



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

(21) International Application Number:
PCT/US2023/084258

(22) International Filing Date:
15 December 2023 (15.12.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/387,767 16 December 2022 (16.12.2022) US

(71) Applicant: PAIRWISE PLANTS SERVICES, INC.
[US/US]; 807 East Main Street, Suite 4-100, Durham, North
Carolina 27701 (US).

(72) Inventor: KIM, Yongjoo; 114 Sir Richard Ln., Chapel Hill,
North Carolina 27517 (US).

(74) Agent: BOBAY, Erin Regal; Myers Bigel, P.A., 4130 Park-
lake Avenue, Ste. 250, Raleigh, North Carolina 27612 (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.

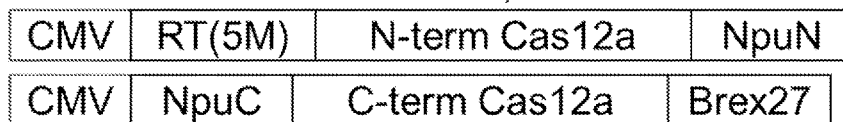
(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, CV,
GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST,
SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ,
RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ,
DE, DK, EE, ES, FI, FR, GB, GR, 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).

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: FUSION PROTEINS COMPRISING AN INTEIN POLYPEPTIDE AND METHODS OF USE THEREOF

FIG. 3



Split site: Cas12a residue: 175, 310, 406, 441, 550

(57) Abstract: Described herein are fusion proteins that include an intein polypeptide along with methods of using of such proteins. Fusion proteins described herein may include a Cas12a polypeptide and an intein polypeptide or a polypeptide of interest (e.g., a reverse transcriptase polypeptide) and an intein polypeptide. Also described herein are compositions and systems for modifying or editing a target nucleic acid.



FUSION PROTEINS COMPRISING AN INTEIN POLYPEPTIDE AND METHODS OF USE THEREOF

STATEMENT REGARDING ELECTRONIC FILING OF A SEQUENCE LISTING

A Sequence Listing in XML format, entitled 1499-116_ST26.xml, 502,046 bytes in size, generated on December 14, 2023, and filed herewith, is hereby incorporated by reference in its entirety for its disclosures.

FIELD

This invention relates to fusion proteins (e.g., engineered proteins) that include a Cas12a polypeptide and an intein polypeptide and to methods of use of such proteins. The invention also relates to fusion proteins (e.g., engineered proteins) that include a polypeptide of interest (e.g., a reverse transcriptase polypeptide) and an intein polypeptide and to methods of use of such proteins. The invention further relates to compositions and systems for modifying or editing a target nucleic acid.

BACKGROUND OF THE INVENTION

Large genome editing agents (e.g., those over 2000 amino acids in length) are usually delivered into tissues using means other than Adeno-associated virus (AAV) vectors. For example, large genome editing agents may be delivered using lipid nanoparticle mediated RNP delivery, which has no size limit, or using mRNA or DNA. However, these approaches prohibit use in applications where the use of AAV is necessary, for example, applications for delivery to the brain of a subject.

Accordingly, new methods for preparing and/or delivering genome editing agents are needed.

SUMMARY OF THE INVENTION

A first aspect of the present invention is directed to a fusion protein comprising an intein polypeptide. In some embodiments, the fusion protein comprises a Cas12a polypeptide fused to an intein polypeptide. In some embodiments, the fusion protein comprises a polypeptide of interest fused to an intein polypeptide. In some embodiments, the fusion protein comprises a reverse transcriptase polypeptide fused to an intein polypeptide. A nucleic acid molecule encoding a fusion protein as described herein is also provided.

Another aspect of the present invention is directed to a complex comprising: a Cas12a protein that is prepared from a first fusion protein and a second fusion protein, wherein the first fusion protein comprises a first Cas12a polypeptide fused to a first intein polypeptide and the second fusion protein comprises a second Cas12a polypeptide fused to a second intein polypeptide; a guide nucleic acid (e.g., a guide RNA); and optionally a deaminase.

An additional aspect of the present invention is directed to a complex comprising: an engineered protein (e.g., a base editor or a templated editor such as a REDRAW editor) that is prepared from a first fusion protein and a second fusion protein, wherein the first fusion protein comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide and the second fusion protein comprises a Cas12a polypeptide fused to a second intein polypeptide; and a guide nucleic acid (e.g., a guide RNA).

A further aspect of the present invention is directed to a method of modifying a target nucleic acid, the method comprising: contacting the target nucleic acid with: a Cas12a protein that is prepared from a first fusion protein and a second fusion protein, wherein the first fusion protein comprises a first Cas12a polypeptide fused to a first intein polypeptide and the second fusion protein comprises a second Cas12a polypeptide fused to a second intein polypeptide; and a guide nucleic acid (e.g., a guide RNA), optionally wherein the Cas12a protein and the guide nucleic acid form a complex or are comprised in a complex, thereby modifying the target nucleic acid.

Another aspect of the present invention is directed to a method of modifying a target nucleic acid, the method comprising: contacting the target nucleic acid with: an engineered protein (e.g., a base editor or a templated editor such as a REDRAW editor) that is prepared from a first fusion protein and a second fusion protein, wherein the first fusion protein comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide and the second fusion protein comprises a Cas12a polypeptide fused to a second intein polypeptide; and a guide nucleic acid (e.g., a guide RNA), optionally wherein the engineered protein and the guide nucleic acid form a complex or are comprised in a complex, thereby modifying the target nucleic acid.

An additional aspect of the present invention is directed to a composition comprising: a first fusion protein that comprises a first Cas12a polypeptide fused to a first intein polypeptide; and a second fusion protein that comprises a second Cas12a polypeptide fused to a second intein polypeptide.

A further aspect of the present invention is directed to a composition comprising: a first fusion protein that comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide)

fused to a first intein polypeptide; and a second fusion protein that comprises a Cas12a polypeptide fused to a second intein polypeptide.

Another aspect of the present invention is directed to a composition comprising: a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a first Cas12a polypeptide fused to a first intein polypeptide; and a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a second Cas12a polypeptide fused to a second intein polypeptide.

An additional aspect of the present invention is directed to a composition comprising: a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide; and a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a Cas12a polypeptide fused to a second intein polypeptide.

A further aspect of the present invention is directed to a kit comprising: a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a first Cas12a polypeptide fused to a first intein polypeptide; and a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a second Cas12a polypeptide fused to a second intein polypeptide.

Another aspect of the present invention is directed to a kit comprising: a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide; and a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a Cas12a polypeptide fused to a second intein polypeptide.

An additional aspect of the present invention is directed to a method of modifying a target nucleic acid, the method comprising: introducing a first nucleic acid molecule and a second nucleic acid molecule into a cell, wherein the first nucleic acid molecule encodes a first fusion protein, the first fusion protein comprising a first Cas12a polypeptide fused to a first intein polypeptide and the second nucleic acid molecule encodes a second fusion protein, the second fusion protein comprising a second Cas12a polypeptide fused to a second intein polypeptide; contacting the target nucleic acid in the cell with a protein comprising at least a portion of the first Cas12a polypeptide and at least a portion of the second Cas12a polypeptide and a guide nucleic acid (e.g., a guide RNA), optionally wherein the protein and the guide nucleic acid form a complex or are comprised in a complex, thereby modifying the target nucleic acid.

Another aspect of the present invention is directed to a method of modifying a target nucleic acid, the method comprising: introducing a first nucleic acid molecule and a second nucleic acid molecule into a cell, wherein the first nucleic acid molecule encodes a first fusion protein, the first fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide and the second nucleic acid molecule encodes a second fusion protein, the second fusion protein comprising a Cas12a polypeptide fused to a second intein polypeptide; contacting the target nucleic acid in the cell with a protein comprising at least a portion of the polypeptide of interest and at least a portion of the Cas12a polypeptide and a guide nucleic acid (e.g., a guide RNA), optionally wherein the protein and the guide nucleic acid form a complex or are comprised in a complex, thereby modifying the target nucleic acid.

The invention further provides expression cassettes and/or vectors comprising a nucleic acid construct of the present invention, and cells comprising a polypeptide, fusion protein and/or nucleic acid construct of the present invention. Additionally, the invention provides kits comprising a nucleic acid construct of the present invention and expression cassettes, vectors and/or cells comprising the same.

It is noted that aspects of the invention described with respect to one embodiment, may be incorporated in a different embodiment although not specifically described relative thereto. That is, all embodiments and/or features of any embodiment can be combined in any way and/or combination. Applicant reserves the right to change any originally filed claim and/or file any new claim accordingly, including the right to be able to amend any originally filed claim to depend from and/or incorporate any feature of any other claim or claims although not originally claimed in that manner. These and other objects and/or aspects of the present invention are explained in detail in the specification set forth below. Further features, advantages and details of the present invention will be appreciated by those of ordinary skill in the art from a reading of the figures and the detailed description of the preferred embodiments that follow, such description being merely illustrative of the present invention.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a schematic illustration of two exemplary split proteins for mCherry.

FIG. 2 is a graph showing the average mCherry fluorescence for different proteins, which can measure the intein splicing activity for a mCherry reconstitution assay. (1): is the expression of wildtype full length mCherry. (2): is only expressing N-terminal mCherry fused

to N-terminal Npu split intein. (3): is only expressing C-terminal mCherry fused to C-terminal Npu split intein with the GEP mutation that enhances intein activity and robustness. (4) is only expressing C-terminal mCherry fused to wildtype (WT) C-terminal Npu split intein. (5) is expressing both N- and C-terminal mCherry parts, using the wildtype variant of the Npu intein. (6) is expressing both N- and C-terminal mCherry parts, using the GEP variant of the Npu intein.

FIG. 3 is a schematic illustration of two split proteins for Redraw Editor 2 (RE2) comprising Cas12a showing exemplary split sites according to some embodiments of the present invention.

FIG. 4 is a graph showing the effect of a two amino acid insertion (e.g., a CF or CA residues) on mCherry fluorescence.

FIG. 5 is a graph showing the percentage of indels generated using non-split and split Redraw editors using crRNAs according to some embodiments of the present invention.

FIG. 6 is a graph showing the percentage of indels generated using non-split and split Redraw editors using stagRNAs according to some embodiments of the present invention.

FIG. 7 is a graph showing the percentage of precise edits generated using non-split and split Redraw editors using stagRNAs according to some embodiments of the present invention.

DETAILED DESCRIPTION

The present invention now will be described hereinafter with reference to the accompanying drawings and examples, in which embodiments of the invention are shown. This description is not intended to be a detailed catalog of all the different ways in which the invention may be implemented, or all the features that may be added to the instant invention. For example, features illustrated with respect to one embodiment may be incorporated into other embodiments, and features illustrated with respect to a particular embodiment may be deleted from that embodiment. Thus, the invention contemplates that in some embodiments of the invention, any feature or combination of features set forth herein can be excluded or omitted. In addition, numerous variations and additions to the various embodiments suggested herein will be apparent to those skilled in the art in light of the instant disclosure, which do not depart from the instant invention. Hence, the following descriptions are intended to illustrate some particular embodiments of the invention, and not to exhaustively specify all permutations, combinations and variations thereof.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention

belongs. The terminology used in the description of the invention herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the invention.

All publications, patent applications, patents and other references cited herein are incorporated by reference in their entireties for the teachings relevant to the sentence and/or paragraph in which the reference is presented.

Unless the context indicates otherwise, it is specifically intended that the various features of the invention described herein can be used in any combination. Moreover, the present invention also contemplates that in some embodiments of the invention, any feature or combination of features set forth herein can be excluded or omitted. To illustrate, if the specification states that a composition comprises components A, B and C, it is specifically intended that any of A, B or C, or a combination thereof, can be omitted and disclaimed singularly or in any combination.

As used in the description of the invention and the appended claims, the singular forms "a," "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise.

Also as used herein, "and/or" refers to and encompasses any and all possible combinations of one or more of the associated listed items, as well as the lack of combinations when interpreted in the alternative ("or").

The term "about," as used herein when referring to a measurable value such as an amount or concentration and the like, is meant to encompass variations of $\pm 10\%$, $\pm 5\%$, $\pm 1\%$, $\pm 0.5\%$, or even $\pm 0.1\%$ of the specified value as well as the specified value. For example, "about X" where X is the measurable value, is meant to include X as well as variations of $\pm 10\%$, $\pm 5\%$, $\pm 1\%$, $\pm 0.5\%$, or even $\pm 0.1\%$ of X. A range provided herein for a measurable value may include any other range and/or individual value therein.

As used herein, phrases such as "between X and Y" and "between about X and Y" should be interpreted to include X and Y. As used herein, phrases such as "between about X and Y" mean "between about X and about Y" and phrases such as "from about X to Y" mean "from about X to about Y."

Recitation of ranges of values herein are merely intended to serve as a shorthand method of referring individually to each separate value falling within the range, unless otherwise indicated herein, and each separate value is incorporated into the specification as if it were individually recited herein. For example, if the range 10 to 15 is disclosed, then 11, 12, 13, and 14 are also disclosed.

The term "comprise," "comprises" and "comprising" as used herein, specify the presence of the stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

As used herein, the transitional phrase "consisting essentially of" means that the scope of a claim is to be interpreted to encompass the specified materials or steps recited in the claim and those that do not materially affect the basic and novel characteristic(s) of the claimed invention. Thus, the term "consisting essentially of" when used in a claim of this invention is not intended to be interpreted to be equivalent to "comprising."

As used herein, the terms "increase," "increasing," "enhance," "enhancing," "improve" and "improving" (and grammatical variations thereof) describe an elevation of at least about 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 100%, 150%, 200%, 300%, 400%, 500% or more such as compared to another measurable property or quantity (e.g., a control value).

As used herein, the terms "reduce," "reduced," "reducing," "reduction," "diminish," and "decrease" (and grammatical variations thereof), describe, for example, a decrease of at least about 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99%, or 100% such as compared to another measurable property or quantity (e.g., a control value). In some embodiments, the reduction can result in no or essentially no (*i.e.*, an insignificant amount, *e.g.*, less than about 10% or even 5%) detectable activity or amount.

A "heterologous nucleotide sequence" or a "recombinant nucleotide sequence" is a nucleotide sequence not naturally associated with a host cell into which it is introduced, including non-naturally occurring multiple copies of a naturally occurring nucleotide sequence.

A "native" or "wild-type" nucleic acid, nucleotide sequence, polypeptide or amino acid sequence refers to a naturally occurring or endogenous nucleic acid, nucleotide sequence, polypeptide or amino acid sequence. Thus, for example, a "native nucleic acid" is a nucleic acid that is naturally occurring in or endogenous to a reference organism. A "homologous" nucleic acid sequence is a nucleotide sequence naturally associated with a host cell into which it is introduced.

As used herein, the terms "nucleic acid," "nucleic acid molecule," "nucleotide sequence" and "polynucleotide" refer to RNA or DNA that is linear or branched, single or double stranded, or a hybrid thereof. The term also encompasses RNA/DNA hybrids. When dsRNA is produced synthetically, less common bases, such as inosine, 5-methylcytosine, 6-

methyladenine, hypoxanthine and others can also be used for antisense, dsRNA, and ribozyme pairing. For example, polynucleotides that contain C-5 propyne analogues of uridine and cytidine have been shown to bind RNA with high affinity and to be potent antisense inhibitors of gene expression. Other modifications, such as modification to the phosphodiester backbone, or the 2'-hydroxy in the ribose sugar group of the RNA can also be made.

As used herein, the term "nucleotide sequence" refers to a heteropolymer of nucleotides or the sequence of these nucleotides from the 5' to 3' end of a nucleic acid molecule and includes DNA or RNA molecules, including cDNA, a DNA fragment or portion, genomic DNA, synthetic (*e.g.*, chemically synthesized) DNA, plasmid DNA, mRNA, and anti-sense RNA, any of which can be single stranded or double stranded. The terms "nucleotide sequence" "nucleic acid," "nucleic acid molecule," "nucleic acid construct," "recombinant nucleic acid," "oligonucleotide" and "polynucleotide" are also used interchangeably herein to refer to a heteropolymer of nucleotides. Nucleic acid molecules and/or nucleotide sequences provided herein are presented herein in the 5' to 3' direction, from left to right and are represented using the standard code for representing the nucleotide characters as set forth in the U.S. sequence rules, 37 CFR §§1.821 - 1.825 and the World Intellectual Property Organization (WIPO) Standard ST.25. A "5' region" as used herein can mean the region of a polynucleotide that is nearest the 5' end of the polynucleotide. Thus, for example, an element in the 5' region of a polynucleotide can be located anywhere from the first nucleotide located at the 5' end of the polynucleotide to the nucleotide located halfway through the polynucleotide. A "3' region" as used herein can mean the region of a polynucleotide that is nearest the 3' end of the polynucleotide. Thus, for example, an element in the 3' region of a polynucleotide can be located anywhere from the first nucleotide located at the 3' end of the polynucleotide to the nucleotide located halfway through the polynucleotide.

As used herein, the term "gene" refers to a nucleic acid molecule capable of being used to produce mRNA, antisense RNA, miRNA, anti-microRNA antisense oligodeoxyribonucleotide (AMO) and the like. Genes may or may not be capable of being used to produce a functional protein or gene product. Genes can include both coding and non-coding regions (*e.g.*, introns, regulatory elements, promoters, enhancers, termination sequences and/or 5' and 3' untranslated regions).

A polynucleotide, gene, or polypeptide may be "isolated" by which is meant a nucleic acid or polypeptide that is substantially or essentially free from components normally found in association with the nucleic acid or polypeptide, respectively, in its natural state. In some embodiments, such components include other cellular material, culture medium from

recombinant production, and/or various chemicals used in chemically synthesizing the nucleic acid or polypeptide.

The term "mutation" refers to point mutations (e.g., missense, or nonsense, or insertions or deletions of single base pairs that result in frame shifts), insertions, deletions, and/or truncations. When the mutation is a substitution of a residue within an amino acid sequence with another residue, or a deletion or insertion of one or more residues within a sequence, the mutations are typically described by identifying the original residue followed by the position of the residue within the sequence and by the identity of the newly substituted residue.

The terms "complementary" or "complementarity," as used herein, refer to the natural binding of polynucleotides under permissive salt and temperature conditions by base-pairing. For example, the sequence "A-G-T" (5' to 3') binds to the complementary sequence "T-C-A" (3' to 5'). Complementarity between two single-stranded molecules may be "partial," in which only some of the nucleotides bind, or it may be complete when total complementarity exists between the single stranded molecules. The degree of complementarity between nucleic acid strands has significant effects on the efficiency and strength of hybridization between nucleic acid strands.

"Complement" as used herein can mean 100% complementarity with the comparator nucleotide sequence or it can mean less than 100% complementarity (e.g., "substantially complementary," such as about 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%, and the like, complementarity).

A "portion" or "fragment" of a nucleotide sequence or polypeptide (including a domain) will be understood to mean a nucleotide sequence or polypeptide of reduced length (e.g., reduced by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more residue(s) (e.g., nucleotide(s) or peptide(s)) relative to a reference nucleotide sequence or polypeptide, respectively, and comprising, consisting essentially of and/or consisting of a nucleotide sequence or polypeptide of contiguous residues, respectively, identical or almost identical (e.g., 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% identical) to the reference nucleotide sequence or polypeptide. In some embodiments, a portion of a reference nucleotide sequence or polypeptide is about 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99%, or more of the full-length reference nucleotide sequence or polypeptide. Such a nucleic acid fragment or portion according to the invention may be, where appropriate, included in a larger

polynucleotide of which it is a constituent. As an example, a repeat sequence of guide nucleic acid of this invention may comprise a portion of a wild-type CRISPR-Cas repeat sequence (e.g., a wild-type Type V CRISPR Cas repeat, e.g., a repeat from the CRISPR Cas system that includes, but is not limited to, Cas12a (Cpf1), Cas12b, Cas12c (C2c3), Cas12d (CasY), Cas12e (CasX), Cas12g, Cas12h, Cas12i, C2c1, C2c4, C2c5, C2c8, C2c9, C2c10, Cas14a, Cas14b, and/or Cas14c, and the like). Similarly a portion of a polypeptide may be included in a larger polypeptide of which it is a constituent.

Different nucleic acids or proteins having homology are referred to herein as "homologues." The term homologue includes homologous sequences from the same and other species and orthologous sequences from the same and other species. "Homology" refers to the level of similarity between two or more nucleic acid and/or amino acid sequences in terms of percent of positional identity (*i.e.*, sequence similarity or identity). Homology also refers to the concept of similar functional properties among different nucleic acids or proteins. Thus, the compositions and methods of the invention further comprise homologues to the nucleotide sequences and polypeptides of this invention. "Orthologous" and "orthologs" as used herein, refers to homologous nucleotide sequences and/ or amino acid sequences in different species that arose from a common ancestral gene during speciation. A homologue or ortholog of a nucleotide sequence of this invention has a substantial sequence identity (e.g., at least about 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%, 99.5% or 100%) to said nucleotide sequence of the invention.

As used herein "sequence identity" refers to the extent to which two optimally aligned polynucleotide or polypeptide sequences are invariant throughout a window of alignment of components, *e.g.*, nucleotides or amino acids. "Identity" can be readily calculated by known methods including, but not limited to, those described in: *Computational Molecular Biology* (Lesk, A. M., ed.) Oxford University Press, New York (1988); *Biocomputing: Informatics and Genome Projects* (Smith, D. W., ed.) Academic Press, New York (1993); *Computer Analysis of Sequence Data, Part I* (Griffin, A. M., and Griffin, H. G., eds.) Humana Press, New Jersey (1994); *Sequence Analysis in Molecular Biology* (von Heinje, G., ed.) Academic Press (1987); and *Sequence Analysis Primer* (Gribskov, M. and Devereux, J., eds.) Stockton Press, New York (1991).

As used herein, the term "percent sequence identity" or "percent identity" refers to the percentage of identical nucleotides in a linear polynucleotide sequence of a reference ("query") polynucleotide molecule (or its complementary strand) as compared to a test ("subject")

polynucleotide molecule (or its complementary strand) when the two sequences are optimally aligned. In some embodiments, "percent identity" can refer to the percentage of identical amino acids in an amino acid sequence as compared to a reference polypeptide.

As used herein, the phrase "substantially identical," or "substantial identity" in the context of two nucleic acid molecules, nucleotide sequences or protein sequences, refers to two or more sequences or subsequences that have at least about 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%, 99.5% or 100% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. In some embodiments of the invention, the substantial identity exists over a region of consecutive nucleotides of a nucleotide sequence of the invention that is about 10 nucleotides to about 20 nucleotides, about 10 nucleotides to about 25 nucleotides, about 10 nucleotides to about 30 nucleotides, about 15 nucleotides to about 25 nucleotides, about 30 nucleotides to about 40 nucleotides, about 50 nucleotides to about 60 nucleotides, about 70 nucleotides to about 80 nucleotides, about 90 nucleotides to about 100 nucleotides, or more nucleotides in length, and any range therein, up to the full length of the sequence. In some embodiments, the nucleotide sequences can be substantially identical over at least about 20 nucleotides (e.g., about 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40 nucleotides). In some embodiments, a substantially identical nucleotide or protein sequence performs substantially the same function as the nucleotide (or encoded protein sequence) to which it is substantially identical.

For sequence comparison, typically one sequence acts as a reference sequence to which one or more test sequence(s) are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for aligning a comparison window are well known to those skilled in the art and may be conducted by tools such as the local homology algorithm of Smith and Waterman, the homology alignment algorithm of Needleman and Wunsch, the search for similarity method of Pearson and Lipman, and optionally by computerized implementations of these algorithms such as GAP, BESTFIT, FASTA, and TFASTA available as part of the GCG® Wisconsin Package® (Accelrys Inc., San Diego, CA) as well as web-based alignment programs such as Clustal Omega, EMBOSS Needle, EMBOSS Stretcher,

EMBOSS Water, LALIGN, GGSEARCH2SEQ, EMBOS Cons, Kalign, MAFFT, MUSCLE, and T-Coffee. In some embodiments, an "optimal alignment" of two sequences (e.g., two polypeptide sequences) is the highest scoring alignment, optionally from an alignment conducted by a tool such as the local homology algorithm of Smith and Waterman, the homology alignment algorithm of Needleman and Wunsch, the search for similarity method of Pearson and Lipman, GAP, BESTFIT, FASTA, and TFASTA available as part of the GCG® Wisconsin Package® (Accelrys Inc., San Diego, CA), Clustal Omega, EMBOSS Needle, EMBOSS Stretcher, EMBOSS Water, LALIGN, GGSEARCH2SEQ, EMBOS Cons, Kalign, MAFFT, MUSCLE, and/or T-Coffee. In some embodiments, an "optimal alignment" of two sequences (e.g., two polypeptide sequences) is the alignment that provides the highest percent sequence identity, optionally allowing for one or more gap(s) to be introduced into one or both sequences. An "identity fraction" for aligned segments of a test sequence and a reference sequence is the number of identical components which are shared by the two aligned sequences divided by the total number of components in the reference sequence segment, e.g., the entire reference sequence or a smaller defined part of the reference sequence. Percent sequence identity is represented as the identity fraction multiplied by 100. The comparison of one or more sequence(s) may be to a full-length sequence or a portion thereof, or to a longer sequence. For purposes of this invention "percent identity" and/or optimal alignment may be determined using Basic Local Alignment Search Tool (BLAST) provided by the National Center for Biotechnology Information such as BLASTX, for translated nucleotide sequences, BLASTN for polynucleotide sequences, and BLASTP for polypeptide sequences.

Two nucleotide sequences may also be considered substantially complementary when the two sequences hybridize to each other under stringent conditions. In some representative embodiments, two nucleotide sequences considered to be substantially complementary hybridize to each other under highly stringent conditions.

"Stringent hybridization conditions" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and Northern hybridizations are sequence dependent, and are different under different environmental parameters. An extensive guide to the hybridization of nucleic acids is found in Tijssen *Laboratory Techniques in Biochemistry and Molecular Biology-Hybridization with Nucleic Acid Probes* part I chapter 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays" Elsevier, New York (1993). Generally, highly stringent hybridization and wash conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH.

The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe. An example of stringent hybridization conditions for hybridization of complementary nucleotide sequences which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formamide with 1 mg of heparin at 42°C, with the hybridization being carried out overnight. An example of highly stringent wash conditions is 0.1 5M NaCl at 72°C for about 15 minutes. An example of stringent wash conditions is a 0.2x SSC wash at 65°C for 15 minutes (*see*, Sambrook, *infra*, for a description of SSC buffer). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example of a medium stringency wash for a duplex of, e.g., more than 100 nucleotides, is 1x SSC at 45°C for 15 minutes. An example of a low stringency wash for a duplex of, e.g., more than 100 nucleotides, is 4-6x SSC at 40°C for 15 minutes. For short probes (e.g., about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.0 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically at least about 30°C. Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization. Nucleotide sequences that do not hybridize to each other under stringent conditions are still substantially identical if the proteins that they encode are substantially identical. This can occur, for example, when a copy of a nucleotide sequence is created using the maximum codon degeneracy permitted by the genetic code.

