

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

(19) World Intellectual Property

Organization

International Bureau

(43) International Publication Date

29 August 2019 (29.08.2019)



(10) International Publication Number

WO 2019/161783 A1

(51) International Patent Classification:

C07K 19/00 (2006.01) C12N 9/22 (2006.01)
C12N 15/62 (2006.01) C12N 15/11 (2006.01)

(21) International Application Number:

PCT/CN2019/075897

(22) International Filing Date:

22 February 2019 (22.02.2019)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

PCT/CN2018/076991
23 February 2018 (23.02.2018) CN
PCT/CN2018/100411
14 August 2018 (14.08.2018) CN

(71) Applicant: SHANGHAITECH UNIVERSITY [CN/CN];
Room 207, Administration Center, No. 393 Middle Huaxia
Road, Pudong New Area, Shanghai 201210 (CN).

(72) Inventors: CHEN, Jia; Room 207, Administration Center,
No. 393 Middle Huaxia Road, Pudong New Area, Shang-
hai 201210 (CN). YANG, Li; Room 207, Administration
Center, No. 393 Middle Huaxia Road, Pudong New Area,
Shanghai 201210 (CN). HUANG, Xingxu; Room 207, Ad-
ministration Center, No. 393 Middle Huaxia Road, Pudong
New Area, Shanghai 201210 (CN). YANG, Bei; Room
207, Administration Center, No. 393 Middle Huaxia Road,
Pudong New Area, Shanghai 201210 (CN). WANG, Xiao;
Room 207, Administration Center, No. 393 Middle Huax-
ia Road, Pudong New Area, Shanghai 201210 (CN). LI,
Jianan; Room 207, Administration Center, No. 393 Middle
Huaxia Road, Pudong New Area, Shanghai 201210 (CN).

(74) Agent: SHANGHAI BESHINING LAW OFFICE; 21st
Floor, SFECO Mansion, 681 Xiaomujiao Road, Shanghai
200032 (CN).

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

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

Declarations under Rule 4.17:

- as to applicant's entitlement to apply for and be granted a
patent (Rule 4.17(ii))
- of inventorship (Rule 4.17(iv))

Published:

- with international search report (Art. 21(3))
- with sequence listing part of description (Rule 5.2(a))

(54) Title: FUSION PROTEINS FOR BASE EDITING

(57) Abstract: Provided are fusion proteins that include an apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A) and a clustered regularly interspaced short palindromic repeats (CRISPR) -associated (Cas) protein, optionally further with uracil glycosylase inhibitor (UGI). Such a fusion protein is able to conduct base editing in DNA by deaminating cytosine to uracil, even when the cytosine is in a GpC context or is methylated.

WO 2019/161783 A1

FUSION PROTEINS FOR BASE EDITING

The present invention claims the priority of the PCT/CN2018/076991 filed on February 23th, 2018, and the PCT/CN2018/100411 filed on August 14th, 2018, the contents of which are incorporated herein by its entirety.

BACKGROUND

Genome editing is a type of genetic engineering in which DNA is inserted, deleted or replaced in the genome of a living organism using engineered nucleases (molecular scissors). Utilizing genome editing tools to genetically manipulate the genome of cells and living organism has broad application interest in life sciences research, biotechnology/agricultural technology development and most importantly pharmaceutical/clinical innovation. For example, genome editing can be used to correct driver mutations underlying genetic diseases and thereby resulting in complete cure of these diseases in a living organism; genome editing can also be applied to engineer the genome of crops, thus increasing the yield of crops and conferring crops resistance to environmental contamination or pathogen infection; likewise, microbial genome transformation through accurate genome editing is of great significance in the development of renewable bio-energy.

CRISPR/Cas (Clustered regularly interspaced short palindromic repeats/CRISPR-associated protein) system has been the most powerful genomic editing tool since its conception for its unparalleled editing efficiency, convenience and the potential applications in living organism. Directed by guide RNA (gRNA), a Cas nuclease can generate DNA double strand breaks (DSBs) at the targeted genomic sites in various cells (both cell lines and cells from living organisms). These DSBs are then repaired by the endogenous DNA repair system, which could be utilized to perform desired genome editing.

