

ENGINEERED DEAMINASES, COMPOSITIONS COMPRISING THE SAME, AND METHODS OF USE THEREOF

RELATED APPLICATION INFORMATION

This application claims the benefit of U.S. Provisional Patent Application Serial No. 63/585,320, filed September 26, 2023, and U.S. Provisional Patent Application Serial No. 63/610,011, filed December 14, 2023, the disclosures of each of which are incorporated herein by reference in their entirety.

STATEMENT REGARDING ELECTRONIC FILING OF A SEQUENCE LISTING

A Sequence Listing in XML format, entitled 1499-136WO_ST26.xml, 1,464,865 bytes in size, generated on September 26, 2024, and filed herewith, is hereby incorporated by reference in its entirety for its disclosures.

FIELD

This invention relates to engineered deaminases along with compositions and systems comprising the same, and methods of use thereof.

BACKGROUND

Certain base editing systems can convert single C•G to T•A base pairs using cytosine base editors (CBEs) and rewrite single or multiple bases through the reverse-transcriptase activity of RNA-template REDRAW editors. In addition, adenine base editors (ABEs) can convert single A•T to C•G base pairs. New base editors may be advantageous.

SUMMARY OF THE INVENTION

A first aspect of the present invention is directed to an engineered protein comprising a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**.

A further aspect of the present invention is directed to a fusion protein comprising: a CRISPR-Cas effector protein; and a polypeptide that has at least about 70%, 75%, 80%, 85%,

90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**.

Another aspect of the present invention is directed to a fusion protein comprising: a first polypeptide having a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:267, 268, and/or 330-332**; and a second polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**.

An additional aspect of the present invention is directed to a fusion protein comprising: a nuclear localization sequence; and a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**.

A further aspect of the present invention is directed to a fusion protein comprising a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:343, 344, and/or 447-498**. In some embodiments, a fusion protein of the present invention comprises a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one of **SEQ ID NOs:343, 344, and/or 447-498**.

An additional aspect of the present invention is directed to a complex comprising an engineered protein of the present invention and a guide nucleic acid.

A further aspect of the present invention is directed to a nucleic acid encoding an engineered protein of the present invention. In some embodiments, a nucleic acid of the present invention comprises a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:345-416 and/or 499-580**. In some embodiments, a nucleic acid of the present invention comprises a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one of **SEQ ID NOs:345-352, 383-416, or 499-580**. Another aspect of the present invention is directed to an expression cassette codon optimized for expression in an organism, the expression cassette comprising a polynucleotide encoding a promoter sequence, and a polynucleotide encoding an engineered protein of the present invention, which may be codon-optimized for expression in the organism.

A further aspect of the present invention is directed to a method of modifying a target nucleic acid in a cell, the method comprising introducing an expression cassette and/or vector of the present invention into the cell, thereby modifying the target nucleic acid in the cell.

An additional aspect of the present invention is directed to a method for producing an engineered protein of the present invention, the method comprising culturing a cell or group of cells that have been transformed with a nucleic acid encoding the engineered protein; and isolating the engineered protein, thereby producing the engineered protein.

Another aspect of the present invention is directed a method of modifying a target nucleic acid, the method comprising contacting the target nucleic acid with an engineered protein of the present invention and a guide nucleic acid, thereby modifying the target nucleic acid.

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 illustrating an optimal alignment of a portion of **SEQ ID NO:590** and a portion of **SEQ ID NO:215** according to some embodiments of the present invention.

FIG. 2 is a schematic illustrating an optimal alignment of **SEQ ID NO:589** and **SEQ ID NO:215** according to some embodiments of the present invention.

FIG. 3 is a graph showing the percentage of A:T to G:C editing for different embodiments that included a tested fusion protein, such as an engineered protein according to some embodiments of the invention, or a control protein at a HEK2 locus associated with spacer sp12 (**SEQ ID NO:311**).

FIG. 4 is a graph showing the percentage of A:T to G:C editing for the embodiments of **FIG. 3**, but displays only the results of editing at A5 in the HEK2 locus associated with spacer sp12 (**SEQ ID NO:311**).

FIG. 5 shows graphs of the percentage of precise editing for embodiments that included a tested fusion protein, such as an engineered protein according to some embodiments of the invention, or a control protein at various sites for different target nucleic acids.

FIG. 6 is a graph showing the percentage of A:T to G:C editing for different embodiments that included a tested fusion protein, such as an engineered protein according to some embodiments of the invention, or a control protein with spacer sp12 (**SEQ ID NO:311**).

FIG. 7 is a graph showing the percentage of A:T to G:C editing for different embodiments that included a tested fusion protein, such as an engineered protein according to some embodiments of the invention, or a control protein with spacer sp90 (**SEQ ID NO:312**).

FIG. 8 is a graph showing the percentage of A:T to G:C editing for different embodiments that included a tested fusion protein, such as an engineered protein according to some embodiments of the invention, or a control protein with spacer sp92 (**SEQ ID NO:313**).

FIG. 9 is a graph showing the percentage of A:T to G:C editing for different embodiments that included a tested fusion protein, such as an engineered protein according to some embodiments of the invention, or a control protein with spacer sp143 (**SEQ ID NO:581**).

FIG. 10 is a graph showing the percentage of A:T to G:C editing for different embodiments that included a tested fusion protein, such as an engineered protein according to some embodiments of the invention, or a control protein with spacer sp454 (**SEQ ID NO:582**).

FIG. 11 and **FIG. 13** are graphs showing the percentage of A:T to G:C editing for different embodiments that included a tested fusion protein, such as an engineered protein according to some embodiments of the invention, or a control protein with spacer sp143 (**SEQ ID NO:581**).

FIG. 12 and **FIG. 14** are graphs showing the percentage of A:T to G:C editing for different embodiments that included a tested fusion protein, such as an engineered protein according to some embodiments of the invention, or a control protein with spacer sp454 (**SEQ ID NO:582**).

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 §§1.831 - 1.835 and the World Intellectual Property Organization (WIPO) Standard ST.26. 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.

As used herein, a "non-natural mutation" refers to a mutation that is generated through human intervention and differs from mutations found in the same gene or polypeptide that have occurred in nature (e.g., occurred naturally and not as a result of a modification made by a human).

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, EMBOSS 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, EMBOSS Cons, Kalign, MAFFT, MUSCLE, and/or T-Coffee. In some embodiments, an "optimal alignment" of two sequences (e.g., two polypeptide sequences or two polynucleotide 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 sequences 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 M 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 polypeptide of the present invention (e.g., an engineered protein and/or 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 deaminase, and/or a reverse transcriptase) 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 CRISPR-Cas effector protein and a peptide tag and/or a polypeptide of interest. 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 be a GS linker. In some embodiments, the peptide linker is a GS 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:1-36**. In some embodiments, the peptide linker may comprise an amino acid sequence of CA, CF, (GGG)_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(GGG)_n (**SEQ ID NO:25**), SGGS (**SEQ ID NO:26**), (GSS)_nG (**SEQ ID NO:36**), or (GGGG)_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, the linker has a sequence of (GSS)_nG (**SEQ ID NO:36**), wherein n is 3. In some embodiments, the linker has a sequence of (GSS)_nG (**SEQ ID NO:36**), wherein n is 5.

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:37** or **SEQ ID NO:38**).

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 RuBisCosmall 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, cetrum 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 RuBisCopromoter 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) *Dev. 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 (*pdk*), 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 and/or is used that comprises an engineered protein of the present invention.

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., Fok1), 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 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., Fok1), 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) 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/or an engineered protein of the present invention, and optionally a deaminase and/or reverse transcriptase). 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/or an engineered protein of the present invention, and optionally a deaminase and/or reverse transcriptase) 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 of an editing system may be contacted with 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 CRISPR-Cas effector protein, an engineered protein of the present invention (e.g., a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**), a guide nucleic acid, and optionally a deaminase and/or reverse transcriptase. In some embodiments, an engineered protein of the present invention comprises a CRISPR-Cas effector protein and the engineered 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, optionally in a composition, complex, ribonucleoprotein, kit, method, system, and/or editing system of the present invention.

In some embodiments, an editing system of the present invention comprises a reverse transcriptase, an extended guide nucleic acid, and an engineered protein of the present invention. In some embodiments, the engineered protein of the present invention, the reverse transcriptase, and the extended guide nucleic acid 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.

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 an engineered protein of the present invention, a polynucleotide encoding CRISPR-Cas effector protein, a polynucleotide encoding a nuclease, a polynucleotide encoding a reverse transcriptase, a polynucleotide encoding a reverse transcriptase 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 an engineered protein of the present invention, and a polynucleotide comprising a CRISPR-Cas effector protein 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 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 an engineered protein of the present invention, 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, 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 comprising, for example, a polynucleotide encoding a CRISPR-Cas effector protein and optionally a guide nucleic acid, under conditions whereby the CRISPR-Cas effector protein is expressed and the CRISPR-Cas effector protein forms a complex with a guide nucleic acid, the complex hybridizes to the target nucleic acid, and optionally the CRISPR-Cas effector protein recruits one or more additional component(s) (e.g., an engineered protein of the present invention, reverse transcriptase, glycosylase inhibitor, and/or deaminase) to it (and thus, to the target nucleic acid) or one or more additional component(s) (e.g., a polypeptide of the present invention) is/are fused to the CRISPR-Cas effector protein, thereby modifying the target nucleic acid. In some embodiments, a CRISPR-Cas effector protein and/or an engineered protein of the present invention localizes at a target nucleic acid, optionally through non-covalent interactions. Methods for recruiting one or more additional component(s) (e.g., an engineered protein of the present invention, reverse transcriptase and/or deaminase) may be used that take advantage of other protein-protein interactions, RNA-protein interactions, and/or chemical interactions. In some embodiments, a target nucleic acid may be contacted with a ribonucleoprotein comprising a CRISPR-Cas effector protein, a guide nucleic acid, and optionally an engineered protein of the present invention and the ribonucleoprotein hybridizes to the target nucleic acid, and optionally the CRISPR-Cas effector protein and/or engineered protein of the present invention may recruit one or more additional component(s) (e.g., a CRISPR-Cas effector protein, engineered protein of the present invention, reverse transcriptase, glycosylase inhibitor, and/or deaminase) to it (and thus, to the target nucleic acid) or one or more additional component(s) is/are fused to the CRISPR-Cas effector protein and/or engineered protein, thereby modifying the target nucleic acid.