A polynucleotide and/or recombinant nucleic acid construct of this invention can be codon optimized for expression. In some embodiments, a polynucleotide, nucleic acid construct, expression cassette, and/or vector of the present invention (e.g., that comprises/encodes a fusion protein, a nucleic acid binding polypeptide (e.g., a DNA binding polypeptide such as a sequence-specific DNA binding domain from a polynucleotide-guided endonuclease, a zinc finger nuclease, a transcription activator-like effector nuclease (TALEN), an Argonaute protein, and/or a CRISPR-Cas effector protein), a guide nucleic acid, a cytosine deaminase and/or adenine deaminase) may be codon optimized for expression in an organism (e.g., an animal such as a human, a plant, a fungus, an archaeon, or a bacterium). In some embodiments, the codon optimized nucleic acid constructs, polynucleotides, expression cassettes, and/or vectors of the invention have about 70% to about 99.9% (e.g., 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%, 99.5%, 99.9% or 100%) identity or more to the reference nucleic acid constructs, polynucleotides, expression cassettes, and/or vectors but which have not been codon optimized.

In any of the embodiments described herein, a polynucleotide or nucleic acid construct of the invention may be operatively associated with a variety of promoters and/or other regulatory elements for expression in an organism or cell thereof (e.g., a mammal and/or a mammalian cell, a plant and/or a cell of a plant, etc.). Thus, in some embodiments, a polynucleotide or nucleic acid construct of this invention may further comprise one or more promoters, introns, enhancers, and/or terminators operably linked to one or more nucleotide sequences. In some embodiments, a promoter may be operably associated with an intron (e.g., Ubi1 promoter and intron). In some embodiments, a promoter associated with an intron may be referred to as a "promoter region" (e.g., Ubi1 promoter and intron).

By "operably linked" or "operably associated" as used herein in reference to polynucleotides, it is meant that the indicated elements are functionally related to each other, and are also generally physically related. Thus, the term "operably linked" or "operably associated" as used herein, refers to nucleotide sequences on a single nucleic acid molecule that are functionally associated. Thus, a first nucleotide sequence that is operably linked to a second nucleotide sequence means a situation when the first nucleotide sequence is placed in a functional relationship with the second nucleotide sequence. For instance, a promoter is operably associated with a nucleotide sequence if the promoter effects the transcription or expression of said nucleotide sequence. Those skilled in the art will appreciate that the control sequences (e.g., promoter) need not be contiguous with the nucleotide sequence to which it is operably associated, as long as the control sequences function to direct the expression thereof. Thus, for example, intervening untranslated, yet transcribed, nucleic acid sequences can be present between a promoter and the nucleotide sequence, and the promoter can still be considered "operably linked" to the nucleotide sequence.

As used herein, the term "linked," or "fused" in reference to polypeptides, refers to the covalent attachment of one polypeptide to another. A polypeptide may be linked or fused to another polypeptide (e.g., at the N-terminus or the C-terminus) directly (e.g., via a peptide bond) or through a linker (e.g., a peptide linker). Two polypeptides being directly fused (e.g., a direct linkage) refers to the covalent attachment of one amino acid residue of a first polypeptide of the two polypeptides to an amino acid residue of a second polypeptide of the two polypeptides without an intervening element between the two amino acid residues. For example, first and second polypeptides may be directly linked via a peptide bond between the

first and second polypeptides without an intervening element (e.g., a linker) between the first and second polypeptides. Two polypeptides being indirectly fused (e.g., an indirect linkage) refers to an intervening element (e.g., a linker such as a peptide linker) that is present between the two polypeptides and is covalently attached to each, optionally the intervening element may attach one end of a first polypeptide of the two polypeptides to an end of the second polypeptide of the two polypeptides.

A “fusion protein” as used herein refers to two or more polypeptides that are covalently attached (e.g., directly or indirectly) so that they are transcribed and translated as a single unit and thereby produce a single polypeptide comprising the two or more polypeptides. In some embodiments, the two or more polypeptides may naturally be encoded by separate genes, but, in the form of a fusion protein, are encoded by a single gene.

The term “linker” is art-recognized and refers to a chemical group or a molecule linking two molecules or moieties, e.g., linking two polypeptides or domains of a fusion protein, such as, for example, a Cas12a polypeptide and an intein polypeptide. A linker may be comprised of a single linking molecule (e.g., a single amino acid) or may comprise more than one linking molecule. In some embodiments, the linker can be an organic molecule, group, polymer, or chemical moiety such as a bivalent organic moiety. In some embodiments, the linker may be an amino acid or it may be a peptide. In some embodiments, the linker is a peptide (e.g., a peptide linker).

In some embodiments, a peptide linker useful with this invention may be about 2 to about 100 or more amino acids in length, for example, about 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 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, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or more amino acids in length (e.g., about 2 to about 40, about 2 to about 50, about 2 to about 60, about 4 to about 40, about 4 to about 50, about 4 to about 60, about 5 to about 40, about 5 to about 50, about 5 to about 60, about 9 to about 40, about 9 to about 50, about 9 to about 60, about 10 to about 40, about 10 to about 50, about 10 to about 60, or about 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 amino acids to about 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, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or more amino acids in length (e.g., about 105, 110, 115, 120, 130, 140 150 or more amino acids in length)). In some embodiments, a peptide linker

may comprise glycine (G) and serine (S) such as a GS linker. In some embodiments, a peptide linker may comprise a cysteine (C) and alanine (A) such as a CA linker. In some embodiments, a peptide linker may comprise a cysteine (C) and phenylalanine (F) such as a CF linker. In some embodiments, the peptide linker is a GS linker, a CA linker, or a CF linker having 2, 3, or 4 amino acid residues, optionally 2 or 4 amino acid residues. In some embodiments, the peptide linker has one of the amino acid sequences of **SEQ ID NOs:1-35**. In some embodiments, the peptide linker may comprise an amino acid sequence of CA, CF, (GGS)_n, GS, SG, GSSG (**SEQ ID NO:31**), GSSGSS (**SEQ ID NO:32**), GSSGSSGS (**SEQ ID NO:33**), (GSS)_n (**SEQ ID NO:34**), (GSS)_nGS (**SEQ ID NO:35**), S(GGS)_n (**SEQ ID NO:25**), SGGS (**SEQ ID NO:26**), (GSS)_nG (**SEQ ID NO:191**), or (GGGGS)_n (**SEQ ID NO:27**), wherein n is an integer of 1-20 (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20). In some embodiments, the peptide linker may comprise the amino acid sequence: SGGSGGSGGS (**SEQ ID NO:28**). In some embodiments, the peptide linker may comprise the amino acid sequence: SGSETPGTSESATPES (**SEQ ID NO:29**), also referred to as the XTEN linker. In some embodiments, the peptide linker may comprise the amino acid sequence: SGGSSGGSSGSETPGTSESATPESSGGSSGGS (**SEQ ID NO:30**), also referred to as the GS-XTEN-GS linker. In some embodiments, a peptide linker has an amino acid sequence of **SEQ ID NOs:189** or **SEQ ID NOs:190**.

As used herein, the term "linked," or "fused" in reference to polynucleotides, refers to the covalent attachment of one polynucleotide to another polynucleotide. In some embodiments, two or more polynucleotide molecules may be linked by a linker that can be an organic molecule, group, polymer, or chemical moiety such as a bivalent organic moiety. A polynucleotide may be linked or fused to another polynucleotide (at the 5' end or the 3' end) via a direct covalent linkage or through one or more linking nucleotides. In some embodiments, a polynucleotide motif of a certain structure may be inserted within another polynucleotide sequence (e.g., extension of the hairpin structure in guide RNA). In some embodiments, the linking nucleotides may be naturally occurring nucleotides. In some embodiments, the linking nucleotides may be non-naturally occurring nucleotides. Two polynucleotides being directly fused (e.g., a direct linkage) refers to the covalent attachment of one nucleotide of a first polynucleotide of the two polynucleotides to a nucleotide of a second polynucleotide of the two polynucleotides without an intervening element between the two polynucleotides. For example, first and second polynucleotides may be directly linked via a phosphodiester bond between the first and second polynucleotides without an intervening element (e.g., a linker) between the first and second polynucleotides. Two polynucleotides being indirectly fused

(e.g., an indirect linkage) refers to an intervening element (e.g., a linker such as a polynucleotide linker) that is present between the two polynucleotides and is covalently attached to each, optionally the intervening element attaches one end of a first polynucleotide of the two polynucleotides to an end of the second polynucleotide of the two polynucleotides.

A "promoter" is a nucleotide sequence that controls or regulates the transcription of a nucleotide sequence (e.g., a coding sequence) that is operably associated with the promoter. The coding sequence controlled or regulated by a promoter may encode a polypeptide and/or a functional RNA. Typically, a "promoter" refers to a nucleotide sequence that contains a binding site for RNA polymerase II and directs the initiation of transcription. In general, promoters are found 5', or upstream, relative to the start of the coding region of the corresponding coding sequence. A promoter may comprise other elements that act as regulators of gene expression; e.g., a promoter region. These include a TATA box consensus sequence, and often a CAAT box consensus sequence (Breathnach and Chambon, (1981) *Annu. Rev. Biochem.* 50:349). In plants, the CAAT box may be substituted by the AGGA box (Messing *et al.*, (1983) in *Genetic Engineering of Plants*, T. Kosuge, C. Meredith and A. Hollaender (eds.), Plenum Press, pp. 211-227). In some embodiments, a promoter region may comprise at least one intron (e.g., **SEQ ID NO:36** or **SEQ ID NO:37**).

Promoters useful with this invention can include, for example, constitutive, inducible, temporally regulated, developmentally regulated, chemically regulated, tissue-preferred and/or tissue-specific promoters for use in the preparation of recombinant nucleic acid molecules, e.g., "synthetic nucleic acid constructs" or "protein-RNA complex." These various types of promoters are known in the art.

The choice of promoter may vary depending on the temporal and spatial requirements for expression, and also may vary based on the host cell to be transformed. Promoters for many different organisms are well known in the art. Based on the extensive knowledge present in the art, the appropriate promoter can be selected for the particular host organism of interest. Thus, for example, much is known about promoters upstream of highly constitutively expressed genes in model organisms and such knowledge can be readily accessed and implemented in other systems as appropriate.

In some embodiments, a promoter functional in a plant may be used with the constructs of this invention. Non-limiting examples of a promoter useful for driving expression in a plant include the promoter of the RubisCo small subunit gene 1 (PrbcS1), the promoter of the actin gene (Pactin), the promoter of the nitrate reductase gene (Pnr) and the promoter of duplicated carbonic anhydrase gene 1 (Pdca1) (See, Walker et al. *Plant Cell Rep.* 23:727-735 (2005); Li

et al. *Gene* 403:132-142 (2007); Li et al. *Mol Biol. Rep.* 37:1143-1154 (2010)). PrbcS1 and Pactin are constitutive promoters and Pnr and Pdca1 are inducible promoters. Pnr is induced by nitrate and repressed by ammonium (Li et al. *Gene* 403:132-142 (2007)) and Pdca1 is induced by salt (Li et al. *Mol Biol. Rep.* 37:1143-1154 (2010)). In some embodiments, a promoter useful with this invention is RNA polymerase II (Pol II) promoter. In some embodiments, a U6 promoter or a 7SL promoter from *Zea mays* may be useful with constructs of this invention. In some embodiments, the U6c promoter and/or 7SL promoter from *Zea mays* may be useful for driving expression of a guide nucleic acid. In some embodiments, a U6c promoter, U6i promoter and/or 7SL promoter from *Glycine max* may be useful with constructs of this invention. In some embodiments, the U6c promoter, U6i promoter and/or 7SL promoter from *Glycine max* may be useful for driving expression of a guide nucleic acid.

Examples of constitutive promoters useful for plants include, but are not limited to, cestrum virus promoter (cmp) (U.S. Patent No. 7,166,770), the rice actin 1 promoter (Wang *et al.* (1992) *Mol. Cell. Biol.* 12:3399-3406; as well as US Patent No. 5,641,876), CaMV 35S promoter (Odell *et al.* (1985) *Nature* 313:810-812), CaMV 19S promoter (Lawton *et al.* (1987) *Plant Mol. Biol.* 9:315-324), nos promoter (Ebert *et al.* (1987) *Proc. Natl. Acad. Sci. USA* 84:5745-5749), Adh promoter (Walker *et al.* (1987) *Proc. Natl. Acad. Sci. USA* 84:6624-6629), sucrose synthase promoter (Yang & Russell (1990) *Proc. Natl. Acad. Sci. USA* 87:4144-4148), and the ubiquitin promoter. The constitutive promoter derived from ubiquitin accumulates in many cell types. Ubiquitin promoters have been cloned from several plant species for use in transgenic plants, for example, sunflower (Binet *et al.*, 1991. *Plant Science* 79: 87-94), maize (Christensen *et al.*, 1989. *Plant Molec. Biol.* 12: 619-632), and arabidopsis (Norris et al. 1993. *Plant Molec. Biol.* 21:895-906). The maize ubiquitin promoter (*UbiP*) has been developed in transgenic monocot systems and its sequence and vectors constructed for monocot transformation are disclosed in the European patent publication EP0342926. The ubiquitin promoter is suitable for the expression of the nucleotide sequences of the invention in transgenic plants, especially monocotyledons. Further, the promoter expression cassettes described by McElroy *et al.* (*Mol. Gen. Genet.* 231: 150-160 (1991)) can be easily modified for the expression of the nucleotide sequences of the invention and are particularly suitable for use in monocotyledonous hosts.

In some embodiments, tissue specific/tissue preferred promoters can be used for expression of a heterologous polynucleotide in a plant cell. Tissue specific or preferred expression patterns include, but are not limited to, green tissue specific or preferred, root specific or preferred, stem specific or preferred, flower specific or preferred or pollen specific

or preferred. Promoters suitable for expression in green tissue include many that regulate genes involved in photosynthesis and many of these have been cloned from both monocotyledons and dicotyledons. In one embodiment, a promoter useful with the invention is the maize PEPC promoter from the phosphoenol carboxylase gene (Hudspeth & Grula, *Plant Molec. Biol.* 12:579-589 (1989)). Non-limiting examples of tissue-specific promoters include those associated with genes encoding the seed storage proteins (such as β -conglycinin, cruciferin, napin and phaseolin), zein or oil body proteins (such as oleosin), or proteins involved in fatty acid biosynthesis (including acyl carrier protein, stearyl-ACP desaturase and fatty acid desaturases (fad 2-1)), and other nucleic acids expressed during embryo development (such as Bce4, *see, e.g.*, Kridl *et al.* (1991) *Seed Sci. Res.* 1:209-219; as well as EP Patent No. 255378). Tissue-specific or tissue-preferential promoters useful for the expression of the nucleotide sequences of the invention in plants, particularly maize, include but are not limited to those that direct expression in root, pith, leaf or pollen. Such promoters are disclosed, for example, in WO 93/07278, incorporated by reference herein for its disclosure of promoters. Other non-limiting examples of tissue specific or tissue preferred promoters useful with the invention the cotton rubisco promoter disclosed in US Patent 6,040,504; the rice sucrose synthase promoter disclosed in US Patent 5,604,121; the root specific promoter described by de Framond (FEBS 290:103-106 (1991); European patent EP 0452269 to Ciba-Geigy); the stem specific promoter described in U.S. Patent 5,625,136 (to Ciba-Geigy) and which drives expression of the maize *trpA* gene; the cestrum yellow leaf curling virus promoter disclosed in WO 01/73087; and pollen specific or preferred promoters including, but not limited to, ProOsLPS10 and ProOsLPS11 from rice (Nguyen *et al.* *Plant Biotechnol. Reports* 9(5):297-306 (2015)), ZmSTK2_USP from maize (Wang *et al.* *Genome* 60(6):485-495 (2017)), LAT52 and LAT59 from tomato (Twell *et al.* *Development* 109(3):705-713 (1990)), Zm13 (U.S. Patent No. 10,421,972), PLA2- δ promoter from arabidopsis (U.S. Patent No. 7,141,424), and/or the ZmC5 promoter from maize (International PCT Publication No. WO 1999/042587).

Additional examples of plant tissue-specific/tissue preferred promoters include, but are not limited to, the root hair-specific *cis*-elements (RHEs) (KIM ET AL. *The Plant Cell* 18:2958-2970 (2006)), the root-specific promoters *RCc3* (Jeong *et al.* *Plant Physiol.* 153:185-197 (2010)) and RB7 (U.S. Patent No. 5459252), the lectin promoter (Lindstrom *et al.* (1990) *Der. Genet.* 11:160-167; and Vodkin (1983) *Prog. Clin. Biol. Res.* 138:87-98), corn alcohol dehydrogenase 1 promoter (Dennis *et al.* (1984) *Nucleic Acids Res.* 12:3983-4000), S-adenosyl-L-methionine synthetase (SAMS) (Vander Mijnsbrugge *et al.* (1996) *Plant and Cell Physiology*, 37(8):1108-1115), corn light harvesting complex promoter (Bansal *et al.* (1992)

Proc. Natl. Acad. Sci. USA 89:3654-3658), corn heat shock protein promoter (O'Dell *et al.* (1985) *EMBO J.* 5:451-458; and Rochester *et al.* (1986) *EMBO J.* 5:451-458), pea small subunit RuBP carboxylase promoter (Cashmore, "Nuclear genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase" pp. 29-39 In: *Genetic Engineering of Plants* (Hollaender ed., Plenum Press 1983; and Poulsen *et al.* (1986) *Mol. Gen. Genet.* 205:193-200), Ti plasmid mannopine synthase promoter (Langridge *et al.* (1989) *Proc. Natl. Acad. Sci. USA* 86:3219-3223), Ti plasmid nopaline synthase promoter (Langridge *et al.* (1989), *supra*), petunia chalcone isomerase promoter (van Tunen *et al.* (1988) *EMBO J.* 7:1257-1263), bean glycine rich protein 1 promoter (Keller *et al.* (1989) *Genes Dev.* 3:1639-1646), truncated CaMV 35S promoter (O'Dell *et al.* (1985) *Nature* 313:810-812), potato patatin promoter (Wenzler *et al.* (1989) *Plant Mol. Biol.* 13:347-354), root cell promoter (Yamamoto *et al.* (1990) *Nucleic Acids Res.* 18:7449), maize zein promoter (Kriz *et al.* (1987) *Mol. Gen. Genet.* 207:90-98; Langridge *et al.* (1983) *Cell* 34:1015-1022; Reina *et al.* (1990) *Nucleic Acids Res.* 18:6425; Reina *et al.* (1990) *Nucleic Acids Res.* 18:7449; and Wandelt *et al.* (1989) *Nucleic Acids Res.* 17:2354), globulin-1 promoter (Belanger *et al.* (1991) *Genetics* 129:863-872), α -tubulin cab promoter (Sullivan *et al.* (1989) *Mol. Gen. Genet.* 215:431-440), PEPCase promoter (Hudspeth & Grula (1989) *Plant Mol. Biol.* 12:579-589), R gene complex-associated promoters (Chandler *et al.* (1989) *Plant Cell* 1:1175-1183), and chalcone synthase promoters (Franken *et al.* (1991) *EMBO J.* 10:2605-2612).

Useful for seed-specific expression is the pea vicilin promoter (Czako *et al.* (1992) *Mol. Gen. Genet.* 235:33-40; as well as the seed-specific promoters disclosed in U.S. Patent No. 5,625,136. Useful promoters for expression in mature leaves are those that are switched at the onset of senescence, such as the SAG promoter from *Arabidopsis* (Gan *et al.* (1995) *Science* 270:1986-1988).

In addition, promoters functional in chloroplasts can be used. Non-limiting examples of such promoters include the bacteriophage T3 gene 9 5' UTR and other promoters disclosed in U.S. Patent No. 7,579,516. Other promoters useful with the invention include but are not limited to the S-E9 small subunit RuBP carboxylase promoter and the Kunitz trypsin inhibitor gene promoter (Kti3).

Additional regulatory elements useful with this invention include, but are not limited to, introns, enhancers, termination sequences and/or 5' and 3' untranslated regions.

An intron useful with this invention can be an intron identified in and isolated from a plant and then inserted into an expression cassette to be used in transformation of a plant. As would be understood by those of skill in the art, introns can comprise the sequences required

for self-excision and are incorporated into nucleic acid constructs/expression cassettes in frame. An intron can be used either as a spacer to separate multiple protein-coding sequences in one nucleic acid construct, or an intron can be used inside one protein-coding sequence to, for example, stabilize the mRNA. If they are used within a protein-coding sequence, they are inserted "in-frame" with the excision sites included. Introns may also be associated with promoters to improve or modify expression. As an example, a promoter/intron combination useful with this invention includes but is not limited to that of the maize Ubi1 promoter and intron.

Non-limiting examples of introns useful with the present invention include introns from the ADHI gene (e.g., *Adh1-S* introns 1, 2 and 6), the ubiquitin gene (Ubi1), the RuBisCO small subunit (*rbcS*) gene, the RuBisCO large subunit (*rbcL*) gene, the actin gene (e.g., *actin-1* intron), the pyruvate dehydrogenase kinase gene (*pdh*), the nitrate reductase gene (*nr*), the duplicated carbonic anhydrase gene 1 (*Tdca1*), the *psbA* gene, the *atpA* gene, or any combination thereof.

An "editing system" as used herein refers to any site-specific (e.g., sequence-specific) nucleic acid editing system now known or later developed, which system can introduce a modification (e.g., a mutation) in a nucleic acid in target specific manner. For example, an editing system (e.g., a site- and/or sequence-specific editing system) can include, but is not limited to, a CRISPR-Cas editing system, a meganuclease editing system, a zinc finger nuclease (ZFN) editing system, a transcription activator-like effector nuclease (TALEN) editing system, a base editing system and/or a prime editing system, each of which may comprise one or more polypeptide(s) and/or one or more polynucleotide(s) that when present and/or expressed together (e.g., as a system) in a composition and/or cell can modify (e.g., mutate) a target nucleic acid in a sequence specific manner. In some embodiments, an editing system (e.g., a site- and/or sequence-specific editing system) can comprise one or more polynucleotide(s) and/or one or more polypeptide(s), including but not limited to a nucleic acid binding polypeptide (e.g., a DNA binding domain), a nuclease, another polypeptide, and/or a polynucleotide. In some embodiments, a CRISPR-Cas editing system is provided, wherein a fusion protein of the present invention is used to provide a Cas12a protein of the CRISPR-Cas editing system. An editing system of the present invention may modify a target nucleic acid that is present in a cell or outside a cell (e.g., a method of the present invention may be carried out *in vitro*, *ex vivo*, and/or *in vivo*).

In some embodiments, an editing system comprises one or more sequence-specific nucleic acid binding polypeptide(s) (e.g., a DNA binding domain) that can be from, for

example, a polynucleotide-guided endonuclease, a CRISPR-Cas endonuclease (e.g., CRISPR-Cas effector protein), a zinc finger nuclease, a transcription activator-like effector nuclease (TALEN) and/or an Argonaute protein. In some embodiments, an editing system comprises one or more cleavage polypeptide(s) (e.g., nucleases) including, but not limited to, an endonuclease (e.g., FokI), a polynucleotide-guided endonuclease, a CRISPR-Cas endonuclease (e.g., CRISPR-Cas effector protein), a zinc finger nuclease, and/or a transcription activator-like effector nuclease (TALEN).

A "nucleic acid binding polypeptide" as used herein refers to a polypeptide or domain that binds and/or is capable of binding a nucleic acid (e.g., a target nucleic acid). A DNA binding domain or DNA binding polypeptide is an exemplary nucleic acid binding polypeptide and may be a site- and/or sequence-specific nucleic acid binding domain. In some embodiments, a nucleic acid binding polypeptide may be a sequence-specific nucleic acid binding polypeptide such as, but not limited to, a sequence-specific binding domain from, for example, a polynucleotide-guided endonuclease, a CRISPR-Cas effector protein (e.g., a CRISPR-Cas endonuclease), a zinc finger nuclease, a transcription activator-like effector nuclease (TALEN) and/or an Argonaute protein. In some embodiments, a nucleic acid binding polypeptide comprises a cleavage domain (e.g., a nuclease domain) such as, but not limited to, an endonuclease (e.g., FokI), a polynucleotide-guided endonuclease, a CRISPR-Cas endonuclease, a zinc finger nuclease, and/or a transcription activator-like effector nuclease (TALEN). In some embodiments, the nucleic acid binding polypeptide associates with and/or is capable of associating with (e.g., forms a complex with) one or more nucleic acid molecule(s) (e.g., forms a complex with a guide nucleic acid as described herein), which may direct and/or guide the nucleic acid binding polypeptide to a specific target nucleotide sequence (e.g., a gene locus of a genome) that is complementary to the one or more nucleic acid molecule(s) (or a portion or region thereof), thereby causing the nucleic acid binding polypeptide to bind to the nucleotide sequence at the specific target site. In some embodiments, the nucleic acid binding polypeptide is a CRISPR-Cas effector protein as described herein.

In some embodiments, an editing system comprises or is a ribonucleoprotein such as an assembled ribonucleoprotein complex (e.g., a ribonucleoprotein that comprises a CRISPR-Cas effector protein, a guide nucleic acid, and optionally a deaminase). In some embodiments, a ribonucleoprotein of an editing system may be assembled together (e.g., a pre-assembled ribonucleoprotein including a CRISPR-Cas effector protein, a guide nucleic acid, and optionally a deaminase) such as when contacted to a target nucleic acid or when introduced into a cell (e.g., a mammalian cell or a plant cell) (e.g., at the time of contacting the components

of the ribonucleoprotein to a target nucleic acid and/or at the time of introducing the components of the ribonucleoprotein into a cell). In some embodiments, a ribonucleoprotein of an editing system may assemble into a complex (e.g., a non-covalently bound complex) while a portion of the ribonucleoprotein is contacting a target nucleic acid and/or may assemble after and/or during introduction into a plant cell. In some embodiments, an editing system may be assembled (e.g., into a non-covalently bound complex) when introduced into a plant cell. In some embodiments, a ribonucleoprotein may comprise a fusion protein of the present invention, a guide nucleic acid, and optionally a deaminase. In some embodiments, a ribonucleoprotein of an editing system may be contacted to a target nucleic acid and/or may be introduced into a plant cell. In some embodiments, an editing system may be assembled (e.g., into a non-covalently bound complex) when introduced into a plant cell. In some embodiments, a ribonucleoprotein may comprise a protein of the present invention (e.g., a protein prepared using a composition and/or method of the present invention), a guide nucleic acid, and optionally a deaminase and/or reverse transcriptase. In some embodiments, a protein of the present invention comprises a CRISPR-Cas effector protein and the protein is used in place of (e.g., substituted for) a CRISPR-Cas effector protein (e.g., in a composition, complex, kit, method, and/or system such as an editing system described herein) and/or functions as a CRISPR-Cas effector protein, templated editor, and/or base editor, optionally in a composition, complex, ribonucleoprotein, kit, method, system, and/or editing system of the present invention.

The terms “transgene” or “transgenic” as used herein refer to at least one nucleic acid sequence that is taken from the genome of one organism or produced synthetically, and which is then introduced into a host cell (e.g., a plant cell) or organism or tissue of interest and which is subsequently integrated into the host’s genome by means of “stable” transformation or transfection approaches. In contrast, the term “transient” transformation or transfection or introduction refers to a way of introducing molecular tools including at least one nucleic acid (DNA, RNA, single-stranded or double-stranded or a mixture thereof) and/or at least one amino acid sequence, optionally comprising suitable chemical or biological agents, to achieve a transfer into at least one compartment of interest of a cell, including, but not restricted to, the cytoplasm, an organelle, including the nucleus, a mitochondrion, a vacuole, a chloroplast, or into a membrane, resulting in transcription and/or translation and/or association and/or activity of the at least one molecule introduced without achieving a stable integration or incorporation into the genome and thus without inheritance of the respective at least one molecule introduced

into the genome of a cell. The term “transgene-free” refers to a condition in which a transgene is not present or found in the genome of a host cell or tissue or organism of interest.

