or completely destroys the function of sequence at or near the targeting site.

[0313] A guide polynucleotide/Cas endonuclease induced targeted mutation can occur in a nucleotide sequence that is located within or outside a genomic target site that is recognized and cleaved by the Cas endonuclease.

[0314] The method for editing a nucleotide sequence in the genome of a cell can be a method without the use of an exogenous selectable marker by restoring function to a non-functional gene product.

[0315] In one embodiment, described herein is a method for modifying a target site in the genome of a cell, the method comprising introducing into a cell at least one PGEN described herein and at least one donor DNA, wherein said donor DNA comprises a polynucleotide of interest, and optionally, further comprising identifying at least one cell that said polynucleotide of interest integrated in or near said target site.

[0316] In one aspect, the methods disclosed herein may employ homologous recombination (HR) to provide integration of the polynucleotide of interest at the target site.

[0317] Various methods and compositions can be employed to produce a cell or organism having a polynucleotide of interest inserted in a target site via activity of a CRISPR-Cas system component described herein. In one method described herein, a polynucleotide of interest is introduced into the organism cell via a donor DNA construct. As used herein, "donor DNA" is a DNA construct that comprises a polynucleotide of interest to be inserted into the target site of a Cas endonuclease. The donor DNA construct further comprises a first and a second region of homology that flank the polynucleotide of interest. The first and second regions of homology of the donor DNA share homology to a first and a second genomic region, respectively, present in or flanking the target site of the cell or organism genome.

[0318] The donor DNA can be tethered to the guide polynucleotide. Tethered donor DNAs can allow for co-localizing target and donor DNA, useful in genome editing, gene insertion, and targeted genome regulation, and can also be useful in targeting post-mitotic cells where function of endogenous HR machinery is expected to be highly diminished (Mali *et al.*, 2013, *Nature Methods* Vol. 10:957-963).

[0319] The amount of homology or sequence identity shared by a target and a donor polynucleotide can vary and includes total lengths and/or regions having unit integral values in the ranges of about 1-20 bp, 20-50 bp, 50-100 bp, 75-150 bp, 100-250 bp, 150-300 bp, 200-400 bp, 250-500 bp, 300-600 bp, 350-750 bp, 400-800 bp, 450-900 bp, 500-1000 bp, 600-1250 bp, 700-1500 bp, 800-1750 bp, 900-2000 bp, 1-2.5 kb, 1.5–3 kb, 2-4 kb, 2.5-5 kb, 3-6 kb, 3.5-7 kb, 4-8 kb, 5-10 kb, or up to and including the total length of the target site. These ranges include every integer within the range, for example, the range of 1-20 bp includes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20 bps. The amount of homology can also be described by percent sequence identity over the full aligned length of the two polynucleotides which includes percent sequence identity of about at least 50%, 55%, 60%, 65%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or 100%. Sufficient homology includes any combination of polynucleotide length, global percent sequence identity, and optionally conserved regions of contiguous nucleotides or local percent sequence identity, for example sufficient homology can be described as a region of 75-150 bp having at least 80% sequence identity to a region of the target locus. Sufficient homology can also be described by the predicted ability of two polynucleotides to specifically hybridize under high stringency conditions, see, for example, Sambrook *et al.*, (1989) *Molecular Cloning: A Laboratory Manual*, (Cold Spring Harbor Laboratory Press, NY); *Current Protocols in Molecular Biology*, Ausubel *et al.*, Eds (1994) Current Protocols, (Greene Publishing Associates, Inc. and John Wiley & Sons,

Inc.); and, Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes*, (Elsevier, New York).

[0320] Episomal DNA molecules can also be ligated into the double-strand break, for example, integration of T-DNAs into chromosomal double-strand breaks (Chilton and Que, (2003) *Plant Physiol* 133:956-65; Salomon and Puchta, (1998) *EMBO J.* 17:6086-95). Once the sequence around the double-strand breaks is altered, for example, by exonuclease activities involved in the maturation of double-strand breaks, gene conversion pathways can restore the original structure if a homologous sequence is available, such as a homologous chromosome in non-dividing somatic cells, or a sister chromatid after DNA replication (Molinier *et al.*, (2004) *Plant Cell* 16:342-52). Ectopic and/or epigenic DNA sequences may also serve as a DNA repair template for homologous recombination (Puchta, (1999) *Genetics* 152:1173-81).

[0321] In one embodiment, the disclosure comprises a method for editing a nucleotide sequence in the genome of a cell, the method comprising introducing into at least one PGEN described herein, and a polynucleotide modification template, wherein said polynucleotide modification template comprises at least one nucleotide modification of said nucleotide sequence, and optionally further comprising selecting at least one cell that comprises the edited nucleotide sequence.

[0322] The guide polynucleotide/Cas endonuclease system can be used in combination with at least one polynucleotide modification template to allow for editing (modification) of a genomic nucleotide sequence of interest. (See also US20150082478, published 19 March 2015 and WO2015026886 published 26 February 2015).

[0323] Polynucleotides of interest and/or traits can be stacked together in a complex trait locus as described in WO2012129373 published 27 September 2012, and in WO2013112686, published 01 August 2013. The guide polynucleotide/Cas9 endonuclease system described herein provides for an efficient system to generate double-strand breaks and allows for traits to be stacked in a complex trait locus.

[0324] A guide polynucleotide/Cas system as described herein, mediating gene targeting, can be used in methods for directing heterologous gene insertion and/or for producing complex trait loci comprising multiple heterologous genes in a fashion similar as disclosed in WO2012129373 published 27 September 2012, where instead of using a double-strand break inducing agent to introduce a gene of interest, a guide polynucleotide/Cas system as disclosed herein is used. By inserting independent transgenes within 0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 2, or even 5 centimorgans (cM) from each other, the transgenes can be bred as a single genetic locus (see, for example, US20130263324 published 03 October 2013 or WO2012129373 published 14

March 2013). After selecting a plant comprising a transgene, plants comprising (at least) one transgene can be crossed to form an F1 that comprises both transgenes. In progeny from these F1 (F2 or BC1) 1/500 progeny would have the two different transgenes recombined onto the same chromosome. The complex locus can then be bred as single genetic locus with both transgene traits. This process can be repeated to stack as many traits as desired.

[0325] Further uses for guide RNA/Cas endonuclease systems have been described (See for example:US20150082478 published 19 March 2015, WO2015026886 published 26 February 2015, US20150059010 published 26 February 2015, WO2016007347 published 14 January 2016, and PCT application WO2016025131 published 18 February 2016) and include but are not limited to modifying or replacing nucleotide sequences of interest (such as a regulatory elements), insertion of polynucleotides of interest, gene knock-out, gene-knock in, modification of splicing sites and/or introducing alternate splicing sites, modifications of nucleotide sequences encoding a protein of interest, amino acid and/or protein fusions, and gene silencing by expressing an inverted repeat into a gene of interest.

[0326] Resulting characteristics from the gene editing compositions and methods described herein may be evaluated. Chromosomal intervals that correlate with a phenotype or trait of interest can be identified. A variety of methods well known in the art are available for identifying chromosomal intervals. The boundaries of such chromosomal intervals are drawn to encompass markers that will be linked to the gene controlling the trait of interest. In other words, the chromosomal interval is drawn such that any marker that lies within that interval (including the terminal markers that define the boundaries of the interval) can be used as a marker for a particular trait. In one embodiment, the chromosomal interval comprises at least one QTL, and furthermore, may indeed comprise more than one QTL. Close proximity of multiple QTLs in the same interval may obfuscate the correlation of a particular marker with a particular QTL, as one marker may demonstrate linkage to more than one QTL. Conversely, e.g., if two markers in close proximity show co-segregation with the desired phenotypic trait, it is sometimes unclear if each of those markers identifies the same QTL or two different QTL. The term “quantitative trait locus” or “QTL” refers to a region of DNA that is associated with the differential expression of a quantitative phenotypic trait in at least one genetic background, e.g., in at least one breeding population. The region of the QTL encompasses or is closely linked to the gene or genes that affect the trait in question. An “allele of a QTL” can comprise multiple genes or other genetic factors within a contiguous genomic region or linkage group, such as a haplotype. An allele of a QTL can denote a haplotype within a specified window wherein said window is a contiguous genomic region that can be defined, and tracked, with a set of one or more polymorphic markers.

A haplotype can be defined by the unique fingerprint of alleles at each marker within the specified window.

[0327] **Introduction of CRISPR-Cas System Components into a Cell**

[0328] The methods and compositions described herein do not depend on a particular method for introducing a sequence into an organism or cell, only that the polynucleotide or polypeptide gains access to the interior of at least one cell of the organism. Introducing includes reference to the incorporation of a nucleic acid into a eukaryotic or prokaryotic cell where the nucleic acid may be incorporated into the genome of the cell and includes reference to the transient (direct) provision of a nucleic acid, protein, or polynucleotide-protein complex (PGEN, RGEN) to the cell.

[0329] Methods for introducing polynucleotides or polypeptides or a polynucleotide-protein complex into cells or organisms are known in the art including, but not limited to, microinjection, electroporation, stable transformation methods, transient transformation methods, ballistic particle acceleration (particle bombardment), whiskers mediated transformation, *Agrobacterium*-mediated transformation, direct gene transfer, viral-mediated introduction, transfection, transduction, cell-penetrating peptides, mesoporous silica nanoparticle (MSN)-mediated direct protein delivery, topical applications, sexual crossing , sexual breeding, and any combination thereof.