In some embodiments, a target nucleic acid may be contacted with a nucleic acid construct of the invention encoding an engineered protein of the present invention and optionally a guide nucleic acid, under conditions whereby the engineered protein is expressed, or a target nucleic acid may be contacted with an engineered protein of the present invention and optionally a guide nucleic acid. The engineered protein of the present invention may form a complex with a guide nucleic acid and/or a CRISPR-Cas effector protein, and the complex may hybridize to a target nucleic acid, and/or an engineered protein of the present invention may be recruited to a target nucleic acid, which may result in modifying the target nucleic acid.

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, 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)); and/or bifunctional ligand approaches (e.g., chemically induced dimerization (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)). In some embodiments, a recruiting method and/or system of the present invention uses a protein-protein interaction and/or nucleic acid-protein interaction (e.g., RNA-protein interactions) to attract a polypeptide or polynucleotide to another polypeptide or polynucleotide (e.g., to a particular location in a genome).

“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 an engineered protein of the present invention, a CRISPR-Cas effector protein, and/or a guide nucleic acid may be introduced into a cell of an organism, thereby transforming the cell with the engineered protein, CRISPR-Cas effector protein, guide nucleic acid, and reverse transcriptase. In some embodiments, an engineered protein of the present invention, a CRISPR-Cas effector protein, and/or a guide nucleic acid may be introduced into a cell of an organism, optionally wherein the engineered protein, the CRISPR-Cas effector protein, and/or 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., a 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 an engineered protein of the present invention, 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, 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:37** or **SEQ ID NO:38**). 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, an engineered protein of the present invention 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, ligase activity, helicase activity, a nuclear localization sequence or activity, 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, 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 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:39-65 or 333-334** and/or a nucleotide sequence of any one of **SEQ ID NOs:66-69**. 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, 55 or 56**. 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:59, 60, or 61**. In some embodiments, a CRISPR-Cas effector protein may be Cas12b and optionally may have an amino acid sequence of **SEQ ID NO:70**. In some embodiments, a CRISPR-Cas effector protein may be Cas12f and optionally may have an amino acid sequence of **SEQ ID NO:71**. In some embodiments, a CRISPR-Cas

effector protein may be Cas12i and optionally may have an amino acid sequence of **SEQ ID NO:72**.

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, Cse1, Cse2, 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 that causes the protein to not have nuclease activity, is commonly referred to as “inactive” or “dead,” e.g., dCas9. 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 CRISPR Cas9 effector protein or Cas9 useful with this invention may be any known or later identified Cas9 nuclease. In some embodiments, a Cas9 may be a protein from, for example, *Streptococcus* spp. (e.g., *S. pyogenes*, *S. thermophilus*), *Lactobacillus* spp., *Bifidobacterium* spp., *Kandleria* spp., *Leuconostoc* spp., *Oenococcus* spp., *Pediococcus* spp., *Weissella* spp., and/or *Olsenella* spp. In some embodiments, a CRISPR-Cas effector protein may be a Cas9 and optionally may have a nucleotide sequence of any one of **SEQ ID NOs:73-87** and/or an amino acid sequence of any one of **SEQ ID NOs:88-97**.

In some embodiments, the CRISPR-Cas effector protein may be a Cas9 derived from *Streptococcus pyogenes* and/or may recognize the PAM sequence motif NGG, NAG, NGA (Mali et al, Science 2013; 339(6121): 823-826). In some embodiments, the CRISPR-Cas effector protein may be a Cas9 derived from *Streptococcus thermophiles* and/or may recognize the PAM sequence motif NGGNG and/or NNAGAAW (W = A or T) (See, e.g., Horvath et al, Science, 2010; 327(5962): 167-170, and Deveau et al, J Bacteriol 2008; 190(4): 1390-1400). In some embodiments, the CRISPR-Cas effector protein may be a Cas9 derived from

Streptococcus mutans and/or may recognize the PAM sequence motif NGG and/or NAAR (R = A or G) (See, e.g., Deveau et al, J BACTERIOL 2008; 190(4): 1390-1400). In some embodiments, the CRISPR-Cas effector protein may be a Cas9 derived from *Streptococcus aureus* and/or may recognize the PAM sequence motif NNGRR (R = A or G). In some embodiments, the CRISPR-Cas effector protein may be a Cas9 derived from *S. aureus* and/or may recognize the PAM sequence motif N GRRT (R = A or G). In some embodiments, the CRISPR-Cas effector protein may be a Cas9 derived from *S. aureus* and/or may recognize the PAM sequence motif N GRRV (R = A or G). In some embodiments, the CRISPR-Cas effector protein may be a Cas9 that is derived from *Neisseria meningitidis* and/or may recognize the PAM sequence motif N GATT or N GCTT (R = A or G, V = A, G or C) (See, e.g., Hou et al, PNAS 2013, 1-6). In the aforementioned embodiments in this paragraph, N in the PAM sequence motif can be any nucleotide residue, e.g., any of A, G, C or T. In some embodiments, the CRISPR-Cas effector protein may be a Cas13a derived from *Leptotrichia shahii* and/or may recognize a protospacer flanking sequence (PFS) (or RNA PAM (rPAM)) sequence motif of a single 3' A, U, or C, which may be located within the target nucleic acid.

A Type V CRISPR-Cas effector protein useful with embodiments of the invention may be any Type V CRISPR-Cas nuclease. Exemplary Type V CRISPR-Cas effector proteins include, but are 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 nuclease. In some embodiments, a Type V CRISPR-Cas effector protein may be a Cas12a. In some embodiments, a Type V CRISPR-Cas effector protein may be a nickase, optionally, a Cas12a nickase.

In some embodiments, 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., a 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 Cas12a may have an amino acid sequence of any one of **SEQ ID NOs:39-65 or 333-334**.

In some embodiments, a CRISPR-Cas effector protein is an engineered protein that has an amino acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:98-189 and/or 335-342**. In some embodiments, a CRISPR-Cas effector protein is an engineered protein (e.g., an engineered CRISPR-Cas effector protein) that comprises (i) a CRISPR-Cas effector polypeptide (e.g., a Cas12a polypeptide and/or domain) and (ii) a polypeptide that is heterologous to the CRISPR-Cas effector polypeptide. The CRISPR-Cas effector polypeptide may be a portion of a CRISPR-Cas effector protein (e.g., a wild-type Cas12a or inactivated Cas12a). For example, an engineered CRISPR-Cas effector protein may comprise (i) a portion of any one of **SEQ ID NOs:39-65 or 333-334** and (ii) a polypeptide that is heterologous to the portion of any one of **SEQ ID NOs:39-65 or 333-334**. In some embodiments, an engineered CRISPR-Cas effector protein comprises at least one non-natural mutation compared to a CRISPR-Cas effector protein (e.g., a wild-type Cas12a or inactivated Cas12a). In some embodiments, an engineered CRISPR-Cas effector protein has a similar structure and/or function compared to a wild-type CRISPR-Cas effector protein and/or comprises all or a portion of a wild-type CRISPR-Cas effector protein or inactivated CRISPR-Cas effector protein. In some embodiments, an engineered CRISPR-Cas effector protein is not found in nature (e.g., comprises a non-natural mutation) and may be modified (e.g., mutated) compared

to a wild-type CRISPR-Cas effector protein and/or inactivated CRISPR-Cas effector protein. In some embodiments, an engineered CRISPR-Cas effector protein comprises a Cas12a polypeptide and is devoid of a portion of a Cas12a. In some embodiments, an engineered CRISPR-Cas effector protein comprises a Cas12a polypeptide and a domain that is not present in a Cas12a. In some embodiments, an engineered CRISPR-Cas effector protein comprises all or a portion of a wedge domain, a Rec1 domain, a Rec2 domain, a PAM-interacting domain, a RuvC domain, a bridge helix, and/or a Nuc domain each of which may be from a Cas12a and/or a protein having a sequence of any one of **SEQ ID NOs:39-65 or 333-334**. In some embodiments, an engineered CRISPR-Cas effector protein comprises all or a portion of a Cas12a domain having a structure as described in Yamano, Takashi, et al., *Mol Cell* 67: 633-645 (2017).

In some embodiments, an engineered CRISPR-Cas effector protein comprises all or a portion of an alpha-helical recognition (REC) lobe, optionally wherein the domains of all or a portion of the REC lobe in the engineered protein may be in a different order and/or structure than in a Cas12a. A REC lobe may contain a Rec1 domain and Rec2 domain. A Rec1 domain may comprise 13 alpha helices and/or a Rec2 domain may comprise 10 alpha helices and two beta strands that may form a small antiparallel sheet. In some embodiments, an engineered CRISPR-Cas effector protein comprises an HNH domain (e.g., an HNH domain of a Cas9, optionally Cas9 having a sequence of any one of **SEQ ID NOs:88-97**) between a first polypeptide comprising all or a portion of a Rec1 domain and a second polypeptide comprising all or a portion of a Rec2 domain, wherein first and second polypeptides may each independently be from a Cas12a and/or a protein having a sequence of any one of **SEQ ID NOs:39-65 or 333-334**. In some embodiments, an engineered CRISPR-Cas effector protein comprises all or a portion of a RuvC domain. In some embodiments, an engineered CRISPR-Cas effector protein comprises an HNH domain (e.g., an HNH domain of a Cas9, optionally Cas9 having a sequence of any one of **SEQ ID NOs:88-97**) between a first polypeptide comprising all or a portion of a Rec1 domain and a second polypeptide comprising all or a portion of a Rec2 domain, wherein first and second polypeptides may each independently be from a Cas12a and/or a protein having a sequence of any one of **SEQ ID NOs:39-65 or 333-334**, and the engineered protein may comprise all or a portion of a RuvC domain. In some embodiments, an engineered CRISPR-Cas effector protein is as described in U.S. Patent Application Publication No. 2002/0112473 and/or PCT/US2023/063398. As one of skill in the art would understand, some domains (e.g., the wedge and RuvC domains of Cas12a) are not continuous in sequence and may be split into two or more (e.g., 2, 3, 4, or more) non-continuous

sequences. In some embodiments, an engineered CRISPR-Cas effector protein comprises a Cas12a domain.