In some embodiments, a polynucleotide and/or a nucleic acid construct of the invention can be an "expression cassette" or can be comprised within an expression cassette. As used herein, "expression cassette" means a recombinant nucleic acid molecule comprising, for example, a nucleic acid construct of the invention (e.g., a polynucleotide encoding a fusion protein of the present invention, a polynucleotide encoding a cytosine deaminase, a polynucleotide encoding an adenine deaminase, a polynucleotide encoding a deaminase fusion protein, a polynucleotide encoding a peptide tag, a polynucleotide encoding an affinity polypeptide, a polynucleotide encoding a glycosylase, and/or a polynucleotide comprising a guide nucleic acid), wherein the nucleic acid construct is operably associated with at least a control sequence (e.g., a promoter). Thus, some embodiments of the invention provide expression cassettes designed to express, for example, a nucleic acid construct of the invention. When an expression cassette comprises more than one polynucleotide, the polynucleotides may be operably linked to a single promoter that drives expression of all of the polynucleotides or the polynucleotides may be operably linked to one or more separate promoters (e.g., three polynucleotides may be driven by one, two or three promoters in any combination). Thus, for example, a polynucleotide encoding a fusion protein, a polynucleotide encoding a deaminase (e.g., an adenine deaminase), and a polynucleotide comprising a guide nucleic acid comprised in an expression cassette may each be operably associated with a single promoter or one or more of the polynucleotide(s) may be operably associated with separate promoters (e.g., two or three promoters) in any combination, which may be the same or different from each other.

In some embodiments, an expression cassette comprising the polynucleotides/nucleic acid constructs of the invention may be optimized for expression in an organism (e.g., an animal, a plant, a bacterium and the like).

An expression cassette comprising a nucleic acid construct of the invention may be chimeric, meaning that at least one of its components is heterologous with respect to at least one of its other components (e.g., a promoter from the host organism operably linked to a polynucleotide of interest to be expressed in the host organism, wherein the polynucleotide of interest is from a different organism than the host or is not normally found in association with that promoter). An expression cassette may also be one that is naturally occurring but has been obtained in a recombinant form useful for heterologous expression.

An expression cassette can optionally include a transcriptional and/or translational termination region (*i.e.*, termination region) and/or an enhancer region that is functional in the

selected host cell. A variety of transcriptional terminators and enhancers are known in the art and are available for use in expression cassettes. Transcriptional terminators are responsible for the termination of transcription and correct mRNA polyadenylation. A termination region and/or the enhancer region may be native to the transcriptional initiation region, may be native to a gene encoding a CRISPR-Cas effector protein or a gene encoding a deaminase, may be native to a host cell, or may be native to another source (*e.g.*, foreign or heterologous to the promoter, to a gene encoding the CRISPR-Cas effector protein or a gene encoding the deaminase, to a host cell, or any combination thereof).

An expression cassette of the invention also can include a polynucleotide encoding a selectable marker, which can be used to select a transformed host cell. As used herein, "selectable marker" means a polynucleotide sequence that when expressed imparts a distinct phenotype to the host cell expressing the marker and thus allows such transformed cells to be distinguished from those that do not have the marker. Such a polynucleotide sequence may encode either a selectable or screenable marker, depending on whether the marker confers a trait that can be selected for by chemical means, such as by using a selective agent (*e.g.*, an antibiotic and the like), or on whether the marker is simply a trait that one can identify through observation or testing, such as by screening (*e.g.*, fluorescence). Many examples of suitable selectable markers are known in the art and can be used in the expression cassettes described herein.

The expression cassettes, the nucleic acid molecules/constructs and polynucleotide sequences described herein can be used in connection with vectors. The term "vector" refers to a composition for transferring, delivering or introducing a nucleic acid (or nucleic acids) into a cell. A vector may comprise a nucleic acid construct comprising one or more nucleotide sequence(s) to be transferred, delivered or introduced into a cell. Vectors for use in transformation of host organisms are well known in the art. Non-limiting examples of general classes of vectors include viral vectors (*e.g.*, Adeno-associated virus (AAV) vectors), plasmid vectors, phage vectors, phagemid vectors, cosmid vectors, fosmid vectors, bacteriophages, artificial chromosomes, minicircles, or *Agrobacterium* binary vectors in double or single stranded linear or circular form which may or may not be self transmissible or mobilizable. In some embodiments, a viral vector can include, but is not limited, to a retroviral, lentiviral, adenoviral, adeno-associated, or herpes simplex viral vector. A vector as defined herein can transform a prokaryotic or eukaryotic host either by integration into the cellular genome or exist extrachromosomally (*e.g.*, autonomous replicating plasmid with an origin of replication). Additionally, included are shuttle vectors by which is meant a DNA vehicle capable, naturally

or by design, of replication in two different host organisms, which may be selected from actinomycetes and related species, bacteria and eukaryotic (e.g., higher plant, mammalian, yeast or fungal cells). In some embodiments, the nucleic acid in the vector is under the control of, and operably linked to, an appropriate promoter or other regulatory elements for transcription in a host cell. The vector may be a bi-functional expression vector which functions in multiple hosts. In the case of genomic DNA, this may contain its own promoter and/or other regulatory elements and in the case of cDNA this may be under the control of an appropriate promoter and/or other regulatory elements for expression in the host cell. Accordingly, a nucleic acid construct of this invention and/or expression cassettes comprising the same may be comprised in vectors as described herein and as known in the art.

As used herein, "contact," "contacting," "contacted," and grammatical variations thereof, refer to placing the components of a desired reaction together under conditions suitable for carrying out the desired reaction (e.g., transformation, transcriptional control, genome editing, nicking, and/or cleavage). Thus, for example, a target nucleic acid may be contacted with a nucleic acid construct of the invention encoding, for example, a nucleic acid binding polypeptide (e.g., a DNA binding polypeptide such as a sequence-specific DNA binding protein (e.g., a polynucleotide-guided endonuclease, a CRISPR-Cas effector protein (e.g., a CRISPR-Cas endonuclease), a zinc finger nuclease, a transcription activator-like effector nuclease (TALEN) and/or an Argonaute protein)), a guide nucleic acid, and optionally a cytosine deaminase and/or adenine deaminase under conditions whereby the nucleic acid binding polypeptide (e.g., a CRISPR-Cas effector protein) is expressed, and the nucleic acid binding polypeptide forms a complex with the guide nucleic acid, the complex hybridizes to the target nucleic acid, and optionally the cytosine deaminase and/or adenine deaminase is/are recruited to the nucleic acid binding polypeptide (and thus, to the target nucleic acid) or the cytosine deaminase and/or adenine deaminase are fused to the nucleic acid binding polypeptide, thereby modifying the target nucleic acid. In some embodiments, the cytosine deaminase and/or adenine deaminase and the nucleic acid binding polypeptide localize at the target nucleic acid, optionally through covalent and/or non-covalent interactions.

In some embodiments, a target nucleic acid may be contacted with a nucleic acid construct of the invention encoding a fusion protein of the present invention, a guide nucleic acid, and optionally a cytosine deaminase and/or adenine deaminase under conditions whereby the fusion protein is produced, or a target nucleic acid may be contacted with a fusion protein of the present invention, a guide nucleic acid, and optionally a cytosine deaminase and/or adenine deaminase. The fusion protein can form a complex with the guide nucleic acid, and

the complex can hybridize to the target nucleic acid, and optionally the cytosine deaminase and/or adenine deaminase is/are recruited to the fusion protein (and thus, to the target nucleic acid) or the cytosine deaminase and/or adenine deaminase are fused to the fusion protein, thereby modifying the target nucleic acid. The cytosine deaminase and/or adenine deaminase and the fusion protein may localize at the target nucleic acid, optionally through covalent and/or non-covalent interactions.

As used herein, "modifying" or "modification" in reference to a target nucleic acid includes editing (e.g., mutating), covalent modification, exchanging/substituting nucleic acids/nucleotide bases, deleting, cleaving, and/or nicking of a target nucleic acid to thereby provide a modified nucleic acid and/or altering transcriptional control of a target nucleic acid to thereby provide a modified nucleic acid. In some embodiments, a modification may include an insertion and/or deletion of any size and/or a single base change (SNP) of any type. In some embodiments, a modification comprises a SNP. In some embodiments, a modification comprises exchanging and/or substituting one or more (e.g., 1, 2, 3, 4, 5, or more) nucleotides. In some embodiments, an insertion or deletion may be about 1 base to about 30,000 bases or more in length (e.g., about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 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, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, 800, 810, 820, 830, 840, 850, 860, 870, 880, 890, 900, 910, 920, 930, 940, 950, 960, 970, 980, 990, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10,000, 10,500, 11,000, 11,500, 12,000, 12,500, 13,000, 13,500, 14,000, 14,500, 15,000, 15,500, 16,000, 16,500, 17,000, 17,500, 18,000, 18,500, 19,000, 19,500, 20,000, 20,500, 21,000, 21,500, 22,000, 22,500, 23,000, 23,500, 24,000, 24,500, 25,000, 25,500, 26,000, 26,500, 27,000, 27,500, 28,000, 28,500, 29,000, 29,500, 30,000 bases in length or more, or any value or range therein). Thus, in some embodiments, an insertion or deletion may be about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 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, 79, 80, 81, 82, 83, 84, 85,

86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300 to about 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, 800, 810, 820, 830, 840, 850, 860, 870, 880, 890, 900, 910, 920, 930, 940, 950, 960, 970, 980, 990, 1000 bases in length, or any range or value therein; about 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, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300 bases to about 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, 800, 810, 820, 830, 840, 850, 860, 870, 880, 890, 900, 910, 920, 930, 940, 950, 960, 970, 980, 990, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000 bases or more in length, or any value or range therein; about 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, 800, 810, 820, 830, 840, 850, 860, 870, 880, 890, 900, 910, 920, 930, 940, 950, 960, 970, 980, 990, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000 bases to about 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, or 10,000 bases or more in length, or any value or range therein; or about 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, or 700 bases to about 710, 720, 730, 740, 750, 760, 770, 780, 790, 800, 810, 820, 830, 840, 850, 860, 870, 880, 890, 900, 910, 920, 930, 940, 950, 960, 970, 980, 990, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2500, 3000, 3500, 4000, 4500, or 5000 bases or more in length, or any value or range therein. In some embodiments, an insertion or deletion may be about 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, or 10,000 bases to about 10,500, 11,000, 11,500, 12,000, 12,500, 13,000, 13,500, 14,000, 14,500, 15,000, 15,500, 16,000, 16,500, 17,000, 17,500, 18,000, 18,500, 19,000, 19,500, 20,000, 20,500, 21,000, 21,500, 22,000, 22,500, 23,000, 23,500, 24,000, 24,500, 25,000, 25,500, 26,000, 26,500, 27,000, 27,500, 28,000, 28,500, 29,000, 29,500, or 30,000 bases or more in length, or any value or range therein.

"Recruit," "recruiting" or "recruitment" as used herein refer to attracting one or more polypeptide(s) or polynucleotide(s) to another polypeptide or polynucleotide (e.g., to a

particular location in a genome) using protein-protein interactions, nucleic acid protein interactions (e.g., RNA-protein interactions), and/or chemical interactions. Protein-protein interactions can include, but are not limited to, peptide tags (epitopes, multimerized epitopes) and corresponding affinity polypeptides, RNA recruiting motifs and corresponding affinity polypeptides, and/or chemical interactions. Example chemical interactions that may be useful with polypeptides and polynucleotides for the purpose of recruitment can include, but are not limited to, rapamycin-inducible dimerization of FRB – FKBP; Biotin-streptavidin interaction; SNAP tag (Hussain et al. *Curr Pharm Des.* 19(30):5437-42 (2013)); Halo tag (Los et al. *ACS Chem Biol.* 3(6):373-82 (2008)); CLIP tag (Gautier et al. *Chemistry & Biology* 15:128–136 (2008)); DmrA-DmrC heterodimer induced by a compound (Tak et al. *Nat Methods* 14(12):1163–1166 (2017)); Bifunctional ligand approaches (fuse two protein-binding chemicals together) (Voß et al. *Curr Opin Chemical Biology* 28:194-201 (2015)) (e.g. dihydrofolate reductase (DHFR) (Kopyteck et al. *Cell Cehm Biol* 7(5):313-321 (2000)).

"Introducing," "introduce," "introduced" (and grammatical variations thereof) in the context of a polynucleotide of interest or editing system means presenting a nucleotide sequence of interest (e.g., polynucleotide, a nucleic acid construct, and/or a guide nucleic acid) and/or editing system (e.g., a polynucleotide, polypeptide, and/or ribonucleoprotein) to a host organism or cell of said organism (e.g., host cell; e.g., a plant cell) in such a manner that the nucleotide sequence and/or editing system gains access to the interior of a cell. Thus, for example, a nucleic acid construct of the invention encoding a fusion protein of the present invention, a guide nucleic acid, and/or a cytosine deaminase and/or adenine deaminase may be introduced into a cell of an organism, thereby transforming the cell with the fusion protein, guide nucleic acid, and/or cytosine deaminase and/or adenine deaminase. In some embodiments, a fusion protein of the present invention and/or a guide nucleic acid may be introduced into a cell of an organism, optionally wherein the fusion protein and guide nucleic acid may be comprised in a complex (e.g., a ribonucleoprotein). In some embodiments, the organism is a eukaryote (e.g., a mammal such as a human).

The term "transformation" as used herein refers to the introduction of a nucleic acid, polypeptide, and/or ribonucleoprotein (e.g., heterologous nucleic acid, polypeptide, and/or ribonucleoprotein) into a cell. Transformation of a cell may be stable or transient. Thus, in some embodiments, a host cell or host organism may be stably transformed with a polynucleotide/nucleic acid molecule of the invention. In some embodiments, a host cell or host organism may be transiently transformed with a nucleic acid construct, a polypeptide, and/or a ribonucleoprotein of the invention.

"Transient transformation" in the context of a polynucleotide, polypeptide, and/or ribonucleoprotein means that a polynucleotide, polypeptide, and/or ribonucleoprotein is introduced into the cell and does not integrate into the genome of the cell.

By "stably introducing" or "stably introduced" in the context of a polynucleotide introduced into a cell is intended that the introduced polynucleotide is stably incorporated into the genome of the cell, and thus the cell is stably transformed with the polynucleotide.

"Stable transformation" or "stably transformed" as used herein means that a nucleic acid molecule is introduced into a cell and integrates into the genome of the cell. As such, the integrated nucleic acid molecule is capable of being inherited by the progeny thereof, more particularly, by the progeny of multiple successive generations. "Genome" as used herein includes the nuclear and the plastid genome, and therefore includes integration of the nucleic acid into, for example, the chloroplast or mitochondrial genome. Stable transformation as used herein can also refer to a transgene that is maintained extrachromasomally, for example, as a minichromosome or a plasmid.

Transient transformation may be detected by, for example, an enzyme-linked immunosorbent assay (ELISA) or Western blot, which can detect the presence of a peptide or polypeptide encoded by one or more transgene introduced into an organism. Stable transformation of a cell can be detected by, for example, a Southern blot hybridization assay of genomic DNA of the cell with nucleic acid sequences which specifically hybridize with a nucleotide sequence of a transgene introduced into an organism (e.g., a mammal, plant, etc.). Stable transformation of a cell can be detected by, for example, a Northern blot hybridization assay of RNA of the cell with nucleic acid sequences which specifically hybridize with a nucleotide sequence of a transgene introduced into a host organism. Stable transformation of a cell can also be detected by, e.g., a polymerase chain reaction (PCR) or other amplification reactions as are well known in the art, employing specific primer sequences that hybridize with target sequence(s) of a transgene, resulting in amplification of the transgene sequence, which can be detected according to standard methods. Transformation can also be detected by direct sequencing and/or hybridization protocols well known in the art.

Accordingly, in some embodiments, nucleotide sequences, polynucleotides, nucleic acid constructs, and/or expression cassettes of the invention may be expressed transiently and/or they can be stably incorporated into the genome of the host organism. Thus, in some embodiments, a nucleic acid construct of the invention may be transiently introduced into a cell with a guide nucleic acid and as such, no DNA maintained in the cell.

A nucleic acid construct, polypeptide, and/or ribonucleoprotein of the invention can be

introduced into a cell by any method known to those of skill in the art. In some embodiments, transformation methods include, but are not limited to, transformation via bacterial-mediated nucleic acid delivery (e.g., via *Agrobacteria*), viral-mediated nucleic acid delivery, silicon carbide and/or nucleic acid whisker-mediated nucleic acid delivery, liposome mediated nucleic acid delivery, microinjection, microparticle bombardment, calcium-phosphate-mediated transformation, cyclodextrin-mediated transformation, electroporation, nanoparticle-mediated transformation, sonication, infiltration, PEG-mediated nucleic acid uptake, as well as any other electrical, chemical, physical (mechanical) and/or biological mechanism that results in the introduction of nucleic acid into the cell (e.g., a plant cell or an animal cell), including any combination thereof. In some embodiments of the invention, transformation of a cell comprises nuclear transformation. In some embodiments, transformation of a cell comprises plastid transformation (e.g., chloroplast transformation). In some embodiments, a recombinant nucleic acid construct of the invention can be introduced into a cell via conventional breeding techniques.

Procedures for transforming both eukaryotic and prokaryotic organisms are well known and routine in the art and are described throughout the literature (*See*, for example, Jiang et al. 2013. *Nat. Biotechnol.* 31:233-239; Ran et al. *Nature Protocols* 8:2281–2308 (2013)). General guides to various plant transformation methods known in the art include Miki et al. (“Procedures for Introducing Foreign DNA into Plants” in *Methods in Plant Molecular Biology and Biotechnology*, Glick, B. R. and Thompson, J. E., Eds. (CRC Press, Inc., Boca Raton, 1993), pages 67-88) and Rakowoczy-Trojanowska (*Cell. Mol. Biol. Lett.* 7:849-858 (2002)).

A nucleotide sequence, polypeptide, and/or ribonucleoprotein therefore can be introduced into a host organism or its cell in any number of ways that are well known in the art. The methods of the invention do not depend on a particular method for introducing one or more nucleotide sequence(s), polypeptide(s), and/or ribonucleoprotein(s) into the organism, only that they gain access to the interior of at least one cell of the organism. Where more than one nucleotide sequence, polypeptide, and/or ribonucleoprotein is to be introduced, they can be assembled as part of a single nucleic acid construct, or as separate nucleic acid constructs, and can be located on the same or different nucleic acid constructs. Accordingly, a nucleotide sequence, polypeptide, and/or ribonucleoprotein can be introduced into the cell of interest in a single transformation event, and/or in separate transformation events, or, alternatively, where relevant, a nucleotide sequence can be incorporated into a plant, for example, as part of a breeding protocol. In some embodiments, the cell is a eukaryotic cell (e.g., a plant cell or a mammalian such as a human cell).

In some embodiments, a nucleic acid construct of the invention (e.g., a polynucleotide encoding a fusion protein of the present invention, a polynucleotide encoding a deaminase, and/or a guide nucleic acid and/or expression cassettes and/or vectors comprising the same) may be operably linked to at least one regulatory sequence, optionally, wherein the at least one regulatory sequence may be codon optimized for expression in a plant. In some embodiments, the at least one regulatory sequence may be, for example, a promoter, an operon, a terminator, or an enhancer. In some embodiments, the at least one regulatory sequence may be a promoter. In some embodiments, the regulatory sequence may be an intron. In some embodiments, the at least one regulatory sequence may be, for example, a promoter operably associated with an intron or a promoter region comprising an intron. In some embodiments, the at least one regulatory sequence may be, for example a ubiquitin promoter and its associated intron (e.g., *Medicago truncatula* and/or *Zea mays* and their associated introns). In some embodiments, the at least one regulatory sequence may be a terminator nucleotide sequence and/or an enhancer nucleotide sequence.

In some embodiments, a nucleic acid construct of the invention may be operably associated with a promoter region, wherein the promoter region comprises an intron, optionally wherein the promoter region may be a ubiquitin promoter and intron (e.g., a *Medicago* or a maize ubiquitin promoter and intron, e.g., **SEQ ID NO:36** or **SEQ ID NO:37**). In some embodiments, the nucleic acid construct of the invention that is operably associated with a promoter region comprising an intron may be codon optimized for expression in a plant.

In some embodiments, a nucleic acid construct of the invention may encode one or more (e.g., 1, 2, 3, 4, or more) polypeptide(s) of interest. The one or more polypeptides of interest may be codon optimized for expression in a eukaryote (e.g., a human or a plant). In some embodiments, a fusion protein may comprise one or more (e.g., 1, 2, 3, 4, or more) polypeptide(s) of interest.

A polypeptide of interest useful with this invention can include, but is not limited to, a polypeptide or protein domain having deaminase activity, nickase activity, recombinase activity, transposase activity, methylase activity, glycosylase (DNA glycosylase) activity, glycosylase inhibitor activity (e.g., uracil-DNA glycosylase inhibitor (UGI)), a reverse transcriptase, a peptide tag (e.g., a GCN4 peptide tag), demethylase activity, transcription activation activity, transcription repression activity, transcription release factor activity, histone modification activity, nuclease activity, single-strand RNA cleavage activity, double-strand RNA cleavage activity, restriction endonuclease activity (e.g., FokI), nucleic acid binding activity, methyltransferase activity, DNA repair activity, DNA damage activity, dismutase

activity, alkylation activity, depurination activity, oxidation activity, pyrimidine dimer forming activity, integrase activity, transposase activity, polymerase activity (e.g., DNA polymerase activity), ligase activity, helicase activity, a nuclear localization sequence or activity, T-DNA processing and/or transfer (e.g., VirD2), an affinity polypeptide, a peptide tag, and/or photolyase activity. In some embodiments, the polypeptide of interest is a FokI nuclease, or a uracil-DNA glycosylase inhibitor. When encoded in a nucleic acid (polynucleotide, expression cassette, and/or vector) the encoded polypeptide or protein domain may be codon optimized for expression in an organism. In some embodiments, a polypeptide of interest may be linked to a fusion protein of the present invention or to a CRISPR-Cas effector protein domain to provide a CRISPR-Cas fusion protein. In some embodiments, a CRISPR-Cas fusion protein that comprises a CRISPR-Cas effector protein domain linked to a recruiting motif (e.g., a peptide tag) may also be linked to a polypeptide of interest (e.g., a CRISPR-Cas effector protein domain may be, for example, linked to both a recruiting motif (e.g., a peptide tag or an affinity polypeptide) and, for example, a polypeptide of interest.

In some embodiments, an editing system of the present invention comprises a CRISPR-Cas effector protein. As used herein, a "CRISPR-Cas effector protein" is a protein or polypeptide that cleaves, cuts, or nicks a nucleic acid; binds a nucleic acid (e.g., a target nucleic acid and/or a guide nucleic acid); and/or that identifies, recognizes, or binds a guide nucleic acid as defined herein. In some embodiments, a CRISPR-Cas effector protein may be an enzyme (e.g., a nuclease, endonuclease, nickase, etc.) and/or may function as an enzyme. In some embodiments, a CRISPR-Cas effector protein refers to a CRISPR-Cas nuclease. In some embodiments, a CRISPR-Cas effector protein comprises nuclease activity and/or nickase activity, comprises a nuclease domain whose nuclease activity and/or nickase activity has been reduced or eliminated, comprises single stranded DNA cleavage activity (ss DNase activity) or which has ss DNase activity that has been reduced or eliminated, and/or comprises self-processing RNase activity or which has self-processing RNase activity that has been reduced or eliminated. A CRISPR-Cas effector protein may bind to a target nucleic acid. A CRISPR-Cas effector protein may be a Type I, II, III, IV, V, or VI CRISPR-Cas effector protein. In some embodiments, a CRISPR-Cas effector protein may be from a Type I CRISPR-Cas system, a Type II CRISPR-Cas system, a Type III CRISPR-Cas system, a Type IV CRISPR-Cas system, Type V CRISPR-Cas system, or a Type VI CRISPR-Cas system. In some embodiments, a CRISPR-Cas effector protein of the invention may be from a Type II CRISPR-Cas system or a Type V CRISPR-Cas system. In some embodiments, a CRISPR-Cas effector protein may be a Type II CRISPR-Cas effector protein, for example, a Cas9 effector protein.

In some embodiments, a CRISPR-Cas effector protein may be Type V CRISPR-Cas effector protein, for example, a Cas12 effector protein. In some embodiments, a CRISPR-Cas effector protein may be Cas12a and optionally may have an amino acid sequence of any one of **SEQ ID NOs:38-60 or 192-195** and/or a nucleotide sequence of any one of **SEQ ID NOs:61-63**. In some embodiments, a CRISPR-Cas effector protein may be an active Cas12a and optionally may have an amino acid sequence of **SEQ ID NO:46 or 55**. In some embodiments, a CRISPR-Cas effector protein may be an inactive (i.e., dead) Cas12a and optionally may have an amino acid sequence of **SEQ ID NO:38**.

Exemplary CRISPR-Cas effector proteins include, but are not limited to, a Cas9, C2c1, C2c3, Cas12a (also referred to as Cpf1), Cas12b, Cas12c, Cas12d, Cas12e, Cas13a, Cas13b, Cas13c, Cas13d, Cas1, Cas1B, Cas2, Cas3, Cas3', Cas3'', Cas4, Cas5, Cas6, Cas7, Cas8, Cas9 (also known as Csn1 and Csx12), Cas10, Csy1, Csy2, Csy3, Cse1, Cse2, Cse3, Cse4, Csa5, Csn2, Csm2, Csm3, Csm4, Csm5, Csm6, Cmr1, Cmr3, Cmr4, Cmr5, Cmr6, Csb1, Csb2, Csb3, Csx17, Csx14, Csx10, Csx16, CsaX, Csx3, Csx1, Csx15, Csf1, Csf2, Csf3, Csf4 (dinG), and/or Csf5 nuclease, optionally wherein the CRISPR-Cas effector protein may be a Cas9, Cas12a (Cpf1), Cas12b, Cas12c (C2c3), Cas12d (CasY), Cas12e (CasX), Cas12g, Cas12h, Cas12i, C2c4, C2c5, C2c8, C2c9, C2c10, Cas14a, Cas14b, and/or Cas14c effector protein.

In some embodiments, a CRISPR-Cas effector protein useful with the invention may comprise a mutation in its nuclease active site and/or nuclease domain (e.g., a RuvC, HNH, e.g., a RuvC site of a Cas12a nuclease domain; e.g., a RuvC site and/or HNH site of a Cas9 nuclease domain). A CRISPR-Cas effector protein having a mutation in its nuclease active site and/or nuclease domain, and therefore, no longer comprising nuclease activity, is commonly referred to as “inactive” or “dead,” e.g., dCas12a. In some embodiments, a CRISPR-Cas effector protein having a mutation in its nuclease active site and/or nuclease domain may have impaired activity or reduced activity (e.g., nickase activity) as compared to the same CRISPR-Cas effector protein without the mutation.