[0330] For example, the guide polynucleotide (guide RNA, crNucleotide + tracrNucleotide, guide DNA and/or guide RNA-DNA molecule) can be introduced into a cell directly (transiently) as a single stranded or double stranded polynucleotide molecule. The guide RNA (or crRNA + tracrRNA) can also be introduced into a cell indirectly by introducing a recombinant DNA molecule comprising a heterologous nucleic acid fragment encoding the guide RNA (or crRNA + tracrRNA), operably linked to a specific promoter that is capable of transcribing the guide RNA (crRNA+tracrRNA molecules) in said cell. The specific promoter can be, but is not limited to, an RNA polymerase III promoter, which allow for transcription of RNA with precisely defined, unmodified, 5'- and 3'-ends (Ma *et al.*, 2014, *Mol. Ther. Nucleic Acids* 3:e161; DiCarlo *et al.*, 2013, *Nucleic Acids Res.* 41:4336-4343; WO2015026887, published 26 February 2015). Any promoter capable of transcribing the guide RNA in a cell can be used and includes a heat shock /heat inducible promoter operably linked to a nucleotide sequence encoding the guide RNA.

[0331] Plant cells differ from animal cells (such as human cells), fungal cells (such as yeast cells) and protoplasts, including for example plant cells comprise a plant cell wall which may act as a barrier to the delivery of components.

[0332] Delivery of the Cas endonuclease, and/or the guide RNA, and/or a ribonucleoprotein complex, and/or a polynucleotide encoding any one or more of the preceding, into plant cells can be achieved through methods known in the art, for example but not limited to: *Rhizobiales*-mediated transformation (*e.g.*, *Agrobacterium*, *Ochrobactrum*), particle mediated delivery (particle bombardment), polyethylene glycol (PEG)-mediated transfection (for example to protoplasts), electroporation, cell-penetrating peptides, or mesoporous silica nanoparticle (MSN)-mediated direct protein delivery.

[0333] The Cas endonuclease, such as the Cas endonuclease described herein, can be introduced into a cell by directly introducing the Cas polypeptide itself (referred to as direct delivery of Cas endonuclease), the mRNA encoding the Cas protein, and/ or the guide polynucleotide/Cas endonuclease complex itself, using any method known in the art. The Cas endonuclease can also be introduced into a cell indirectly by introducing a recombinant DNA molecule that encodes the Cas endonuclease. The endonuclease can be introduced into a cell transiently or can be incorporated into the genome of the host cell using any method known in the art. Uptake of the endonuclease and/or the guided polynucleotide into the cell can be facilitated with a Cell Penetrating Peptide (CPP) as described in WO2016073433 published 12 May 2016. Any promoter capable of expressing the Cas endonuclease in a cell can be used and includes a heat shock /heat inducible promoter operably linked to a nucleotide sequence encoding the Cas endonuclease.

[0334] Direct delivery of a polynucleotide modification template into plant cells can be achieved through particle mediated delivery, and any other direct method of delivery, such as but not limiting to, polyethylene glycol (PEG)-mediated transfection to protoplasts, whiskers mediated transformation, electroporation, particle bombardment, cell-penetrating peptides, or mesoporous silica nanoparticle (MSN)-mediated direct protein delivery can be successfully used for delivering a polynucleotide modification template in eukaryotic cells, such as plant cells.

[0335] The donor DNA can be introduced by any means known in the art. The donor DNA may be provided by any transformation method known in the art including, for example, *Agrobacterium*-mediated transformation or biolistic particle bombardment. The donor DNA may be present transiently in the cell or it could be introduced via a viral replicon. In the presence of the Cas endonuclease and the target site, the donor DNA is inserted into the transformed plant's genome.

[0336] Direct delivery of any one of the guided Cas system components can be accompanied by direct delivery (co-delivery) of other mRNAs that can promote the enrichment and/or visualization of cells receiving the guide polynucleotide/Cas endonuclease complex

components. For example, direct co-delivery of the guide polynucleotide/Cas endonuclease components (and/or guide polynucleotide/Cas endonuclease complex itself) together with mRNA encoding phenotypic markers (such as but not limiting to transcriptional activators such as CRC (Bruce *et al.* 2000 *The Plant Cell* 12:65-79) can enable the selection and enrichment of cells without the use of an exogenous selectable marker by restoring function to a non-functional gene product as described in WO2017070032 published 27 April 2017.

[0337] Introducing a guide RNA/Cas endonuclease complex described herein, (representing the cleavage ready complex described herein) into a cell includes introducing the individual components of said complex either separately or combined into the cell, and either directly (direct delivery as RNA for the guide and protein for the Cas endonuclease and protein subunits, or functional fragments thereof) or via recombination constructs expressing the components (guide RNA, Cas endonuclease, protein subunits, or functional fragments thereof). Introducing a guide RNA/Cas endonuclease complex (RGEN) into a cell includes introducing the guide RNA/Cas endonuclease complex as a ribonucleotide-protein into the cell. The ribonucleotide-protein can be assembled prior to being introduced into the cell as described herein. The components comprising the guide RNA/Cas endonuclease ribonucleotide protein (at least one Cas endonuclease, at least one guide RNA, at least one protein subunit) can be assembled *in vitro* or assembled by any means known in the art prior to being introduced into a cell (targeted for genome modification as described herein).

[0338] Direct delivery of the RGEN ribonucleoprotein, allows for genome editing at a target site in the genome of a cell which can be followed by rapid degradation of the complex, and only a transient presence of the complex in the cell. This transient presence of the RGEN complex may lead to reduced off-target effects. In contrast, delivery of RGEN components (guide RNA, Cas9 endonuclease) via plasmid DNA sequences can result in constant expression of RGENs from these plasmids which can intensify off target effects (Cradick, T. J. *et al.* (2013) *Nucleic Acids Res* 41:9584-9592; Fu, Y *et al.* (2014) *Nat. Biotechnol.* 31:822-826).

[0339] Direct delivery can be achieved by combining any one component of the guide RNA/Cas endonuclease complex (RGEN), representing the cleavage ready complex described herein, (such as at least one guide RNA, at least one Cas protein, and optionally one additional protein), with a delivery matrix comprising a microparticle (such as but not limited to of a gold particle, tungsten particle, and silicon carbide whisker particle) (see also WO2017070032 published 27 April 2017). The delivery matrix may comprise any one of the components, such as the Cas endonuclease, that is attached to a solid matrix (*e.g.*, a particle for bombardment).

[0340] In one aspect the guide polynucleotide/Cas endonuclease complex, is a complex wherein the guide RNA and Cas endonuclease protein forming the guide RNA /Cas endonuclease complex are introduced into the cell as RNA and protein, respectively.

[0341] In one aspect the guide polynucleotide/Cas endonuclease complex, is a complex wherein the guide RNA and Cas endonuclease protein and the at least one protein subunit of a complex forming the guide RNA/Cas endonuclease complex are introduced into the cell as RNA and proteins, respectively.

[0342] In one aspect the guide polynucleotide/Cas endonuclease complex, is a complex wherein the guide RNA and Cas endonuclease protein and the at least one protein subunit of a complex forming the guide RNA /Cas endonuclease complex (cleavage ready complex) are preassembled *in vitro* and introduced into the cell as a ribonucleotide-protein complex.

[0343] Protocols for introducing polynucleotides, polypeptides or polynucleotide-protein complexes (PGEN, RGEN) into eukaryotic cells, such as plants or plant cells are known and include microinjection (Crossway *et al.*, (1986) *Biotechniques* 4:320-34 and U.S. Patent No. 6,300,543), meristem transformation (U.S. Patent No. 5,736,369), electroporation (Riggs *et al.*, (1986) *Proc. Natl. Acad. Sci. USA* 83:5602-6, *Agrobacterium*-mediated transformation (U.S. Patent Nos. 5,563,055 and 5,981,840), whiskers mediated transformation (Ainley *et al.* 2013, *Plant Biotechnology Journal* 11:1126-1134; Shaheen A. and M. Arshad 2011 Properties and Applications of Silicon Carbide (2011), 345-358 Editor(s):Gerhardt, Rosario. Publisher: InTech, Rijeka, Croatia. CODEN:69PQBP; ISBN:978-953-307-201-2), direct gene transfer (Paszkowski *et al.*, (1984) *EMBO J* 3:2717-22), and ballistic particle acceleration (U.S. Patent Nos. 4,945,050; 5,879,918; 5,886,244; 5,932,782; Tomes *et al.*, (1995) "Direct DNA Transfer into Intact Plant Cells via Microprojectile Bombardment" in *Plant Cell, Tissue, and Organ Culture: Fundamental Methods*, ed. Gamborg & Phillips (Springer-Verlag, Berlin); McCabe *et al.*, (1988) *Biotechnology* 6:923-6; Weissinger *et al.*, (1988) *Ann Rev Genet* 22:421-77; Sanford *et al.*, (1987) *Particulate Science and Technology* 5:27-37 (onion); Christou *et al.*, (1988) *Plant Physiol* 87:671-4 (soybean); Finer and McMullen, (1991) *In vitro Cell Dev Biol* 27P:175-82 (soybean); Singh *et al.*, (1998) *Theor Appl Genet* 96:319-24 (soybean); Datta *et al.*, (1990) *Biotechnology* 8:736-40 (rice); Klein *et al.*, (1988) *Proc. Natl. Acad. Sci. USA* 85:4305-9 (maize); Klein *et al.*, (1988) *Biotechnology* 6:559-63 (maize); U.S. Patent Nos. 5,240,855; 5,322,783 and 5,324,646; Klein *et al.*, (1988) *Plant Physiol* 91:440-4 (maize); Fromm *et al.*, (1990) *Biotechnology* 8:833-9 (maize); Hooykaas-Van Slogteren *et al.*, (1984) *Nature* 311:763-4; U.S. Patent No. 5,736,369 (cereals); Bytebier *et al.*, (1987) *Proc. Natl. Acad. Sci. USA* 84:5345-9 (*Liliaceae*); De Wet *et al.*, (1985) in *The Experimental Manipulation of Ovule Tissues,*