In general, two major DNA repair pathways could be activated by DSBs, non-homologous end joining (NHEJ) and homology-directed repair (HDR). NHEJ can introduce random insertions/deletions (indels) in the genomic DNA region around the DSBs, thereby leading to open reading frame (ORF) shift and ultimately gene inactivation. In contrast, when HDR is triggered, the genomic DNA sequence at target site could be replaced by the sequence of the

exogenous donor DNA template through a homologous recombination mechanism, which can result in the correction of genetic mutation.

However, the practical efficiency of HDR-mediated gene correction is low (normally < 5%) because the occurrence of homologous recombination is both cell type-specific and cell cycle-dependent and NHEJ is triggered more frequently than HDR is. The relatively low efficiency of HDR therefore limited the translation of CRISPR/Cas genome editing tools in the field of precision gene therapy (diseases-driven gene correction).

Base editors (BE), which integrate the CRISPR/Cas system with the APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) cytosine deaminase family, were recently invented that greatly enhanced the efficiency of CRISPR/Cas9-mediated gene correction. Through fusion with Cas9 nickase (nCas9), the cytosine (C) deamination activity of rat APOBEC1 (rA1) can be purposely directed to the target bases in genome and to catalyze C to Thymine (T) substitutions at these bases.

However, current rA1-based BEs cannot efficiently edit C that follows a G (*i.e.*, C of GpC), thereby limiting the genome targeting breadth. Therefore, creating new BEs that can efficiently edit C of GpC is highly desirable. Such new BEs will enable us to perform efficient base editing in a broader genomic space of various living organisms. Importantly, the high efficiency of such BEs on C of GpC will promote clinical translation, particularly in gene therapies that involve restoring disease-related GpT-to-GpC mutations.

SUMMARY

The present disclosure demonstrates that when an apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A or A3A) is fused to a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, optionally further with uracil glycosylase inhibitor (UGI), the resulting fusion protein is able to efficiently deaminate cytosine's to uracil's resulting in C to T substitution. Such base editing, surprisingly and unexpectedly, was effective even when the C follows a G (*i.e.*, in a GpC dinucleotide context) or when the C is methylated. The editing efficiency can be further increased when the A3A includes a few tested mutations. This has significant clinical significance as cytosine methylation is common in living cells.

In conventional base editors, Cas9 is the commonly used DNA endonuclease. The Cas12a (Cpf1) has the advantage of recognizing A/T rich sequence when used together with APOBEC1 in base editors. In another surprising discovery, when APOBEC1 was replaced with A3A, the editing efficiency was greatly increased. Yet, the editing efficiency of such a Cas12a-A3A can be further increased when the A3A includes a few tested mutations.

Accordingly, in one embodiment, the present disclosure provides a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein. In some embodiments, the fusion protein further comprises a uracil glycosylase inhibitor (UGI).

Preferably, the fusion protein has fewer than 3000, 2500, 2200, 2100, 2000, 1900, 1800, 1700, 1600, or 1500 amino acid residues in total.

In some embodiments, the APOBEC3A is a wildtype human APOBEC3A or a mutant of human APOBEC3A having a mutation selected from the group consisting of Y130F, D131Y, D131E, Y132D, W104A, W98Y, P134Y and combinations thereof, according to residue numbering in SEQ ID NO:1, wherein the mutant retains cytidine deaminase activity.

In some embodiments, the APOBEC3A is a mutant human APOBEC3A having mutations selected from the group consisting of Y130F+D131E+Y132D, Y130F+D131Y+Y132D, W98Y+W104A, W98Y+P134Y, W104A+P134Y, W104A+Y130F, W104A+Y132D, W98Y+W104A+Y130F, W98Y+W104A+Y132D, W104A+Y130F+P134Y, and W104A+Y132D+P134Y, according to residue numbering in SEQ ID NO:1.

In some embodiments, the APOBEC3A comprises the amino acid sequence of SEQ ID NO:1 or has at least 90% sequence identity to amino acid residues 29-199 of SEQ ID NO:1 and retains cytidine deaminase activity. In some embodiments, the APOBEC3A comprises an amino acid sequence selected from the group consisting of SEQ ID NO:1-10 and 22-36.