A Type V CRISPR-Cas effector protein useful with embodiments of the invention may be a Cas12a. The CRISPR-Cas effector protein may be a Type V Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas nuclease. Cas12a differs in several respects from the more well-known Type II CRISPR Cas9 nuclease. For example, Cas9 recognizes a G-rich protospacer-adjacent motif (PAM) that is 3' to its guide RNA (gRNA, sgRNA, crRNA, crDNA, CRISPR array) binding site (protospacer, target nucleic acid, target DNA) (3'-NGG), while Cas12a recognizes a T-rich PAM that is located 5' to the target nucleic acid (5'-TTN, 5'-TTTN). In fact, the orientations in which Cas9 and Cas12a bind their guide

RNAs are very nearly reversed in relation to their N and C termini. Furthermore, Cas12a enzymes use a single guide RNA (gRNA, CRISPR array, crRNA) rather than the dual guide RNA (sgRNA (e.g., crRNA and tracrRNA)) found in natural Cas9 systems, and Cas12a processes its own gRNAs. Additionally, Cas12a nuclease activity produces staggered DNA double stranded breaks instead of blunt ends produced by Cas9 nuclease activity, and Cas12a relies on a single RuvC domain to cleave both DNA strands, whereas Cas9 utilizes an HNH domain and a RuvC domain for cleavage.

A CRISPR Cas12a effector protein useful with this invention may be any known or later identified Cas12a (previously known as Cpf1) (see, e.g., U.S. Patent No. 9,790,490, which is incorporated by reference for its disclosures of Cpf1 (Cas12a) sequences). The term "Cas12a" refers to an RNA-guided protein that can have nuclease activity, the protein comprising a guide nucleic acid binding domain and an active, inactive, or partially active DNA cleavage domain, thereby the RNA-guided nuclease activity of the Cas12a may be active, inactive or partially active, respectively. In some embodiments, a Cas12a useful with the invention may comprise a mutation in the nuclease active site (e.g., RuvC site of the Cas12a domain). A Cas12a having a mutation in its nuclease domain and/or nuclease active site, and therefore, no longer comprising nuclease activity, is commonly referred to as deadCas12a (e.g., dCas12a). In some embodiments, a Cas12a having a mutation in its nuclease domain and/or nuclease active site may have impaired activity, e.g., may have reduced nickase activity.

In some embodiments, a CRISPR-Cas effector protein (e.g., Cas12a) may be optimized for expression in an organism, for example, in an animal (e.g., a mammal such as a human), a plant, a fungus, an archaeon, or a bacterium. In some embodiments, a CRISPR-Cas effector protein (e.g., Cas12a) may be optimized for expression in a plant.

Any deaminase domain/polypeptide useful for base editing may be used with this invention. A "cytosine deaminase" and "cytidine deaminase" as used herein refer to a polypeptide or domain thereof that catalyzes or is capable of catalyzing cytosine deamination in that the polypeptide or domain catalyzes or is capable of catalyzing the removal of an amine group from a cytosine base. Thus, a cytosine deaminase may result in conversion of cytosine to a thymidine (through a uracil intermediate), causing a C to T conversion, or a G to A conversion in the complementary strand in the genome. Thus, in some embodiments, the cytosine deaminase encoded by the polynucleotide of the invention generates a C→T conversion in the sense (e.g., "+"; template) strand of the target nucleic acid or a G→A conversion in antisense (e.g., "-", complementary) strand of the target nucleic acid. In some

embodiments, a cytosine deaminase encoded by a polynucleotide of the invention generates a C to T, G, or A conversion in the complementary strand in the genome.

A cytosine deaminase useful with this invention may be any known or later identified cytosine deaminase from any organism (see, e.g., U.S. Patent No. 10,167,457 and Thuronyi et al. *Nat. Biotechnol.* 37:1070–1079 (2019), each of which is incorporated by reference herein for its disclosure of cytosine deaminases). Cytosine deaminases can catalyze the hydrolytic deamination of cytidine or deoxycytidine to uridine or deoxyuridine, respectively. Thus, in some embodiments, a deaminase or deaminase domain useful with this invention may be a cytidine deaminase domain, catalyzing the hydrolytic deamination of cytosine to uracil. In some embodiments, a cytosine deaminase may be a variant of a naturally-occurring cytosine deaminase, including, but not limited to, a primate (e.g., a human, monkey, chimpanzee, gorilla), a dog, a cow, a rat or a mouse. Thus, in some embodiments, an cytosine deaminase useful with the invention may be about 70% to about 100% identical to a wild-type cytosine deaminase (e.g., about 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% identical, and any range or value therein, to a naturally occurring cytosine deaminase).

In some embodiments, a cytosine deaminase useful with the invention may be an apolipoprotein B mRNA-editing complex (APOBEC) family deaminase. In some embodiments, the cytosine deaminase may be an APOBEC1 deaminase, an APOBEC2 deaminase, an APOBEC3A deaminase, an APOBEC3B deaminase, an APOBEC3C deaminase, an APOBEC3D deaminase, an APOBEC3F deaminase, an APOBEC3G deaminase, an APOBEC3H deaminase, an APOBEC4 deaminase, a human activation induced deaminase (hAID), an rAPOBEC1, FERNY, and/or a CDA1, optionally a pmCDA1, an atCDA1 (e.g., At2g19570), and evolved versions of the same. Evolved deaminases are disclosed in, for example, U.S. Patent No. 10,113,163, Gaudelli et al. *Nature* 551(7681):464-471 (2017)) and Thuronyi et al. (*Nature Biotechnology* 37: 1070-1079 (2019)), each of which are incorporated by reference herein for their disclosure of deaminases and evolved deaminases. In some embodiments, the cytosine deaminase may be an APOBEC1 deaminase having the amino acid sequence of **SEQ ID NO:64**. In some embodiments, the cytosine deaminase may be an APOBEC3A deaminase having the amino acid sequence of **SEQ ID NO:65**. In some embodiments, the cytosine deaminase may be an CDA1 deaminase, optionally a CDA1 having the amino acid sequence of **SEQ ID NO:66**. In some embodiments, the cytosine deaminase may be a FERNY deaminase, optionally a FERNY having the amino

acid sequence of **SEQ ID NO:67**. In some embodiments, the cytosine deaminase may be a rAPOBEC1 deaminase, optionally a rAPOBEC1 deaminase having the amino acid sequence of **SEQ ID NO:68**. In some embodiments, the cytosine deaminase may be a hAID deaminase, optionally a hAID having the amino acid sequence of **SEQ ID NO:69** or **SEQ ID NO:70**. In some embodiments, a cytosine deaminase useful with the invention may be about 70% to about 100% identical (e.g., 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%, 99.5% or 100% identical) to the amino acid sequence of a naturally occurring cytosine deaminase (e.g., "evolved deaminases") (see, e.g., **SEQ ID NO:71**, **SEQ ID NO:72**, **SEQ ID NO:73**). In some embodiments, a cytosine deaminase useful with the invention may be about 70% to about 99.5% identical (e.g., about 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 99.5% identical) to the amino acid sequence of any one of **SEQ ID NOs:64-73** (e.g., at least 80%, at least 85%, at least 90%, at least 92%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or at least 99.5% identical to the amino acid sequence of any one of **SEQ ID NOs:64-73**). In some embodiments, a polynucleotide encoding a cytosine deaminase may be codon optimized for expression in a plant and the codon optimized polypeptide may be about 70% to 99.5% identical to the reference polynucleotide.

An "adenine deaminase" and "adenosine deaminase" as used herein refer to a polypeptide or domain thereof that catalyzes or is capable of catalyzing the hydrolytic deamination (e.g., removal of an amine group from adenine) of adenine or adenosine. In some embodiments, an adenine deaminase may catalyze the hydrolytic deamination of adenosine or deoxyadenosine to inosine or deoxyinosine, respectively. In some embodiments, the adenosine deaminase may catalyze the hydrolytic deamination of adenine or adenosine in DNA. In some embodiments, an adenine deaminase encoded by a nucleic acid construct of the invention may generate an A→G conversion in the sense (e.g., "+", template) strand of the target nucleic acid or a T→C conversion in the antisense (e.g., "-", complementary) strand of the target nucleic acid. An adenine deaminase useful with this invention may be any known or later identified adenine deaminase from any organism (see, e.g., U.S. Patent No. 10,113,163, which is incorporated by reference herein for its disclosure of adenine deaminases).

In some embodiments, an adenosine deaminase may be a variant of a naturally-occurring adenine deaminase. Thus, in some embodiments, an adenosine deaminase may be about 70% to 100% identical to a wild-type adenine deaminase (e.g., about 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% identical, and any range or value therein, to a naturally occurring adenine deaminase). In some embodiments, the deaminase or deaminase does not occur in nature and may be referred to as an engineered, mutated or evolved adenosine deaminase. Thus, for example, an engineered, mutated or evolved adenine deaminase polypeptide or an adenine deaminase domain may be about 70% to 99.9% identical to a naturally occurring adenine deaminase polypeptide/domain (e.g., about 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%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8% or 99.9% identical, and any range or value therein, to a naturally occurring adenine deaminase polypeptide or adenine deaminase domain). In some embodiments, the adenosine deaminase may be from a bacterium, (e.g., *Escherichia coli*, *Staphylococcus aureus*, *Haemophilus influenzae*, *Caulobacter crescentus*, and the like). In some embodiments, a polynucleotide encoding an adenine deaminase polypeptide/domain may be codon optimized for expression in a plant.

In some embodiments, an adenine deaminase domain may be a wild-type tRNA-specific adenosine deaminase domain, e.g., a tRNA-specific adenosine deaminase (TadA) and/or a mutated/evolved adenosine deaminase domain, e.g., mutated/evolved tRNA-specific adenosine deaminase domain (TadA*). In some embodiments, a TadA domain may be from *E. coli*. In some embodiments, the TadA may be modified, e.g., truncated, missing one or more N-terminal and/or C-terminal amino acids relative to a full-length TadA (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 6, 17, 18, 19, or 20 N-terminal and/or C terminal amino acid residues may be missing relative to a full length TadA. In some embodiments, a TadA polypeptide or TadA domain does not comprise an N-terminal methionine. In some embodiments, a wild-type *E. coli* TadA comprises the amino acid sequence of **SEQ ID NO:74**. In some embodiments, a mutated/evolved *E. coli* TadA* comprises the amino acid sequence of any one of **SEQ ID NOs:75-78**. In some embodiments, a polynucleotide encoding a TadA/TadA* may be codon optimized for expression in a plant. In some embodiments, an adenine deaminase may comprise all or a portion of an amino acid sequence of any one of **SEQ ID NOs:79-84**. In some embodiments, an adenine deaminase may comprise all or a portion of an amino acid sequence of any one of **SEQ ID NOs:74-84**.

In some embodiments, a nucleic acid construct of this invention may further encode a glycosylase inhibitor (e.g., a uracil glycosylase inhibitor (UGI) such as uracil-DNA glycosylase inhibitor). In some embodiments, the invention provides fusion proteins comprising a UGI

and/or one or more polynucleotides encoding the same, optionally wherein the one or more polynucleotides may be codon optimized for expression in a plant.

A “uracil glycosylase inhibitor” useful with the invention may be any protein or polypeptide that is capable of inhibiting a uracil-DNA glycosylase base-excision repair enzyme. In some embodiments, a UGI domain comprises a wild-type UGI or a fragment thereof. In some embodiments, a UGI domain useful with the invention may be about 70% to about 100% identical (e.g., 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%, 99.5% or 100% identical and any range or value therein) to the amino acid sequence of a naturally occurring UGI domain. In some embodiments, a UGI domain may comprise the amino acid sequence of **SEQ ID NO:85** or a polypeptide having about 70% to about 99.5% identity to the amino acid sequence of **SEQ ID NO:85** (e.g., at least 80%, at least 85%, at least 90%, at least 92%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or at least 99.5% identical to the amino acid sequence of **SEQ ID NO:85**). For example, in some embodiments, a UGI domain may comprise a fragment of the amino acid sequence of **SEQ ID NO:85** that is 100% identical to a portion of consecutive nucleotides (e.g., 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80 consecutive nucleotides; e.g., about 10, 15, 20, 25, 30, 35, 40, 45, to about 50, 55, 60, 65, 70, 75, 80 consecutive nucleotides) of the amino acid sequence of **SEQ ID NO:85**. In some embodiments, a UGI domain may be a variant of a known UGI (e.g., **SEQ ID NO:85**) having about 70% to about 99.5% identity (e.g., 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%, 99.5% identity, and any range or value therein) to the known UGI. In some embodiments, a polynucleotide encoding a UGI may be codon optimized for expression in a plant (e.g., a plant) and the codon optimized polypeptide may be about 70% to about 99.5% identical to the reference polynucleotide.

A fusion protein of the present invention may be used in combination with a guide nucleic acid (e.g., guide RNA (gRNA), CRISPR array, CRISPR RNA, crRNA) that is designed to function with the fusion protein to modify a target nucleic acid. A guide nucleic acid useful with this invention may comprise at least one spacer sequence and at least one repeat sequence. The guide nucleic acid is capable of forming a complex with the fusion protein (e.g., with a nuclease domain of the fusion protein) and the spacer sequence is capable of hybridizing to a target nucleic acid, thereby guiding the complex to the target nucleic acid, wherein the target nucleic acid may be modified (e.g., cleaved or edited) and/or modulated (e.g., modulating

transcription) by a deaminase (e.g., a cytosine deaminase and/or adenine deaminase) or reverse transcriptase, optionally present in and/or recruited to the complex.

A “guide nucleic acid,” “guide RNA,” “gRNA,” “CRISPR RNA/DNA” “crRNA” or “crDNA” as used herein means a nucleic acid that comprises at least one spacer sequence, which is complementary to (and hybridizes to) a target nucleic acid (e.g., a target DNA and/or protospacer), and at least one repeat sequence (e.g., a repeat of a Type V Cas12a CRISPR-Cas system, or a fragment or portion thereof, wherein the repeat sequence may be linked to the 5’ end and/or the 3’ end of the spacer sequence. In some embodiments, the guide nucleic acid comprises DNA. In some embodiments, the guide nucleic acid comprises RNA (e.g., is a guide RNA). The design of a gRNA of this invention may be based on a Type I, Type II, Type III, Type IV, Type V, or Type VI CRISPR-Cas system. In some embodiments, a Cas12a gRNA may comprise, from 5’ to 3’, a repeat sequence (full length or portion thereof (“handle”); e.g., pseudoknot-like structure) and a spacer sequence.

In some embodiments, a guide nucleic acid may comprise more than one repeat sequence-spacer sequence (e.g., 2, 3, 4, 5, 6, 7, 8, 9, 10, or more repeat-spacer sequences) (e.g., repeat-spacer-repeat, e.g., repeat-spacer-repeat-spacer-repeat-spacer-repeat-spacer-repeat-spacer, and the like). The guide nucleic acids of this invention are synthetic, human-made and not found in nature. A gRNA can be quite long and may be used as an aptamer (like in the MS2 recruitment strategy) or other RNA structures hanging off the spacer.

A “repeat sequence” as used herein, refers to, for example, any repeat sequence of a wild-type CRISPR Cas locus (e.g., a Cas9 locus, a Cas12a locus, a C2c1 locus, etc.) or a repeat sequence of a synthetic repeat sequence (e.g., a synthetic crRNA) that is functional with the CRISPR-Cas effector protein encoded by the nucleic acid constructs of the invention. A repeat sequence useful with this invention can be any known or later identified repeat sequence of a CRISPR-Cas locus (e.g., Type I, Type II, Type III, Type IV, Type V or Type VI) or it can be a synthetic repeat designed to function in a Type I, II, III, IV, V or VI CRISPR-Cas system. A repeat sequence may comprise a hairpin structure and/or a stem loop structure. In some embodiments, a repeat sequence may form a pseudoknot-like structure at its 5’ end (i.e., “handle”). Thus, in some embodiments, a repeat sequence can be identical to or substantially identical to a repeat sequence from wild-type Type I CRISPR-Cas loci, Type II, CRISPR-Cas loci, Type III, CRISPR-Cas loci, Type IV CRISPR-Cas loci, Type V CRISPR-Cas loci and/or Type VI CRISPR-Cas loci. A repeat sequence from a wild-type CRISPR-Cas locus may be determined through established algorithms, such as using the CRISPRfinder offered through CRISPRdb (see, Grissa et al. *Nucleic Acids Res.* 35(Web Server issue):W52-7). In some

embodiments, a repeat sequence or portion thereof is linked at its 3' end to the 5' end of a spacer sequence, thereby forming a repeat-spacer sequence (e.g., guide nucleic acid, guide RNA/DNA, crRNA, crDNA).

In some embodiments, a repeat sequence comprises, consists essentially of, or consists of at least 10 nucleotides depending on the particular repeat and whether the guide nucleic acid comprising the repeat is processed or unprocessed (e.g., about 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 to 100 or more nucleotides, or any range or value therein; e.g., about). In some embodiments, a repeat sequence comprises, consists essentially of, or consists of about 10 to about 20, about 10 to about 30, about 10 to about 45, about 10 to about 50, about 15 to about 30, about 15 to about 40, about 15 to about 45, about 15 to about 50, about 20 to about 30, about 20 to about 40, about 20 to about 50, about 30 to about 40, about 40 to about 80, about 50 to about 100 or more nucleotides.

A repeat sequence linked to the 5' end of a spacer sequence can comprise a portion of a repeat sequence (e.g., 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 or more contiguous nucleotides of a wild-type repeat sequence). In some embodiments, a portion of a repeat sequence linked to the 5' end of a spacer sequence can be about five to about ten consecutive nucleotides in length (e.g., about 5, 6, 7, 8, 9, 10 nucleotides) and have at least 90% sequence identity (e.g., at least about 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or more) to the same region (e.g., 5' end) of a wild-type CRISPR Cas repeat nucleotide sequence. In some embodiments, a portion of a repeat sequence may comprise a pseudoknot-like structure at its 5' end (e.g., "handle").

A "spacer sequence" as used herein is a nucleotide sequence that is complementary to a target nucleic acid (e.g., target DNA) (e.g., protospacer). The spacer sequence can be fully complementary or substantially complementary (e.g., at least about 70% complementary (e.g., about 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 more)) to a target nucleic acid. Thus, in some embodiments, the spacer sequence can have one, two, three, four, or five mismatches as compared to the target nucleic acid, which mismatches can be contiguous or noncontiguous. In some embodiments, the spacer sequence can have 70% complementarity to a target nucleic acid. In other embodiments, the spacer nucleotide sequence can have 80% complementarity to a target nucleic acid. In still other embodiments, the spacer nucleotide sequence can have 85%, 90%, 95%, 96%, 97%, 98%, 99% or 99.5% complementarity, and the like, to the target nucleic acid (protospacer). In some embodiments,

the spacer sequence is 100% complementary to the target nucleic acid. A spacer sequence may have a length from about 15 nucleotides to about 30 nucleotides (e.g., 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides, or any range or value therein). Thus, in some embodiments, a spacer sequence may have complete complementarity or substantial complementarity over a region of a target nucleic acid (e.g., protospacer) that is at least about 15 nucleotides to about 30 nucleotides in length. In some embodiments, the spacer is about 20 nucleotides in length. In some embodiments, the spacer is about 21, 22, or 23 nucleotides in length.

In some embodiments, the 5' region of a spacer sequence of a guide nucleic acid may be fully complementary to a target nucleic acid, while the 3' region of the spacer may be substantially complementary to the target nucleic acid (such as for a spacer in a Type V CRISPR-Cas system), or the 3' region of a spacer sequence of a guide nucleic acid may be fully complementary to a target nucleic acid, while the 5' region of the spacer may be substantially complementary to the target nucleic acid (such as for a spacer in a Type II CRISPR-Cas system), and therefore, the overall complementarity of the spacer sequence to the target nucleic acid may be less than 100%. Thus, for example, in a guide nucleic acid for a Type V CRISPR-Cas system, the first 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 nucleotides in the 5' region (i.e., seed region) of, for example, a 20 nucleotide spacer sequence may be 100% complementary to the target nucleic acid, while the remaining nucleotides in the 3' region of the spacer sequence are substantially complementary (e.g., at least about 70% complementary) to the target nucleic acid. In some embodiments, the first 1 to 8 nucleotides (e.g., the first 1, 2, 3, 4, 5, 6, 7, 8, nucleotides, and any range therein) of the 5' end of the spacer sequence may be 100% complementary to the target nucleic acid, while the remaining nucleotides in the 3' region of the spacer sequence are substantially complementary (e.g., at least about 50% complementary (e.g., 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 more)) to the target nucleic acid.

As a further example, in a guide nucleic acid for a Type II CRISPR-Cas system, the first 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 nucleotides in the 3' region (i.e., seed region) of, for example, a 20 nucleotide spacer sequence may be 100% complementary to the target nucleic acid, while the remaining nucleotides in the 5' region of the spacer sequence are substantially complementary (e.g., at least about 70% complementary) to the target nucleic acid. In some embodiments, the first 1 to 10 nucleotides (e.g., the first 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 nucleotides, and any range therein) of the 3' end of the spacer sequence may be 100% complementary to

the target nucleic acid, while the remaining nucleotides in the 5' region of the spacer sequence are substantially complementary (e.g., at least about 50% complementary (e.g., at least 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%, 99%, or more or any range or value therein)) to the target nucleic acid.

In some embodiments, a seed region of a spacer may be about 8 to about 10 nucleotides in length, about 5 to about 6 nucleotides in length, or about 6 nucleotides in length.

In some embodiments, an editing system of the present invention comprises an extended guide nucleic acid, a fusion protein of the present invention, and optionally a reverse transcriptase. In some embodiments, a fusion protein of the present invention comprises all or a portion of a reverse transcriptase. In some embodiments, a fusion protein of the present invention, an extended guide nucleic acid, and optionally a reverse transcriptase may form a complex or may be comprised in a complex that is capable of interacting with a target nucleic acid.

In some embodiments, a guide nucleic acid further comprises a reverse transcriptase template and may be referred to as an extended guide nucleic acid. An "extended guide nucleic acid" as used herein is a guide nucleic acid as described herein that further comprises a reverse transcriptase template (RTT) and/or a primer binding site (PBS). In some embodiments, an extended guide nucleic acid is an engineered prime editing guide RNA (pegRNA). An extended guide nucleic acid may be a targeted allele guide RNA (tagRNA) or a stabilized targeted allele guide RNA (stagRNA). A "tagRNA" as used herein refers to an extended guide nucleic acid that comprises a PBS and a RTT and has target strand complementarity. A "stagRNA" as used herein refers to a tagRNA that comprises a stabilization motif. A stabilization motif may be present at the 3' and/or 5' end of a tagRNA. In some embodiments, a stabilization motif is present at the 3' end of a tagRNA. Exemplary stabilization motifs include, but are not limited to, recruiting motifs, RNA hairpins, pseudoknot sequences, and/or PP7 motifs (e.g., a PP7 RNA hairpin sequence). In some embodiments, a stagRNA is a tagRNA that comprises a PP7 RNA hairpin sequence. In some embodiments, a CRISPR-Cas effector protein (e.g., a Type II or Type V CRISPR-Cas effector protein), a reverse transcriptase, and an extended guide nucleic acid can form a complex or are comprised in a complex.

In some embodiments, an extended guide nucleic acid comprises an extended portion that includes a primer binding site and a reverse transcriptase template, wherein the reverse transcriptase template comprises the modification (e.g., edit) to be incorporated into a target

nucleic acid. In some embodiments, an extended guide nucleic acid comprises, at its 3' end, a primer binding site and a modification (e.g., an edit) to be incorporated into the target nucleic acid (e.g., a reverse transcriptase template). In some embodiments, an extended guide nucleic acid comprises: (1) a sequence that interacts (e.g., recruits and/or binds) with a CRISPR-Cas effector protein (e.g., a CRISPR-Cas nuclease), (2) a spacer having substantial complementarity to a first site on a target nucleic acid (e.g., a CRISPR RNA (crRNA) (a first crRNA) and/or tracrRNA+crRNA (sgRNA)), and (3) a nucleic acid encoded repair template (e.g., an RNA encoded repair template) comprising a primer binding site and an RNA template (e.g., that encodes the modification to be incorporated into the target nucleic acid). In some embodiments, an extended guide nucleic acid (e.g., an extended guide RNA) may comprise, 5'-3', a spacer sequence, a repeat sequence, and an extended portion, the extended portion comprising, 5' to 3', a reverse transcriptase template and a primer binding site. In some embodiments, an extended guide nucleic acid may comprise, 5'-3', a spacer sequence, a repeat sequence and an extended portion, the extended portion comprising, 5' to 3', a primer binding site and a reverse transcriptase template. In some embodiments, an extended guide nucleic acid may comprise, 5'-3', an extended portion, a spacer sequence, and a repeat sequence, wherein the extended portion comprises, 5' to 3', a reverse transcriptase template and a primer binding site. In some embodiments, an extended guide nucleic acid may comprise, 5'-3', an extended portion, a spacer sequence, and a repeat sequence, wherein the extended portion comprises, 5' to 3', a primer binding site and a reverse transcriptase template.

According to some embodiments, an extended guide nucleic acid (e.g., a pegRNA) may have a structure and/or be designed as described in Anzalone et al., *Nature*, 2019 Dec; 576(7785): 149–157. In some embodiments, an extended guide nucleic acid comprises a primer binding site (PBS) optionally having a sequence of 1, 2, 3, 4, or 5 to 6, 7, 8, 9, 10, 11, 12, 13, 14, or 15 nucleotides and a reverse transcriptase template (RT template) sequence optionally having a sequence of 65 nucleotides or more. In some embodiments, a PBS of an extended guide nucleic acid has a sequence of less than 15 nucleotides and has a sequence of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, or 14 nucleotides (e.g., a sequence of 5 or 6 nucleotides in length). The RT template sequence may be after the PBS sequence in the 5' to 3' direction. In some embodiments, the RT template sequence of the extended guide nucleic acid has a length of greater than 65 nucleotides and may comprise about 50 or more nucleotides of heterology relative to the target site (e.g., target nucleic acid), followed by about 15 or more nucleotides of homology relative to the target site. In some embodiments, the RT template sequence of the extended guide nucleic acid is after the PBS sequence and the RT template

sequence has a length of greater than 65 nucleotides with the sequence including more than 50 nucleotides of heterology relative to the target site, followed by more than 15 nucleotides of homology relative to the target site. Accordingly, in some embodiments, when the extended guide nucleic acid is reverse transcribed, the resulting newly transcribed sequence may hybridize and/or is configured to hybridize with the unnicked strand of the target site, which may thereby create a heteroduplex DNA with a large insertion into the newly synthesized strand. Upon repair of this mismatched DNA, the resultant repaired DNA may contain a large insertion (e.g., greater than 50 nucleotides) of DNA sequence. In some embodiments, the method may provide a large deletion (e.g., greater than 50 nucleotides) of DNA sequence. In some embodiments, the PBS and the 15 or more nucleotides of homology to the target site may comprise homology arms, which may serve to insert the heterology into the target site optionally using homology directed repair. The inserted DNA may correspond to any functional sequence of DNA such as, but not limited to: a functional transgene; a fragment of DNA that is inserted into a gene in a way that, when the gene is transcribed, would produce a hairpin RNA that is sufficient to silence homologous genes through RNAi; and/or one or more functional site-specific recombination sites, e.g. lox, frt, which could then be used in subsequent Cre or Flp mediated site-specific recombination processes. In some embodiments, an extended guide nucleic acid may be too large to produce using a PolIII promoter *in vivo*. In some embodiments, an extended guide nucleic acid may be operatively associated with and/or produced using a PolII promoter. In some embodiments, a DNA binding polypeptide (e.g., a DNA binding domain) and/or DNA endonuclease may have a structure and/or be designed as described in Anzalone et al., Nature, 2019 Dec; 576(7785): 149–157. In some embodiments, a DNA binding domain and/or DNA endonuclease is a CRISPR Cas polypeptide such as a Cas9 nickase, a nicking variant of another CRISPR Cas polypeptide, or Cas12a.