ed. Chapman *et al.*, (Longman, New York), pp. 197-209 (pollen); Kaeppler *et al.*, (1990) *Plant Cell Rep* 9:415-8) and Kaeppler *et al.*, (1992) *Theor Appl Genet* 84:560-6 (whisker-mediated transformation); D'Halluin *et al.*, (1992) *Plant Cell* 4:1495-505 (electroporation); Li *et al.*, (1993) *Plant Cell Rep* 12:250-5; Christou and Ford (1995) *Annals Botany* 75:407-13 (rice) and Osjoda *et al.*, (1996) *Nat Biotechnol* 14:745-50 (maize via *Agrobacterium tumefaciens*).

[0344] Alternatively, polynucleotides may be introduced into plant or plant cells by contacting cells or organisms with a virus or viral nucleic acids. Generally, such methods involve incorporating a polynucleotide within a viral DNA or RNA molecule. In some examples a polypeptide of interest may be initially synthesized as part of a viral polyprotein, which is later processed by proteolysis *in vivo* or *in vitro* to produce the desired recombinant protein. Methods for introducing polynucleotides into plants and expressing a protein encoded therein, involving viral DNA or RNA molecules, are known, see, for example, U.S. Patent Nos. 5,889,191, 5,889,190, 5,866,785, 5,589,367 and 5,316,931.

[0345] The polynucleotide or recombinant DNA construct can be provided to or introduced into a prokaryotic and eukaryotic cell or organism using a variety of transient transformation methods. Such transient transformation methods include, but are not limited to, the introduction of the polynucleotide construct directly into the plant.

[0346] Nucleic acids and proteins can be provided to a cell by any method including methods using molecules to facilitate the uptake of anyone or all components of a guided Cas system (protein and/or nucleic acids), such as cell-penetrating peptides and nanocarriers. See also US20110035836 published 10 February 2011, and EP2821486A1 published 07 January 2015.

[0347] Other methods of introducing polynucleotides into a prokaryotic and eukaryotic cell or organism or plant part can be used, including plastid transformation methods, and the methods for introducing polynucleotides into tissues from seedlings or mature seeds.

[0348] Stable transformation is intended to mean that the nucleotide construct introduced into an organism integrates into a genome of the organism and is capable of being inherited by the progeny thereof. Transient transformation is intended to mean that a polynucleotide is introduced into the organism and does not integrate into a genome of the organism or a polypeptide is introduced into an organism. Transient transformation indicates that the introduced composition is only temporarily expressed or present in the organism.

[0349] A variety of methods are available to identify those cells having an altered genome at or near a target site without using a screenable marker phenotype. Such methods can be viewed as directly analyzing a target sequence to detect any change in the target sequence, including but

not limited to PCR methods, sequencing methods, nuclease digestion, Southern blots, and any combination thereof.

[0350] Cells and Plants.

[0351] The presently disclosed polynucleotides and polypeptides can be introduced into a cell. Cells include, but are not limited to, human, non-human, animal, mammalian, bacterial, fungal, insect, yeast, non-conventional yeast, and plant cells as well as plants and seeds produced by the methods described herein. Any plant can be used with the compositions and methods described herein, including monocot and dicot plants, and plant elements.

[0352] Examples of monocot plants that can be used include, but are not limited to, corn (*Zea mays*), rice (*Oryza sativa*), rye (*Secale cereale*), sorghum (*Sorghum bicolor*, *Sorghum vulgare*), millet (e.g., pearl millet (*Pennisetum glaucum*), proso millet (*Panicum miliaceum*), foxtail millet (*Setaria italica*), finger millet (*Eleusine coracana*)), wheat (*Triticum* species, for example *Triticum aestivum*, *Triticum monococcum*), sugarcane (*Saccharum spp.*), oats (*Avena*), barley (*Hordeum*), switchgrass (*Panicum virgatum*), pineapple (*Ananas comosus*), banana (*Musa spp.*), palm, ornamentals, turfgrasses, and other grasses.

[0353] Examples of dicot plants that can be used include, but are not limited to, soybean (*Glycine max*), *Brassica* species (for example but not limited to:oilseed rape or Canola) (*Brassica napus, B. campestris, Brassica rapa, Brassica. juncea),* alfalfa *(Medicago sativa),*), tobacco (*Nicotiana tabacum*), *Arabidopsis* (*Arabidopsis thaliana*), sunflower (*Helianthus annuus*), cotton (*Gossypium arboreum, Gossypium barbadense*), and peanut (*Arachis hypogaea*), tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*.

[0354] Additional plants that can be used include safflower (*Carthamus tinctorius*), sweet potato (*Ipomoea batatus*), cassava (*Manihot esculenta*), coffee (*Coffea spp.*), coconut (*Cocos nucifera*), citrus trees (*Citrus spp.*), cocoa (*Theobroma cacao*), tea (*Camellia sinensis*), banana (*Musa spp.*), avocado (*Persea americana*), fig (*Ficus casica*), guava (*Psidium guajava*), mango (*Mangifera indica*), olive (*Olea europaea*), papaya (*Carica papaya*), cashew (*Anacardium occidentale*), macadamia (*Macadamia integrifolia*), almond (*Prunus amygdalus*), sugar beets (*Beta vulgaris*), vegetables, ornamentals, and conifers.

[0355] Vegetables that can be used include tomatoes (*Lycopersicon esculentum*), lettuce (*e.g.*, *Lactuca sativa*), green beans (*Phaseolus vulgaris*), lima beans (*Phaseolus limensis*), peas (*Lathyrus spp.*), and members of the genus *Cucumis* such as cucumber (*C. sativus*), cantaloupe (*C. cantalupensis*), and musk melon (*C. melo*). Ornamentals include azalea (*Rhododendron spp.*), hydrangea (*Macrophylla hydrangea*), hibiscus (*Hibiscus rosasanensis*), roses (*Rosa spp.*),

tulips (*Tulipa spp.*), daffodils (*Narcissus spp.*), petunias (*Petunia hybrida*), carnation (*Dianthus caryophyllus*), poinsettia (*Euphorbia pulcherrima*), and chrysanthemum.

[0356] Conifers that may be used include pines such as loblolly pine (*Pinus taeda*), slash pine (*Pinus elliotii*), ponderosa pine (*Pinus ponderosa*), lodgepole pine (*Pinus contorta*), and Monterey pine (*Pinus radiata*); Douglas fir (*Pseudotsuga menziesii*); Western hemlock (*Tsuga canadensis*); Sitka spruce (*Picea glauca*); redwood (*Sequoia sempervirens*); true firs such as silver fir (*Abies amabilis*) and balsam fir (*Abies balsamea*); and cedars such as Western red cedar (*Thuja plicata*) and Alaska yellow cedar (*Chamaecyparis nootkatensis*).

[0357] In certain embodiments of the disclosure, a fertile plant is a plant that produces viable male and female gametes and is self-fertile. Such a self-fertile plant can produce a progeny plant without the contribution from any other plant of a gamete and the genetic material comprised therein. Other embodiments of the disclosure can involve the use of a plant that is not self-fertile because the plant does not produce male gametes, or female gametes, or both, that are viable or otherwise capable of fertilization.

[0358] The present disclosure finds use in the breeding of plants comprising one or more introduced traits, or edited genomes.

[0359] A non-limiting example of how two traits can be stacked into the genome at a genetic distance of, for example, 5 cM from each other is described as follows: A first plant comprising a first transgenic target site integrated into a first DSB target site within the genomic window and not having the first genomic locus of interest is crossed to a second transgenic plant, comprising a genomic locus of interest at a different genomic insertion site within the genomic window and the second plant does not comprise the first transgenic target site. About 5% of the plant progeny from this cross will have both the first transgenic target site integrated into a first DSB target site and the first genomic locus of interest integrated at different genomic insertion sites within the genomic window. Progeny plants having both sites in the defined genomic window can be further crossed with a third transgenic plant comprising a second transgenic target site integrated into a second DSB target site and/or a second genomic locus of interest within the defined genomic window and lacking the first transgenic target site and the first genomic locus of interest. Progeny are then selected having the first transgenic target site, the first genomic locus of interest and the second genomic locus of interest integrated at different genomic insertion sites within the genomic window. Such methods can be used to produce a transgenic plant comprising a complex trait locus having at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,15, 16, 17, 19, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31 or more transgenic

target sites integrated into DSB target sites and/or genomic loci of interest integrated at different sites within the genomic window. In such a manner, various complex trait loci can be generated.