In some embodiments, the Cas protein is selected from the group consisting of SpCas9, FnCas9, St1Cas9, St3Cas9, NmCas9, SaCas9, AsCpf1, LbCpf1, FnCpf1, VQR SpCas9, EQR SpCas9, VRER SpCas9, RHA FnCas9, KKH SaCas9, NmeCas9, StCas9, CjCas9, AsCpf1, FnCpf1, SsCpf1, PcCpf1, BpCpf1, CmtCpf1, LiCpf1, PmCpf1, Pb3310Cpf1, Pb4417Cpf1, BsCpf1, EeCpf1, BhCas12b, AkCas12b, EbCas12b, LsCas12b, RfCas13d, LwaCas13a, PspCas13b,

PguCas13b, RanCas13b, CasX, and CasY. In some embodiments, the Cas protein is a mutant of protein selected from the group consisting of SpCas9, FnCas9, St1Cas9, St3Cas9, NmCas9, SaCas9, AsCpf1, LbCpf1, FnCpf1, VQR SpCas9, EQR SpCas9, VRER SpCas9, RHA FnCas9, KKH SaCas9, NmeCas9, StCas9, CjCas9, AsCpf1, FnCpf1, SsCpf1, PcCpf1, BpCpf1, CmtCpf1, LiCpf1, PmCpf1, Pb3310Cpf1, Pb4417Cpf1, BsCpf1, EeCpf1, BhCas12b, AkCas12b, EbCas12b, LsCas12b, RfCas13d, LwaCas13a, PspCas13b, PguCas13b, RanCas13b, CasX, and CasY, wherein the mutant retains the DNA-binding capability but does not introduce double strand DNA breaks. In some embodiments, the mutant is capable of introducing a nick to one of the strands of a double stranded DNA bound by the mutant. In some embodiments, the Cas protein comprises the amino acid sequence of any one of SEQ ID NO:11 and 37-39.

In some embodiments, the UGI comprises the amino acid sequence of SEQ ID NO:12 or has at least at least 90% sequence identity to SEQ ID NO:12 and retains the uracil glycosylase inhibition activity.

In some embodiments, the first fragment is at the N-terminal side of the second fragment. In some embodiments, the first fragment is at the N-terminal side of the second fragment which is at the N-terminal side of the UGI.

In some embodiments, the fusion protein further comprises a peptide linker between the first fragment and the second fragment. In some embodiments, the peptide linker has from 1 to 100 amino acid residues. In some embodiments, at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% or 90% of the amino acid residues of peptide linker are amino acid residues selected from the group consisting of alanine, glycine, cysteine, and serine. In some embodiments, the peptide linker has an amino acid sequence of SEQ ID NO:13 or 14. In some embodiments, the fusion protein further comprises a nuclear localization sequence.

Non-limiting examples of fusion proteins include those having an amino acid sequence selected from the group consisting of SEQ ID NO:16-20 and 40-50.

In another embodiment, a fusion protein is provided that comprises a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A) and a second fragment comprising a CRISPR-associated endonuclease in *Prevotella* and *Francisella* 1 (Cpf1). In some embodiments, the Cpf1 is catalytically inactive.

The Cpf1 (Cas12a) can be selected from the group consisting of AsCpf1, LbCpf1, and FnCpf1, in some embodiments. In a specific embodiment, the Cpf1 is a catalytically inactive *Lachnospiraceae* bacterium Cpf1 (dLbCpf1).

In some embodiments, the APOBEC3A is a wildtype human APOBEC3A or a mutant of human APOBEC3A having a mutation selected from the group consisting of Y130F, D131Y, D131E, Y132D, W104A, W98Y, P134Y and combinations thereof, according to residue numbering in SEQ ID NO:1, wherein the mutant retains cytidine deaminase activity.

Also provided is a polynucleotide that encodes a fusion protein of the present disclosure. Still, in another embodiment, provided is a composition comprising a fusion protein of the present disclosure and a pharmaceutically acceptable carrier. In some embodiments, the composition further comprises a guide RNA.