In some embodiments, two extended guide nucleic acids (e.g., pegRNAs) may be used (e.g., an editing system may comprise two extended guide nucleic acids). One or both of the two extended guide nucleic acids may have a structure and/or be designed as described in Anzalone et al., Nature, 2019 Dec; 576(7785): 149–157. The two extended guide nucleic acids may comprise a primer binding site (PBS) optionally having a sequence of 1, 2, 3, 4, or 5 to 6, 7, 8, 9, 10, 11, 12, 13, 14, or 15 nucleotides and a reverse transcriptase template (RT template) sequence optionally having a sequence of 50 nucleotides or more. The RT template sequences of the two extended guide nucleic acids may be complementary to each other and as such the polynucleotides that are respectively reverse transcribed from each the RT templates will be complementary to each other and will be able to hybridize with each other. This may allow for

the intermediates that are produced by this system and/or method to join together two sections of DNA that are otherwise separated by more than 50 nucleotides, e.g. within a chromosome, or that are positioned on two separate pieces of DNA, e.g. on two different chromosomes. After repair of the intermediates, the resultant products may produce, depending on the design of the RT template, large deletions, large inversions, or inter-chromosomal recombinations. Since all of these products are produced by homology directed repair, the products may be predictably precise and/or reproducible. In some embodiments, a DNA binding polypeptide (e.g., a DNA binding domain) and/or DNA endonuclease may have a structure and/or be designed as described in Anzalone et al., *Nature*, 2019 Dec; 576(7785): 149–157. In some embodiment, a DNA binding polypeptide and/or DNA endonuclease is a CRISPR Cas polypeptide such as a Cas9 nickase, a similar nicking variant of another CRISPR Cas polypeptide, or Cas12a. In some embodiments, a DNA binding polypeptide and/or DNA endonuclease is a Cas9 nuclease, a similar nuclease from another CRISPR Cas polypeptide, or Cas12a. Using a nuclease (rather than a nickase) may facilitate the intra- or interchromosomal recombination processes through single-strand annealing of the more than 50 nucleotide 3' overhangs that would be produced at each of the two target sites corresponding to the two pegRNA target nucleic acids. In some embodiments, an editing system comprises one extended guide nucleic acid and a guide nucleic acid that is devoid of a reverse transcriptase template and/or primer binding site.

An extended guide nucleic acid may comprise a CRISPR nucleic acid (e.g., CRISPR RNA, CRISPR DNA, crRNA, crDNA) and/or a CRISPR nucleic acid and a tracr nucleic acid; and (b) an extended portion comprising a primer binding site and a reverse transcriptase template (RT template), wherein the RT template encodes a modification to be incorporated into the target nucleic acid. The CRISPR nucleic acid may be a Type II or Type V CRISPR nucleic acid and/or the tracr nucleic acid may be any tracr corresponding to the appropriate Type II or Type V CRISPR nucleic acid. In some embodiments, an extended guide nucleic acid comprises: (i) a Type V CRISPR nucleic acid or a Type II CRISPR nucleic acid (e.g., a Type II or Type V CRISPR RNA, Type II or Type V CRISPR DNA, Type II or Type V crRNA, or Type II or Type V crDNA) and/or a CRISPR nucleic acid and a tracr nucleic acid (e.g., a Type II or Type V tracrRNA, Type II or Type V tracrDNA); and (ii) an extended portion comprising a primer binding site and a reverse transcriptase template (RT template), wherein the Type V CRISPR nucleic acid or Type II CRISPR nucleic acid comprises a spacer that binds to a first strand (e.g., the target strand) of a target nucleic acid (e.g., the spacer is complementary to a portion of consecutive nucleotides in the first strand of the target nucleic acid) and the primer binding site binds to the first strand (e.g., target strand). In some embodiments, the

extended portion can be fused to either the 5' end or 3' end of the CRISPR nucleic acid (e.g., from 5' to 3': repeat-spacer-extended portion or extended portion-repeat-spacer) and/or to the 5' or 3' end of the tracr nucleic acid. In some embodiments, the extended portion of an extended guide nucleic acid comprises, 5' to 3', an RT template (RTT) and a primer binding site (PBS) (e.g., 5'-crRNA-spacer-RTT(edit encoded)-PBS-3') or comprises 5' to 3' a PBS and RTT, depending on the location of the extended portion relative to the CRISPR nucleic acid of the extended guide nucleic acid (e.g., 5'-crRNA-spacer-PBS-RTT(edit encoded)-3'). For example, in some embodiments, an extended portion of the extended guide nucleic acid may comprise, 5' to 3', an RT template and a primer binding site (when the extended guide is linked to the 3' end of the CRISPR nucleic acid). In some embodiments, an extended portion of the extended guide may comprise, 5' to 3', a primer binding site and an RT template (when the extended guide is linked to the 5' end of the CRISPR nucleic acid).

In some embodiments, a target nucleic acid is double stranded and comprises a first strand and a second strand and a primer binding site of an extended guide nucleic acid binds to the second strand (e.g., the non-target, top strand) of the target nucleic acid. In some embodiments, a target nucleic acid is double stranded and comprises a first strand and a second strand and a primer binding site of an extended guide nucleic acid binds to the first strand (e.g., binds to the target strand, optionally the same strand to which a CRISPR-Cas effector protein is recruited, bottom strand) of the target nucleic acid. In some embodiments, a target nucleic acid is double stranded and comprises a first strand and a second strand and the primer binding site of an extended guide nucleic acid binds to the second strand (e.g., the non-target strand, optionally the opposite strand from that to which the CRISPR-Cas effector protein is recruited) of the target nucleic acid. In some embodiments, a reverse transcriptase (RT) may add to the target strand of a target nucleic acid (e.g., the strand to which the spacer of the CRISPR nucleic acid of the extended guide nucleic acid is complementary and to which the CRISPR-Cas effector protein is recruited). In some embodiments, the reverse transcriptase (RT) adds to the non-target strand of a target nucleic acid (e.g., the strand that is complementary to the strand to which the spacer of the CRISPR nucleic acid is complementary and to which the CRISPR-Cas effector protein is recruited). Example methods and editing systems are described in International Patent Publication No. WO 2021/092130, International Patent Publication No. WO 2022/098993, and U.S. Patent Application Publication Nos. 2021/0147862, 2021/0130835, 2021/0147862, and 2022/0145334, each of which are incorporated herein by reference in their entirety.

The RT template of an extended guide nucleic acid may encode one or more modification(s) (e.g., edit(s)) to be incorporated into a target nucleic acid. The one or more modification(s) may be located in any position within an RT template (e.g., where the position location may be relative to the position of a protospacer adjacent motif (PAM) of the target nucleic acid). In some embodiments, an RT template has a modification at one or more positions from -1 to 23 (e.g., -1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, or 23) relative to the position of a protospacer adjacent motif (PAM) (e.g., TTTG) in a target nucleic acid. In some embodiments, an RT template may comprise a modification located at nucleotide position -1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, or 23. In some embodiments, an RT template may comprise a modification located at nucleotide position 4 to nucleotide position 17 (e.g., position 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, or 17) of the RT template relative to the position of a PAM of a target nucleic acid. In some embodiments, an RT template may comprise a modification located at nucleotide position 10 to nucleotide position 17 (e.g., position 10, 11, 12, 13, 14, 15, 16, or 17) of the RT template relative to the position of a PAM of a target nucleic acid. In some embodiments, an RT template may comprise a modification located at nucleotide position 12 to nucleotide position 15 (e.g., position 12, 13, 14, or 15) of the RT template relative to the position of a PAM of a target nucleic acid.

In some embodiments, an extended portion of an extended guide nucleic acid may comprise, 5' to 3', an RT template and a primer binding site (e.g., when the extended portion is linked to the 3' end of a CRISPR nucleic acid). In some embodiments, an extended portion of an extended guide nucleic acid may comprise, 5' to 3', a primer binding site and an RT template (RTT) (e.g., when the extended portion is linked to the 5' end of the CRISPR nucleic acid). In some embodiments, an RT template may have a length of about 1 nucleotide to about 100 nucleotides (e.g., about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 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, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or more nucleotides, and any range or value therein), e.g., about 1 nucleotide to about 10 nucleotides, about 1 nucleotide to about 15 nucleotides, about 1 nucleotide to about 20 nucleotides, about 1 nucleotide to about 25 nucleotides, about 1 nucleotide to about 30 nucleotides, about 1 nucleotide to about 35, 36, 37, 38, 39 or 40 nucleotides, about 1 nucleotide to about 50 nucleotides, about 5 nucleotides to about 15 nucleotides, about 5 nucleotides to about 20 nucleotides, about 5 nucleotides to about 25 nucleotides, about 5 nucleotides to about

30 nucleotides, about 5 nucleotides to about 35, 36, 37, 38, 39 or 40 nucleotides, about 5 nucleotides to about 50 nucleotides, about 8 nucleotides to about 15 nucleotides, about 8 nucleotide to about 20 nucleotides, about 8 nucleotide to about 25 nucleotides, about 8 nucleotide to about 30 nucleotides, about 8 nucleotide to about 35, 36, 37, 38, 39 or 40 nucleotides, about 8 nucleotide to about 50 nucleotides in length, about 8 nucleotides to about 100 nucleotides, about 10 nucleotide to about 15 nucleotides, about 10 nucleotide to about 20 nucleotides, about 10 nucleotide to about 25 nucleotides, about 10 nucleotide to about 30 nucleotides, about 10 nucleotide to about 36 nucleotides, about 10 nucleotide to about 40 nucleotides, about 10 nucleotide to about 50 nucleotides, about 10 nucleotides to about 100 nucleotides in length and any range or value therein. In some embodiments, the length of an RT template may be at least 8 nucleotides, optionally about 8 nucleotides to about 100 nucleotides. In some embodiments, the length of an RT template is 36, 37, 38, 39 or 40 nucleotides or less (e.g., about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40 nucleotides in length, or any value or range therein (e.g., about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, or 15 nucleotides in length to about 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40 nucleotides in length)). In some embodiments, the length of an RT template may be at least 30 nucleotides, optionally about 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40 nucleotides in length to about to about 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, 79, or 80 nucleotides in length, or any range or value therein. In some embodiments, the length of an RT template may be about 36, 40, 44, 47, 50, 52, 55, 63, 72 or 74 nucleotides. Within the length of the RTT one or more modification(s) may be present. The one or more modification(s) may be located anywhere within the RTT, wherein the position of the modification may be described relative to the position of a protospacer adjacent motif (PAM) of a target nucleic acid. In some embodiments, an RT template may comprise a modification located at nucleotide position -1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, or 23. In some embodiments, an RT template may comprise a modification located at nucleotide position 4 to nucleotide position 17 (e.g., position 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, or 17) of the RT template relative to the position of a protospacer adjacent motif (PAM) of a target nucleic acid. In some embodiments, an RT template may comprise a modification located at nucleotide position 10 to nucleotide position 17 (e.g., position 10, 11, 12, 13, 14, 15, 16, or 17) of the RT template relative to the position of a protospacer adjacent motif (PAM) of a target nucleic acid. In some embodiments, an RT template may comprise a

modification located at nucleotide position 12 to nucleotide position 15 (e.g., position 12, 13, 14, or 15) of the RT template relative to the position of a protospacer adjacent motif (PAM) of a target nucleic acid.

As used herein, a "primer binding site" (PBS) of an extended portion of an extended guide nucleic acid (e.g., a tagRNA) refers to a sequence of consecutive nucleotides that can bind to a region or "primer" on a target nucleic acid, e.g., is complementary to the target nucleic acid primer. As an example, a CRISPR Cas effector protein (e.g., a Type II or Type V, e.g., Cas 9 or Cas12a) may nick/cut the DNA and the 3' end of the cut DNA acts as a primer for the PBS portion of the extended guide nucleic acid. The PBS may be complementary to the 3' end of a strand of the target nucleic acid and may bind and/or may be configured to bind to either the target strand or non-target strand. A primer binding site can be fully complementary to the primer or it may be substantially complementary (e.g., at least 70% complementary (e.g., 70% or about 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%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9% or more)) to the primer of a target nucleic acid. In some embodiments, the length of a primer binding site of an extended portion may be about 1 nucleotide to about 100 nucleotides in length (e.g., about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 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, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or more nucleotides, or any value or range therein), or about 4 nucleotide to about 85 nucleotides, about 10 nucleotide to about 80 nucleotides, about 20 nucleotide to about 80 nucleotides, about 25 nucleotides to about 80 nucleotides about 30 nucleotide to about 80 nucleotides, about 40 nucleotide to about 80 nucleotides, about 45 nucleotide to about 80 nucleotides, about 45 nucleotide to about 75 nucleotides, or about 45 nucleotide to about 60 nucleotides, or any range or value therein. In some embodiments, the length of a PBS may be at least 30 nucleotides, optionally about 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40 nucleotides to about 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, 79, or 80 nucleotides in length, or any range or value therein. In some embodiments, the length of a PBS may be about 8, 16, 24, 32, 40, 48, 56, 64, 72, or 80 nucleotides.

In some embodiments, an RTT may have a length of about 35 nucleotides to about 75 nucleotides and a PBS may have a length of about 30 nucleotides to about 80 nucleotides, optionally wherein the PBS may comprise a length of about 8, 16, 24, 32, 40, 48, 56, 64, 72,

or 80 nucleotides and the RTT may comprise a length of about 36, 40, 44, 47, 50, 52, 55, 63, 72 or 74 nucleotides, or any combination thereof of the RTT length and/or PBS length.

In some embodiments, an extended portion of an extended guide nucleic acid may be fused to either the 5' end or 3' end of a Type II or a Type V CRISPR nucleic acid (e.g., 5' to 3': repeat-spacer-extended portion, or extended portion-repeat-spacer) and/or to the 5' or 3' end of the tracr nucleic acid. In some embodiments, when an extended portion is located 5' of the crRNA, a Type V CRISPR-Cas effector protein is modified to reduce (or eliminate) self-processing RNase activity.

In some embodiments, the extended portion of an extended guide nucleic acid may be linked to the Type II or Type V CRISPR nucleic acid and/or the Type II or Type V tracrRNA via a linker. In some embodiments, a linker have a length of about 1 to about 100 nucleotides or more (e.g., about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 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, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or more nucleotides in length, and any range therein (e.g., about 2 to about 40, about 2 to about 50, about 2 to about 60, about 4 to about 40, about 4 to about 50, about 4 to about 60, about 5 to about 40, about 5 to about 50, about 5 to about 60, about 9 to about 40, about 9 to about 50, about 9 to about 60, about 10 to about 40, about 10 to about 50, about 10 to about 60, about 40 to about 100, about 50 to about 100, or about 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 nucleotides to about 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, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or more nucleotides in length (e.g., about 105, 110, 115, 120, 130, 140 150 or more nucleotides in length).

A guide nucleic acid and/or an extended guide nucleic acid may comprise one or more recruiting motifs as described herein, which may be linked to the 5' end and/or the 3' end of the guide nucleic acid and/or it may be inserted into the guide nucleic acid (e.g., within a hairpin loop of the guide nucleic acid). In some embodiments, an extended guide nucleic acid may be linked to an RNA recruiting motif. An extended guide nucleic acid and/or guide nucleic acid may be linked to one or to two or more RNA recruiting motifs (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more motifs; e.g., at least 10 to about 25 motifs), optionally wherein the two or more RNA recruiting motifs may be the same RNA recruiting motif or different RNA recruiting motifs. In some embodiments, an RNA recruiting motif may be located on the 3' end of the extended

portion of an extended guide nucleic acid (e.g., 5'-3', repeat-spacer-extended portion (RT template-primer binding site)-RNA recruiting motif). In some embodiments, an RNA recruiting motif may be embedded in the extended portion of an extended guide nucleic acid.

In some embodiments, an editing system comprises an extended guide nucleic acid that is linked to an RNA recruiting motif and a reverse transcriptase that is a reverse transcriptase fusion protein, wherein the reverse transcriptase fusion protein comprises a reverse transcriptase polypeptide fused to an affinity polypeptide that binds to the RNA recruiting motif, wherein the extended guide nucleic acid binds to a target nucleic acid and the RNA recruiting motif binds to the affinity polypeptide, thereby recruiting the reverse transcriptase fusion protein to the extended guide nucleic acid and contacting the target nucleic acid with the reverse transcriptase. In some embodiments, two or more reverse transcriptase fusion proteins may be recruited to an extended guide nucleic acid, thereby contacting the target nucleic acid with two or more reverse transcriptase fusion proteins.

A “target nucleic acid”, “target DNA,” “target nucleotide sequence,” “target region,” and “target region in the genome” are used interchangeably herein and refer to a region of an organism’s (e.g., a plant’s) genome that comprises a sequence that is fully complementary (100% complementary) or substantially complementary (e.g., at least 70% complementary (e.g., 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 more)) to a spacer sequence in a guide nucleic acid as defined herein. A target nucleic acid is targeted by an editing system (or a component thereof) as described herein. A target region useful for a CRISPR-Cas system may be located immediately 3’ (e.g., Type V CRISPR-Cas system) or immediately 5’ (e.g., Type II CRISPR-Cas system) to a PAM sequence in the genome of the organism (e.g., a plant genome or mammalian (e.g., human) genome). A target region may be selected from any region of at least 15 consecutive nucleotides (e.g., 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 nucleotides, and the like) located immediately adjacent to a PAM sequence.

A “protospacer sequence” or “protospacer” as used herein refer to a sequence that is fully or substantially complementary to (and can hybridize to) a spacer sequence of a guide nucleic acid. In some embodiments, the protospacer is all or a portion of a target nucleic acid as defined herein that is fully or substantially complementary (and hybridizes) to the spacer sequence of the CRISPR repeat-spacer sequences (e.g., guide nucleic acids, CRISPR arrays, crRNAs).

In the case of Type V CRISPR-Cas (e.g., Cas12a) systems and Type II CRISPR-Cas (Cas9) systems, the protospacer sequence is flanked by (e.g., immediately adjacent to) a protospacer adjacent motif (PAM). For Type IV CRISPR-Cas systems, the PAM is located at the 5' end on the non-target strand and at the 3' end of the target strand (see below, as an example).

5'-NNNNNNNNNNNNNNNNNNNN-3' RNA Spacer

3'AAANNNNNNNNNNNNNNNNNNNNN-5' Target strand

|||

5'TTTNNNNNNNNNNNNNNNNNNNN-3' Non-target strand

In the case of Type II CRISPR-Cas (e.g., Cas9) systems, the PAM is located immediately 3' of the target region. The PAM for Type I CRISPR-Cas systems is located 5' of the target strand. There is no known PAM for Type III CRISPR-Cas systems. Makarova et al. describes the nomenclature for all the classes, types and subtypes of CRISPR systems (*Nature Reviews Microbiology* 13:722–736 (2015)). Guide structures and PAMs are described in R. Barrangou (*Genome Biol.* 16:247 (2015)).

Canonical Cas12a PAMs are T rich. In some embodiments, a canonical Cas12a PAM sequence may be 5'-TTN, 5'-TTTN, or 5'-TTTV. In some embodiments, canonical Cas9 (e.g., *S. pyogenes*) PAMs may be 5'-NGG-3'. In some embodiments, non-canonical PAMs may be used but may be less efficient.

Additional PAM sequences may be determined by those skilled in the art through established experimental and computational approaches. Thus, for example, experimental approaches include targeting a sequence flanked by all possible nucleotide sequences and identifying sequence members that do not undergo targeting, such as through the transformation of target plasmid DNA (Esvelt et al. 2013. *Nat. Methods* 10:1116-1121; Jiang et al. 2013. *Nat. Biotechnol.* 31:233-239). In some aspects, a computational approach can include performing BLAST searches of natural spacers to identify the original target DNA sequences in bacteriophages or plasmids and aligning these sequences to determine conserved sequences adjacent to the target sequence (Briner and Barrangou. 2014. *Appl. Environ. Microbiol.* 80:994-1001; Mojica et al. 2009. *Microbiology* 155:733-740).

In some embodiments, the present invention provides expression cassettes and/or vectors comprising the nucleic acid constructs of the invention (e.g., one or more components of an editing system of the invention). In some embodiments, expression cassettes and/or vectors comprising the nucleic acid constructs of the invention and/or one or more guide

nucleic acids may be provided. In some embodiments, a nucleic acid construct of the invention encodes a fusion protein, and/or a deaminase, and each may be comprised on the same or on a separate expression cassette or vector from that comprising the one or more guide nucleic acids. When the nucleic acid construct encoding a fusion protein or the components of an editing system is/are comprised on separate expression cassette(s) or vector(s) from that comprising the guide nucleic acid, a target nucleic acid may be contacted with (e.g., provided with) the expression cassette(s) or vector(s) encoding the fusion protein or components of an editing system in any order from one another and the guide nucleic acid, e.g., prior to, concurrently with, or after the expression cassette comprising the guide nucleic acid is provided (e.g., contacted with the target nucleic acid).

Methods of recruiting one or more components of an editing system to each other and/or to a target nucleic acid are known in the art and may include the use of a peptide tag or an affinity polypeptide that interacts with the peptide tag. In some embodiments, a guide nucleic acid may be linked to an RNA recruiting motif and a deaminase may be linked to an affinity polypeptide capable of interacting with the RNA recruiting motif, thereby recruiting the deaminase to the target nucleic acid. Alternatively, chemical interactions may be used to recruit a polypeptide (e.g., a deaminase) to a target nucleic acid.

A “recruiting motif” as used herein refers to one half of a binding pair that may be used to recruit a compound to which the recruiting motif is bound to another compound that includes the other half of the binding pair (i.e., a “corresponding motif”). The recruiting motif and corresponding motif may bind noncovalently. In some embodiments, a recruiting motif is an RNA recruiting motif (e.g., an RNA recruiting motif that is capable of binding and/or configured to bind to an affinity polypeptide), an affinity polypeptide (e.g., an affinity polypeptide that is capable of binding and/or configured to bind an RNA recruiting motif and/or a peptide tag), or a peptide tag (e.g., a peptide tag that is capable of binding and/or configured to bind an affinity polypeptide). For example, when a recruiting motif is an RNA recruiting motif, the corresponding motif for the RNA recruiting motif may be an affinity polypeptide that binds the RNA recruiting motif. A further example is that when a recruiting motif is a peptide tag, the corresponding motif for the peptide tag may be an affinity polypeptide that binds the peptide tag. Thus, a compound comprising a recruiting motif (e.g., an affinity polypeptide) may be recruited to another compound (e.g., a guide nucleic acid) comprising a corresponding motif for the recruiting motif (e.g., an RNA recruiting motif).

A peptide tag (e.g., epitope) useful with this invention may include, but is not limited to, a GCN4 peptide tag (e.g., Sun-Tag), a c-Myc affinity tag, an HA affinity tag, a His affinity

tag, an S affinity tag, a methionine-His affinity tag, an RGD-His affinity tag, a FLAG octapeptide, a strep tag or strep tag II, a V5 tag, and/or a VSV-G epitope. Any epitope that may be linked to a polypeptide and for which there is a corresponding affinity polypeptide that may be linked to another polypeptide may be used with this invention as a peptide tag. In some embodiments, a peptide tag may comprise 1 or 2 or more copies of a peptide tag (e.g., repeat unit, multimerized epitope (e.g., tandem repeats)) (e.g., 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 or more repeat units. In some embodiments, an affinity polypeptide that interacts with/binds to a peptide tag may be an antibody. In some embodiments, the antibody may be a scFv antibody. In some embodiments, an affinity polypeptide that binds to a peptide tag may be synthetic (e.g., evolved for affinity interaction) including, but not limited to, an affibody, an anticalin, a monobody and/or a DARPin (see, e.g., Sha et al., *Protein Sci.* 26(5):910-924 (2017)); Gilbreth (*Curr Opin Struc Biol* 22(4):413-420 (2013)), U.S. Patent No. 9,982,053, each of which are incorporated by reference in their entireties for the teachings relevant to affibodies, anticalins, monobodies and/or DARPins.

In some embodiments, a guide nucleic acid may be linked to an RNA recruiting motif, and a polypeptide to be recruited (e.g., a deaminase) may be fused to an affinity polypeptide that binds to the RNA recruiting motif, wherein the guide binds to the target nucleic acid and the RNA recruiting motif binds to the affinity polypeptide, thereby recruiting the polypeptide to the guide and contacting the target nucleic acid with the polypeptide (e.g., deaminase). In some embodiments, two or more polypeptides may be recruited to a guide nucleic acid, thereby contacting the target nucleic acid with two or more polypeptides (e.g., deaminases).

In some embodiments of the invention, a guide RNA may be linked to one or to two or more RNA recruiting motifs (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more motifs; e.g., at least 10 to about 25 motifs), optionally wherein the two or more RNA recruiting motifs may be the same RNA recruiting motif or different RNA recruiting motifs. In some embodiments, an RNA recruiting motif and corresponding affinity polypeptide may include, but is not limited, to a telomerase Ku binding motif (e.g., Ku binding hairpin) and the corresponding affinity polypeptide Ku (e.g., Ku heterodimer), a telomerase Sm7 binding motif and the corresponding affinity polypeptide Sm7, an MS2 phage operator stem-loop and the corresponding affinity polypeptide MS2 Coat Protein (MCP), a PP7 phage operator stem-loop and the corresponding affinity polypeptide PP7 Coat Protein (PCP), an SfMu phage Com stem-loop and the corresponding affinity polypeptide Com RNA binding protein, a PUF binding site (PBS) and the affinity polypeptide Pumilio/fem-3 mRNA binding factor (PUF), and/or a synthetic RNA-aptamer and the aptamer ligand as the corresponding affinity polypeptide. In some

embodiments, the RNA recruiting motif and corresponding affinity polypeptide may be an MS2 phage operator stem-loop and the affinity polypeptide MS2 Coat Protein (MCP). In some embodiments, the RNA recruiting motif and corresponding affinity polypeptide may be a PUF binding site (PBS) and the affinity polypeptide Pumilio/fem-3 mRNA binding factor (PUF). Exemplary RNA recruiting motifs and corresponding affinity polypeptides that may be useful with this invention can include, but are not limited to, **SEQ ID NOs:86-96**.

In some embodiments, the components for recruiting polypeptides and nucleic acids may include those that function through chemical interactions that may include, but are not limited to, rapamycin-inducible dimerization of FRB – FKBP; Biotin-streptavidin; SNAP tag; Halo tag; CLIP tag; DmrA-DmrC heterodimer induced by a compound; bifunctional ligand (e.g., chemically induced dimerization).