[0360] Cells and Animals.

[0361] The presently disclosed polynucleotides and polypeptides can be introduced into an animal cell. Animal cells can include, but are not limited to, an organism of a phylum including chordates, arthropods, mollusks, annelids, cnidarians, or echinoderms; or an organism of a class including mammals, insects, birds, amphibians, reptiles, or fishes. In some aspects, the animal is human, mouse, C. elegans, rat, fruit fly (*Drosophila* spp.), zebrafish, chicken, dog, cat, guinea pig, hamster, chicken, Japanese ricefish, sea lamprey, pufferfish, tree frog (*e.g.*, *Xenopus* spp.), monkey, or chimpanzee. Particular cell types that are contemplated include haploid cells, diploid cells, reproductive cells, neurons, muscle cells, endocrine or exocrine cells, epithelial cells, muscle cells, tumor cells, embryonic cells, hematopoietic cells, bone cells, germ cells, somatic cells, stem cells, pluripotent stem cells, induced pluripotent stem cells, progenitor cells, meiotic cells, and mitotic cells. In some aspects, a plurality of cells from an organism may be used.

[0362] The novel engineered Cas polypeptides disclosed may be used to edit the genome of an animal cell in various ways. In one aspect, it may be desirable to delete one or more nucleotides. In another aspect, it may be desirable to insert one or more nucleotides. In one aspect, it may be desirable to replace one or more nucleotides. In another aspect, it may be desirable to modify one or more nucleotides via a covalent or non-covalent interaction with another atom or molecule.

[0363] Genome modification via a disclosed engineered Cas polypeptide may be used to effect a genotypic and/or phenotypic change on the target organism. Such a change is preferably related to an improved phenotype of interest or a physiologically-important characteristic, the correction of an endogenous defect, or the expression of some type of expression marker. In some aspects, the phenotype of interest or physiologically-important characteristic is related to the overall health, fitness, or fertility of the animal, the ecological fitness of the animal, or the relationship or interaction of the animal with other organisms in its environment. In some aspects, the phenotype of interest or physiologically-important characteristic is selected from the group consisting of: improved general health, disease reversal, disease modification, disease stabilization, disease prevention, treatment of parasitic infections, treatment of viral infections, treatment of retroviral infections, treatment of bacterial infections, treatment of neurological disorders (for example but not limited to: multiple sclerosis), correction of endogenous genetic defects (for example but not limited to: metabolic disorders, Achondroplasia, Alpha-1 Antitrypsin Deficiency, Antiphospholipid Syndrome, Autism, Autosomal Dominant Polycystic

Kidney Disease, Barth syndrome, Breast cancer, Charcot-Marie-Tooth, Colon cancer, Cri du chat, Crohn's Disease, Cystic fibrosis, Dercum Disease, Down Syndrome, Duane Syndrome, Duchenne Muscular Dystrophy, Factor V Leiden Thrombophilia, Familial Hypercholesterolemia, Familial Mediterranean Fever, Fragile X Syndrome, Gaucher Disease, Hemochromatosis, Hemophilia, Holoprosencephaly, Huntington's disease, Klinefelter syndrome, Marfan syndrome, Myotonic Dystrophy, Neurofibromatosis, Noonan Syndrome, Osteogenesis Imperfecta, Parkinson's disease, Phenylketonuria, Poland Anomaly, Porphyria, Progeria, Prostate Cancer, Retinitis Pigmentosa, Severe Combined Immunodeficiency (SCID), Sickle cell disease, Skin Cancer, Spinal Muscular Atrophy, Tay-Sachs, Thalassemia, Trimethylaminuria, Turner Syndrome, Velocardiofacial Syndrome, WAGR Syndrome, and Wilson Disease), treatment of innate immune disorders (for example but not limited to: immunoglobulin subclass deficiencies), treatment of acquired immune disorders (for example but not limited to: AIDS and other HIV-related disorders), treatment of cancer, as well as treatment of diseases, including rare or "orphan" conditions, that have eluded effective treatment options with other methods.

[0364] Cells that have been genetically modified using the compositions or methods disclosed herein may be transplanted to a subject for purposes such as gene therapy, *e.g.* to treat a disease, or as an antiviral, antipathogenic, or anticancer therapeutic, for the production of genetically modified organisms in agriculture, or for biological research.

[0365] In vitro Polynucleotide Detection, Binding, and Modification

[0366] The compositions disclosed herein may further be used as compositions for use in *in vitro* methods, in some aspects with isolated polynucleotide sequence(s). Said isolated polynucleotide sequence(s) may comprise one or more target sequence(s) for modification. In some aspects, said isolated polynucleotide sequence(s) may be genomic DNA, a PCR product, or a synthesized oligonucleotide.

[0367] Compositions

[0368] Modification of a target sequence may be in the form of a nucleotide insertion, a nucleotide deletion, a nucleotide substitution, the addition of an atom molecule to an existing nucleotide, a nucleotide modification, or the binding of a heterologous polynucleotide or polypeptide to said target sequence. The insertion of one or more nucleotides may be accomplished by the inclusion of a donor polynucleotide in the reaction mixture: said donor polynucleotide is inserted into a double-strand break created by an engineered Cas endonuclease disclosed herein. The insertion may be via non-homologous end joining or via homologous recombination.

[0369] In one aspect, the sequence of the target polynucleotide is known prior to modification and compared to the sequence(s) of polynucleotide(s) that result from treatment with the engineered Cas endonuclease. In one aspect, the sequence of the target polynucleotide is not known prior to modification, and the treatment with the engineered Cas endonuclease is used as part of a method to determine the sequence of said target polynucleotide.

[0370] Polynucleotide modification with an engineered Cas polypeptide may be accomplished by usage of a full-length polypeptide sharing at least 80% identity with any of SEQ ID NOs:19-39. In some aspects, said Cas polypeptide variant is a functional variant of any of SEQ ID NOs:19-39. In some aspects, said Cas polypeptide variant is one of SEQ ID NOs:19-39. In some aspects, the Cas polypeptide variant is provided by way of a Cas polypeptide polynucleotide encoding the Cas polypeptide variant. In some aspects, said polynucleotide encodes an engineered Cas polypeptide selected from the group consisting of SEQID NOs: 19-39 or a sequence sharing at least 80%, 85%, 90%, 95%, 97%, 99%, or 100% with any one of SEQID NOs: 19-39.

[0371] In some aspects, the engineered Cas polypeptide may be selected from the group consisting of: an engineered or modified wild type Cas endonuclease, a functional Cas endonuclease ortholog variant, a functional engineered Cas polypeptide fragment, a fusion protein comprising an active or deactivated engineered Cas polypeptide variant, an engineered Cas polypeptide disclosed herein further comprising one or more nuclear localization sequences (NLS) on the C-terminus or on the N-terminus or on both the N- and C-termini, a biotinylated engineered Cas polypeptide, an engineered Cas polypeptide further comprising a Histidine tag, and a mixture of any two or more of the preceding.

[0372] In some aspects, the engineered Cas polypeptide is a fusion protein further comprising a nuclease domain, a transcriptional activator domain, a transcriptional repressor domain, an epigenetic modification domain, a cleavage domain, a nuclear localization signal, a cell-penetrating domain, a translocation domain, a marker, or a transgene that is heterologous to the target polynucleotide sequence or to the cell from which said target polynucleotide sequence is obtained or derived.

[0373] In some aspects, a plurality of engineered Cas polypeptides may be desired. In some aspects, said plurality may comprise engineered Cas polypeptides derived from different source organisms or from different loci within the same organism. In some aspects, said plurality may comprise engineered Cas polypeptides with different binding specificities to the target polynucleotide. In some aspects, said plurality may comprise engineered Cas endonucleases with different cleavage efficiencies. In some aspects, said plurality may comprise engineered Cas

polypeptides with different PAM specificities. In some aspects, said plurality may comprise engineered Cas polypeptide of different molecular compositions, i.e., a polynucleotide encoding an engineered Cas polypeptide and a polypeptide that is an engineered Cas polypeptide.

[0374] The guide polynucleotide may be provided as a single guide RNA (sgRNA), a chimeric molecule comprising a tracrRNA, a chimeric molecule comprising a crRNA, a chimeric RNA-DNA molecule, a DNA molecule, or a polynucleotide comprising one or more chemically modified nucleotides.

[0375] The storage conditions of the engineered Cas polypeptide and/or the guide polynucleotide include parameters for temperature, state of matter, and time. In some aspects, the Cas polypeptide and/or the guide polynucleotide is stored at about -80 degrees Celsius, at about -20 degrees Celsius, at about 4 degrees Celsius, at about 20-25 degrees Celsius, or at about 37 degrees Celsius. In some aspects, the Cas polypeptide and/or the guide polynucleotide is stored as a liquid, a frozen liquid, or as a lyophilized powder. In some aspects, the Cas polypeptide and/or the guide polynucleotide is stable for at least one day, at least one week, at least one month, at least one year, or even greater than one year.