Methods of using the fusion proteins and compositions are also provided. In one embodiment, a method for editing a target polynucleotide is provided, comprising contacting to the target polynucleotide a fusion protein of the present disclosure and a guide RNA having at least partial sequence complementarity to the target polynucleotide, wherein the editing comprises deamination of a cytosine (C) in the target polynucleotide. In some embodiments, the C is in a GpC context. In some embodiments, the C is methylated. In some embodiments, the contacting is *in vitro*, *ex vivo*, or *in vivo*. In some embodiments, the method further comprises contacting to the target polynucleotide with a uracil glycosylase inhibitor (UGI) not fused to a Cas protein.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1A-B. Construction and performance of hA3A-BE. Panel A: Schematic diagram illustrating the co-expression of BE3/sgRNA or hA3A-BE/sgRNA. Panel B: Comparing to the co-expression of BE3/sgRNA, the co-expression of hA3A-BE/sgRNA achieved more efficient base editing on the C of GpC in the sgRNA targeted genomic regions (sgFANCF-M-L6 and sgSITE4). Dashed boxes represent the cytosine's locating in the context of GpC.

FIG. 2A-B. Construction and performance of hA3A-BE-Y130F and hA3A-BE-Y132D. Panel A: Schematic diagram illustrating the co-expression of hA3A-BE/sgRNA, hA3A-BE-Y130F/sgRNA or hA3A-BE-Y132D/sgRNA. Panel B: Comparing to the co-expression of hA3A-BE/sgRNA, the co-expression of hA3A-BE-Y130F/sgRNA or hA3A-BE-Y132D/sgRNA

induced base editing in more narrowed windows in the sgRNA targeted genomic regions (sgSITE3 and sgEMX1). Dashed boxes represent the base editing windows.

FIG. 3A-B. Construction and performance of hA3A-BE-W104A and hA3A-BE-D131Y. Panel A: Schematic diagram illustrating the co-expression of hA3A-BE/sgRNA, hA3A-BE-W104A/sgRNA or hA3A-BE-D131Y/sgRNA. Panel B: Comparing to the co-expression of hA3A-BE/sgRNA, the co-expression of hA3A-BE-W104A/sgRNA or hA3A-BE-D131Y/sgRNA induced more efficient base editing in the sgRNA targeted genomic regions (sgFANCF and sgSITE2). Dashed boxes represent the edited cytosine's.

FIG. 4A-B. Construction and performance of hA3A-BE-Y130F-D131E-Y132D and hA3A-BE-Y130F-D131Y-Y132D. Panel A: Schematic diagram illustrating the co-expression of hA3A-BE/sgRNA, hA3A-BE-Y130F-D131E-Y132D/sgRNA or hA3A-BE-Y130F-D131Y-Y132D/sgRNA. Panel B: Comparing to the co-expression of hA3A-BE/sgRNA, the co-expression of hA3A-BE-Y130F-D131E-Y132D/sgRNA or hA3A-BE-Y130F-D131Y-Y132D/sgRNA induced base editing in more narrowed windows in the sgRNA targeted genomic regions (sgFANCF and sgSITE3). Dashed boxes represent the edited cytosine's.

FIG. 5a-h. hA3A-BE3 induces efficient base editing in methylated region and in GpC context. (a) Distribution of BE-editable T-to-C (or A-to-G) variants. Potentially editable cytosines (underlined) are sub-classified according to their 3' adjacent bases. (b) Screening of BEs for efficient base editing in a high-methylation background. A series of new BEs were constructed by fusing different APOBEC/AID deaminases with Cas9 nickase (nCas9) and uracil DNA glycosylase inhibitor (UGI). (c) Cumulative base editing frequencies induced by different BEs in unmethylated and methylated vectors. A commonly used rA1-based BE3 was chosen for comparison. Means $\pm$ s.d. were from three (six for hA3A-BE3) independent experiments. (d) Immunoblots of BE3 and hA3A-BE3 co-transfected with unmethylated or methylated vectors. Tubulin was used as a loading control and immunoblot images are representative of three independent experiments. (e) Comparison of base editing efficiencies induced by BE3 and hA3A-BE3 in genomic regions with natively high levels of DNA methylation. C-to-T editing frequencies of indicated cytosines were determined individually. Target site sequences are shown with the BE3 editing window (position 4-8, setting the base distal to the PAM as position 1) in pink, PAM in cyan and CpG site in capital. Shaded gray, guanines at 5' end of editable cytosines. NT, native HEK293T cells with no treatment. (f) Statistical analysis of normalized C-to-T editing frequencies in regions with natively high levels of DNA methylation shown in (e),