In some embodiments, an engineered CRISPR-Cas effector protein comprises, optionally in the N- to C-terminus direction, all or a portion of one or more of the following: first portion of the wedge domain (WED-1), Rec1 domain, Rec2 domain, second portion of the wedge domain (WED-2), PAM-interacting domain (PI), third portion of the wedge domain (WED-3), first portion of the RuvC domain (RuvC-1), bridge helix, second portion of the RuvC domain (RuvC-2), Nuc domain, and/or third portion of the RuvC domain (RuvC-3). For example, based on **SEQ ID NO:69**, an engineered CRISPR-Cas effector protein may be encoded by all or a portion of one or more of the following: nucleotides 1-69 making up WED-1, nucleotides 70-1,560 making up the REC lobe, nucleotides 1,561-1,755 making up WED-2, nucleotides 1,756-2,031 making up PI, nucleotides 2,032-2,421 WED-3, nucleotides 2,422-2,613 making up RuvC-1, nucleotides 2,614-2,667 making up the bridge helix, nucleotides 2,668-2,988 making up RuvC-2, nucleotides 2,989-3,534 making up Nuc domain, and/or nucleotides 3,535-3,681 making up RuvC-3. As another example, based on **SEQ ID NO:56**, an engineered CRISPR-Cas effector protein may comprise, optionally in the N- to C-terminus direction, all or a portion of one or more of the following: amino acid residues 1-23 making up WED-1, amino acid residues 24-520 making up the REC lobe, amino acid residues 521-585 making up WED-2, amino acid residues 586-677 making up PI, amino acid residues 678-807 WED-3, amino acid residues 808-871 making up RuvC-1, amino acid residues 872-889 making up the bridge helix, amino acid residues 890-996 making up RuvC-2, amino acid residues 997-1,178 making up Nuc domain, and/or amino acid residues 1,179-1,227 making up RuvC-3.

In some embodiments, an engineered CRISPR-Cas effector protein comprises all or a portion of an active RuvC domain. In some embodiments, an engineered CRISPR-Cas effector protein comprises all or a portion of an inactivated RuvC domain, optionally all or a portion of an inactivated RuvC domain that has a D10A mutation. In some embodiments, an engineered CRISPR-Cas effector protein comprises all or a portion of an inactivated RuvC domain and has an alanine at a position corresponding to amino acid residue 831 of **SEQ ID NO:59** when the engineered protein is optimally aligned to **SEQ ID NO:59**, optionally wherein the mutation is referred to as a D10A and/or D832A mutation. In some embodiments, an engineered CRISPR-Cas effector protein comprises a polypeptide that comprises all or a portion of an inactivated RuvC domain and has a mutation of: D832A and/or E925A with reference to position numbering of **SEQ ID NO:44, 45, 55, 56, and/or 59** (LbCas12a), D908A and/or E993A with reference to position numbering of **SEQ ID NO:39** (AsCas12a), D917A and/or

E1006A with reference to position numbering of **SEQ ID NO:43 and/or 58** (FnCas12a), or a mutation corresponding thereto.

In some embodiments, a CRISPR-Cas effector protein (e.g., an engineered CRISPR-Cas effector protein) is a target strand nickase and/or a non-target strand nickase. In some embodiments, a CRISPR-Cas effector protein (e.g., an engineered CRISPR-Cas effector protein) is a non-target strand nickase.

In some embodiments, a CRISPR-Cas effector protein 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 polypeptide/domain or a Cas9 polypeptide/domain) may be optimized for expression in a plant.

A polypeptide (e.g., a fusion protein and/or engineered 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, or extended guide nucleic acid) that is designed to function with a CRISPR-Cas effector protein to modify a target nucleic acid. In some embodiments, an engineered protein comprises a CRISPR-Cas effector protein. 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 may be capable of forming a complex with a CRISPR-Cas effector protein (e.g., with a nuclease domain of the protein) and/or an engineered protein of the present invention 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 the CRISPR-Cas effector protein and/or engineered protein, optionally present in and/or recruited to the complex.

In some embodiments, a CRISPR-Cas effector protein comprising a Cas9 domain (or a nucleic acid construct encoding the same) may be used in combination with a Cas9 guide nucleic acid to modify a target nucleic acid, and may be in or may form a complex.

Likewise, a CRISPR-Cas effector protein may comprise a Cas12a domain (or other selected CRISPR-Cas nuclease, e.g., C2c1, C2c3, Cas12b, Cas12c, Cas12d, Cas12f, Cas12i, 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, Csel, Cse2, Csel, Csc2, 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), which may be used in combination with a Cas12a guide nucleic acid

(or the guide nucleic acid for the other selected CRISPR-Cas nuclease) to modify a target nucleic acid, thereby editing the target nucleic acid.

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 a 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; a repeat of a Type II Cas9 CRISPR-Cas system, or fragment thereof; a repeat of a Type V C2c1 CRISPR Cas system, or a fragment thereof; a repeat of a CRISPR-Cas system of, for example, C2c3, Cas12a (also referred to as Cpf1), Cas12b, Cas12c, Cas12d, Cas12e, Cas12f, Cas12i, Cas13a, Cas13b, Cas13c, Cas13d, Cas1B, Cas2, Cas3, Cas3', Cas3", Cas4, Cas5, Cas6, Cas7, Cas8, Cas9 (also known as Csn1 and Csx12), Cas10, Csy1, Csy2, Csy3, Cse1, Cse2, Cse1, Csc2, 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, or a fragment 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 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). 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. A recruiting guide RNA further comprises one or more recruiting motifs as described herein, which may be linked to the 5' end of the guide or the 3' end or it may be inserted into the recruiting guide nucleic acid (e.g., within the hairpin loop).

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, a guide nucleic acid further comprises a reverse transcriptase template and may be referred to as an extended guide nucleic acid.

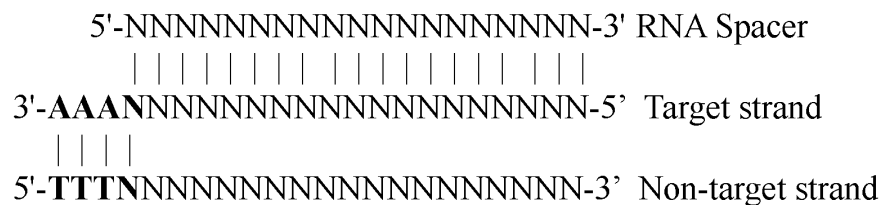
A guide nucleic acid and/or an extended guide nucleic acid may comprise one or more recruiting motifs as described herein, which may be linked to the 5' end and/or the 3' end of the guide nucleic acid and/or it may be inserted into the guide nucleic acid (e.g., within a hairpin loop of the guide nucleic acid).

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 V 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).



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 by 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 an engineered protein of the present invention and one or more guide nucleic acids, which 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 an engineered protein of the present invention 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 engineered protein of the present invention 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 an engineered protein of the present invention may be linked to an affinity polypeptide capable of interacting with the RNA recruiting motif, thereby recruiting the engineered protein of the invention to the target nucleic acid. Alternatively, chemical interactions may be used to recruit a polypeptide (e.g., an engineered protein of the invention) 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., an engineered protein of the present invention) 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., an engineered protein of the present invention). 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., one or more engineered protein(s) of the present invention).

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:190-200**.

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:201** and **SEQ ID NO:202**. An affinity polypeptide useful with peptide tags includes, but is not limited to, **SEQ ID NO:203**.

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 and/or polypeptide of the invention may be 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 an engineered protein of the present invention.

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 an engineered protein of the present invention. 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 with 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 engineered proteins. An engineered protein of the present invention comprises a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, an engineered protein of the present invention has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, an engineered protein of the present invention has 100% sequence identity to one of **SEQ ID NOs:241-266, 314-329, or 589-590**. In some embodiments, an engineered protein of the present invention has the sequence of one

of **SEQ ID NOs:241-266, 314-329, or 589-590**. In some embodiments, an engineered protein of the present invention comprises two or more copies of a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, a polypeptide of the present invention that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590** is an adenosine deaminase and/or has at least one function of an adenosine deaminase. In some embodiments, an "engineered protein" refers to a polypeptide or protein that is not found naturally in nature.

In some embodiments, a polypeptide of the present invention that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**, optionally when optimally aligned to one of **SEQ ID NOs:215-225**, has less than 70% (e.g., less than 65%, 60%, 55%, 50%, or 45%) sequence identity to the one of **SEQ ID NOs:215-225**. For example, in some embodiments, a polypeptide and/or engineered protein of the present invention having at least about 70% sequence identity to one of **SEQ ID NOs:241-266, 314-329, or 589-590** has less than 70% (e.g., less than 65%, 60%, 55%, 50%, or 45%) sequence identity to one of **SEQ ID NOs:215-225** when optimally aligned thereto. In some embodiments, a polypeptide and/or engineered protein of the present invention having at least about 70% sequence identity to one of **SEQ ID NOs:241-266, 314-329, or 589-590**, optionally when optimally aligned to one of **SEQ ID NOs:215-225**, has less than 70%, 65%, 60%, 55%, 50%, 45%, or 40% sequence identity to the one of **SEQ ID NOs:215-225**. In some embodiments, a polypeptide and/or engineered protein of the present invention having at least about 70% sequence identity to one of **SEQ ID NOs:241-266, 314-329, or 589-590**, optionally when optimally aligned to one of **SEQ ID NOs:215-225**, has less than 60% or less than 55% sequence identity to the one of **SEQ ID NOs:215-225**. In some embodiments, a polypeptide having a sequence of one of **SEQ ID NOs:241-266, 314-329, or 589-590** has less than 70%, 65%, 60%, 55%, 50%, 45%, or 40% sequence identity to one of **SEQ ID NOs:215-225** when optimally aligned thereto. In some embodiments, one of **SEQ ID NOs:215-225** is the reference sequence for a sequence alignment. In some embodiments, the reference sequence for a sequence alignment is **SEQ ID NO:215**. In some embodiments, a polypeptide and/or engineered protein of the present invention having at least about 70% sequence identity to one of **SEQ ID NOs:241-266, 314-329, or 589-590**, has less than 60% or less than 55% sequence identity to **SEQ ID NO:215** when optimally aligned thereto.