[0376] Any or all of the possible polynucleotide components of the reaction (e.g., guide polynucleotide, donor polynucleotide, optionally a *Cas polypeptide* polynucleotide) may be provided as part of a vector, a construct, a linearized or circularized plasmid, or as part of a chimeric molecule. Each component may be provided to the reaction mixture separately or together. In some aspects, one or more of the polynucleotide components are operably linked to a heterologous noncoding regulatory element that regulates its expression.

[0377] The method for modification of a target polynucleotide comprises combining the minimal elements into a reaction mixture comprising: an engineered Cas polypeptide (or variant, fragment, or other related molecule as described above), a guide polynucleotide comprising a sequence that is substantially complementary to, or selectively hybridizes to, the target polynucleotide sequence of the target polynucleotide, and a target polynucleotide for modification. In some aspects, the engineered Cas polypeptide is provided as a polypeptide. In some aspects, the engineered Cas polypeptide is provided as a *Cas polypeptide* polynucleotide. In some aspects, the guide polynucleotide is provided as an RNA molecule, a DNA molecule, an RNA:DNA hybrid, or a polynucleotide molecule comprising a chemically modified nucleotide.

[0378] The storage buffer of any one of the components, or the reaction mixture, may be optimized for stability, efficacy, or other parameters. Additional components of the storage buffer or the reaction mixture may include a buffer composition, Tris, EDTA, dithiothreitol

(DTT), phosphate-buffered saline (PBS), sodium chloride, magnesium chloride, HEPES, glycerol, BSA, a salt, an emulsifier, a detergent, a chelating agent, a redox reagent, an antibody, nuclease-free water, a proteinase, and/or a viscosity agent. In some aspects, the storage buffer or reaction mixture further comprises a buffer solution with at least one of the following components: HEPES, MgCl2, NaCl, EDTA, a proteinase, Proteinase K, glycerol, nuclease-free water.

[0379] Incubation conditions will vary according to desired outcome. The temperature is preferably at least 10 degrees Celsius, between 10 and 15, at least 15, between 15 and 17, at least 17, between 17 and 20, at least 20, between 20 and 22, at least 22, between 22 and 25, at least 25, between 25 and 27, at least 27, between 27 and 30, at least 30, between 30 and 32, at least 32, between 32 and 35, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, or even greater than 40 degrees Celsius. The time of incubation is at least 1 minute, at least 2 minutes, at least 3 minutes, at least 4 minutes, at least 5 minutes, at least 6 minutes, at least 7 minutes, at least 8 minutes, at least 9 minutes, at least 10 minutes, or even greater than 10 minutes.

[0380] The sequence(s) of the polynucleotide(s) in the reaction mixture prior to, during, or after incubation may be determined by any method known in the art. In one aspect, modification of a target polynucleotide may be ascertained by comparing the sequence(s) of the polynucleotide(s) purified from the reaction mixture to the sequence of the target polynucleotide prior to combining with the Cas endonuclease ortholog.

[0381] Any one or more of the compositions disclosed herein, useful for *in vitro* or *in vivo* polynucleotide detection, binding, and/or modification, may be comprised within a kit. A kit comprises a Cas endonuclease ortholog or a polynucleotide Cas endonuclease ortholog encoding such, optionally further comprising buffer components to enable efficient storage, and one or more additional compositions that enable the introduction of said Cas endonuclease ortholog or Cas endonuclease ortholog to a heterologous polynucleotide, wherein said Cas endonuclease ortholog or Cas endonuclease ortholog is capable of effecting a modification, addition, deletion, or substitution of at least one nucleotide of said heterologous polynucleotide. In an additional aspect, a Cas endonuclease ortholog disclosed herein may be used for the enrichment of one or more polynucleotide target sequences from a mixed pool. In an additional aspect, a Cas endonuclease ortholog disclosed herein may be immobilized on a matrix for use in *in vitro* target polynucleotide detection, binding, and/or modification.

[0382] A Cas endonuclease may be attached, associated with, or affixed to a solid matrix for the purposes of storage, purification, and/or characterization. Examples of a solid matrix include,

but are not limited to: a filter, a chromatography resin, an assay plate, a test tube, a cryogenic vial, etc. A Cas endonuclease may be substantially purified and stored in an appropriate buffer solution or lyophilized.

[0383] Methods of Detection

[0384] Methods of detecting the engineered Cas polypeptide:guide polynucleotide complex bound to the target polynucleotide may include any known in the art, including but not limited to microscopy, chromatographic separation, electrophoresis, immunoprecipitation, filtration, nanopore separation, microarrays, as well as those described below.

[0385] A DNA Electrophoretic Mobility Shift Assay (EMSA): studies proteins binding to known DNA oligonucleotide probes and assesses the specificity of the interaction. The technique is based on the principle that protein-DNA complexes migrate more slowly than free DNA molecules when subjected to polyacrylamide or agarose gel electrophoresis. Because the rate of DNA migration is retarded upon protein binding, the assay is also called a gel retardation assay. Adding a protein-specific antibody to the binding components creates an even larger complex (antibody-protein-DNA) which migrates even slower during electrophoresis, this is known as a supershift and can be used to confirm protein identities.

[0386] DNA Pull-down Assays use a DNA probe labelled with a high affinity tag, such as biotin, which allows the probe to be recovered or immobilized. A DNA probe can be complexed with a protein from a cell lysate in a reaction similar to that used in the EMSA and then used to purify the complex using agarose or magnetic beads. The proteins are then eluted from the DNA and detected by Western blot or identified by mass spectrometry. Alternatively, the protein may be labelled with an affinity tag or the DNA-protein complex may be isolated using an antibody against the protein of interest (similar to a supershift assay). In this case, the unknown DNA sequence bound by the protein is detected by Southern blotting or through PCR analysis.

[0387] Reporter assays provide a real-time in vivo read-out of translational activity for a promoter of interest. Reporter genes are fusions of a target promoter DNA sequence and a reporter gene DNA sequence which is customized by the researcher and the DNA sequence codes for a protein with detectable properties like firefly /Renilla luciferase or alkaline phosphatase. These genes produce enzymes only when the promoter of interest is activated. The enzyme, in turn, catalyzes a substrate to produce either light or a color change that can be detected by spectroscopic instrumentation. The signal from the reporter gene is used as an indirect determinant for the translation of endogenous proteins driven from the same promoter.

[0388] Microplate Capture and Detection Assays use immobilized DNA probes to capture specific protein-DNA interactions and confirm protein identities and relative amounts with target

specific antibodies. Typically, a DNA probe is immobilized on the surface of 96- or 384-well microplates coated with streptavidin. A cellular extract is prepared and added to allow the binding protein to bind to the oligonucleotide. The extract is then removed and each well is washed several times to remove non-specifically bound proteins. Finally, the protein is detected using a specific antibody labelled for detection. This method can be extremely sensitive , detecting less than 0.2pg of the target protein per well. This method may also be utilized for oligonucleotides labelled with other tags, such as primary amines that can be immobilized on microplates coated with an amine-reactive surface chemistry.

[0389] DNA Footprinting is one of the most widely used methods for obtaining detailed information on the individual nucleotides in protein–DNA complexes, even inside living cells. In such an experiment, chemicals or enzymes are used to modify or digest the DNA molecules. When sequence specific proteins bind to DNA, they can protect the binding sites from modification or digestion. This can subsequently be visualized by denaturing gel electrophoresis, where unprotected DNA is cleaved more or less at random. Therefore it appears as a 'ladder' of bands and the sites protected by proteins have no corresponding bands and look like footprints in the pattern of bands. The footprints there by identify specific nucleosides at the protein–DNA binding sites.

[0390] Microscopic techniques include optical, fluorescence, electron, and atomic force microscopy (AFM).

[0391] Chromatin immunoprecipitation analysis (ChIP) causes proteins to bind covalently to their DNA targets, after which they are unlinked and characterized separately.

[0392] Systematic Evolution of Ligands by EXponential enrichment (SELEX) exposes target proteins to a random library of oligonucleotides. Those genes that bind are separated and amplified by PCR.

[0393] While the invention has been particularly shown and described with reference to a preferred embodiment and various alternate embodiments, it will be understood by persons skilled in the relevant art that various changes in form and details can be made therein without departing from the spirit and scope of the invention. For instance, while the particular examples below may illustrate the compositions, methods, and embodiments described herein using a specific target site or target organism, the principles in these examples may be applied to any target site or target organism. Therefore, it will be appreciated that the scope of this invention is encompassed by the embodiments of the inventions described in the claims and specification herein, rather than the specific examples below. All cited patents, applications, and publications referred to in this application are herein incorporated by reference in their entirety, for all

purposes, to the same extent as if each were individually and specifically incorporated by reference, except for any definitions, subject matter disclaimers or disavowals, and except to the extent that the incorporated material is inconsistent with the express disclosure herein, in which case the language in this disclosure controls.

EXAMPLES

[0394] The following are Examples of specific embodiments of some aspects of the invention. The Examples are offered for illustrative purposes only, and are not intended to limit the scope of the invention in any way. Efforts have been made to ensure accuracy with respect to numbers used (e.g., amounts, temperatures, etc.), but some experimental error and deviation should, of course, be allowed for.

[0395] **Example 1: *Saccharomyces cerevisiae* DNA expression cassettes.** In this Example, methods for generating Cas-alpha endonuclease and guide RNA expression cassettes for use in *Saccharomyces cerevisiae* cells are described.