setting the ones induced by BE3 as 100%. n = 48 samples from three independent experiments. (g) Comparison of base editing efficiencies induced by BE3 and hA3A-BE3 at C of GpC in genomic regions with natively low levels of DNA methylation. (h) Statistical analysis of normalized C-to-T editing frequencies at GpC sites in regions with natively low levels of DNA methylation shown in (g), setting the ones induced by BE3 as 100%. n = 24 samples from three independent experiments. (e,g) Means $\pm$ s.d. were from three independent experiments. (f,h) *P* value, one-tailed Student's *t* test. The median and interquartile range (IQR) are shown.

FIG. 6a-i. Improvements in hA3A-BE3. (a) Comparison of base editing efficiencies induced by BE3, hA3A-BE3, hA3A-BE3-Y130F and hA3A-BE3-Y132D in genomic regions with natively high levels of DNA methylation. Target site sequences are shown with the overlapped editing window (position 4-7) in pink, PAM in cyan and CpG site in capital. NT, native HEK293T cells with no treatment. (b) Statistical analysis of normalized C-to-T editing frequencies in the overlapped editing window shown in (a), setting the ones induced by BE3 as 100%. n = 12 samples from three independent experiments. (c) Comparison of base editing efficiencies induced by BE3, hA3A-BE3, hA3A-BE3-Y130F and hA3A-BE3-Y132D at C of GpC in the overlapped editing window in genomic regions with natively low levels of DNA methylation. (d) Statistical analysis of normalized C-to-T editing frequencies shown in (c), setting the ones induced by BE3 as 100%. n = 9 samples from three independent experiments. (e) Immunoblots of BEs transfected into HEK293T cells. Tubulin was used as a loading control and immunoblot images are representative of three independent experiments. (f) Comparison of base editing efficiencies induced by hA3A-BE3-Y130F, hA3A-eBE-Y130F, hA3A-BE3-Y132D and hA3A-eBE-Y132D at C of GpC in the overlapped editing window in genomic regions with natively low levels of DNA methylation. (g) Statistical analysis of normalized C-to-T editing frequencies shown in (f), setting the ones induced by hA3A-BE3-Y130F (left) or hA3A-BE3-Y132D (right) as 100%. n = 9 samples from three independent experiments. (h,i) Comparison of product purity (h) and indels (i) yielded by hA3A-BE3-Y130F, hA3A-eBE-Y130F, hA3A-BE3-Y132D and hA3A-eBE-Y132D in genomic DNA regions with natively low levels of DNA methylation. Asterisk denotes an unusually high basal indel frequency (or amplification, sequencing or alignment artifact) at the examined VEGFA-M-c site in NT. (a,c,f,i) Means $\pm$ s.d. were from three independent experiments. (b,d,g) *P* value, one-tailed Student's *t* test. The median and IQR are shown.

FIG. 7A-B and 8A-B show the vector structures of each of the tested base editors and charting showing their editing efficiencies on the target DYRK1A gene.

FIG. 9A-B and 10A-B show the vector structures of each of the tested base editors and charting showing their editing efficiencies on the target SITE6 gene.

FIG. 11A-B and 12A-B show the vector structures of each of the tested base editors and charting showing their editing efficiencies on the target RUNX1 gene.

FIG. 13-18 show the sequencing results for Examples 3-5.

DETAILED DESCRIPTION

Definitions

It is to be noted that the term “a” or “an” entity refers to one or more of that entity; for example, “an antibody,” is understood to represent one or more antibodies. As such, the terms “a” (or “an”), “one or more,” and “at least one” can be used interchangeably herein.