As described herein, a “peptide tag” may be employed to recruit one or more polypeptides. A peptide tag may be any polypeptide that is capable of being bound by a corresponding motif such as an affinity polypeptide. A peptide tag may also be referred to as an "epitope" and when provided in multiple copies, a "multimerized epitope." Example peptide tags can include, but are not limited to, a GCN4 peptide tag (e.g., Sun-Tag), a c-Myc affinity tag, an HA affinity tag, a His affinity tag, an S affinity tag, a methionine-His affinity tag, an RGD-His affinity tag, a FLAG octapeptide, a strep tag or strep tag II, a V5 tag, and/or a VSV-G epitope. In some embodiments, a peptide tag may also include phosphorylated tyrosines in specific sequence contexts recognized by SH2 domains, characteristic consensus sequences containing phosphoserines recognized by 14-3-3 proteins, proline rich peptide motifs recognized by SH3 domains, PDZ protein interaction domains or the PDZ signal sequences, and an AGO hook motif from plants. Peptide tags are disclosed in WO2018/136783 and U.S. Patent Application Publication No. 2017/0219596, which are incorporated by reference for their disclosures of peptide tags. Peptide tags that may be useful with this invention can include, but are not limited to, **SEQ ID NO:97** and **SEQ ID NO:98**. An affinity polypeptide useful with peptide tags includes, but is not limited to, **SEQ ID NO:99**.

A peptide tag may comprise or be present in one copy or in 2 or more copies of the peptide tag (e.g., multimerized peptide tag or multimerized epitope) (e.g., about 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 9, 20, 21, 22, 23, 24, or 25 or more peptide tags). When multimerized, the peptide tags may be fused directly to one another or they may be linked to one another via one or more amino acids (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more amino acids, optionally about 3 to about 10, about 4 to about 10, about 5 to about 10, about 5 to about 15, or about 5 to about 20 amino acids, and the like, and any

value or range therein. Thus, in some embodiments, a CRISPR-Cas effector protein of the invention may comprise a CRISPR-Cas effector protein fused to one peptide tag or to two or more peptide tags, optionally wherein the two or more peptide tags are fused to one another via one or more amino acid residues. In some embodiments, a peptide tag useful with the invention may be a single copy of a GCN4 peptide tag or epitope or may be a multimerized GCN4 epitope comprising about 2 to about 25 or more copies of the peptide tag (e.g., about 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 or more copies of a GCN4 epitope or any range therein).

In some embodiments, a peptide tag may be fused to a CRISPR-Cas polypeptide or domain. In some embodiments, a peptide tag may be fused or linked to the C-terminus of a CRISPR-Cas effector protein to form a CRISPR-Cas fusion protein. In some embodiments, a peptide tag may be fused or linked to the N-terminus of a CRISPR-Cas effector protein to form a CRISPR-Cas fusion protein. In some embodiments, a peptide tag may be fused within a CRISPR-Cas effector protein (e.g., a peptide tag may be in a loop region of a CRISPR-Cas effector protein). In some embodiments, peptide tag may be fused to a cytosine deaminase and/or to an adenine deaminase.

An "affinity polypeptide" (e.g., "recruiting polypeptide") refers to any polypeptide that is capable of binding to its corresponding peptide tag, peptide tag, or RNA recruiting motif. An affinity polypeptide for a peptide tag may be, for example, an antibody and/or a single chain antibody that specifically binds the peptide tag, respectively. In some embodiments, an antibody for a peptide tag may be, but is not limited to, an scFv antibody. In some embodiments, an affinity polypeptide may be fused or linked to the N-terminus of a deaminase (e.g., a cytosine deaminase or an adenine deaminase). In some embodiments, the affinity polypeptide is stable under the reducing conditions of a cell or cellular extract.

The nucleic acid constructs of the invention and/or guide nucleic acids may be comprised in one or more expression cassettes as described herein. In some embodiments, a nucleic acid construct of the invention may be comprised in the same or in a separate expression cassette or vector from that comprising a guide nucleic acid and/or an extended guide nucleic acid.

In some embodiments, a nucleic acid construct, expression cassette, or vector of the invention that is optimized for expression in an organism (e.g., a human or plant) may be about 70% to 100% identical (e.g., about 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%, 99.5% or 100%) to a nucleic acid construct, expression cassette or

vector comprising the same polynucleotide(s) but which have not been codon optimized for expression in the organism.

When used in combination a guide nucleic acid, a nucleic acid construct of the invention (and expression cassette and/or vector comprising the same) may be used to modify a target nucleic acid and/or its expression. A target nucleic acid may be contacted with a nucleic acid construct of the invention and/or expression cassettes and/or vectors comprising the same prior to, concurrently with or after contacting the target nucleic acid with the guide nucleic acid/recruiting guide nucleic acid (and/or expression cassettes and vectors comprising the same).

According to embodiments of the present invention, provided herein are fusion proteins (e.g., engineered proteins) that include an intein polypeptide. In some embodiments, a fusion protein of the present invention includes a Cas12a polypeptide and an intein polypeptide. In some embodiments, a fusion protein of the present invention includes a polypeptide of interest and an intein polypeptide. In some embodiments, a fusion protein of the present invention includes a reverse transcriptase polypeptide and an intein polypeptide. An “engineered protein” as used herein is a polypeptide or protein that is not found naturally in nature. A fusion protein of the present invention may comprise a Cas12a polypeptide and/or a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to (e.g., linked and/or attached to) an intein polypeptide. A Cas12a polypeptide and an intein polypeptide may be directly fused (e.g., no amino acid residue or linker between the two polypeptides) or indirectly fused (e.g., a linker (e.g., an amino acid or peptide) or another polypeptide is between the two polypeptides). Similarly, a polypeptide of interest (e.g., a reverse transcriptase polypeptide) and an intein polypeptide may be directly fused or indirectly fused. In some embodiments, a Cas12a polypeptide is directly fused (e.g., via a peptide bond) to an intein polypeptide. In some embodiments, a Cas12a polypeptide is indirectly fused (e.g., via a peptide linker) to an intein polypeptide. A Cas12a polypeptide and an intein polypeptide may be fused in any orientation. For example, in some embodiments, the N-terminus of the Cas12a polypeptide is fused to the C-terminus of the intein polypeptide or to the N-terminus of the intein polypeptide. In some embodiments, the C-terminus of the Cas12a polypeptide is fused to the C-terminus of the intein polypeptide or to the N-terminus of the intein polypeptide. A polypeptide of interest (e.g., a reverse transcriptase polypeptide) and an intein polypeptide may be fused in any orientation. In some embodiments, a fusion protein of the present invention has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:100-109 or 187-188**. According to some embodiments, provided is a nucleic acid

molecule that encodes a polypeptide that has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:100-109 or 187-188**. In some embodiments, a nucleic acid molecule comprises a polynucleotide that has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:185-186**.

In some embodiments, an intein polypeptide is a portion (e.g., a fragment such as a N-terminal intein fragment or a C-terminal intein fragment) of an intein such as a portion of a molecular scaffold formed from two corresponding portions (e.g., two corresponding fragments or a pair of intein polypeptides) that together can or are configured to catalyze both the cleavage and formation of a peptide bond. An “intein” as used herein refers to a catalytically active complex formed from two polypeptides (e.g., a pair of intein polypeptides) that are associated with each other, wherein the complex can or is configured to excise itself (e.g., the two polypeptides) from a larger precursor polypeptide and can or is configured to ligate the ends flanking the two polypeptides with a peptide bond, optionally wherein the excising and ligating occur concurrently. In some embodiments, an intein polypeptide is a portion of a split intein such as a trans-splicing splint intein. A “split intein” as used herein can perform protein trans-splicing in which two fragments of the intein (e.g., two intein polypeptides or a pair of intein polypeptides) associate (e.g., non-covalently bind) to form a catalytically competent complex or molecular scaffold that catalyzes excision of the two intein fragments and the ligation of their flanking sequences. In some embodiments, an intein polypeptide is an auto-catalytic polypeptide that together with a corresponding intein polypeptide to form an intein (e.g., a split intein) is capable of excising the intein polypeptide from a larger precursor protein (e.g., a fusion protein of the present invention) and enable the flanking polypeptide sequences (e.g., the sequence adjacent to the excised intein polypeptide) to be ligated through the formation of a new peptide bond. A fusion protein of the present invention may include an intein polypeptide that is one part of two total parts such that the intein polypeptide together with another intein polypeptide (e.g., the second part) together form an intein such as a trans-splicing split intein. In some embodiments, a split intein and/or an intein polypeptide thereof may be able to function (e.g., perform protein trans-splicing) without any assistance and/or conditions other than the two portions of the split intein (e.g., the two intein polypeptides that together provide the split intein). For example, two intein polypeptides may spontaneously associate to form the intein and may spontaneously catalyze their own excision and the ligation of their flanking sequences without assistance (e.g., an external condition and/or cofactor). In some embodiments, an intein may be used for which protein

trans-splicing is controlled (e.g., the intein undergoes conditional trans-splicing). For example, certain conditions (e.g., light and/or a cofactor) may be required for an intein to function. In some embodiments, a split intein and/or an intein polypeptide is a light inducible intein (e.g., as described in Wong S, et al. (2015) An Engineered Split Intein for Photoactivated Protein Trans-Splicing. PLoS ONE 10(8): e0135965) that uses light in order to control the association of the two intein polypeptides that together provide the intein (e.g., the catalytically active complex). In some embodiments, a cofactor (e.g., a small molecule) and/or activator is used to bring together two intein polypeptides that together provide the intein and thereby control protein trans-splicing such as described in Gramespacher, Josef A., et al. *J Am Chem Soc.* 2019 Sep 4; 141(35): 13708–13712. In some embodiments, an intein polypeptide of a fusion protein of the present invention may be configured to be removed (e.g., excised) from the fusion protein and fused with another intein polypeptide (e.g., an intein polypeptide that is a portion of a different fusion protein of the present invention) *in situ* and/or *in vivo*.

In some embodiments, an intein of the present invention is an intein present in a DNA polymerase III gene (DnaE) in cyanobacteria and/or an intein as described in Pinto, F., Thornton, E.L. & Wang, B. An expanded library of orthogonal split inteins enables modular multi-peptide assemblies. *Nat Commun* 11, 1529 (2020). Further exemplary inteins include, but are not limited, *Nostoc punctiforme* (Npu) inteins and mutants thereof (e.g., NpuGEP, a mutant that contains 3 amino acid residue mutations). In some embodiments, an intein polypeptide is a portion of a *Nostoc punctiforme* (Npu) intein and/or a portion of a mutant Npu intein (e.g., a portion of NpuGEP). An intein polypeptide may be an N-terminal portion of an intein in that the intein polypeptide includes the N-terminus of the full-length intein and/or active complex. In some embodiments, an intein polypeptide may be a C-terminal portion of an intein in that the intein polypeptide includes the C-terminus of the full-length intein and/or active complex. An intein polypeptide that is an N-terminal portion of an intein and an intein polypeptide that contains the remaining portion of the intein (e.g., the C-terminal portion of the intein) together are an intein pair and form the intein and/or the active complex. An intein polypeptide of the present invention may have at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:110-112**.

A Cas12a polypeptide of the present invention may be a portion of a Cas12a protein, optionally a portion of a Cas12a fusion protein. In some embodiments, a Cas12a protein and/or Cas12a fusion protein may be a protein as described in U.S. Patent Application Publication No. 2022/0112473, the contents of which are incorporated herein by reference in its entirety. In some embodiments, a Cas12a polypeptide is portion of a sequence of **SEQ ID NO:38-60, 113-**

149, 192-195, or 196-259. In some embodiments, a Cas12a polypeptide is about 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, or more consecutive amino acids of the sequence of **SEQ ID NO:38-60, 113-149, 192-195, or 196-259.** In some embodiments, a Cas12a polypeptide is one part of two total parts that together form the Cas12a protein. For example, a first Cas12a polypeptide that is present in a first fusion protein of the present invention together with a second Cas12a polypeptide that is present in a second fusion protein of the present invention together form a Cas12a protein. In some embodiments, a Cas12a polypeptide is a N-terminal portion of a Cas12a protein in that the Cas12a polypeptide includes the N-terminus of the full-length Cas12a protein. In some embodiments, a Cas12a polypeptide is a C-terminal portion of a Cas12a protein in that the Cas12a polypeptide includes the C-terminus of the full-length Cas12a protein. Two Cas12a polypeptides that are individually present in two different fusion proteins of the present invention may, upon fusing together of the two Cas12a polypeptides, provide a Cas12a protein that is part of an editing system as described herein such as a CRISPR-Cas editing system. The editing system may be used to modify a target nucleic acid. In some embodiments, a fusion protein of the present invention comprises a Cas12a polypeptide that is a N-terminal portion of a Cas12a protein and an intein polypeptide that is a N-terminal portion of an intein, wherein the intein polypeptide is at the C-terminus of the fusion protein and/or at the C-terminus of the Cas-12a polypeptide. In some embodiments, a fusion protein of the present invention comprises a Cas12a polypeptide that is a C-terminal portion of a Cas12a protein and an intein polypeptide that is a C-terminal portion of an intein, wherein the intein polypeptide is at the N-terminus of the fusion protein and/or at the N-terminus of the Cas-12a polypeptide.

In some embodiments, a Cas12a protein is split into two portions and a fusion protein of the present invention comprises one of the portions. For example, a Cas12a protein may be split into two portions between amino acid residues 173 and 174, 174 and 175, 175 and 176, 309 and 310, 310 and 311, 405 and 406, 406 and 407, 440 and 441, 441 and 442, 549 and 550, or 550 and 551 such that a Cas12a polypeptide includes 173, 174, 175, 309, 310, 405, 406, 440, 441, 549, or 550 consecutive amino acids of a Cas12a protein and another Cas12a polypeptide includes the remaining portion of the Cas12a protein (e.g., from amino acid residue 174, 175, 176, 310, 311, 406, 407, 441, 442, 549, or 551 to the end of the protein). In some embodiments, a Cas12a polypeptide has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:150-159 and 175-184.**

In some embodiments, a fusion protein of the present invention comprises all or a portion of a polypeptide of interest. In some embodiments, a polypeptide of interest is fused

(directly or via a linker) to a Cas12a polypeptide and/or to an intein polypeptide. In some embodiments, a polypeptide of interest is fused (directly or via a linker) to a Cas12a polypeptide and intein polypeptide such that the polypeptide of interest is between the Cas12a polypeptide and the intein polypeptide. In some embodiments, a polypeptide of interest is fused (directly or via a linker) to a Cas12a polypeptide and intein polypeptide such that the Cas12a polypeptide is between the polypeptide of interest and the intein polypeptide. In some embodiments, a fusion protein of the present invention comprises a polypeptide of interest that is fused (directly or via a linker) to an intein polypeptide and the fusion protein is devoid of a Cas12a polypeptide. In some embodiments, a polypeptide of interest is fused (directly or via a linker) to the N-terminus of an intein polypeptide. In some embodiments, a polypeptide of interest is fused (directly or via a linker) to the C-terminus of an intein polypeptide.

A fusion protein of the present invention may comprise a reverse transcriptase. A reverse transcriptase polypeptide of the present invention may be all or a portion of a reverse transcriptase. In some embodiments, a reverse transcriptase has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:160-171**. In some embodiments, a reverse transcriptase is fused (directly or via a linker) to a Cas12a polypeptide such that the Cas12a polypeptide is between the reverse transcriptase and the intein polypeptide. In some embodiments, a reverse transcriptase is fused (directly or via a linker) to a Cas12a polypeptide and intein polypeptide such that the reverse transcriptase is between the Cas12a polypeptide and the intein polypeptide. In some embodiments, a reverse transcriptase is fused (directly or via a linker) to the N-terminus of a Cas12a polypeptide. In some embodiments, a reverse transcriptase is fused (directly or via a linker) to the C-terminus of a Cas12a polypeptide. In some embodiments, a reverse transcriptase polypeptide is fused (directly or via a linker) to an intein polypeptide to provide a fusion protein. In some embodiments, a fusion protein comprising a reverse transcriptase polypeptide and an intein polypeptide is devoid of a Cas12a polypeptide.

A fusion protein of the present invention may comprise a nuclear localization signal. In some embodiments, a nuclear localization signal has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:172-174**.

In some embodiments, a nucleic acid molecule encoding a fusion protein of the present invention is provided. The nucleic acid molecule may be operably associated with a promoter. In some embodiments, an expression cassette or vector comprising a nucleic acid molecule encoding a fusion protein of the present invention is provided. In some embodiments, an AAV

vector comprising a nucleic acid molecule encoding a fusion protein of the present invention is provided.

A complex comprising a Cas12a protein, a guide nucleic acid (e.g., a guide RNA and/or an extended guide nucleic acid), optionally a reverse transcriptase, and optionally a deaminase may be provided according to embodiments of the present invention. In some embodiments, a complex comprises a Cas12a protein, an extended guide nucleic acid, and a reverse transcriptase. In some embodiments, a complex comprises a Cas12a protein, a guide nucleic acid, and a deaminase. The Cas12a protein of a complex of the present invention may be prepared from a first fusion protein of the present invention and a second fusion protein of the present invention, wherein the first fusion protein comprises a first Cas12a polypeptide fused to a first intein polypeptide and the second fusion protein comprises a second Cas12a polypeptide fused to a second intein polypeptide. Upon contact of the first fusion protein and the second fusion protein (e.g., the first and second fusion proteins being provided together (e.g., in the same composition or cell) under conditions suitable for carrying out the excision of the first and second intein polypeptides, association of the first and second intein polypeptides, and fusion of the first and second Cas12a polypeptides), the first and second intein polypeptides may associate to form an intein (e.g., an active complex), and the intein may excise the intein (e.g., the first and second intein polypeptides) and fuse the first and second Cas12a polypeptides together optionally with a linker (e.g., a peptide linker) between the first and second Cas12a polypeptides.

In some embodiments, a complex of the present invention comprises an engineered protein (e.g., a fusion protein, a base editor, a templated editor, etc.) and a guide nucleic acid (e.g., a guide RNA). A base editor may comprise a CRISPR-Cas effector protein (e.g., a Cas12a) and a deaminase. In some embodiments, a templated editor may comprise a CRISPR-Cas effector protein (e.g., a Cas12a) and a reverse transcriptase. In some embodiments, a templated editor may be referred to as a REDRAW editor. In some embodiments, a complex of the present invention comprises an engineered protein that is prepared from a first fusion protein of the present invention and a second fusion protein of the present invention, wherein the first fusion protein comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide and the second fusion protein comprises a Cas12a polypeptide fused to a second intein polypeptide; and a guide nucleic acid (e.g., a guide RNA). In some embodiments, a first fusion protein of the present invention and a second fusion protein of the present invention together provide and/or form an engineered protein, wherein the first fusion protein comprises a polypeptide of interest (e.g., a reverse transcriptase

polypeptide) fused to a first intein polypeptide and the second fusion protein comprises a Cas12a polypeptide fused to a second intein polypeptide. In some embodiments, a fusion protein of the present invention comprises, optionally in the N- to C-direction, all or a portion of a polypeptide of interest (e.g., a reverse transcriptase polypeptide), a linker, and an intein polypeptide, optionally wherein the linker comprises a sequence of **SEQ ID NO:189**. In some embodiments, a fusion protein of the present invention comprises, optionally in the N- to C-direction, an intein polypeptide, a linker, and all or a portion of a Cas12a, optionally wherein the linker comprises a sequence of **SEQ ID NO:190**.

In some embodiments, a composition is provided that comprises: a first fusion protein of the present invention that comprises a first Cas12a polypeptide fused to a first intein polypeptide; and a second fusion protein of the present invention that comprises a second Cas12a polypeptide fused to a second intein polypeptide. The first fusion protein and the second fusion protein may be different from each other. In some embodiments, the first intein polypeptide of the first fusion protein and the second intein polypeptide of the second fusion protein together form an intein and/or are two parts that make up a full-length intein and/or the first Cas12a polypeptide of the first fusion protein and the second Cas12a polypeptide of the second fusion protein together form a Cas12a protein and/or are two parts that make up a full-length Cas12a protein. In some embodiments, the first and second fusion proteins are present in the same cell and may optionally be delivered to the cell using separate compositions or a composition comprising both the first and second fusion proteins.

In some embodiments, a composition of the present invention comprises: a first fusion protein of the present invention that comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide; and a second fusion protein of the present invention that comprises a Cas12a polypeptide fused to a second intein polypeptide. The first fusion protein and the second fusion protein may be different from each other. In some embodiments, the first intein polypeptide of the first fusion protein and the second intein polypeptide of the second fusion protein together form an intein and/or are two parts that make up a full-length intein and/or the polypeptide of interest of the first fusion protein and the Cas12a polypeptide of the second fusion protein together form a fusion protein (e.g., an engineered protein and/or a templated editor) and/or are two parts that make up a full-length fusion protein (e.g., an engineered protein and/or a templated editor). In some embodiments, the first and second fusion proteins are present in the same cell and may optionally be delivered to the cell using separate compositions or a composition comprising both the first and second fusion proteins.

In some embodiments, a composition of the present invention comprises a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a first Cas12a polypeptide fused to a first intein polypeptide; and a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a second Cas12a polypeptide fused to a second intein polypeptide. The first nucleic acid molecule may encode a fusion protein of the present invention and the second nucleic acid molecule may encode a fusion protein of the present invention. In some embodiments, the first nucleic acid molecule is present in a first expression cassette and/or vector and the second nucleic acid molecule is present in a second expression cassette and/or vector, wherein the first and second expression cassettes and/or vectors are separate from each other and/or are different.

In some embodiments, a composition of the present invention comprises a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide; and a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a Cas12a polypeptide fused to a second intein polypeptide. The first nucleic acid molecule may encode a fusion protein of the present invention and the second nucleic acid molecule may encode a fusion protein of the present invention. In some embodiments, the first nucleic acid molecule is present in a first expression cassette and/or vector and the second nucleic acid molecule is present in a second expression cassette and/or vector, wherein the first and second expression cassettes and/or vectors are separate from each other and/or are different.

A kit may be provided according to some embodiments of the present invention. A kit of the present invention may comprise a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a first Cas12a polypeptide fused to a first intein polypeptide; and a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a second Cas12a polypeptide fused to a second intein polypeptide. In some embodiments, a kit of the present invention comprises a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide; and a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a Cas12a polypeptide fused to a second intein polypeptide. In some embodiments, the first nucleic acid molecule of the present invention is present in a first expression cassette and/or vector and the second nucleic acid molecule of the present invention is present in a second

expression cassette and/or vector, wherein the first and second expression cassettes and/or vectors are separate from each other and/or are different.

A complex and/or method of the present invention may use and/or include a Cas12a protein that is provided by (e.g., prepared from) two different fusion proteins of the present invention. For example, when the two different fusion proteins are in contact (with each fusion protein comprising a Cas12a polypeptide), the intein polypeptides of the two fusion proteins may associate to form an intein (e.g., an active complex), and the intein may excise the intein (e.g., the first and second intein polypeptides) and fuse the first and second Cas12a polypeptides together optionally with a linker (e.g., a peptide linker) between the first and second Cas12a polypeptides, to form the Cas12a protein. The two different fusion proteins may be provided by and/or be present in a composition and/or kit of the present invention. In some embodiments, a method of the present invention uses an editing system (e.g., a CRISPR-Cas editing system) in which a Cas12a protein of the present invention (e.g., a Cas12a protein formed by a composition and/or method of the present invention) is part of the editing system and is provided by (e.g., prepared from) two different fusion proteins of the present invention. The editing system may be used to modify a target nucleic acid.

In some embodiments, a complex and/or method of the present invention may use and/or include a fusion protein (e.g., an engineered protein) that is provided by (e.g., prepared from) two different fusion proteins of the present invention. For example, when a first fusion protein that comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide) and a first intein polypeptide is in contact with a second fusion protein that comprises a Cas12a polypeptide and a second intein polypeptide, the intein polypeptides of the two fusion proteins may associate to form an intein (e.g., an active complex), and the intein may excise the intein (e.g., the first and second intein polypeptides) and fuse the polypeptide of interest and Cas12a polypeptide together optionally with a linker (e.g., a peptide linker) between the polypeptide of interest and Cas12a polypeptide, to form a fusion protein (e.g., an engineered protein). The two different fusion proteins may be provided by and/or be present in a composition and/or kit of the present invention. In some embodiments, a method of the present invention uses an editing system (e.g., a CRISPR-Cas editing system) in which a fusion protein of the present invention (e.g., an engineered protein (e.g., a templated editor) formed by a composition and/or method of the present invention) is part of the editing system and is provided by (e.g., prepared from) two different fusion proteins of the present invention. The editing system may be used to modify a target nucleic acid.

According to some embodiments, provided is a method of modifying a target nucleic acid, the method comprising contacting the target nucleic acid with: a Cas12a protein prepared from a first fusion protein of the present invention and a second fusion protein of the present invention, wherein the first fusion protein comprises a first Cas12a polypeptide fused to a first intein polypeptide and the second fusion protein comprises a second Cas12a polypeptide fused to a second intein polypeptide; and a guide nucleic acid (e.g., a guide RNA). The Cas12a protein and the guide nucleic acid may form a complex or may be comprised in a complex. The target nucleic acid may be present in a cell (e.g., a eukaryotic cell). In some embodiments, the target nucleic acid is present in a plant cell or a human cell. A method of modifying a target nucleic acid may comprise introducing a first nucleic acid molecule encoding the first fusion protein into the cell and introducing a second nucleic acid molecule encoding the second fusion protein into the cell, and expressing the first fusion protein and the second fusion protein in the cell. The first and second nucleic acid molecules may be present in the same composition such that the first and second nucleic acid molecules may be introduced together. In some embodiments, the first and second nucleic acid molecules are in different compositions such that the first and second nucleic acid molecules are introduced together in different, separate compositions or are introduced sequentially in any order. In some embodiments, the first nucleic acid molecule and/or the second nucleic acid molecule is/are present in an expression cassette and/or vector. In some embodiments, the expression cassette and/or vector is an AAV vector. In some embodiments, the target nucleic acid is present in a cell, optionally in a cell of an organism (e.g., a plant, human, etc.). In some embodiments, the method of modifying a target nucleic acid is carried out *in vitro*, *in vivo*, or *ex vivo*.

In some embodiments, a method of modifying a target nucleic acid of the present invention comprises contacting the target nucleic acid with: an engineered protein (e.g., a templated editor) that is prepared from a first fusion protein and a second fusion protein, wherein the first fusion protein comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide and the second fusion protein comprises a Cas12a polypeptide fused to a second intein polypeptide; and a guide nucleic acid (e.g., a guide RNA). The engineered protein and the guide nucleic acid may form a complex or may be comprised in a complex. The target nucleic acid may be present in a cell (e.g., a eukaryotic cell). In some embodiments, the target nucleic acid is present in a plant cell or a human cell. A method of modifying a target nucleic acid may comprise introducing a first nucleic acid molecule encoding the first fusion protein into the cell and introducing a second nucleic acid molecule encoding the second fusion protein into the cell, and expressing the first fusion

protein and the second fusion protein in the cell. The first and second nucleic acid molecules may be present in the same composition such that the first and second nucleic acid molecules may be introduced together. In some embodiments, the first and second nucleic acid molecules are in different compositions such that the first and second nucleic acid molecules are introduced together in different, separate compositions or are introduced sequentially in any order. In some embodiments, the first nucleic acid molecule and/or the second nucleic acid molecule is/are present in an expression cassette and/or vector. In some embodiments, the expression cassette and/or vector is an AAV vector. In some embodiments, the target nucleic acid is present in a cell, optionally in a cell of an organism (e.g., a plant, human, etc.). In some embodiments, the method of modifying a target nucleic acid is carried out *in vitro*, *in vivo*, or *ex vivo*.