In some embodiments, an engineered protein and/or polypeptide of the present invention has at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to **SEQ ID NO:242**. In some embodiments, an engineered protein and/or polypeptide of the present invention has at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to **SEQ ID NO:590**. In some embodiments, an engineered protein and/or polypeptide of the present invention has at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to **SEQ ID NO:241**. In some embodiments, an engineered protein and/or polypeptide of the present invention has at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to **SEQ ID NO:589**. In some embodiments, one of **SEQ ID NOs:241-266, 314-329, or 589-590** is the reference sequence for a sequence alignment. In some embodiments, the reference sequence for a sequence alignment is **SEQ ID NO:241**. In some embodiments, the reference sequence for a sequence alignment is **SEQ ID NO:589**. In some embodiments, the reference sequence for a sequence alignment is **SEQ ID NO:242**. In some embodiments, the reference sequence for a sequence alignment is **SEQ ID NO:590**.

In some embodiments, a polypeptide and/or engineered protein of the present invention, optionally when optimally aligned to **SEQ ID NO:215**, has: a valine at an amino acid residue corresponding to amino acid residue 105 of **SEQ ID NO:215**, an asparagine at an amino acid residue corresponding to amino acid residue 107 of **SEQ ID NO:215**, an arginine at an amino acid residue corresponding to amino acid residue 110 of **SEQ ID NO:215**, a tyrosine at an amino acid residue corresponding to amino acid residue 148 of **SEQ ID NO:215**, and/or a proline at an amino acid residue corresponding to amino acid residue 151 of **SEQ ID NO:215**. For example, in some embodiments, a polypeptide and/or engineered protein of the present invention, when optimally aligned to **SEQ ID NO:215**, has: a valine at amino acid residue 105 with reference to the residue numbering of **SEQ ID NO:215** (i.e., amino acid residue 105 of **SEQ ID NO:215** corresponds to a respective amino acid residue in the polypeptide and/or engineered protein based on an optimal alignment), an asparagine at an amino acid residue 107 with reference to the residue numbering of **SEQ ID NO:215**, an arginine at an amino acid residue 110 with reference to the residue numbering of **SEQ ID NO:215**, a tyrosine at an amino acid residue 148 with reference to the residue numbering of **SEQ ID NO:215**, and/or a proline at an amino acid residue 151 with reference to the residue numbering of **SEQ ID NO:215**. In some embodiments, a polypeptide and/or engineered protein of the present invention, optionally when optimally aligned to **SEQ ID NO:215**, has: a valine at an amino acid residue corresponding to amino acid residue 105 of **SEQ ID NO:215**, an asparagine at an amino acid

residue corresponding to amino acid residue 107 of **SEQ ID NO:215**, an arginine at an amino acid residue corresponding to amino acid residue 110 of **SEQ ID NO:215**, a tyrosine at an amino acid residue corresponding to amino acid residue 148 of **SEQ ID NO:215**, and a proline at an amino acid residue corresponding to amino acid residue 151 of **SEQ ID NO:215**. For example, as shown in **FIG. 1**, when **SEQ ID NO:215** and **SEQ ID NO:590** are optimally aligned, amino acid residue 101 of **SEQ ID NO:590** is valine and corresponds to amino acid residue 105 of **SEQ ID NO:215**, amino acid residue 103 of **SEQ ID NO:590** is asparagine and corresponds to amino acid residue 107 of **SEQ ID NO:215**, amino acid residue 106 of **SEQ ID NO:590** is arginine and corresponds to amino acid residue 110 of **SEQ ID NO:215**, amino acid residue 144 of **SEQ ID NO:590** is tyrosine and corresponds to amino acid residue 148 of **SEQ ID NO:215**, and amino acid residue 147 of **SEQ ID NO:590** is proline and corresponds to amino acid residue 151 of **SEQ ID NO:215**. Similarly, as shown in **FIG. 2**, **SEQ ID NO:215** and **SEQ ID NO:589** are optimally aligned, and **SEQ ID NO:589** has a valine at an amino acid residue corresponding to amino acid residue 105 of **SEQ ID NO:215**, an asparagine at an amino acid residue corresponding to amino acid residue 107 of **SEQ ID NO:215**, an arginine at an amino acid residue corresponding to amino acid residue 110 of **SEQ ID NO:215**, a tyrosine at an amino acid residue corresponding to amino acid residue 148 of **SEQ ID NO:215**, and a proline at an amino acid residue corresponding to amino acid residue 151 of **SEQ ID NO:215**.

In some embodiments, a polypeptide and/or engineered protein of the present invention, optionally when optimally aligned to **SEQ ID NO:590**, has: a valine at an amino acid residue corresponding to amino acid residue 101 of **SEQ ID NO:590**, an asparagine at an amino acid residue corresponding to amino acid residue 103 of **SEQ ID NO:590**, an arginine at an amino acid residue corresponding to amino acid residue 106 of **SEQ ID NO:590**, a tyrosine at an amino acid residue corresponding to amino acid residue 144 of **SEQ ID NO:590**, and/or a proline at an amino acid residue corresponding to amino acid residue 147 of **SEQ ID NO:590**. In some embodiments, a polypeptide and/or engineered protein of the present invention, optionally when optimally aligned to **SEQ ID NO:590**, has: an alanine at an amino acid residue corresponding to amino acid residue 9 of **SEQ ID NO:590**, an alanine at an amino acid residue corresponding to amino acid residue 44 of **SEQ ID NO:590**, a serine at an amino acid residue corresponding to amino acid residue 78 of **SEQ ID NO:590**, a phenylalanine at an amino acid residue corresponding to amino acid residue 80 of **SEQ ID NO:590**, a serine at an amino acid residue corresponding to amino acid residue 105 of **SEQ ID NO:590**, and/or an arginine at an amino acid residue corresponding to amino acid residue 150 of **SEQ ID NO:590**. In some

embodiments, a polypeptide and/or engineered protein of the present invention, optionally when optimally aligned to **SEQ ID NO:590**, has: a valine at an amino acid residue corresponding to amino acid residue 101 of **SEQ ID NO:590**, an asparagine at an amino acid residue corresponding to amino acid residue 103 of **SEQ ID NO:590**, an arginine at an amino acid residue corresponding to amino acid residue 106 of **SEQ ID NO:590**, a tyrosine at an amino acid residue corresponding to amino acid residue 144 of **SEQ ID NO:590**, a proline at an amino acid residue corresponding to amino acid residue 147 of **SEQ ID NO:590**, an alanine at an amino acid residue corresponding to amino acid residue 9 of **SEQ ID NO:590**, an alanine at an amino acid residue corresponding to amino acid residue 44 of **SEQ ID NO:590**, a serine at an amino acid residue corresponding to amino acid residue 78 of **SEQ ID NO:590**, a phenylalanine at an amino acid residue corresponding to amino acid residue 80 of **SEQ ID NO:590**, a serine at an amino acid residue corresponding to amino acid residue 105 of **SEQ ID NO:590**, and/or an arginine at an amino acid residue corresponding to amino acid residue 150 of **SEQ ID NO:590**.

In some embodiments, an engineered protein of the present invention is a fusion protein. A fusion protein of the present invention may comprise a first polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**; and a second polypeptide. In some embodiments, the second polypeptide is heterologous to the first polypeptide. In some embodiments, the first polypeptide and the second polypeptide are directly fused (e.g., linked via a peptide bond) together without an intervening element between the first polypeptide and the second polypeptide. In some embodiments, a linker (e.g., a peptide linker) fuses/joins the first polypeptide and the second polypeptide. The linker may be a peptide linker that has a length of 1, 5, 10, or 15 amino acid residue(s) to 20, 25, 30, 35, 45, or 50 amino acid residues. In some embodiments, a fusion protein of the present invention comprises a peptide linker that has a length of 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, or 50 amino acid residue(s). A fusion protein of the present invention may comprise one or more linker(s) (e.g., 1, 2, 3, 4, or more) (e.g., one or more peptide linker(s)). In some embodiments, a fusion protein of the present invention comprises two linkers (e.g., two peptide linkers). In some embodiments, a fusion protein of the present invention comprises a peptide linker that comprises glycine and/or serine. In some embodiments, a fusion protein of the present invention comprises a peptide linker that has an amino acid sequence of one of **SEQ ID NOs:1-36** or a sequence selected from: CA, CF, (GGG)_n, GS, SG, 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, a fusion protein of the present invention comprises a first polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**, and one or more additional polypeptide(s) (e.g., one or more second polypeptides that may be the same or different from each other), which optionally are selected from the sequence of one or more of **SEQ ID NOs:267, 268, and/or 330-332**, a CRISPR-Cas effector protein, a polypeptide of interest, and/or a nuclear localization sequence.

An engineered protein of the present invention may comprise a first polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590** and a second polypeptide comprising the sequence of **SEQ ID NO:267**. In some embodiments, an engineered protein of the present invention comprises a first polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590** and a second polypeptide comprising the sequence of **SEQ ID NO:268**. In some embodiments, an engineered protein of the present invention comprises a first polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590** and a second polypeptide comprising the sequence of **SEQ ID NO:330**. In some embodiments, an engineered protein of the present invention comprises a first polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590** and a second polypeptide comprising the sequence of **SEQ ID NO:331**. In some embodiments, an engineered protein of the present invention comprises a first polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590** and a second polypeptide comprising the sequence of **SEQ ID NO:332**. The second polypeptide may be fused to the N-terminus or to the C-terminus of the first polypeptide. In some embodiments, an engineered protein of the present invention comprises a first polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**, a second polypeptide comprising the sequence of **SEQ ID NO:267**, and a third polypeptide comprising the sequence of **SEQ ID NO:268**, which may be fused together in any orientation (e.g., both the second and third polypeptides at the C-terminus or N-terminus of the

first polypeptide or one of the second and third polypeptides at the C-terminus of the first polypeptide and the other at the N-terminus of the first polypeptide).

In some embodiments, an engineered protein of the present invention comprises a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590** and a CRISPR-Cas effector protein. In some embodiments, the CRISPR-Cas effector protein has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one of **SEQ ID NOs:39-65, 70-72, 88-189, or 333-342**. The CRISPR-Cas effector protein and the polypeptide may be directly fused (e.g., linked via a peptide bond) together without an intervening element between the CRISPR-Cas effector protein and the polypeptide. In some embodiments, a linker (e.g., a peptide linker) fuses/joins the CRISPR-Cas effector protein and the polypeptide. The linker may be a peptide linker that has a length of 1, 5, 10, or 15 amino acid residue(s) to 20, 25, 30, 35, 45, or 50 amino acid residues. In some embodiments, the peptide linker has a length of 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, or 50 amino acid residue(s) and/or has an amino acid sequence of one of **SEQ ID NOs:1-36** or a sequence selected from: CA, CF, (GGG)_n, GS, SG, 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, a fusion protein of the present invention may comprise a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590** that is fused (directly or indirectly) to the N- and/or C-terminus of a CRISPR-Cas effector protein. In some embodiments, the polypeptide is fused to the N-terminus of the CRISPR-Cas effector protein. In some embodiments, the polypeptide is fused to the C-terminus of the CRISPR-Cas effector protein. In some embodiments, the polypeptide is fused to the N-terminus of the CRISPR-Cas effector protein.