As used herein, the term “polypeptide” is intended to encompass a singular “polypeptide” as well as plural “polypeptides,” and refers to a molecule composed of monomers (amino acids) linearly linked by amide bonds (also known as peptide bonds). The term “polypeptide” refers to any chain or chains of two or more amino acids, and does not refer to a specific length of the product. Thus, peptides, dipeptides, tripeptides, oligopeptides, “protein,” “amino acid chain,” or any other term used to refer to a chain or chains of two or more amino acids, are included within the definition of “polypeptide,” and the term “polypeptide” may be used instead of, or interchangeably with any of these terms. The term “polypeptide” is also intended to refer to the products of post-expression modifications of the polypeptide, including without limitation glycosylation, acetylation, phosphorylation, amidation, derivatization by known protecting/blocking groups, proteolytic cleavage, or modification by non- naturally occurring amino acids. A polypeptide may be derived from a natural biological source or produced by recombinant technology, but is not necessarily translated from a designated nucleic acid sequence. It may be generated in any manner, including by chemical synthesis.

The term “isolated” as used herein with respect to cells, nucleic acids, such as DNA or RNA, refers to molecules separated from other DNAs or RNAs, respectively, that are present in the

natural source of the macromolecule. The term “isolated” as used herein also refers to a nucleic acid or peptide that is substantially free of cellular material, viral material, or culture medium when produced by recombinant DNA techniques, or chemical precursors or other chemicals when chemically synthesized. Moreover, an “isolated nucleic acid” is meant to include nucleic acid fragments which are not naturally occurring as fragments and would not be found in the natural state. The term “isolated” is also used herein to refer to cells or polypeptides which are isolated from other cellular proteins or tissues. Isolated polypeptides is meant to encompass both purified and recombinant polypeptides.

As used herein, the term “recombinant” as it pertains to polypeptides or polynucleotides intends a form of the polypeptide or polynucleotide that does not exist naturally, a non-limiting example of which can be created by combining polynucleotides or polypeptides that would not normally occur together.

“Homology” or “identity” or “similarity” refers to sequence similarity between two peptides or between two nucleic acid molecules. Homology can be determined by comparing a position in each sequence which may be aligned for purposes of comparison. When a position in the compared sequence is occupied by the same base or amino acid, then the molecules are homologous at that position. A degree of homology between sequences is a function of the number of matching or homologous positions shared by the sequences. An “unrelated” or “non-homologous” sequence shares less than 40% identity, though preferably less than 25% identity, with one of the sequences of the present disclosure.

A polynucleotide or polynucleotide region (or a polypeptide or polypeptide region) has a certain percentage (for example, 60 %, 65 %, 70 %, 75 %, 80 %, 85 %, 90 %, 95 %, 98 % or 99 %) of “sequence identity” to another sequence means that, when aligned, that percentage of bases (or amino acids) are the same in comparing the two sequences. This alignment and the percent homology or sequence identity can be determined using software programs known in the art, for example those described in Ausubel *et al.* eds. (2007) Current Protocols in Molecular Biology. Preferably, default parameters are used for alignment. One alignment program is BLAST, using default parameters. In particular, programs are BLASTN and BLASTP, using the following default parameters: Genetic code = standard; filter = none; strand = both; cutoff = 60; expect = 10; Matrix = BLOSUM62; Descriptions = 50 sequences; sort by = HIGH SCORE; Databases = non-redundant, GenBank + EMBL + DDBJ + PDB + GenBank CDS translations + SwissProtein + SPupdate + PIR. Biologically equivalent polynucleotides are those having the

above-noted specified percent homology and encoding a polypeptide having the same or similar biological activity.

The term “an equivalent nucleic acid or polynucleotide” refers to a nucleic acid having a nucleotide sequence having a certain degree of homology, or sequence identity, with the nucleotide sequence of the nucleic acid or complement thereof. A homolog of a double stranded nucleic acid is intended to include nucleic acids having a nucleotide sequence which has a certain degree of homology with or with the complement thereof. In one aspect, homologs of nucleic acids are capable of hybridizing to the nucleic acid or complement thereof. Likewise, “an equivalent polypeptide” refers to a polypeptide having a certain degree of homology, or sequence identity, with the amino acid sequence of a reference polypeptide. In some aspects, the sequence identity is at least about 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99%. In some aspects, the equivalent polypeptide or polynucleotide has one, two, three, four or five addition, deletion, substitution and their combinations thereof as compared to the reference polypeptide or polynucleotide. In some aspects, the equivalent sequence retains the activity (*e.g.*, epitope-binding) or structure (*e.g.*, salt-bridge) of the reference sequence.