[0396] In one method, to confer efficient expression in *S. cerevisiae*, the gene encoding the Cas-alpha endonuclease was yeast codon optimized. To facilitate nuclear localization of the optimized Cas-alpha endonuclease protein, a nucleotide sequence encoding the Simian virus 40 (SV40) monopartite nuclear localization signal (NLS) (PKKKRKV (SEQ ID NO:17)) was optionally added to either the 5' or 3' ends. The nucleotide sequences of the optimized *cas-alpha* endonuclease gene and NLS variants were then synthesized and operably cloned into a 2 micron yeast plasmid DNA between the ROX3 promoter and CYC1 terminator (GenScript).

[0397] The Cas-alpha endonuclease is directed by small RNAs (referred to herein as guide RNAs) to cleave double-stranded DNA in the presence of a 5' protospacer adjacent motif (PAM) (Bigelyte *et al.* (2021), *Nature Communications*. 12: 6191, Karvelis *et al.* (2020), *Nucleic Acids Research*. 48, 5016-5023 and U.S. Patent Application Publication No. US20200190494A1). These guide RNAs comprise a sequence that aids recognition by Cas-alpha (referred to as Cas-alpha recognition domain) and a sequence that serves to direct Cas-alpha cleavage by base pairing with one strand of the DNA target site (Cas-alpha variable targeting domain). To transcribe small RNAs necessary for directing Cas-alpha endonuclease cleavage activity in *S. cerevisiae cells*, DNA sequences encoding the Hepatitis Delta Virus ribozyme was first appended to the 3' end of a DNA sequence encoding a Cas-alpha single guide RNA (sgRNA) with a variable targeting domain capable of targeting the yeast *ADE2* gene. Next, the SNR52 promoter and SUP4 terminator were operably linked to the ends of the ribozyme and Cas-alpha encoding sgRNA incorporating a G bp onto the 3' end of SNR52 to promote transcription of the

Cas-alpha sgRNA. DNA fragments were then synthesized and cloned into the *S. cerevisiae* 2 micron vector containing the *cas-alpha* gene (GenScript).

[0398] A schema showing a yeast optimized Cas-alpha nuclease expression cassette with the sequences described herein is provided in FIG. 1.

[0399] **Example 2: *Saccharomyces cerevisiae* transformation.** In this Example, methods for transforming Cas-alpha endonuclease and guide RNA expression cassettes into *Saccharomyces cerevisiae* cells are described.

[0400] Several methods (lithium acetate, polyethylene glycol (PEG), heat shock, electroporation, biolistic, and others) can be used to transform *S. cerevisiae* (Kawai *et al.* (2010) *Bioengineered Bugs.* 1:395-403). Here, an approach similar to a lithium cation-based method using the Frozen-EZ yeast Transformation II™ kit from Zymo Research (Irvine, CA USA) was used. Per the manufacture's instruction, *S. cerevisiae* competent cells were produced. This was accomplished by growing *S. cerevisiae* (BY4742 (Baker *et al.* (1998) *Yeast.* 14, 115-132) (ATCC)) in yeast extract-peptone-dextrose (YPD) broth (Gibco) to mid-log phase corresponding to an OD 600 nm of 0.8-1.0. Next, the cells were pelleted by centrifugation (500xg for 4 minutes), media decanted, and the pellet gently washed with 10 ml of EZ 1 solution spinning down the cells again prior to removing the wash solution. The cells were then resuspended in 1 ml of EZ 2 solution and aliquoted and either stored at -70°C or used in the next step. Transformation was performed next by adding 0.5-1 µg (in less than 5 µl) of 2 micron yeast plasmid DNA to 50 µl of competent cells. Optionally, double-stranded DNA repair template with homology flanking the expected Cas-alpha double-strand break site was also included (0.5 µl at 50 µM). After gently mixing in the DNA, 500 µl of EZ 3 solution was added. Next, cells were incubated at 30°C for 60-90 min. flicking or vortexing the cells 3-4 times over the duration of the incubation. After transformation, cells were grown-out in YPD broth for ~3 hours, pelleted, washed once with 1 ml of sterile water, resuspended in 1 ml of sterile water, and then ~200 µl plated onto selective media (for example but not limited to 6.7 g/L yeast nitrogen base without amino acids (Becton Dickinson), 20 g/L glucose (Phytotechnology Labs), 1.92 g/L yeast histidine dropout media (MP Biomedicals) and 20 g/L Bacto™ Agar (Becton Dickinson)). To determine the culture conditions optimal for activity, cells were incubated at 30°C until colonies formed, or at a range of temperatures (typically 37°C and 45°C) overnight and then placed back at 30°C until colony growth was visible.

[0401] **Example 3: Selecting for improved cellular DNA target cleavage.** In this Example, methods for selecting for Cas-alpha endonuclease or guide RNA variants with improved double-stranded DNA target cleavage are described.

[0402] In one method, the *ADE2* gene in *Saccharomyces cerevisiae* (BY4742, genotype - MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0) was targeted for Cas-alpha endonuclease target cleavage (FIG. 2). Here, target cleavage and cellular repair that results in the formation of a non-functional *ade2* gene results in adenine auxotrophy and the switch from a white to a red (pink) cellular phenotype (Ugolini *et al.* (1996) *Curr. Genet.* 30:485-492 and U.S. Patent Application Publication No. US20200190494A1). This alteration in color was used to select cells expressing a Cas nuclease variant and/or associated guide RNA (gRNA) with improved targeted DSB activity. When viewed as a colony and depending on how quickly the *ADE2* gene was disrupted, red coloration varied from entirely red to the observance of smaller red sector(s) within the otherwise white colony. These phenotypic patterns were also used to quantify differences in activity among variants. Colonies scoring as completely red, those containing several sectors or just one red sector were counted and divided by the total number of colonies present. In some instances, instead of counting colonies, images of yeast colonies were captured with a Nikon Digital Sight Ds-Fi1 camera (Nikon, Japan) and NIS-Elements BR software (version 4.00.07) (Nikon, Japan) and analyzed by first determining the total yeast area (as pixels) and then calculating the total percentage of red using custom scripts.

[0403] Once an improved Cas variant was identified, it was transferred to 5 mls of yeast broth without histidine (for example but not limited to 6.7 g/L yeast nitrogen base without amino acids (Becton Dickinson), 20 g/L glucose (Phytotechnology Labs) and 1.92 g/L yeast histidine dropout media (MP Biomedicals)) and incubated overnight at 30°C with shaking. The 2 micron plasmid encoding the variant(s) was then isolated using a Yeast Plasmid Miniprep 96 kit (Zymo Research). After purification, the plasmid was next transformed into TransforMax™ EPI300 (Lucigen) *E. coli* competent cells per the manufacture's instruction and plated on selective media (for example but not limited to 6.7 g/L yeast nitrogen base without amino acids (Becton Dickinson), 20 g/L glucose (Phytotechnology Labs), 1.92 g/L yeast histidine dropout media (MP Biomedicals) and 20 g/L Bacto Agar (Becton Dickinson)). Since multiple 2 micron vectors can be maintained in a single yeast cell, 6 colonies from each *E. coli* transformation were selected and each used to inoculate 2 mls of 2X YT medium (Sigma-Aldrich) or equivalent containing carbenicillin. Cultures were grown overnight at 37°C with shaking and then half of the culture was subject to rolling circle amplification and sanger sequencing (Eurofins Scientific) using primers specific to the regions immediately adjacent to or within the *cas-alpha* gene. After sequencing, plasmid DNA was isolated from *E. coli* cultures (the remaining half) shown to contain different variants using a Qiagen Spin Miniprep Kit (Qiagen). Finally, ~1 μg of each plasmid was re-transformed back into *S. cerevisiae* and evaluated for a red cellular phenotype to

confirm improvements. To make a comparison among improved variants and the original (wildtype) Cas-alpha endonuclease, an additional yeast transformation was typically performed with all plasmids and their ability to recognize and cleave the *ADE2* target site was compared.

[0404] **Example 4: Library design and generation.** In this Example, methods for designing and generating Cas endonuclease variants for improved double-stranded DNA target cleavage are described.

[0405] In one method, saturation mutagenesis was performed introducing all other amino acids (19 in total) at each position of the entire protein (for example but not limited to Cas-alpha 8 (FIG. 3)). In a second method, beneficial changes identified herein were incorporated (either individually or in combination) into variants already containing one or more beneficial changes and assayed for their combined effects.

[0406] For all library approaches, codons within the *cas-alpha* nuclease gene in the yeast expression plasmid shown in FIG. 1 (SEQ ID NO:3) were altered to encode for different amino acids using GenPlus gene synthesis technology (GenScript).

[0407] **Example 5: Cas endonuclease variants with improved DNA target cleavage**. In this Example, Cas endonuclease variants with improved double-stranded DNA cleavage activity are described. All variants generated in Example 4 were tested. These included variants that had all possible amino acid substitutions across the entire length of the protein (positions 1-422, relative to SEQ ID NO:18) and variants containing a combination of beneficial alterations. Assays were conducted to capture variants that had improved desirable activity.