In some embodiments, a fusion protein of the present invention comprises a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one of **SEQ ID NOs:241-266, 314-329, or 589-590**, wherein the polypeptide is fused to the C-terminus of a CRISPR-Cas effector protein having a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:39-65, 70-72, 88-189, and/or 333-342**. In some embodiments, a fusion protein of the present invention comprises a polypeptide that has

at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one of **SEQ ID NOs:241-266, 314-329, or 589-590**, wherein the polypeptide is fused to the C-terminus of a CRISPR-Cas effector protein having a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one of **SEQ ID NOs:39-65, 70-72, 88-97 or 333-334**. In some embodiments, a fusion protein of the present invention comprises a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one of **SEQ ID NOs:241-266, 314-329, or 589-590**, wherein the polypeptide is fused to the C-terminus of a CRISPR-Cas effector protein having a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to of **SEQ ID NOs:98-189 or 335-342**.

A fusion protein of the present invention may comprise a CRISPR-Cas effector protein having a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:39-65, 70-72, 88-189, and/or 333-342**. In some embodiments, the CRISPR-Cas effector protein has a sequence of one of **SEQ ID NOs:39-65, 70-72, 88-189, or 333-342**. In some embodiments, a fusion protein of the present invention comprises one or more (e.g., 1, 2, 3, or more) polypeptide(s) that have a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, a fusion protein of the present invention comprises two or more polypeptides that have a sequence of at least about 70% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590** and/or two or more copies of a polypeptide that has a sequence of at least about 70% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, a fusion protein of the present invention comprises a CRISPR-Cas effector protein having a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:39-65, 70-72, 88-189, and/or 333-342** that is fused (directly or indirectly) to a polypeptide that has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, a fusion protein of the present invention comprises a CRISPR-Cas effector protein that is fused to a polypeptide having a sequence of one of **SEQ ID NOs:241-266, 314-329, or 589-590**.

A fusion protein of the present invention may comprise a polypeptide of interest that is fused (directly or indirectly) to the N- or C-terminus of a polypeptide that has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence

identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, a fusion protein of the present invention comprises two or more (e.g., 2, 3, 4, or more) polypeptides of interest that are the same or different from each other.

In some embodiments, an engineered protein of the present invention comprises a deaminase (e.g., an adenosine deaminase and/or a cytidine deaminase). Two or more (e.g., 2, 3, 4 or more) deaminases that are the same or different may be present in an engineered protein of the present invention. A deaminase may be fused (directly or indirectly) to the N- or C-terminus of a polypeptide that has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, an engineered protein of the present invention comprises an adenosine deaminase and a cytidine deaminase. In some embodiments, an engineered protein of the present invention is fused to a cytosine deaminase (e.g., all or a portion of a cytosine deaminase having at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one of **SEQ ID NOs:204-213**, optionally wherein the cytosine deaminase is fused (directly or indirectly) to the N- or C-terminus of a polypeptide that has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, an engineered protein of the present invention is fused to a glycosylase inhibitor, optionally wherein the glycosylase inhibitor is a uracil glycosylase inhibitor (UGI), e.g., all or a portion of a UGI having at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to **SEQ ID NO:214**. In some embodiments, an engineered protein of the present invention is fused (directly or indirectly) to a cytosine deaminase and a glycosylase inhibitor in any order. In some embodiments, an engineered protein of the present invention is fused to an adenine deaminase (e.g., all or a portion of an adenine deaminase having at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one of **SEQ ID NOs:215-225**), optionally wherein the adenine deaminase is fused (directly or indirectly) to the N- or C-terminus of a polypeptide that has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, an engineered protein of the present invention is fused (directly or indirectly) to an adenine deaminase and a cytosine deaminase in any order, optionally wherein the adenine deaminase is fused (directly or indirectly) to one of the N- or C-terminus of a polypeptide that has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-**

329, and/or 589-590 and the cytosine deaminase is fused (directly or indirectly) to the other of the N- or C-terminus of the polypeptide that has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, an engineered protein of the present invention comprises a reverse transcriptase, optionally wherein the reverse transcriptase is fused (directly or indirectly) to the N- or C-terminus of a polypeptide that has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, an engineered protein of the present invention is fused to a nuclear localization signal (NLS). An engineered protein of the present invention may comprise one or more (e.g., 1, 2, 3, or more) nuclear localization signal(s). In some embodiments, a nuclear localization signal 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:226-228**. In some embodiments, an engineered protein of the present invention comprises a NLS that is fused (directly or indirectly) to the N- or C-terminus of a polypeptide that has a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, a NLS is present at the N-terminus and/or C-terminus of an engineered protein of the present invention.

An engineered protein of the present invention may provide at least one improvement (e.g., modified editing window, increased nuclease activity, increased deaminase activity, increased stability, increased binding to a nucleic acid, etc.) compared to an editing system that is devoid of a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, an engineered protein of the present invention has at least one improvement (e.g., increased activity and/or binding) compared to a protein having an amino acid sequence of any one of **SEQ ID NOs:215-255**. In some embodiments, an engineered protein of the present invention has improved (e.g., increased) DNA base editing activity and/or specificity compared to the DNA base editing activity and/or specificity of an adenosine deaminase of one of **SEQ ID NOs:215-255** when compared under the same conditions. In some embodiments, an engineered protein of the present invention has improved (e.g., increased) deaminase activity compared to the deaminase activity of an adenosine deaminase of one of **SEQ ID NOs:215-255** when compared under the same conditions. In some embodiments, an engineered protein of the present invention has improved (e.g.,

increased) RNA base editing activity and/or specificity compared to the RNA base editing activity and/or specificity of an adenosine deaminase of one of **SEQ ID NOs:215-255** when compared under the same conditions. In some embodiments, an engineered protein of the present invention is used (e.g., in an editing system and/or complex) for A to U base editing (e.g., A to U in mRNA base editing) and/or for A to G DNA base editing.

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:204**. In some embodiments, the cytosine deaminase may be an APOBEC3A deaminase having the amino acid sequence of **SEQ ID NO:205**. In some embodiments, the cytosine deaminase may be an CDA1 deaminase, optionally a CDA1 having the nucleotide sequence of **SEQ ID NO:206**. In some embodiments, the cytosine deaminase may be a FERNY deaminase, optionally a FERNY having the amino acid sequence of **SEQ ID NO:209**. In some embodiments, the cytosine deaminase may be a rAPOBEC1 deaminase, optionally a rAPOBEC1 deaminase having the amino acid sequence of **SEQ ID NO:210**. In some embodiments, the cytosine deaminase may be a hAID deaminase, optionally a hAID having the amino acid sequence of **SEQ ID NO:208** or **SEQ ID NO:211**. 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:212**, **SEQ ID NO:213**, **SEQ ID NO:214**). 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:204-213**(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:204-213**). 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:215**. In some embodiments, a mutated/evolved *E. coli* TadA* comprises the amino acid sequence of any one of **SEQ ID NOs:216-219**. 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:215-225**. In some embodiments, an adenine deaminase may comprise all or a portion of an amino acid sequence of any one of **SEQ ID NOs: 215-225**.

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 an engineered protein and 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:214** or a polypeptide having about 70% to about 99.5% identity to the amino acid sequence of **SEQ ID NO:214** (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:214**). For example, in some embodiments, a UGI domain may comprise a fragment of the amino acid sequence of **SEQ ID NO:214** 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:214**. In some embodiments, a UGI domain may be a variant of a known UGI (e.g., **SEQ ID NO:214**) 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 “reverse transcriptase” as used herein includes, but is not limited to, a naturally occurring reverse transcriptase, an engineered reverse transcriptase, or a commercially available reverse transcriptase. Exemplary reverse transcriptases include, but are not limited to, Moloney Murine Leukemia Virus Reverse Transcriptase (MMLV-RT or M-MuLV-RT), mutated MMLV (e.g., 5M-MMLV), Avian Myeloblastosis Virus Reverse Transcriptase (AMV-RT), Human Immunodeficiency Virus Reverse Transcriptase (HIV-RT). In some embodiments, a reverse transcriptase may be MMLV and/or 5M-MMLV. In some embodiments, a reverse transcriptase may have a sequence that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:229-240**.

A nucleic acid molecule may encode an engineered protein of the present invention, which may be present in an expression cassette and/or vector. 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 an engineered protein of the present invention), and/or a guide nucleic acid may be codon optimized for expression in an organism (e.g., an animal, a plant, a fungus, an archaeon, or a bacterium). In some embodiments, an expression cassette and/or vector of the present invention comprises a polynucleotide that encodes a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, an expression cassette and/or vector of the present invention comprises a first polynucleotide that encodes a promoter sequence and a second polynucleotide that encodes an engineered protein that comprises at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. A polynucleotide of the present invention may be codon-optimized for expression in a particular organism (e.g., a human or plant). In

some embodiments, the organism is an animal (e.g., a human), a plant, a fungus, an archaeon, or a bacterium.

Methods of modifying a target nucleic acid in a cell are provided, and may comprise introducing an expression cassette and/or vector of the present invention into the cell to provide a modified target nucleic acid. In some embodiments, the cell is a plant cell and the method further comprises regenerating the plant cell comprising the modified target nucleic acid to produce a plant comprising the modified target nucleic acid. In some embodiments, the introducing of the expression cassette is carried out at a temperature of about 20°C to about 42°C. In some embodiments, the cell is a mammalian cell (e.g., a human cell).

Methods for producing an engineered protein of the present invention are provided herein and may comprise culturing a cell or a plurality of cells that have been transformed with a nucleic acid encoding an engineered protein of the present invention; and isolating the engineered protein of the present invention to thereby produce the engineered protein. In some embodiments, the step of isolating is performed by dialysis, centrifugation, column purification, and/or the like.