Hybridization reactions can be performed under conditions of different “stringency”. In general, a low stringency hybridization reaction is carried out at about 40°C in about 10 x SSC or a solution of equivalent ionic strength/temperature. A moderate stringency hybridization is typically performed at about 50°C in about 6 x SSC, and a high stringency hybridization reaction is generally performed at about 60°C in about 1 x SSC. Hybridization reactions can also be performed under “physiological conditions” which is well known to one of skill in the art. A non-limiting example of a physiological condition is the temperature, ionic strength, pH and concentration of Mg^{2+} normally found in a cell.

A polynucleotide is composed of a specific sequence of four nucleotide bases: adenine (A); cytosine (C); guanine (G); thymine (T); and uracil (U) for thymine when the polynucleotide is RNA. Thus, the term “polynucleotide sequence” is the alphabetical representation of a polynucleotide molecule. This alphabetical representation can be input into databases in a computer having a central processing unit and used for bioinformatics applications such as functional genomics and homology searching. The term “polymorphism” refers to the coexistence of more than one form of a gene or portion thereof. A portion of a gene of which there are at least two different forms, *i.e.*, two different nucleotide sequences, is referred to as a

“polymorphic region of a gene”. A polymorphic region can be a single nucleotide, the identity of which differs in different alleles.

The terms “polynucleotide” and “oligonucleotide” are used interchangeably and refer to a polymeric form of nucleotides of any length, either deoxyribonucleotides or ribonucleotides or analogs thereof. Polynucleotides can have any three-dimensional structure and may perform any function, known or unknown. The following are non-limiting examples of polynucleotides: a gene or gene fragment (for example, a probe, primer, EST or SAGE tag), exons, introns, messenger RNA (mRNA), transfer RNA, ribosomal RNA, ribozymes, cDNA, dsRNA, siRNA, miRNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes and primers. A polynucleotide can comprise modified nucleotides, such as methylated nucleotides and nucleotide analogs. If present, modifications to the nucleotide structure can be imparted before or after assembly of the polynucleotide. The sequence of nucleotides can be interrupted by non-nucleotide components. A polynucleotide can be further modified after polymerization, such as by conjugation with a labeling component. The term also refers to both double- and single-stranded molecules. Unless otherwise specified or required, any embodiment of this disclosure that is a polynucleotide encompasses both the double-stranded form and each of two complementary single-stranded forms known or predicted to make up the double-stranded form.

The term “encode” as it is applied to polynucleotides refers to a polynucleotide which is said to “encode” a polypeptide if, in its native state or when manipulated by methods well known to those skilled in the art, it can be transcribed and/or translated to produce the mRNA for the polypeptide and/or a fragment thereof. The antisense strand is the complement of such a nucleic acid, and the encoding sequence can be deduced therefrom.

Fusion Proteins

The current rA1-based BEs (base editors) cannot efficiently edit C in methylated regions or in the context of GpC, which limits the use of base editing. The present disclosure provides fusion molecules that combine an apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A or A3A) and a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, optionally further with uracil glycosylase inhibitor (UGI).

The resulting fusion protein is able to efficiently deaminate cytosine's to uracil's resulting in C to T substitution. Such base editing, surprisingly and unexpectedly, was effective even when the C follows a G (*i.e.*, in a GpC dinucleotide context) and/or even when it is in a methylated region. This has significant clinical significance as cytosine methylation is common in living cells.

In accordance with one embodiment of the present disclosure, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein.

APOBEC3A, also referred to as apolipoprotein B mRNA editing enzyme catalytic subunit 3A or A3A, is a protein of the APOBEC3 family found in humans, non-human primates, and some other mammals. The APOBEC3A protein lacks the zinc binding activity of other family members. In human, isoform a (NP_663745.1; SEQ ID NO:1) and isoform b (NP_001257335.1; SEQ ID NO:6) both are active, while isoform a includes a few more residues close to the N-terminus. The term "APOBEC3A" also encompasses variants and mutants that have certain level (*e.g.*, 70%, 75%, 80%, 85%, 90%, 95%, 98%, 99%) of sequence identity to a wildtype mammalian APOBEC3A and retains its cytidine deaminating activity.