According to some embodiments, provided is a method of modifying a target nucleic acid, the method comprising: introducing a first nucleic acid molecule and a second nucleic acid molecule into a cell, wherein the first nucleic acid molecule encodes a first fusion protein, the first fusion protein comprising a first Cas12a polypeptide fused to a first intein polypeptide and the second nucleic acid molecule encodes a second fusion protein, the second fusion protein comprising a second Cas12a polypeptide fused to a second intein polypeptide; contacting the target nucleic acid in the cell with a protein comprising at least a portion of the first Cas12a polypeptide and at least a portion of the second Cas12a polypeptide and a guide nucleic acid (e.g., a guide RNA and/or an extended guide nucleic acid). In some embodiments, the guide nucleic acid and the protein comprising at least a portion of the first Cas12a polypeptide and at least a portion of the second Cas12a polypeptide form a complex or are comprised in a complex. The method may comprises expressing the first fusion protein and the second fusion protein in the cell. In some embodiments, following the introducing step, the method comprises cleaving (e.g., excising) the first intein polypeptide from the first fusion protein and cleaving (e.g., excising) the second intein polypeptide from the second fusion protein. The method may also comprise prior to, during, and/or after cleaving, associating the first intein polypeptide and the second intein polypeptide to form an intein. The cleaving step may cleave the first Cas12a polypeptide from the first fusion protein and the second Cas12a polypeptide from the second fusion protein and/or the method may further comprise cleaving the first Cas12a polypeptide from the first fusion protein and the second Cas12a polypeptide from the second fusion protein. Prior to, during, and/or after cleaving of the first and second intein polypeptides, the method may comprise fusing the first and second Cas12a polypeptides together (e.g., via a peptide bond between the first and second Cas12a polypeptides) to form a

Cas12a protein, wherein the protein that contacts the target nucleic acid in the cell is the Cas12a protein. In some embodiments, the intein fuses the first and second Cas12a polypeptides together. In some embodiments, cleaving of the first intein polypeptide from the first fusion protein and cleaving the second intein polypeptide from the second fusion protein occurs concurrently with fusing the first and second Cas12a polypeptides together. The introducing step may comprise introducing a first expression cassette and/or vector that comprises the first nucleic acid molecule and introducing a second expression cassette and/or vector that comprises the second nucleic acid molecule into the cell. The first and/or second expression cassette and/or vector may comprise the guide nucleic acid, or the method may comprise introducing into the cell a third expression cassette and/or vector that comprises the guide nucleic acid. In some embodiments, a first, second, and/or third expression cassette and/or vector is an AAV vector.

In some embodiments, a method of modifying a target nucleic acid of the present invention comprises introducing a first nucleic acid molecule and a second nucleic acid molecule into a cell, wherein the first nucleic acid molecule encodes a first fusion protein, the first fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide, and the second nucleic acid molecule encodes a second fusion protein, the second fusion protein comprising a Cas12a polypeptide fused to a second intein polypeptide; contacting the target nucleic acid in the cell with a protein (e.g., a templated editor) comprising at least a portion of the polypeptide of interest and at least a portion of the Cas12a polypeptide and a guide nucleic acid (e.g., a guide RNA). In some embodiments, the guide nucleic acid and the protein comprising at least a portion of the polypeptide of interest and at least a portion of the Cas12a polypeptide form a complex or are comprised in a complex. The method may comprise expressing the first fusion protein and the second fusion protein in the cell. In some embodiments, following the introducing step, the method comprises cleaving (e.g., excising) the first intein polypeptide from the first fusion protein and cleaving (e.g., excising) the second intein polypeptide from the second fusion protein. The method may also comprise prior to, during, and/or after cleaving, associating the first intein polypeptide and the second intein polypeptide to form an intein. The cleaving step may cleave the polypeptide of interest from the first fusion protein and the Cas12a polypeptide from the second fusion protein and/or the method may further comprise cleaving the polypeptide of interest from the first fusion protein and the Cas12a polypeptide from the second fusion protein. Prior to, during, and/or after cleaving of the first and second intein polypeptides, the method may comprise fusing the polypeptide of interest and the Cas12a polypeptide

together (e.g., via a peptide bond that is between the polypeptide of interest and the Cas12a polypeptide) to form a fusion protein (optionally wherein the fusion protein is a templated editor), wherein the fusion protein contacts the target nucleic acid in the cell. In some embodiments, the intein fuses the polypeptide of interest and Cas12a polypeptide together. In some embodiments, cleaving of the first intein polypeptide from the first fusion protein and cleaving the second intein polypeptide from the second fusion protein occurs concurrently with fusing the polypeptide of interest and Cas12a polypeptide together. The introducing step may comprise introducing a first expression cassette and/or vector that comprises the first nucleic acid molecule and introducing a second expression cassette and/or vector that comprises the second nucleic acid molecule into the cell. The first and/or second expression cassette and/or vector may comprise the guide nucleic acid, or the method may comprise introducing into the cell a third expression cassette and/or vector that comprises the guide nucleic acid. In some embodiments, a first, second, and/or third expression cassette and/or vector is an AAV vector.

In some embodiments, a method of the present invention has increased efficiency in modifying a target nucleic acid compared to the efficiency of a control method. An exemplary control method includes a method that contacts a target nucleic acid with a wild-type CRISPR-Cas effector protein that is not fused together via protein splicing (e.g., from two different proteins) optionally using an intein. Another exemplary control method includes a method that contacts a target nucleic acid with a fusion protein (e.g., an engineered protein and/or a templated editor) that is not fused together via protein splicing (e.g., from two different proteins) optionally using an intein. A method of the present invention may generate increased indels and/or increased levels of modification (e.g., precise modifications) compared to a control method. In some embodiments, an editing system used in a method of the present invention is a Redraw editing system such as described in U.S. Patent Application Publication No. 2021/0130835 and/or in U.S. Patent Application Publication No. 2022/0145334, the contents of each of which are incorporated herein by reference in their entirety, but optionally wherein the CRISPR-Cas effector protein is a Cas12a protein that is fused together via protein splicing from two fusion proteins of the present invention and/or wherein the templated editor is a protein that is fused together via protein splicing from two fusion proteins of the present invention.

According to embodiments of the present invention, a Cas12a protein may be split into two parts (e.g., two Cas12a polypeptides) and each part may individually be fused to a portion (e.g., a fragment) of a trans-splicing split intein (e.g., an intein polypeptide) to provide two different fusion proteins. Each of the two fusion proteins may be separately packaged in an

AAV vector. For example, a nucleic acid molecule encoding a fusion portion comprising a Cas12a polypeptide and an intein polypeptide may be provided in an AAV vector. In some embodiments, the two fusion proteins (the Cas12a polypeptide of each form the full-length Cas12a protein and the intein polypeptide of each form the full-length intein) are individually packaged in separate AAV vectors and when both AAV vectors are introduced into (e.g., infect) the same cell, both fusion proteins can be expressed and the two Cas12a polypeptides can be fused (e.g., spliced) together *in situ*.

In some embodiments, an engineered protein (e.g., a base editor, templated editor, etc.) may be split into two parts (e.g., one part comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) and a second part comprising a Cas12a polypeptide) and each part may individually be fused to a portion (e.g., a fragment) of a trans-splicing split intein (e.g., an intein polypeptide) to provide two different fusion proteins. Each of the two fusion proteins may be separately packaged in an AAV vector. For example, a nucleic acid molecule encoding a fusion portion comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) and an intein polypeptide may be provided in an AAV vector. In some embodiments, the two fusion proteins (of which together form the engineered protein and the intein polypeptides of each form a full-length intein) are individually packaged in separate AAV vectors and when both AAV vectors are introduced into (e.g., infect) the same cell, both fusion proteins can be expressed and the polypeptide of interest and Cas12a polypeptide can be fused (e.g., spliced) together *in situ*.

In some embodiments, an editing system of the present invention utilizes the Redraw editing system. Further details on the Redraw editing system can be found in U.S. Patent Application Publication No. 2021/0130835 and/or in U.S. Patent Application Publication No. 2022/0145334, the contents of each of which are incorporated herein by reference in their entirety.

As described herein, the fusion proteins, nucleic acids, expression cassettes, and/or vectors of the present invention may be codon optimized for expression in an organism. An organism useful with this invention may be any organism or cell thereof for which nucleic acid modification may be useful. An organism can include, but is not limited to, any animal (e.g., a mammal), any plant, any fungus, any archaeon, or any bacterium. In some embodiments, the organism may be a plant or cell thereof. In some embodiments, the organism is an animal such as a mammal (e.g., a human).

The target nucleic acid may be a genomic sequence from any organism (e.g., eukaryote such as a mammal or a plant). In some embodiments, the target nucleic acid is a genomic

sequence from a model organism such as, but not limited to, *Escherichia coli*, an immortalized human cell line (e.g., HEK293, HeLa, etc.), *Caenorhabditis elegans*, *Arabidopsis thaliana*, and/or *Drosophila Melanogaster*. In some embodiments, the target nucleic acid is a genomic sequence from a non-model organism. Exemplary non-model organisms include, but are not limited to crop plants (e.g., fruit crop plants, vegetable crop plants, and/or field crop plants) and/or animals such as humans, primates and/or mice. In some embodiments, the non-model organism is a crop plant such as corn, soybean, wheat, or canola. In some embodiments, the non-model organism is an animal for testing and/or use of a human therapeutic.

A target nucleic acid of any plant or plant part may be modified using the nucleic acid constructs of the invention. Any plant (or groupings of plants, for example, into a genus or higher order classification) may be modified using a fusion protein of the invention including an angiosperm, a gymnosperm, a monocot, a dicot, a C3, C4, CAM plant, a bryophyte, a fern and/or fern ally, a microalgae, and/or a macroalgae. A plant and/or plant part useful with this invention may be a plant and/or plant part of any plant species/variety/cultivar. The term "plant part," as used herein, includes but is not limited to, embryos, pollen, ovules, seeds, leaves, stems, shoots, flowers, branches, fruit, kernels, ears, cobs, husks, stalks, roots, root tips, anthers, plant cells including plant cells that are intact in plants and/or parts of plants, plant protoplasts, plant tissues, plant cell tissue cultures, plant calli, plant clumps, and the like. As used herein, "shoot" refers to the above ground parts including the leaves and stems. Further, as used herein, "plant cell" refers to a structural and physiological unit of the plant, which comprises a cell wall and also may refer to a protoplast. A plant cell can be in the form of an isolated single cell or can be a cultured cell or can be a part of a higher-organized unit such as, for example, a plant tissue or a plant organ.

Non-limiting examples of plants useful with the present invention include turf grasses (e.g., bluegrass, bentgrass, ryegrass, fescue), feather reed grass, tufted hair grass, miscanthus, arundo, switchgrass, vegetable crops, including artichokes, kohlrabi, arugula, leeks, asparagus, lettuce (e.g., head, leaf, romaine), malanga, melons (e.g., muskmelon, watermelon, crenshaw, honeydew, cantaloupe), cole crops (e.g., brussels sprouts, cabbage, cauliflower, broccoli, collards, kale, Chinese cabbage, bok choy), cardoni, carrots, napa, okra, onions, celery, parsley, chick peas, parsnips, chicory, peppers, potatoes, cucurbits (e.g., marrow, cucumber, zucchini, squash, pumpkin, honeydew melon, watermelon, cantaloupe), radishes, dry bulb onions, rutabaga, eggplant, salsify, escarole, shallots, endive, garlic, spinach, green onions, squash, greens, beet (sugar beet and fodder beet), sweet potatoes, chard, horseradish, tomatoes, turnips, and spices; a fruit crop such as apples, apricots, cherries, nectarines, peaches, pears, plums,

prunes, cherry, quince, fig, nuts (e.g., chestnuts, pecans, pistachios, hazelnuts, pistachios, peanuts, walnuts, macadamia nuts, almonds, and the like), citrus (e.g., clementine, kumquat, orange, grapefruit, tangerine, mandarin, lemon, lime, and the like), blueberries, black raspberries, boysenberries, cranberries, currants, gooseberries, loganberries, raspberries, strawberries, blackberries, grapes (wine and table), avocados, bananas, kiwi, persimmons, pomegranate, pineapple, tropical fruits, pomes, melon, mango, papaya, and lychee, a field crop plant such as clover, alfalfa, timothy, evening primrose, meadow foam, corn/maize (field, sweet, popcorn), hops, jojoba, buckwheat, safflower, quinoa, wheat, rice, barley, rye, millet, sorghum, oats, triticale, sorghum, tobacco, kapok, a leguminous plant (beans (e.g., green and dried), lentils, peas, soybeans), an oil plant (rape, canola, mustard, poppy, olive, sunflower, coconut, castor oil plant, cocoa bean, groundnut, oil palm), duckweed, *Arabidopsis*, a fiber plant (cotton, flax, hemp, jute), *Cannabis* (e.g., *Cannabis sativa*, *Cannabis indica*, and *Cannabis ruderalis*), lauraceae (cinnamon, camphor), or a plant such as coffee, sugar cane, tea, and natural rubber plants; and/or a bedding plant such as a flowering plant, a cactus, a succulent and/or an ornamental plant (e.g., roses, tulips, violets), as well as trees such as forest trees (broad-leaved trees and evergreens, such as conifers; e.g., elm, ash, oak, maple, fir, spruce, cedar, pine, birch, cypress, eucalyptus, willow), as well as shrubs and other nursery stock. In some embodiments, the nucleic acid constructs of the invention and/or expression cassettes and/or vectors encoding the same may be used to modify maize, soybean, wheat, canola, rice, tomato, pepper, sunflower, raspberry, blackberry, black raspberry and/or cherry.

In some embodiments, the invention provides cells (e.g., plant cells, animal cells, bacterial cells, archaeon cells, and the like) comprising the polypeptides, polynucleotides, nucleic acid constructs, expression cassettes or vectors of the invention.

The present invention further comprises a kit or kits to carry out the methods of this invention. A kit of this invention can comprise reagents, buffers, and apparatus for mixing, measuring, sorting, labeling, etc., as well as instructions and the like as would be appropriate for modifying a target nucleic acid.

In some embodiments, the invention provides a kit for comprising one or more fusion proteins of the present invention, nucleic acid constructs of the present invention, and/or expression cassettes and/or vectors and/or cells comprising the same as described herein, with optional instructions for the use thereof. In some embodiments, a kit may further comprise a CRISPR-Cas guide nucleic acid (corresponding to a Cas12a protein as provided herein, which may be encoded by a polynucleotide of the invention) and/or expression cassettes and/or vectors and or cells comprising the same. In some embodiments, a guide nucleic acid may be

provided on the same expression cassette and/or vector as one or more nucleic acid constructs of the invention. In some embodiments, the guide nucleic acid may be provided on a separate expression cassette or vector from that comprising the one or more nucleic acid constructs of the invention.

Accordingly, in some embodiments, kits are provided comprising a nucleic acid construct comprising (a) a polynucleotide(s) as provided herein and (b) a promoter that drives expression of the polynucleotide(s) of (a). In some embodiments, the kit may further comprise a nucleic acid construct encoding a guide nucleic acid, wherein the construct comprises a cloning site for cloning of a nucleic acid sequence identical or complementary to a target nucleic acid sequence into backbone of the guide nucleic acid.

In some embodiments, the nucleic acid construct of the invention may be an mRNA that may encode one or more introns within the encoded polynucleotide(s). In some embodiments, the nucleic acid constructs of the invention, and/or an expression cassettes and/or vectors comprising the same, may further encode one or more selectable markers useful for identifying transformants (e.g., a nucleic acid encoding an antibiotic resistance gene, herbicide resistance gene, and the like).

A polypeptide, polynucleotide, nucleic acid construct, expression cassette, vector, composition, kit, system and/or cell of the present invention may comprise all or a portion of a sequence of one or more of **SEQ ID NOs:1-259**. In some embodiments, a polypeptide, polynucleotide, nucleic acid construct, expression cassette, vector, composition, kit, system and/or cell of the present invention may comprise at least about 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more consecutive amino acids of a sequence of one or more of **SEQ ID NOs:1-259**.

The invention will now be described with reference to the following examples. It should be appreciated that these examples are not intended to limit the scope of the claims to the invention, but are rather intended to be exemplary of certain embodiments. Any variations in the exemplified methods that occur to the skilled artisan are intended to fall within the scope of the invention.

EXAMPLES

Example 1: Validation of split intein protein reconstitution with mCherry

The trans-splicing activity of wildtype Npu (**SEQ ID NOs:110** and **111**, the N- and C-terminal portions, respectively) and NpuGEP (a mutant that contains 3 residue mutations; **SEQ ID NOs:110** and **112**, the N- and C-terminal portions, respectively) in HEK293T cells was

evaluated. For each of the wildtype (WT) Npu and NpuGEP, the N-terminal half of the mCherry protein was fused to the N-terminal portion of the Npu intein (NpuN) and the C-terminal half of the mCherry was fused to the C-terminal portion of the Npu intein (NpuC) (**FIG. 1**). After transfecting plasmids encoding two halves of mCherry, their splicing and reconstitution was measured via flow cytometry after 3 days (**FIG. 2**). Robust fluorescence was detected from dual plasmid setups approaching the fluorescence intensity of native mCherry (**FIG. 2**). Cells where only one half of the mCherry fragments was provided did not show any fluorescence, which demonstrated that both halves of the protein must be present for function (**FIG. 2**). This shows that the split intein is functional in human cells and allows rapid splicing of proteins from transient plasmid transfection of each component.

Example 2: Split intein protein reconstitution with Redraw editors

Fusion proteins were prepared using a Redraw editor (RE2; **SEQ ID NO:113**) and the NpuGEP system. Several sites within RE2 were used to introduce a split site (**FIG. 3**). In particular, RE2 includes a reverse transcriptase (RT) and Cas12a sequence and fusion proteins were prepared in which a portion of RE2 was included with the split occurring in the Cas12a sequence. Accordingly, some fusion proteins included an N-terminal portion of the Cas12a sequence, which included amino acid residues 1-175, 1-310, 1-406, 1-441, or 1-550 of the Cas12 sequence, and other fusion proteins included the remaining C-terminal portion of the Cas-12a sequence (e.g., amino acid residues 176, 311, 407, 442, or 551 to the end of Cas12a). Fusion proteins having a sequence of **SEQ ID NOs:100-109** were prepared. In some cases, two amino acids (cysteine and alanine (CA) or cysteine and phenylalanine (CF)) were inserted between the C-terminal portion of the Npu intein (NpuC) and the C-terminal portion of RE2 (e.g., a C terminal portion of Cas12a), which will leave a “CA” or “CF” scar within the reconstituted RE2 protein. Insertion of the 2 amino acid “scar” residues (here, “CA” or “CF”) was confirmed in mature mCherry protein and it was confirmed that the scar did not affect the protein’s ability to fluoresce (**FIG. 4**).

Example 3: Intein-mediated reconstitution of the Redraw editor in HEK293T cells

HEK293T cells were transfected with a plasmid encoding a fusion protein with an N-terminal component of RE2, a plasmid encoding a fusion protein with a C-terminal component of RE2, and a plasmid encoding a stagRNA or crRNA targeting endogenous sites as described in Example 2. After 3 days, high throughput amplicon sequencing was conducted to quantify the Redraw activity at the target sites. It was observed that several split RE2 forms are capable

of being spliced together in cotransfected HEK293T cells to enable Redraw activity. Robust indel activity was observed (**FIGS. 5-6**) as well as precise editing activity (**FIG. 7**) that was comparable to the single polypeptide RE2 construct. In addition, it was observed that both halves of RE2 must be expressed in the cell to enable indel and Redraw activity.

These results demonstrate that the Redraw editor can be separated into two halves (e.g., one half that comprises a reverse transcriptase polypeptide (such as RT(5M)-NpuN) and another half that comprises a Cas12a polypeptide (such as NpuC-Cas12a-Brex27) or e.g., one half that comprises a first Cas12a polypeptide and another half that comprises a second Cas12a polypeptide (such as a RE2 split, for example at 175) and reconstituted in a cell effectively. This can allow the Redraw editor to be separately packaged into viral delivery vehicles (such as adeno associated virus (AAV) vectors) and co-infected to effect Redraw editing in cells targeted by the viral delivery vehicles.

The foregoing is illustrative of the present invention, and is not to be construed as limiting thereof. The invention is defined by the following claims, with equivalents of the claims to be included therein.

THAT WHICH IS CLAIMED IS:

1. A fusion protein comprising a Cas12a polypeptide fused to an intein polypeptide.
2. The fusion protein of claim 1, wherein the intein polypeptide is a first portion of a trans-splicing split intein, optionally wherein the intein polypeptide is one part of two total parts that together form the trans-splicing split intein.
3. The fusion protein of claim 1 or 2, wherein the intein polypeptide is a portion of a *Nostoc punctiforme* (Npu) intein and/or a portion of a mutant Npu intein.
4. The fusion protein of any one of the preceding claims, wherein the intein polypeptide has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:110-112**.
5. The fusion protein of any one of the preceding claims, wherein the intein polypeptide is fused to the N-terminus of the Cas12a polypeptide, optionally wherein a linker (e.g., 1, 2, 3, 4, or more amino acid residues) is present between the intein polypeptide and the Cas12a polypeptide.
6. The fusion protein of any one of claims 1-4, wherein the intein polypeptide is fused to the C-terminus of the Cas12a polypeptide, optionally wherein a linker (e.g., 1, 2, 3, 4, or more amino acid residues) is present between the intein polypeptide and the Cas12a polypeptide.
7. The fusion protein of any one of the preceding claims, wherein the intein polypeptide is configured to be removed from the fusion protein and fused with another intein polypeptide *in situ* and/or *in vivo*.
8. The fusion protein of any one of the preceding claims, wherein the Cas12a polypeptide is a first portion of a Cas12a protein, optionally wherein the Cas12a polypeptide is one part of two total parts that together form the Cas12a protein.

9. The fusion protein of any one of the preceding claims, wherein the Cas12a polypeptide is a N-terminal portion of a Cas12a protein and the intein polypeptide is a N-terminal portion of an intein, optionally wherein the intein polypeptide is fused to the C-terminus of the Cas12a polypeptide.
10. The fusion protein of any one of claims 1-8, wherein the Cas12a polypeptide is a C-terminal portion of a Cas12a protein and the intein polypeptide is a C-terminal portion of an intein, optionally wherein the intein polypeptide is fused to the N-terminus of the Cas12a polypeptide.
11. The fusion protein of any one of the preceding claims, further comprising a reverse transcriptase that is fused to the Cas12a polypeptide and the intein polypeptide, optionally wherein the reverse transcriptase is fused to the N-terminus of the Cas12a polypeptide and the intein is fused to the C-terminus of the Cas12a polypeptide.
12. The fusion protein of any one of the preceding claims, wherein the Cas12a polypeptide comprises about 175, 310, 406, 441, or 550 consecutive amino acids of a Cas12a protein.
13. The fusion protein of any one of the preceding claims, wherein the Cas12a polypeptide has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:150-159** and **175-184**.
14. The fusion protein of any one of the preceding claims, wherein the fusion protein has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:100-109** or **188**.
15. A nucleic acid molecule encoding the fusion protein of any one of claims 1-14.
16. An expression cassette or vector comprising the nucleic acid molecule of claim 15.
17. A complex comprising:
 - a Cas12a protein prepared from a first fusion protein and a second fusion protein, wherein the first fusion protein comprises a first Cas12a polypeptide fused to a first intein

polypeptide and the second fusion protein comprises a second Cas12a polypeptide fused to a second intein polypeptide;

a guide nucleic acid (e.g., a guide RNA); and
optionally a deaminase.

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

a Cas12a protein prepared from a first fusion protein and a second fusion protein, wherein the first fusion protein comprises a first Cas12a polypeptide fused to a first intein polypeptide and the second fusion protein comprises a second Cas12a polypeptide fused to a second intein polypeptide; and

a guide nucleic acid (e.g., a guide RNA),
optionally wherein the Cas12a protein and the guide nucleic acid form a complex or are comprised in a complex, thereby modifying the target nucleic acid.

19. The method of claim 18, wherein the first fusion protein and/or the second fusion protein is the fusion protein of any one of claims 1-14.

20. The method of claim 18 or 19, wherein the target nucleic acid is present in a cell (e.g., a eukaryotic cell), optionally wherein the target nucleic acid is present in a plant cell or a human cell.

21. The method of claim 20, further comprising introducing a first nucleic acid molecule encoding the first fusion protein into the cell and introducing a second nucleic acid molecule encoding the second fusion protein into the cell, and producing the first fusion protein and the second fusion protein in the cell.

22. The method of claim 21, wherein the first nucleic acid molecule and/or the second nucleic acid molecule are present in an expression cassette and/or vector, optionally wherein the expression cassette and/or vector is an Adeno-associated virus vector.

23. A composition comprising:

a first fusion protein that comprises a first Cas12a polypeptide fused to a first intein polypeptide; and

a second fusion protein that comprises a second Cas12a polypeptide fused to a second intein polypeptide.

24. The composition of claim 23, wherein the first fusion protein and/or the second fusion protein is the fusion protein of any one of claims 1-14, and wherein the first fusion protein and second fusion protein are different.

25. A composition comprising:

a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a first Cas12a polypeptide fused to a first intein polypeptide; and

a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a second Cas12a polypeptide fused to a second intein polypeptide.

26. A kit comprising:

a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a first Cas12a polypeptide fused to a first intein polypeptide; and

a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a second Cas12a polypeptide fused to a second intein polypeptide.

27. The composition of claim 25 or kit of claim 26, wherein a first expression cassette and/or vector comprises the first nucleic acid molecule and a second expression cassette and/or vector comprises the second nucleic acid molecule, optionally wherein the first expression cassette and/or vector is different than the second expression cassette and/or vector.

28. The composition or kit of any one of claims 25-27, wherein the first nucleic acid molecule and/or the second nucleic acid molecule is the nucleic acid molecule of claim 15, and wherein the first nucleic acid molecule is different than the second nucleic acid molecule.

29. A method of modifying a target nucleic acid, the method comprising:

introducing a first nucleic acid molecule and a second nucleic acid molecule into a cell comprising the target nucleic acid, wherein the first nucleic acid molecule encodes a first fusion protein, the first fusion protein comprising a first Cas12a polypeptide fused to a first intein polypeptide and the second nucleic acid molecule encodes a second fusion protein, the

second fusion protein comprising a second Cas12a polypeptide fused to a second intein polypeptide;

contacting the target nucleic acid in the cell with a guide nucleic acid (e.g., a guide RNA) and a protein comprising at least a portion of the first Cas12a polypeptide and at least a portion of the second Cas12a polypeptide,

optionally wherein the protein and the guide nucleic acid form a complex or are comprised in a complex, thereby modifying the target nucleic acid.

30. The method of claim 29, further comprising producing the first fusion protein and the second fusion protein in the cell.

31. The method of claim 29 or 30, wherein the first fusion protein and/or the second fusion protein is the fusion protein of any one of claims 1-14 and the first fusion protein and second fusion protein are different.

32. The method of any one of claims 29-31, further comprising cleaving the first intein polypeptide from the first fusion protein and cleaving the second intein polypeptide from the second fusion protein, optionally wherein, prior to, during, and/or after cleaving, the first intein polypeptide and the second intein polypeptide associate to form an intein.