Methods for performing reverse transcription are provided herein and may comprise contacting a target nucleic acid with an engineered protein of the present invention. In some embodiments, all or a portion of the engineered protein has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. An engineered protein of the present invention optionally together with one or more components may reverse transcribe the target nucleic acid to provide a DNA (e.g., a cDNA). In some embodiments, the target nucleic acid is in a plant and/or in a plant cell that includes a cell wall. In some embodiments, the target nucleic acid is in a mammal and/or is in a mammalian cell.

Methods of modifying a target nucleic acid are provided herein and may comprise contacting the target nucleic acid with an engineered protein of the present invention and a guide nucleic acid. In some embodiments, all or a portion of the engineered protein has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or more sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**. In some embodiments, the target nucleic acid is in a plant and/or in a plant cell that includes a cell wall. In some embodiments, the target nucleic acid is in a mammal and/or is in a mammalian cell. In some embodiments, the method of modifying the target nucleic acid is a templated editing method. In some embodiments, the method further comprises introducing a nucleic acid molecule encoding the engineered protein into the cell and producing the engineered protein in the cell.

A method of the present invention may comprise contacting the target nucleic acid with an extended guide nucleic acid. In some embodiments, the extended guide nucleic acid comprises a primer binding site and the target nucleic acid is double stranded and comprises a first strand and a second strand. In some embodiments, the primer binding site of the extended guide nucleic acid binds to the first strand or to the second strand of the target nucleic acid. In some embodiments, the second strand is the non-target strand of the target nucleic acid. In some embodiments, the target nucleic acid is double stranded and comprises a first strand and a second strand and the primer binding site binds to the first strand of the target nucleic acid. The first strand may be the target strand of the target nucleic acid and/or an engineered protein of the present invention may be recruited to the first strand. In some embodiments, the target nucleic acid is double stranded and comprises a first strand and a second strand and the primer binding site of the extended guide nucleic acid binds to the second strand of the target nucleic acid. In some embodiments, the second strand is the non-target strand of the target nucleic acid and/or the engineered protein is recruited to the second strand. In some embodiments, the engineered protein of the present invention is a double stranded nuclease that cuts the first strand and the second strand of the target nucleic acid resulting in a double stranded break. In some embodiments, the engineered protein and the extended guide nucleic acid form a complex or are comprised in a complex.

A method of the present invention may comprise contacting a target nucleic acid with and/or introducing into a cell an extended guide nucleic acid that comprises (i) a CRISPR nucleic acid and/or a CRISPR nucleic acid and a tracr nucleic acid; and (ii) an extended portion comprising a primer binding site and a reverse transcriptase template (RT template). In some embodiments, the extended portion of the extended guide nucleic acid is fused to either the 5' end or 3' end of the 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, the extended portion of the extended guide nucleic acid comprises, 5' to 3', an RT template and a primer binding site. In some embodiments, the extended portion of the extended guide nucleic acid is located 5' of the crRNA.

In some embodiments, the primer binding site has a length of about one nucleotide to about 100 nucleotides. In some embodiments, the primer binding site is at least 45 nucleotides in length or about 45 nucleotides to about 100 nucleotides, e.g., 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, or 100 nucleotides in length.

In some embodiments, the RT template has a length of about one to about 100 nucleotides, or the RT template can have a length of about 40 nucleotides or less, e.g., 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 nucleotide.

In some embodiments, the extended portion of the extended guide nucleic acid is linked to the CRISPR nucleic acid and/or the tracrRNA via a linker. The linker may be about 1 to about 100 nucleotides in length, 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, 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, or 100 nucleotides in length.

In some embodiments, in a method of the present invention, an engineered protein of the present invention is fused to one or more components that recruit the engineered protein to a CRISPR-Cas effector protein, deaminase, reverse transcriptase, guide nucleic acid, and/or target nucleic acid. In some embodiments, an engineered protein of the present invention is fused, e.g. linked, to a peptide tag, for example, an epitope or a multimerized epitope, and a CRISPR-Cas effector protein is fused, e.g., linked, to an affinity polypeptide that binds to the peptide tag. In some embodiments, an engineered protein of the present invention is fused, e.g., linked, to a peptide tag, for example, an epitope or a multimerized epitope, and a CRISPR-Cas effector protein is fused (linked) to an affinity polypeptide that binds to the peptide tag. In some embodiments, the target nucleic acid is contacted with two or more reverse transcriptases.

In some embodiments, an extended guide nucleic acid is linked to an RNA recruiting motif, and a reverse transcriptase is fused, e.g., linked, to an affinity polypeptide that binds to the RNA recruiting motif. In some embodiments, the target nucleic acid is contacted with two or more reverse transcriptases. In some embodiments, the extended guide nucleic acid, e.g. extended guide RNA, is linked to two or more RNA recruiting motifs, optionally wherein the two or more RNA recruiting motifs are the same RNA recruiting motif or different RNA recruiting motifs. In some embodiments, at least one of the two or more RNA recruiting motifs is located on the 3' end of the extended portion of the extended guide nucleic acid or is embedded in the extended portion.

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 (e.g., a method that uses a control adenosine deaminase (e.g., an adenosine deaminase having a sequence of one of **SEQ ID NOs:215-225** and is performed under the same conditions). In some

embodiments, a control method uses an adenine base editor that optionally comprises a sequence of **SEQ ID NOs:220, 221, 269, and/or 274-276** and/or is encoded by a sequence of **SEQ ID NOs:270-273 and/or 277-280**. 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.

An engineered protein of the present invention may be present in a complex with a guide nucleic acid. The guide nucleic acid may be designed to function with a CRISPR-Cas effector protein that is optionally present in the complex. In some embodiments, a complex and/or method of the present invention may be a complex and/or method 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 may include an engineered protein of the present invention optionally in place of an adenosine deaminase.

In some embodiments, an editing system of the present invention is used in prime editing. “Prime editing” and grammatical variants thereof as used herein refer to a nucleic acid editing technology that uses a Cas9 nickase domain fused to a reverse transcriptase and modifies a target nucleic acid without a double strand break or a donor DNA template. In Prime editing, the Cas9 nickase domain cuts the non-complementary strand of DNA upstream of the PAM site, thereby providing a 3’ flap that is extended with the extension including a modification. Further details on Prime editing can be found in Anzalone et al. (2019) *Nature* 576, 149–157 and/or U.S. Patent Application Publication No. 2021/0147862, the contents of each of which are incorporated herein by reference in their entirety.

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, polypeptides, 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*, 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 an engineered protein of the present 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 polypeptides of the present invention (e.g., engineered proteins), 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 CRISPR-Cas

effector protein as provided herein, which may be encoded by a polynucleotide), a CRISPR-Cas effector protein, 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:1-590**. In some embodiments, if not present in an amino acid sequence described herein and/or encoded in a nucleotide sequence described herein, the amino acid sequence and/or the nucleotide sequence may further include 1, 2, 3, 4, or 5 additional amino acid(s) or corresponding nucleotides such as a methionine at amino acid residue 1 of the amino acid sequence or the corresponding nucleotides and/or 1, 2, 3, or 4 additional amino acid(s) such as glycine or alanine or the corresponding nucleotides, which may aid in expression or enable the construction or cloning as needed for construction of a plasmid or construct or vector. In some embodiments, if present in an amino acid sequence described herein and/or encoded in a nucleotide sequence described herein, the amino acid sequence and/or the nucleotide sequence may be devoid of 1, 2, 3, 4, or 5 amino acid(s) present at amino acid residues 1-5 at the N-terminus of the amino acid sequence or corresponding nucleotides such as a methionine at amino acid residue 1 of the amino acid sequence or the corresponding nucleotides and/or 1, 2, 3, or 4 additional amino acid(s) such as glycine or alanine or the corresponding nucleotides, which may aid in expression or enable the

construction or cloning as needed for construction of a plasmid or construct or vector. 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:1-590**.

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:

HEK293T cells were seeded into 48-well collagen-coated plates (Corning) in the absence of an antibiotic using DMEM media. At 70-80% confluency, cells were transfected with 1.5 μ L of LTX (ThermoFisher Scientific) using 500 ng of plasmid including an adenine base editor (ABE) (as a control), which is a fusion of an adenosine deaminase and a CRISPR-Cas effector protein, or a tested fusion protein (a putative adenosine deaminase and a CRISPR-Cas effector protein) and 500 ng of guide RNA plasmid according to manufacturer's protocol. After 3 days, cells were lysed with a crude extraction method using Triton X buffer.

Table 1: Spacer sequences used in the guide RNAs.

Spacer	Target	Spacer	PAM	Direction
sp12	HEK2	GAACACAAAGCATAGACTGC (SEQ ID NO:311)	GGG	+
sp90	FANCF	GCACTACCTACGTCAGCACC (SEQ ID NO:312)	TGG	+
sp92	HEK2	CTGTCACAGTTAGCTCAGCC (SEQ ID NO:313)	AGG	-

The average observed frequency of adenine to guanine editing for each adenine base editor (control) or tested fusion protein at edited positions within each of the tested spacers was determined. The results are shown in **FIGS. 3-5** and demonstrate that certain tested fusion

proteins have activity as an adenosine base editor. In reference to **FIGS. 3-5**, the controls were: ABE8e which has an amino acid sequence of **SEQ ID NO:275** (which is encoded by **SEQ ID NO:278**) and includes **SEQ ID NO:221** (which is encoded by **SEQ ID NO: 271**), ABEmax which has an amino acid sequence of **SEQ ID NO:274** (which is encoded by **SEQ ID NO:277**) and includes **SEQ ID NO:220** (which is encoded by **SEQ ID NO:270**), ABEmax(chiA) which has an amino acid sequence of **SEQ ID NO:276** (which is encoded by **SEQ ID NO:279**) and includes **SEQ ID NO:269** (which is encoded by **SEQ ID NO:272**), and ABE8e(chiA) which has an amino acid sequence of **SEQ ID NO:275** (which is encoded by **SEQ ID NO:280**) and includes **SEQ ID NO:221** (which is encoded by **SEQ ID NO:273**). The tested fusion proteins included a polypeptide having an amino acid sequence as identified in **Table 2** fused to a CRISPR-Cas effector protein, and each of the tested fusion proteins included the same CRISPR-Cas effector protein and linker that joined the polypeptide and CRISPR-Cas effector protein. For example, as shown in **Table 2**, the pWISE8839 vector encodes a fusion protein that includes a CRISPR-Cas effector protein and a polypeptide having an amino acid sequence of **SEQ ID NO:242**, and the pWISE8826 vector encodes a fusion protein that includes the CRISPR-Cas effector protein and a polypeptide having an amino acid sequence of **SEQ ID NO:241**. These results demonstrate that the polypeptides having an amino acid sequence of **SEQ ID NO:241** or **SEQ ID NO:242** have adenosine deaminase activity.