[0408] Table 1 lists the amino acid substitutions that improved double-stranded DNA target cleavage activity. *S. cerevisiae* experiments were performed with an overnight 37°C incubation and then placed back at 30°C until colony growth was visible. Photographs were then taken of the *S. cerevisiae* colonies and total percentage of *ade2* red phenotype calculated by image analysis with custom scripts. The fold improvement in targeted DNA cleavage was calculated by dividing the total percentage of red *S. cerevisiae* produced with each variant by that observed from the unmodified wildtype Cas-alpha 8 protein (SEQ ID NO:18 (FIG. 3)).

Table 1

Amino Acid Alteration	% Red Yeast	Fold Improvement in Target DNA Cleavage	SEQ ID NO:
None (wildtype Cas-alpha 8)	0.41	n/a	18
I123Y	28.78	69.53	19
L226Q	8.78	21.22	20
A231E	21.71	52.44	21
A231T	51.40	124.17	22
A231Y	35.72	86.30	23

R266T	68.35	165.12	24
A295P	40.96	98.94	25
T301R	52.62	127.12	26
Y305H	49.86	120.45	27
R335D	33.39	80.65	28
R335E	37.29	90.07	29
R335P	14.40	34.80	30
R335Q	26.49	63.99	31
F336D	37.44	90.43	32
F336E	46.51	112.36	33
F336V	34.02	82.19	34
L337I	33.48	80.87	35
L337T	35.10	84.79	36
L337V	40.09	96.84	37
T341P	18.32	44.25	38

[0409] The alterations listed in Table 1 yielded between a 21 to 165-fold enhancement in targeted DNA cleavage activity relative to the starting wildtype protein.

[0410] Combinations of the changes listed in Table 1 were then tested for additional improvement in targeted DNA cleavage using the method and calculations based on *ade2* red phenotype described above. One combination (SEQ ID NO:39) that combined amino acid substitutions I23Y+R266T+A295P demonstrated a significant further improvement of targeted DNA cleavage activity (Table 2).

Table 2

Amino Acid Alteration	**% Red Yeast**	**Fold Improvement in Target DNA Cleavage**	**SEQ ID NO:**
None (wildtype Cas-alpha 8)	0.39	na	18
I123Y, R266T+A295P	77.36	198.36	39

[0411] Each of the modified Cas polypeptides disclosed in Table 1 and Table 2 is an example of a novel engineered Cas polypeptide disclosed herein and can be used in the methods disclosed herein.

WE CLAIM:

1. An engineered Cas polypeptide comprising a sequence having 90% amino acid sequence identity to SEQ ID NO:18 and one or more of the following amino acids at positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341, wherein the engineered Cas polypeptide is capable of site specifically binding to a target site of a polynucleotide.

2. An engineered Cas polypeptide, comprising:

(a) A C-terminal tri-split RuvC domain and a zinc finger motif; and

(b) one or more of the following amino acids at positions relative to an alignment with SEQ ID NO:18: Tyrosine at 123, Glutamine at 226; Glutamate or Threonine at 231, Tyrosine at 231, Threonine at 266, Proline at 295, Arginine at 301, Histidine at 305, Aspartate or Glutamate or Proline or Glutamine at 335, Aspartate or Glutamate or Valine at 336, Isoleucine or Threonine or Valine at 337, and Proline at 341,

wherein the engineered Cas polypeptide is capable of site specifically binding to a target site of a polynucleotide.

3. The engineered Cas polypeptide of claim 1 or 2 comprising a sequence having at least 95% amino acid sequence identity to any one of SEQ ID NOs:19 to 39.

4. The engineered Cas polypeptide of claim 1 or 2 comprising the amino acid sequence of any one of SEQ ID NOs:19 to 39.

5. The engineered Cas polypeptide of any one of Claims 1-4, wherein the engineered Cas polypeptide has greater cleavage activity than SEQ ID NO:18 on the same DNA substrate.

6. The engineered Cas polypeptide of Claim 5, wherein the engineered wherein the engineered Cas polypeptide has at least 10 times the DNA cleavage activity relative to the cleavage activity of SEQ ID NO:18 on the same DNA substrate.

7. The engineered Cas polypeptide of Claim 5, wherein the engineered Cas polypeptide has at least 100 times the DNA cleavage activity relative to the cleavage activity of SEQ ID NO:18 on the same DNA substrate.

8. The engineered Cas polypeptide of any one of Claims 5-7, wherein the engineered Cas polypeptide has greater cleavage activity than SEQ ID NO:18 at 30°C or less.

9. The engineered Cas polypeptide of any one of Claims 1-8, wherein the polypeptide has fewer than about 500 amino acids in length.

10. The engineered Cas polypeptide of any one of Claims 1-9, wherein the polypeptide is in a complex, the complex comprising a target site on a double-stranded DNA polynucleotide.

11. The engineered Cas polypeptide of any one of Claims 1-10, further comprising a guide polynucleotide comprising a variable targeting domain that comprises a region of complementarity to the target site of a polynucleotide.

12. The engineered Cas polypeptide of Claim 11, wherein the guide polynucleotide variable targeting domain comprises fewer than 20 nucleotides.

13. The engineered Cas polypeptide of Claim 11 or 12, wherein the engineered Cas polypeptide recognizes a PAM sequence on a target polynucleotide, and wherein the guide polynucleotide and the Cas polypeptide form a complex that binds the target site on a double-stranded DNA polynucleotide.

14. The engineered Cas polypeptide of any one of Claims 1-13, wherein the Cas polypeptide is an endonuclease that cleaves a double-stranded DNA polynucleotide.

15. The engineered Cas polypeptide of any one of Claims 1-3**Error! Reference source not found.** or 9-13, wherein the engineered Cas polypeptide is catalytically inactive for endonuclease activity.

16. The engineered Cas polypeptide of any one of Claims 1-15, wherein the engineered Cas polypeptide recognizes a PAM sequence that comprises thymine dinucleotide (TT).

17. The engineered Cas polypeptide of any one of Claims 1-16, wherein the Cas polypeptide is part of a fusion protein.

18. The engineered Cas polypeptide of any one of Claims 17, wherein the fusion protein further comprises a heterologous nuclease domain.

19. The engineered Cas polypeptide of any one of Claims 1-18, further comprising a deaminase.

20. A synthetic composition comprising the engineered Cas polypeptide of any one of Claims 1-19, further comprising a heterologous polynucleotide.

21. The synthetic composition of Claim 20, wherein the heterologous polynucleotide is an expression element, transgene, donor DNA molecule or polynucleotide modification template.

22. The synthetic composition of Claim 20, wherein the heterologous polynucleotide is a temperature-inducible promoter.

23. A synthetic composition comprising:

(a) An engineered Cas polypeptide in accordance with any of Claims 1-10;

(b) a target double-stranded DNA polynucleotide; and

(c) a guide polynucleotide comprising a variable targeting domain that comprises a region of complementarity to a target double-stranded DNA polynucleotide;

wherein the Cas polypeptide recognizes a PAM sequence on the target double-stranded DNA polynucleotide, wherein the guide polynucleotide and the Cas polypeptide form a complex that binds the target double-stranded DNA polynucleotide.

24. A polynucleotide encoding the engineered Cas polypeptide of any one of Claims 1-19 or the synthetic composition of any one of Claims 20-23.

25. The polynucleotide of Claim 24, wherein the polynucleotide encodes the engineered Cas polypeptide and at least one expression element.

26. The polynucleotide of Claim 24, wherein the polynucleotide encodes the engineered Cas polypeptide and a gene.

27. The engineered Cas polypeptide of any one of Claims 1-19, wherein the Cas polypeptide is attached to a solid matrix or the Cas polypeptide is complexed with a guide polynucleotide and the Cas polypeptide /guide polynucleotide complex is attached to a solid matrix.

28. A eukaryotic cell comprising the engineered Cas polypeptide of any one of claims 1-19, the synthetic composition of any one of Claims 20-23, or the polynucleotide of any one of Claims 24-26.

29. The eukaryotic cell of Claim 28, wherein the eukaryotic cell is a plant cell, an animal cell, or a fungal cell.

30. The eukaryotic cell of Claim 28, wherein the eukaryotic cell is a monocot plant cell or a dicot plant cell.

31. The eukaryotic cell of Claim 29, wherein the plant cell is a cell from maize, soybean, cotton, wheat, canola, oilseed rape, sorghum, rice, rye, barley, millet, oats, sugarcane, turfgrass, switchgrass, alfalfa, sunflower, tobacco, peanut, potato, *Arabidopsis*, safflower, or tomato.

32. The eukaryotic cell of any one of Claims 28-31, wherein the eukaryotic cell is at a temperature of about 37 degrees or less, about 35 degrees Celsius or less, about 30 degrees Celsius or less, about 25 degrees Celsius or less, or about 20 degrees or less.

33. A method of introducing a targeted edit in a target polynucleotide, the method comprising:

(a) providing the Cas polypeptide and guide polynucleotide of any one of claims 10-12, wherein the Cas polypeptide/guide polynucleotide form a complex that recognizes a PAM sequence on the target polynucleotide; and

(b) contacting the Cas polypeptide/guide polynucleotide complex with the target; and

(c) introducing a targeted edit in the target polynucleotide.