33. The method of claim 32, wherein the intein is a *Nostoc punctiforme* (Npu) intein and/or a portion of a mutant Npu intein.

34. The method of claim 32 or 33, wherein the first intein polypeptide and/or the second polypeptide has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:110-112**.

35. The method of any one of claims 29-35, further comprising cleaving the first Cas12a polypeptide from the first fusion protein and cleaving the second Cas12a polypeptide from the second fusion protein, optionally wherein, prior to, during, and/or after cleaving, the first Cas12a polypeptide and the second Cas12a polypeptide are fused to form a Cas12a protein, wherein the protein that contacts the target nucleic acid in the cell is the Cas12a protein.

36. The method of claim 35, wherein the protein has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:38-60, 113-149, 192-195, and 196-259**.

37. The method of any one of claims 29-36, wherein the first fusion protein has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one of **SEQ ID NOs:100, 102, 104, 106, and 108** and the second fusion protein has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one of **SEQ ID NOs:101, 103, 105, 107, and 109**, respectively, wherein the first fusion protein and the second fusion protein are different.

38. The method of any one of claims 29-37, wherein introducing the first nucleic acid molecule and the second nucleic acid molecule into the cell comprises introducing a first expression cassette and/or vector that comprises the first nucleic acid molecule and introducing a second expression cassette and/or vector that comprises the second nucleic acid molecule into the cell, optionally wherein the first expression cassette and/or vector and/or the second expression cassette and/or vector is the expression cassette and/or vector of claim 16.

39. The method of any one of claims 29-38, further comprising introducing a third expression cassette and/or vector that comprises the guide nucleic acid.

40. The method of claim 38 or 39, wherein the first expression cassette and/or vector, the second expression cassette and/or vector, and/or the third expression cassette and/or vector is an Adeno-associated virus vector.

41. The method of any one of claims 18, 19, and 29-40, wherein the method has increased efficiency in modifying the target nucleic acid compared to the efficiency of a control method (e.g., a method comprising contacting the target nucleic acid with a wild-type CRISPR-Cas effector protein that is not fused together via protein splicing optionally using an intein).

42. An engineered protein comprising an amino acid sequence that has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:100-109 and 187-188**.

43. A nucleic acid molecule encoding the engineered protein of claim 42.
44. A nucleic acid molecule comprising a polynucleotide that has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:187-188**.
45. A fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to an intein polypeptide.
46. The fusion protein of claim 45, wherein the intein polypeptide is a first portion of a trans-splicing split intein, optionally wherein the intein polypeptide is one part of two total parts that together form the trans-splicing split intein.
47. The fusion protein of claim 45 or 46, wherein the intein polypeptide is a portion of a *Nostoc punctiforme* (Npu) intein and/or a portion of a mutant Npu intein.
48. The fusion protein of any one of claims 45-47, wherein the intein polypeptide has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:110-112**.
49. The fusion protein of any one of claims 45-48, wherein the intein polypeptide is fused to the N-terminus of the polypeptide of interest, optionally wherein a linker (e.g., 1, 2, 3, 4, or more amino acid residues) is present between the intein polypeptide and the polypeptide of interest.
50. The fusion protein of any one of claims 45-48, wherein the intein polypeptide is fused to the C-terminus of the polypeptide of interest, optionally wherein a linker (e.g., 1, 2, 3, 4, or more amino acid residues) is present between the intein polypeptide and the polypeptide of interest.
51. The fusion protein of any one of claims 45-50, wherein the intein polypeptide is configured to be removed from the fusion protein and fused with another intein polypeptide *in situ* and/or *in vivo*.

52. The fusion protein of any one of claims 45-51, wherein the polypeptide of interest has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:160-171**.

53. The fusion protein of any one of claims 45-52, wherein the intein polypeptide is a N-terminal portion of an intein, optionally wherein the intein polypeptide is fused to the C-terminus of the polypeptide of interest.

54. The fusion protein of any one of claims 45-53, further comprising a linker, optionally wherein the linker is between the polypeptide of interest and the intein polypeptide.

55. The fusion protein of any one of claims 1-14 or 45-54, further comprising a nuclear localization signal, optionally wherein the nuclear localization signal has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:172-174**.

56. A nucleic acid molecule encoding the fusion protein of any one of claims 45-55.

57. An expression cassette or vector comprising the nucleic acid molecule of claim 56.

58. A complex comprising:

a fusion protein (e.g., a templated editor) prepared from a first fusion protein and a second fusion protein, wherein the first fusion protein comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide and the second fusion protein comprises a Cas12a polypeptide fused to a second intein polypeptide; and a guide nucleic acid (e.g., a guide RNA).

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

a fusion protein prepared from a first fusion protein and a second fusion protein, wherein the first fusion protein comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide and the second fusion protein comprises a Cas12a polypeptide fused to a second intein polypeptide; and

a guide nucleic acid (e.g., a guide RNA),
optionally wherein the fusion protein and the guide nucleic acid form a complex or are comprised in a complex, thereby modifying the target nucleic acid.

60. The method of claim 59, wherein the first fusion protein is the fusion protein of any one of claims 45-55 and/or the second fusion protein is the fusion protein of any one of claims 1-14.

61. The method of claim 59 or 60, wherein the target nucleic acid is present in a cell (e.g., a eukaryotic cell), optionally wherein the target nucleic acid is present in a plant cell or a human cell.

62. The method of claim 61, further comprising introducing a first nucleic acid molecule encoding the first fusion protein into the cell and introducing a second nucleic acid molecule encoding the second fusion protein into the cell, and producing the first fusion protein and the second fusion protein in the cell.

63. The method of claim 62, wherein the first nucleic acid molecule and/or the second nucleic acid molecule are present in an expression cassette and/or vector, optionally wherein the expression cassette and/or vector is an Adeno-associated virus vector.

64. A composition comprising:
a first fusion protein that comprises a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide; and
a second fusion protein that comprises a Cas12a polypeptide fused to a second intein polypeptide.

65. The composition of claim 64, wherein the first fusion protein is the fusion protein of any one of claims 45-55 and/or the second fusion protein is the fusion protein of any one of claims 1-14.

66. A composition comprising:

a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide; and

a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a Cas12a polypeptide fused to a second intein polypeptide.

67. A kit comprising:

a first nucleic acid molecule encoding a first fusion protein, the first fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide; and

a second nucleic acid molecule encoding a second fusion protein, the second fusion protein comprising a Cas12a polypeptide fused to a second intein polypeptide.

68. The composition of claim 66 or kit of claim 67, wherein a first expression cassette and/or vector comprises the first nucleic acid molecule and a second expression cassette and/or vector comprises the second nucleic acid molecule, optionally wherein the first expression cassette and/or vector is different than the second expression cassette and/or vector.

69. The composition or kit of any one of claims 66-68, wherein the first nucleic acid molecule and/or the second nucleic acid molecule is the nucleic acid molecule of claim 43 or 44.

70. A method of modifying a target nucleic acid, the method comprising:

introducing a first nucleic acid molecule and a second nucleic acid molecule into a cell comprising the target nucleic acid, wherein the first nucleic acid molecule encodes a first fusion protein, the first fusion protein comprising a polypeptide of interest (e.g., a reverse transcriptase polypeptide) fused to a first intein polypeptide and the second nucleic acid molecule encodes a second fusion protein, the second fusion protein comprising a Cas12a polypeptide fused to a second intein polypeptide;

contacting the target nucleic acid in the cell with a guide nucleic acid (e.g., a guide RNA) and a protein comprising at least a portion of the polypeptide of interest and at least a portion of the Cas12a polypeptide,

optionally wherein the protein and the guide nucleic acid form a complex or are comprised in a complex, thereby modifying the target nucleic acid.

71. The method of claim 70, further comprising producing the first fusion protein and the second fusion protein in the cell.

72. The method of claim 70 or 71, wherein the first fusion protein is the fusion protein of any one of claims 45-55 and/or the second fusion protein is the fusion protein of any one of claims 1-14.

73. The method of any one of claims 70-72, further comprising cleaving the first intein polypeptide from the first fusion protein and cleaving the second intein polypeptide from the second fusion protein, optionally wherein, prior to, during, and/or after cleaving, the first intein polypeptide and the second intein polypeptide associate to form an intein.

74. The method of claim 73, wherein the intein is a *Nostoc punctiforme* (Npu) intein and/or a portion of a mutant Npu intein.

75. The method of claim 73 or 74, wherein the first intein polypeptide and/or the second polypeptide has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:110-112**.

76. The method of any one of claims 70-85, further comprising cleaving the polypeptide of interest from the first fusion protein and cleaving the Cas12a polypeptide from the second fusion protein, optionally wherein, prior to, during, and/or after cleaving, the polypeptide of interest and the Cas12a polypeptide are fused to form a fusion protein, wherein the protein that contacts the target nucleic acid in the cell is the fusion protein.

77. The method of claim 76, wherein the fusion protein has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:38-60, 113-149, 192-195, and 196-259**.

78. The method of any one of claims 70-77, wherein the first fusion protein has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to **SEQ**

ID NO:187 and the second fusion protein has at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to **SEQ ID NO:188**.

79. The method of any one of claims 70-78, wherein introducing the first nucleic acid molecule and the second nucleic acid molecule into the cell comprises introducing a first expression cassette and/or vector that comprises the first nucleic acid molecule and introducing a second expression cassette and/or vector that comprises the second nucleic acid molecule into the cell.

80. The method of any one of claims 70-79, further comprising introducing a third expression cassette and/or vector that comprises the guide nucleic acid.

81. The method of claim 79 or 89, wherein the first expression cassette and/or vector, the second expression cassette and/or vector, and/or the third expression cassette and/or vector is an Adeno-associated virus vector.

82. The method of any one of claims 59-63 and 70-81, wherein the method has increased efficiency in modifying the target nucleic acid compared to the efficiency of a control method (e.g., a method comprising contacting the target nucleic acid with a templated editor that is not fused together via protein splicing optionally using an intein).

FIG. 1

CMV	N-term mCherry	NpuN
CMV	NpuC	C-term mCherry

FIG. 2

Average P4 mCherry Fluorescence (Median EDA)

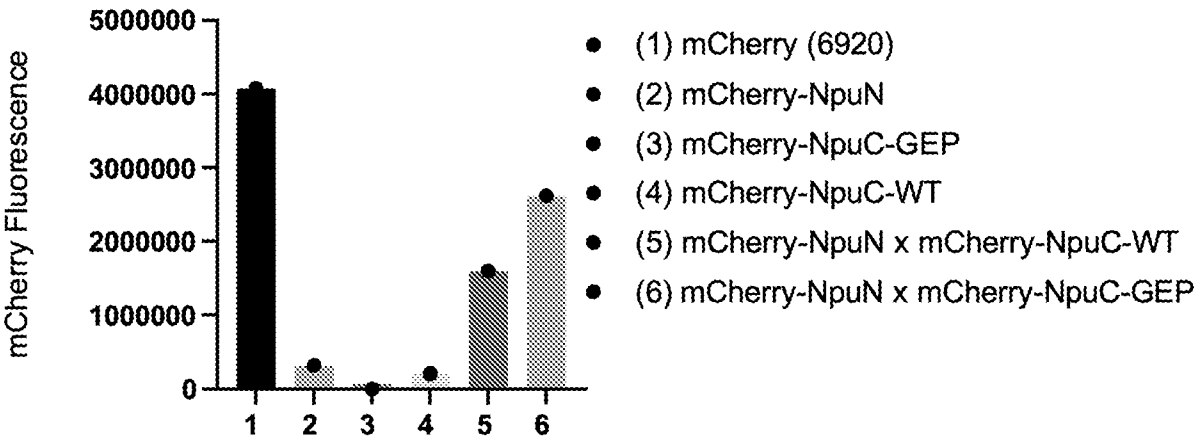


FIG. 3

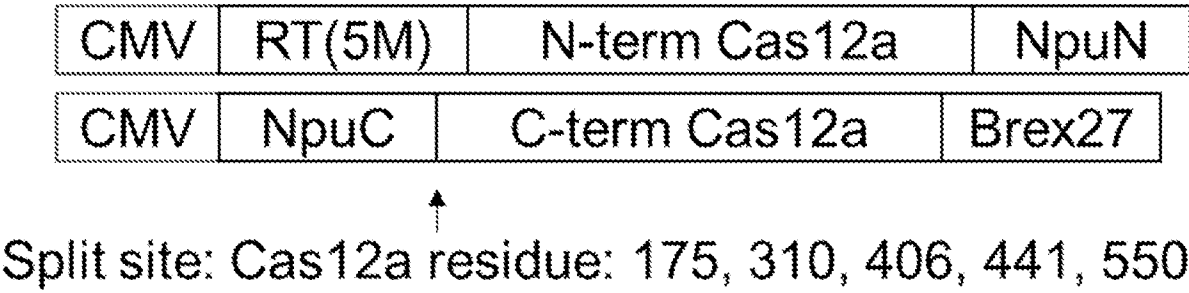


FIG. 4

Comparison of mCherry Fluorescence of mCherry v. mCherry Control Inteins

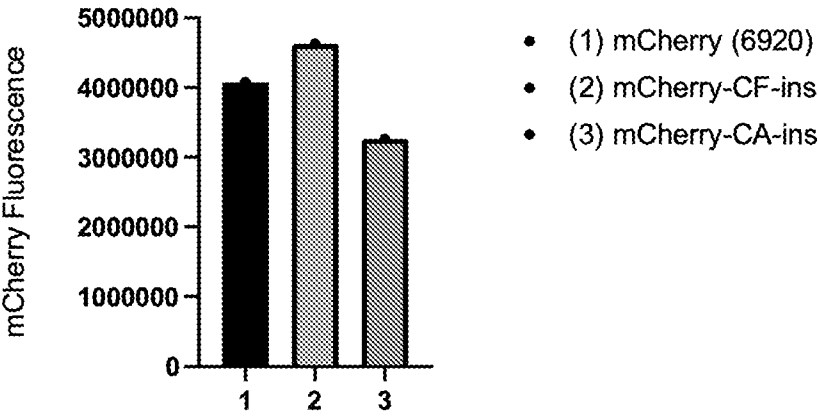
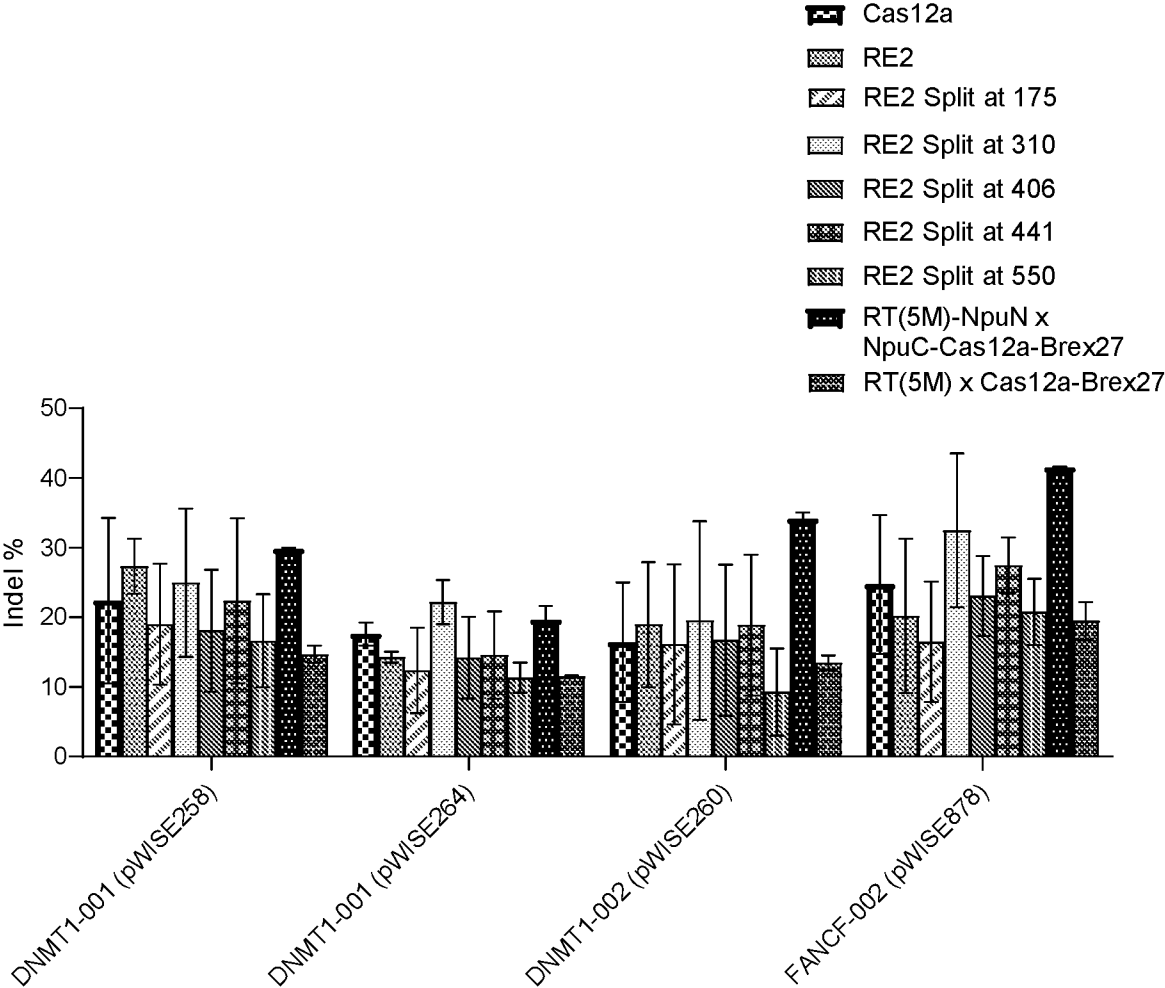


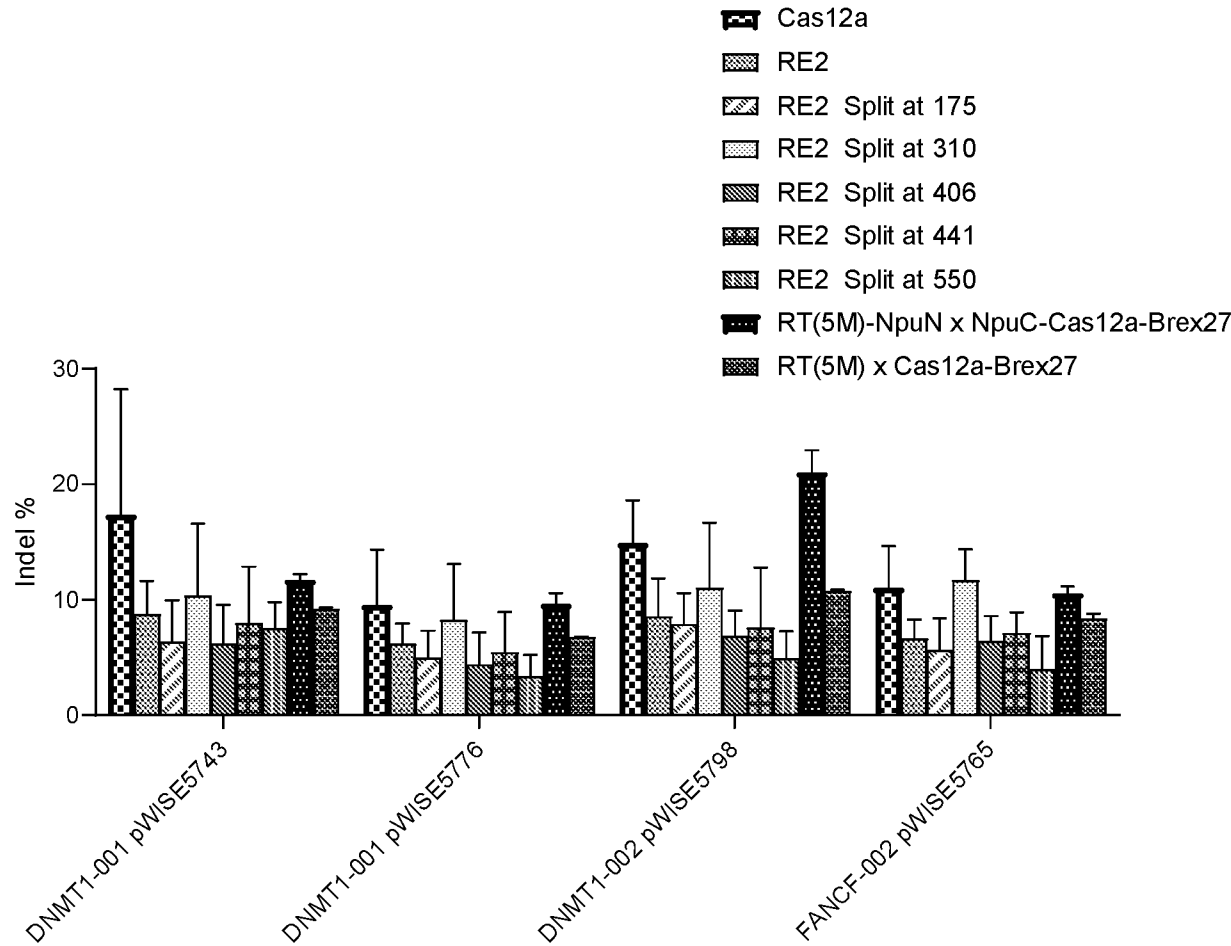
FIG. 5



Description:

- LbCpf1 x pUC19
- ▨ RT(5M) - LbCpf1-Brex27 x pUC19
- ▩ RT(5M)-Cas12a175-NpuN x NpuC-176Cas12a-Brex27
- ▧ RT(5M)-Cas12a310-NpuN x NpuC-311Cas12a-Brex27
- ▦ RT(5M)-Cas12a406-NpuN x NpuC-407Cas12a-Brex27
- ▥ RT(5M)-Cas12a441-NpuN x NpuC-442Cas12a-Brex27
- ▤ RT(5M)-Cas12a550-NpuN x NpuC-551Cas12a-Brex27
- ▣ RT(5M)-NpuN x NpuC-Cas12a-Brex27
- ▢ RT(5M) x Cas12a-Brex27

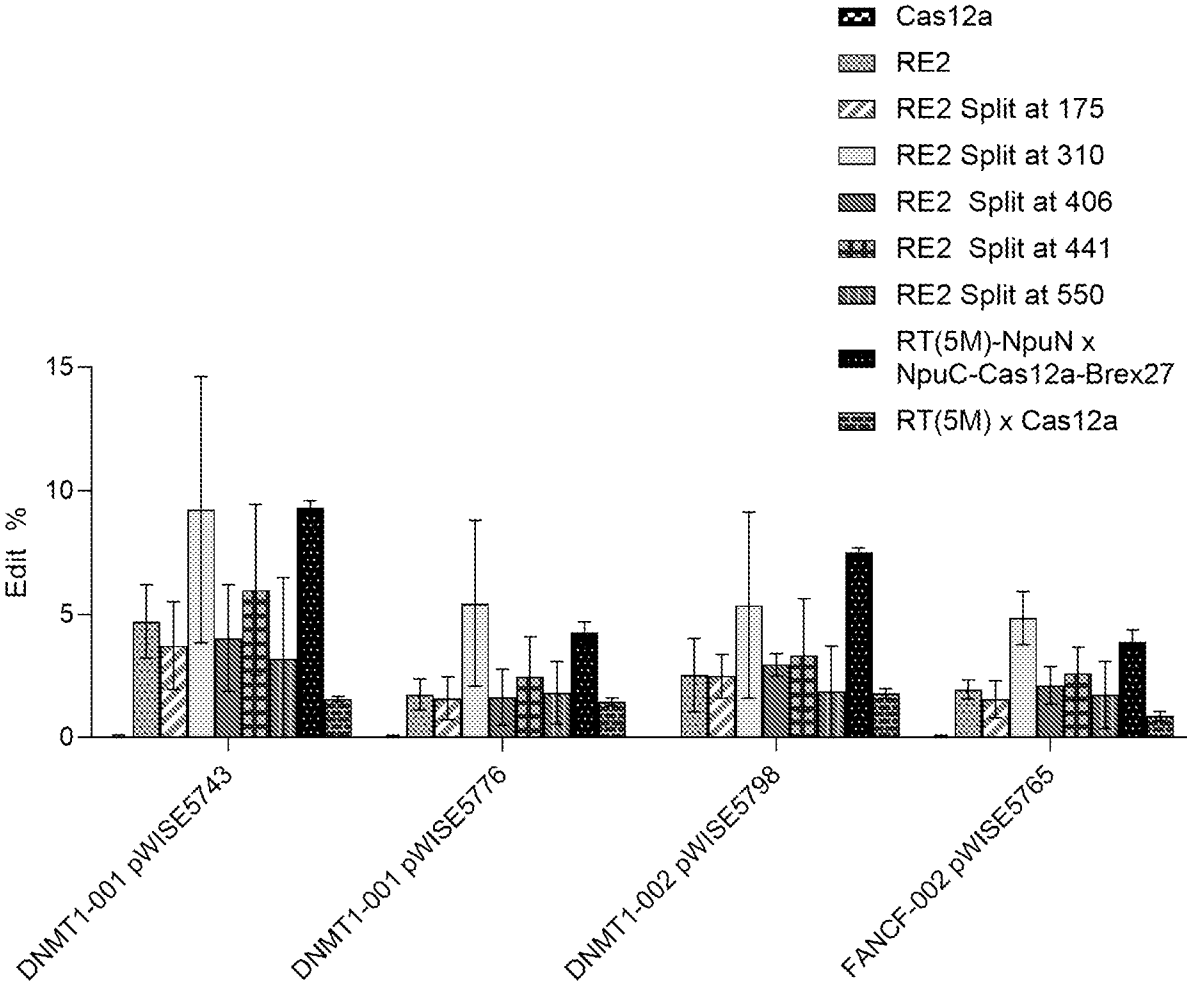
FIG. 6



Description:

- LbCpf1 x pUC19
- RT(5M) - LbCpf1-Brex27 x pUC19
- RT(5M)-Cas12a175-NpuN x NpuC-176Cas12a-Brex27
- RT(5M)-Cas12a310-NpuN x NpuC-311Cas12a-Brex27
- RT(5M)-Cas12a406-NpuN x NpuC-407Cas12a-Brex27
- RT(5M)-Cas12a441-NpuN x NpuC-442Cas12a-Brex27
- RT(5M)-Cas12a550-NpuN x NpuC-551Cas12a-Brex27
- RT(5M)-NpuN x NpuC-Cas12a-Brex27
- RT(5M) x Cas12a-Brex27

FIG. 7



Description:

- LbCpf1 x pUC19
- RT(5M) - LbCpf1-Brex27 x pUC19
- RT(5M)-Cas12a175-NpuN x NpuC-176Cas12a-Brex27
- RT(5M)-Cas12a310-NpuN x NpuC-311Cas12a-Brex27
- RT(5M)-Cas12a406-NpuN x NpuC-407Cas12a-Brex27
- RT(5M)-Cas12a441-NpuN x NpuC-442Cas12a-Brex27
- RT(5M)-Cas12a550-NpuN x NpuC-551Cas12a-Brex27
- RT(5M)-NpuN x NpuC-Cas12a-Brex27
- RT(5M) x Cas12a-Brex27