Table 2: Vectors encoding the tested fusion proteins.

Vector	Polypeptide encoded by the vector	Nucleotide sequence encoding the polypeptide	Fusion protein encoded by the vector	Nucleotide sequence encoding the fusion protein
pWISE8808	SEQ ID NO:281	SEQ ID NO:353	SEQ ID NO:417	SEQ ID NO:499
pWISE8809	SEQ ID NO:282	SEQ ID NO:354	SEQ ID NO:418	SEQ ID NO:500
pWISE8810	SEQ ID NO:283	SEQ ID NO:355	SEQ ID NO:419	SEQ ID NO:501
pWISE8811	SEQ ID NO:284	SEQ ID NO:356	SEQ ID NO:420	SEQ ID NO:502
pWISE8812	SEQ ID NO:285	SEQ ID NO:357	SEQ ID NO:421	SEQ ID NO:503
pWISE8813	SEQ ID NO:286	SEQ ID NO:358	SEQ ID NO:422	SEQ ID NO:504
pWISE8814	SEQ ID NO:287	SEQ ID NO:359	SEQ ID NO:423	SEQ ID NO:505
pWISE8815	SEQ ID NO:288	SEQ ID NO:360	SEQ ID NO:424	SEQ ID NO:506
pWISE8816	SEQ ID NO:289	SEQ ID NO:361	SEQ ID NO:425	SEQ ID NO:507
pWISE8817	SEQ ID NO:290	SEQ ID NO:362	SEQ ID NO:426	SEQ ID NO:508
pWISE8818	SEQ ID NO:291	SEQ ID NO:363	SEQ ID NO:427	SEQ ID NO:509
pWISE8819	SEQ ID NO:292	SEQ ID NO:364	SEQ ID NO:428	SEQ ID NO:510
pWISE8820	SEQ ID NO:293	SEQ ID NO:365	SEQ ID NO:429	SEQ ID NO:511
pWISE8821	SEQ ID NO:294	SEQ ID NO:366	SEQ ID NO:430	SEQ ID NO:512
pWISE8822	SEQ ID NO:295	SEQ ID NO:367	SEQ ID NO:431	SEQ ID NO:513

pWISE8823	SEQ ID NO:296	SEQ ID NO:368	SEQ ID NO:432	SEQ ID NO:514
pWISE8824	SEQ ID NO:297	SEQ ID NO:369	SEQ ID NO:433	SEQ ID NO:515
pWISE8825	SEQ ID NO:298	SEQ ID NO:370	SEQ ID NO:434	SEQ ID NO:516
pWISE8826	SEQ ID NO:241	SEQ ID NO:383	SEQ ID NO:448	SEQ ID NO:530
pWISE8827	SEQ ID NO:299	SEQ ID NO:371	SEQ ID NO:435	SEQ ID NO:517
pWISE8828	SEQ ID NO:300	SEQ ID NO:372	SEQ ID NO:436	SEQ ID NO:518
pWISE8829	SEQ ID NO:301	SEQ ID NO:373	SEQ ID NO:437	SEQ ID NO:519
pWISE8830	SEQ ID NO:302	SEQ ID NO:374	SEQ ID NO:438	SEQ ID NO:520
pWISE8831	SEQ ID NO:303	SEQ ID NO:375	SEQ ID NO:439	SEQ ID NO:521
pWISE8832	SEQ ID NO:304	SEQ ID NO:376	SEQ ID NO:440	SEQ ID NO:522
pWISE8833	SEQ ID NO:305	SEQ ID NO:377	SEQ ID NO:441	SEQ ID NO:523
pWISE8834	SEQ ID NO:306	SEQ ID NO:378	SEQ ID NO:442	SEQ ID NO:524
pWISE8835	SEQ ID NO:307	SEQ ID NO:379	SEQ ID NO:443	SEQ ID NO:525
pWISE8836	SEQ ID NO:308	SEQ ID NO:380	SEQ ID NO:444	SEQ ID NO:526
pWISE8837	SEQ ID NO:309	SEQ ID NO:381	SEQ ID NO:445	SEQ ID NO:527
pWISE8838	SEQ ID NO:310	SEQ ID NO:382	SEQ ID NO:446	SEQ ID NO:528
pWISE8839	SEQ ID NO:242	SEQ ID NO:384	SEQ ID NO:447	SEQ ID NO:529

Example 2:

HEK293T cells were seeded into 48-well collagen-coated plates (Corning) in the absence of an antibiotic using DMEM media. At 70-80% confluency, cells were transfected with 1.5 μ L of LTX (ThermoFisher Scientific) using 500 ng of plasmid including an ABE (as a control; **SEQ ID NOs:583 and 584**), which is a fusion of an adenosine deaminase and a CRISPR-Cas effector protein, or a tested fusion protein (a putative adenosine deaminase and a CRISPR-Cas effector protein) and 500 ng of guide RNA plasmid according to manufacturer's protocol. After 3 days, cells were lysed with a crude extraction method using Triton X buffer.

The fusion proteins listed in **Table 3** were tested with each of the spacers in **Table 1**. The results for the tested fusion proteins of **Table 3** are provided in **FIGS. 6-8**. Fusion proteins including the sequence of **SEQ ID NO:242, 243, 245, or 247** were particularly active. **SEQ ID NOs:243-249** each demonstrated activity and have greater than 65% sequence identity to **SEQ ID NO:242**.

Table 3: Tested fusion proteins

Vector	Polypeptide encoded by the vector	Nucleotide sequence encoding the polypeptide	Fusion protein encoded by the vector	Nucleotide sequence encoding the fusion protein
pWISE67			SEQ ID NO:583	SEQ ID NO:585
pWISE2910			SEQ ID NO:584	SEQ ID NO:586
pWISE8839	SEQ ID NO:242	SEQ ID NO:384	SEQ ID NO:447	SEQ ID NO:529
pWISE9829	SEQ ID NO:250	SEQ ID NO:352	SEQ ID NO:465	SEQ ID NO:547

pWISE9830	SEQ ID NO:249	SEQ ID NO:351	SEQ ID NO:466	SEQ ID NO:548
pWISE9831	SEQ ID NO:248	SEQ ID NO:350	SEQ ID NO:467	SEQ ID NO:549
pWISE9832	SEQ ID NO:247	SEQ ID NO:349	SEQ ID NO:468	SEQ ID NO:550
pWISE9833	SEQ ID NO:246	SEQ ID NO:348	SEQ ID NO:469	SEQ ID NO:551
pWISE9834	SEQ ID NO:245	SEQ ID NO:347	SEQ ID NO:470	SEQ ID NO:552
pWISE9835	SEQ ID NO:244	SEQ ID NO:346	SEQ ID NO:471	SEQ ID NO:553
pWISE9836	SEQ ID NO:243	SEQ ID NO:345	SEQ ID NO:472	SEQ ID NO:554

Additional fusion proteins as listed in **Table 4** were tested with each of the spacers in **Table 5**. The fusion proteins of **Table 4** included a polypeptide having a sequence of one of **SEQ ID NOs:254-256, 259, 260, 263, 264, or 314-325** at the N-terminus of a CRISPR-Cas effector protein. The results for the tested fusion proteins of **Table 4** are provided in **FIGS. 9-10**. **Tables 6-7** provide the editing frequency for certain fusion proteins at specified nucleotides and the improvement in editing compared to pWISE9652.

Table 4: Additional tested fusion proteins

Vector	Polypeptide encoded by the vector	Nucleotide sequence encoding the polypeptide	Fusion protein encoded by the vector	Nucleotide sequence encoding the fusion protein
pWISE9650	SEQ ID NO:254	SEQ ID NO:388	SEQ ID NO:452	SEQ ID NO:534
pWISE9651	SEQ ID NO:255	SEQ ID NO:389	SEQ ID NO:453	SEQ ID NO:535
pWISE9652	SEQ ID NO:256	SEQ ID NO:390	SEQ ID NO:454	SEQ ID NO:536
pWISE9655	SEQ ID NO:259	SEQ ID NO:393	SEQ ID NO:457	SEQ ID NO:539
pWISE9656	SEQ ID NO:260	SEQ ID NO:394	SEQ ID NO:458	SEQ ID NO:540
pWISE9659	SEQ ID NO:263	SEQ ID NO:397	SEQ ID NO:461	SEQ ID NO:543
pWISE9660	SEQ ID NO:264	SEQ ID NO:398	SEQ ID NO:462	SEQ ID NO:544
pWISE10410	SEQ ID NO:314	SEQ ID NO:401	SEQ ID NO:481	SEQ ID NO:563
pWISE10411	SEQ ID NO:315	SEQ ID NO:402	SEQ ID NO:482	SEQ ID NO:564
pWISE10412	SEQ ID NO:316	SEQ ID NO:403	SEQ ID NO:483	SEQ ID NO:565
pWISE10413	SEQ ID NO:317	SEQ ID NO:404	SEQ ID NO:484	SEQ ID NO:566
pWISE10414	SEQ ID NO:318	SEQ ID NO:405	SEQ ID NO:485	SEQ ID NO:567
pWISE10415	SEQ ID NO:319	SEQ ID NO:406	SEQ ID NO:486	SEQ ID NO:568
pWISE10416	SEQ ID NO:320	SEQ ID NO:407	SEQ ID NO:487	SEQ ID NO:569
pWISE10417	SEQ ID NO:321	SEQ ID NO:408	SEQ ID NO:488	SEQ ID NO:570
pWISE10418	SEQ ID NO:322	SEQ ID NO:409	SEQ ID NO:489	SEQ ID NO:571
pWISE10419	SEQ ID NO:323	SEQ ID NO:410	SEQ ID NO:490	SEQ ID NO:572
pWISE10420	SEQ ID NO:324	SEQ ID NO:411	SEQ ID NO:491	SEQ ID NO:573
pWISE10421	SEQ ID NO:325	SEQ ID NO:412	SEQ ID NO:492	SEQ ID NO:574

Table 5: Spacer sequences used in the guide RNAs.