34. The method of Claim 33, wherein the target polynucleotide is a target genomic sequence of a cell and the method comprises:

(i) delivering the Cas polypeptide/guide polynucleotide complex to the cell;

(ii) incubating the cell at a temperature of about 37 degrees or less, about 35 degrees Celsius or less, about 30 degrees Celsius or less, about 25 degrees Celsius or less, or about 20 degrees or less;

(iii) modifying at least one nucleotide in the target genomic sequence of the cell to generate a modified genomic sequence as compared to the target genomic sequence of the cell prior to the delivering the Cas polypeptide/guide polynucleotide complex; and

(iv) generating a whole organism from the cell, wherein the organism comprises the modified genomic sequence.

35. The method of Claim 34, wherein the cell is a eukaryotic cell.

36. The method of Claim 35, wherein the eukaryotic cell is derived or obtained from an animal, a fungus, or a plant.

37. The method of Claim 36, wherein the eukaryotic cell is from a plant that is a monocot or a dicot.

38. The method of Claim 37, wherein the plant is selected from the group consisting of: maize, soybean, cotton, wheat, canola, oilseed rape, sorghum, rice, rye, barley, millet, oats, sugarcane, turfgrass, switchgrass, alfalfa, sunflower, tobacco, peanut, potato, *Arabidopsis*, safflower, and tomato.

39. The method of any one of Claims 33-38, wherein the guide polynucleotide variable targeting domain comprises fewer than 20 nucleotides.

40. The method of Claims 33-39, further comprising providing a heterologous polynucleotide

41. The method of Claim 40, wherein the heterologous polynucleotide is a donor DNA molecule.

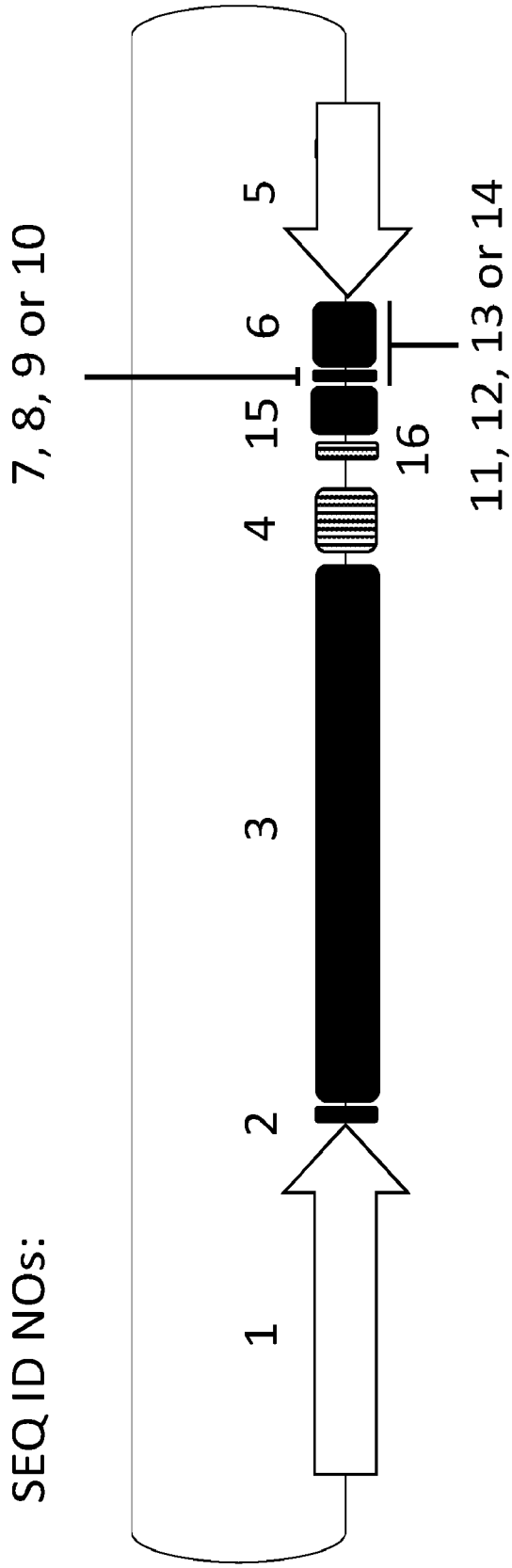
42. The method of Claim 40, wherein the heterologous polynucleotide is a polynucleotide modification template that comprises a sequence at least 50% identical to a sequence in the cell.

43. The method of Claim 40, wherein the heterologous polynucleotide is an inducible promoter.

44. The method of anyone of Claims 33-43, wherein the targeted edit is introduced at a temperature of about 40 degrees Celsius or less, about 37 degrees or less, about 35 degrees Celsius or less, about 30 degrees Celsius or less, about 25 degrees Celsius or less, or about 20 degrees or less.

45. A kit comprising the Cas polypeptide of any one of claims 1-19, the polynucleotide of any one of claims 24-26, or the Cas endonuclease and solid matrix of claim 27.

FIG. 1



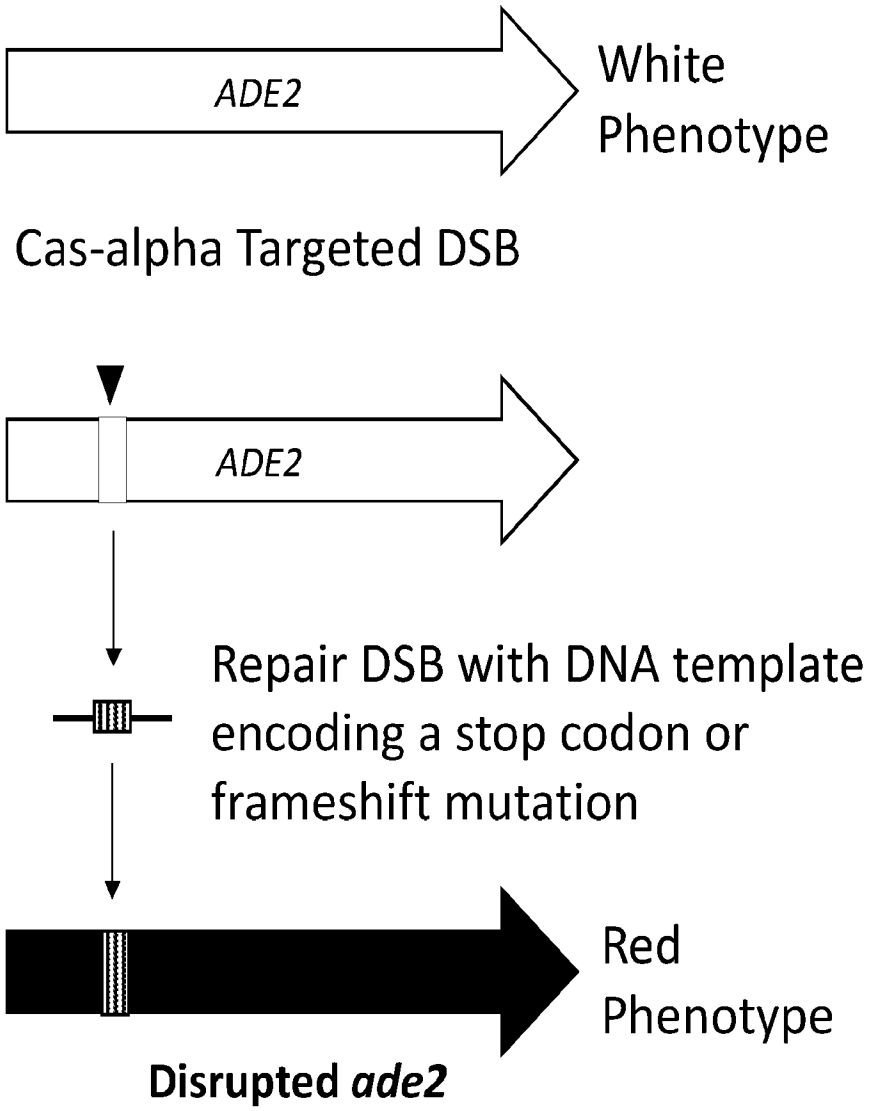

ADE2
White
Phenotype
Cas-alpha Targeted DSB
ADE2
Repair DSB with DNA template
encoding a stop codon or
frameshift mutation
Red
Phenotype
Disrupted ade2

FIG. 2

FIG. 3

SEQ ID NO: 18

1 MIKVYRYETVKPLDLDWKEFGTILRQLQQETRFALNKATQLAWEWMGFSS 50

51 DYKDNHGEYPKSKDILGYTNVHGYAYHTIKTKAYRLNSGNLSQTIKRATD 100

101 RFKAYQKEILRGDMSIPSYKRDIPLDLIKENISVNRMNHGDYIASLSLLS 150

151 NPAKQEMNVKRKISVIIIVRGAGKTIMDRILSGEYQVSASQIIHDDRKNK 200

201 WYLNISYDFEPQTRVLDLNKIMGI**D**LGVAVAVYMAFQHTPARYKLEGGEI 250

251 ENFRRQVESRRISMLRQGKYAGGARGGHGRDKRIKPIEQLRDKIANFRDT 300

301 TNHRYSRYIVDMAIKEGCGTIQM**E**DLTNIRDIGSRFLQNWTYYDLQQKII 350

351 YKAEEAGIKVIKIDPQYTSQRCSECGNIDSGNRIGQAIFKCRACGYEANA 400

401 **D**YNAARNIAIPNIDKIIAESIK 422