Spacer	Target	Spacer	PAM	Direction
sp143	DNMT1	GCTCAGCAGGCACCTGCCTCAGC (SEQ ID NO:581)	TTTG	+
sp454	RNF2	CACGTCTCATATGCCCCTTGGCA (SEQ ID NO:582)	TTTA	-

Table 6: Editing results for specific fusion proteins with the sp143 spacer.

pWISE	A12	A12 Improvement
pWISE9652	0.24%	--
pWISE9660	0.29%	1.2x
pWISE10416	2.19%	9.1x

Table 7: Editing results for specific fusion proteins with the sp454 spacer.

pWISE	A11	A11 Improvement
pWISE9652	0.58%	--
pWISE9660	1.03%	1.8x
pWISE10416	1.47%	2.5x

Example 3:

HEK293T cells were seeded into 48-well collagen-coated plates (Corning) in the absence of an antibiotic using DMEM media. At 70-80% confluency, cells were transfected with 1.5 μ L of LTX (ThermoFisher Scientific) using 500 ng of plasmid including an adenine base editor (as a control), which is a fusion of an adenosine deaminase and a CRISPR-Cas effector protein, or a tested fusion protein (a putative adenosine deaminase and a CRISPR-Cas effector protein) and 500 ng of guide RNA plasmid according to manufacturer's protocol. After 3 days, cells were lysed with a crude extraction method using Triton X buffer.

The fusion proteins listed in **Table 8** were tested with each of the spacers in **Table 5**. The results for the tested fusion proteins of **Table 8** are provided in **FIGS. 11-14**. Fusions proteins of **Table 8** that included a polypeptide having a sequence of **SEQ ID NO:256** or **264** at the C-terminus of a CRISPR-Cas effector protein demonstrated high activity. **Table 9** provides the editing frequency for certain fusion proteins at specified nucleotides and the improvement in editing compared to pWISE9652.

Table 8: Tested fusion proteins

Vector	Polypeptide encoded by the vector	Nucleotide sequence encoding the polypeptide	Fusion protein encoded by the vector	Nucleotide sequence encoding the fusion protein
pWISE6854			SEQ ID NO:587	SEQ ID NO:588
pWISE9652	SEQ ID NO:256	SEQ ID NO:390	SEQ ID NO:454	SEQ ID NO:536
pWISE9660	SEQ ID NO:264	SEQ ID NO:398	SEQ ID NO:462	SEQ ID NO:544
pWISE10165	SEQ ID NO:256	SEQ ID NO:390	SEQ ID NO:473	SEQ ID NO:555
pWISE10166	SEQ ID NO:264	SEQ ID NO:398	SEQ ID NO:474	SEQ ID NO:556
pWISE10167	SEQ ID NO:256	SEQ ID NO:390	SEQ ID NO:475	SEQ ID NO:557
pWISE10168	SEQ ID NO:264	SEQ ID NO:398	SEQ ID NO:476	SEQ ID NO:558
pWISE10169	SEQ ID NO:256	SEQ ID NO:390	SEQ ID NO:477	SEQ ID NO:559
pWISE10170	SEQ ID NO:264	SEQ ID NO:398	SEQ ID NO:478	SEQ ID NO:560

Table 9: Editing results for specific fusion proteins with the sp143 and sp454 spacers.

	sp143		sp454	
	A12	fold-improvement	A11	fold-improvement
pWISE9652	0.20%	--	0.31%	--
pWISE10165	0.12%	0.6	0.15%	0.5
pWISE10167	0.21%	1.1	0.33%	1.1
pWISE10169	8.87%	44.4	4.07%	13.1
pWISE9660	0.50%	--	0.57%	--
pWISE10166	0.35%	0.7	0.51%	0.9
pWISE10168	n.d.	n.d.	n.d.	n.d.
pWISE10170	10.00%	20.0	4.33%	7.6

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. An engineered protein comprising a polypeptide that has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**.
2. The engineered protein of claim 1, wherein the polypeptide, when optimally aligned to **SEQ ID NO:215**, has:
 - a valine at an amino acid residue corresponding to amino acid residue 105 of **SEQ ID NO:215**,
 - an asparagine at an amino acid residue corresponding to amino acid residue 107 of **SEQ ID NO:215**,
 - an arginine at an amino acid residue corresponding to amino acid residue 110 of **SEQ ID NO:215**,
 - a tyrosine at an amino acid residue corresponding to amino acid residue 148 of **SEQ ID NO:215**, and
 - a proline at an amino acid residue corresponding to amino acid residue 151 of **SEQ ID NO:215**.
3. The engineered protein of claim 1 or 2, wherein the polypeptide, optionally when optimally aligned to one of **SEQ ID NOs:215-225**, has less than 70% (e.g., less than 65%, 60%, 55%, 50%, or 45%) sequence identity to the one of **SEQ ID NOs:215-225**.
4. The engineered protein of any one of the preceding claims, wherein the polypeptide has at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to **SEQ ID NO:242**.
5. The engineered protein of any one of the preceding claims, wherein the polypeptide has at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to **SEQ ID NO:241**.
6. The engineered protein of any one of the preceding claims, wherein the polypeptide, optionally when optimally aligned to **SEQ ID NO:590**, has:

a valine at an amino acid residue corresponding to amino acid residue 101 of **SEQ ID NO:590**,

an asparagine at an amino acid residue corresponding to amino acid residue 103 of **SEQ ID NO:590**,

an arginine at an amino acid residue corresponding to amino acid residue 106 of **SEQ ID NO:590**,

a tyrosine at an amino acid residue corresponding to amino acid residue 144 of **SEQ ID NO:590** and/or

a proline at an amino acid residue corresponding to amino acid residue 147 of **SEQ ID NO:590**.

7. The engineered protein of any one of the preceding claims, wherein the polypeptide, optionally when optimally aligned to **SEQ ID NO:590**, has:

an alanine at an amino acid residue corresponding to amino acid residue 9 of **SEQ ID NO:590**,

an alanine at an amino acid residue corresponding to amino acid residue 44 of **SEQ ID NO:590**,

a serine at an amino acid residue corresponding to amino acid residue 78 of **SEQ ID NO:590**,

a phenylalanine at an amino acid residue corresponding to amino acid residue 80 of **SEQ ID NO:590**,

a serine at an amino acid residue corresponding to amino acid residue 105 of **SEQ ID NO:590** and/or

an arginine at an amino acid residue corresponding to amino acid residue 150 of **SEQ ID NO:590**.

8. The engineered protein of any one of the preceding claims, further comprising CRISPR-Cas effector protein, optionally wherein the CRISPR-Cas effector protein has a sequence of one of **SEQ ID NOs:39-65, 70-72, 88-189, or 333-342**.

9. The engineered protein of claim 8, further comprising a linker (e.g., a peptide linker) that is between the CRISPR-Cas effector protein and the polypeptide.

10. The engineered protein of claim 9, wherein the linker is a peptide linker that has a length of 1, 5, 10, or 15 amino acid residue(s) to 20, 25, 30, 35, or 40 amino acid residues, optionally wherein the linker is a peptide linker that comprises glycine and/or serine.
11. The engineered protein of claim 9 or 10, wherein the linker is a peptide linker that has an amino acid sequence of one of **SEQ ID NOs:1-36** or a sequence selected from: CA, CF, (GGS)_n, GS, SG, 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).
12. The engineered protein of claim 8, wherein the CRISPR-Cas effector protein and the polypeptide are directly fused together without an intervening element between the CRISPR-Cas effector protein and the polypeptide.
13. The engineered protein of any one of the preceding claims, wherein the engineered protein comprises two or more copies of the polypeptide.
14. The engineered protein of any one of claims 8-13, wherein the polypeptide is fused to the N-terminus of the CRISPR-Cas effector protein.
15. The engineered protein of any one of claims 8-13, wherein the polypeptide is fused to the C-terminus of the CRISPR-Cas effector protein.
16. The engineered protein of any one of claims 1-15, further comprising a nuclear localization sequence, optionally wherein the engineered protein comprises two or more nuclear localization sequences.
17. The engineered protein of any one of claims 1-7, wherein the engineered protein has at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:241-266, 314-329, and/or 589-590**.
18. A complex comprising:
the engineered protein of any one of claims 1-17; and
a guide nucleic acid.

19. The complex of claim 18, wherein the guide nucleic acid is an extended guide nucleic acid or further comprising an extended guide nucleic acid.
20. The complex of claim 18 or 19, comprised in an expression cassette, optionally wherein the expression cassette is comprised in a vector.
21. A nucleic acid encoding the engineered protein of any one of claims 1-17, optionally wherein the nucleic acid is codon-optimized for expression in an organism (e.g., a plant or a mammal).
22. An expression cassette codon optimized for expression in an organism comprising:
a polynucleotide encoding a promoter sequence, and
a polynucleotide encoding the engineered protein of any one of claims 1-17,
optionally wherein the polynucleotide encoding the engineered protein is codon-optimized for expression in the organism.
23. The expression cassette of claim 22, wherein the organism is an animal, a plant, a fungus, an archaeon, or a bacterium.
24. A method of modifying a target nucleic acid in a cell, the method comprising:
introducing the expression cassette of claim 22 or 23 into the cell, thereby modifying the target nucleic acid in the cell.
25. The method of claim 24, wherein the cell is a plant cell (e.g., a plant cell that includes a cell wall) and the method further comprises regenerating the plant cell comprising the modified target nucleic acid to produce a plant comprising the modified target nucleic acid.
26. The method of any one of claims 24 or 25, wherein the introducing is carried out at a temperature of about 20°C to about 42°C.
27. A method for producing an engineered protein, the method comprising:
culturing a cell or group of cells that have been transformed with a nucleic acid encoding the engineered protein of any one of claims 1-17; and

isolating said polypeptide (e.g., isolating by dialysis, centrifugation, column purification, and/or the like), thereby producing the polypeptide.

28. A method of modifying a target nucleic acid, the method comprising:
contacting the target nucleic acid with:
the engineered protein of any one of claims 1-17; and
a guide nucleic acid, thereby modifying the target nucleic acid.
29. The method of claim 28, 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.
30. The method of claim 28 or 29, further comprising introducing a nucleic acid molecule encoding the engineered protein into the cell and producing the engineered protein in the cell.
31. A protein comprising a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:343, 344, and/or 447-498**.
32. A nucleic acid comprising a sequence having at least about 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to one or more of **SEQ ID NOs:345-352, 383-416, and/or 499-580**.