As demonstrated in the experimental examples, certain mutants (*e.g.*, Y130F (SEQ ID NO:2), Y132D (SEQ ID NO:3), W104A (SEQ ID NO:4), D131Y (SEQ ID NO:5), D131E (SEQ ID NO:22), W98Y (SEQ ID NO:24), W104A (SEQ ID NO:25), and P134Y (SEQ ID NO:26)) even outperformed the wildtype human APOBEC3A. Furthermore, a number of tested combinations of these mutations also exhibited great performances. Moreover, although not specifically tested, the same mutations are believed to also work in the isoform b of A3A. Examples of such variants and mutants are provided in **Table 1** below.

Table 1. Examples of APOBEC3A Sequences

Name	Sequence	SEQ ID NO:
Human APOBEC3A isoform a wildtype	1 MEASPASGPR HLMDFHITS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV FLRIFAARIY DYDPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTEVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	1

Human APOBEC3A isoform a Y130F	1 MEASPASGPR HLMDFPHITS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARIY DYDPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CFFQPWDGLD EHSQALSGRL RAILQNQGN	2
Human APOBEC3A isoform a Y132D	1 MEASPASGPR HLMDFPHITS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARIY <u>DD</u> DPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CFFQPWDGLD EHSQALSGRL RAILQNQGN	3
Human APOBEC3A isoform W104A	1 MEASPASGPR HLMDFPHITS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFS <u>A</u> GCAGEV RAFLQENTHV RLRIFAARIY DYDPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CFFQPWDGLD EHSQALSGRL RAILQNQGN	4
Human APOBEC3A isoform a D131Y	1 MEASPASGPR HLMDFPHITS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARIY <u>Y</u> DPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CFFQPWDGLD EHSQALSGRL RAILQNQGN	5
Human APOBEC3A isoform wildtype	1 MEASPASGPR HKTYLCYEVE RLDNGTSVKM DQHRGFLHNQ AKNLLCGFYG 51 RHAELRFLDL VPSLQLDPAQ IYRVTWFISW SPCFSWGCAG EVRAFLQENT 101 HVRLRIFAAR IYDYDPLYKE ALQMLRDAGA QVSIMTYDEF KHCWDTFVDH 151 QGCPFPWDG LDEHSQALSG RLRAILQNQG N	6
Human APOBEC3A isoform b Y112F	1 MEASPASGPR HKTYLCYEVE RLDNGTSVKM DQHRGFLHNQ AKNLLCGFYG 51 RHAELRFLDL VPSLQLDPAQ IYRVTWFISW SPCFSWGCAG EVRAFLQENT 101 HVRLRIFAAR I <u>F</u> DYDPLYKE ALQMLRDAGA QVSIMTYDEF KHCWDTFVDH 151 QGCPFPWDG LDEHSQALSG RLRAILQNQG N	7
Human APOBEC3A isoform b Y114D	1 MEASPASGPR HKTYLCYEVE RLDNGTSVKM DQHRGFLHNQ AKNLLCGFYG 51 RHAELRFLDL VPSLQLDPAQ IYRVTWFISW SPCFSWGCAG EVRAFLQENT 101 HVRLRIFAAR IY <u>DD</u> DPLYKE ALQMLRDAGA QVSIMTYDEF KHCWDTFVDH 151 QGCPFPWDG LDEHSQALSG RLRAILQNQG N	8
Human APOBEC3A isoform b W86A	1 MEASPASGPR HKTYLCYEVE RLDNGTSVKM DQHRGFLHNQ AKNLLCGFYG 51 RHAELRFLDL VPSLQLDPAQ IYRVTWFISW SPCFS <u>A</u> GCAG EVRAFLQENT 101 HVRLRIFAAR IYDYDPLYKE ALQMLRDAGA QVSIMTYDEF KHCWDTFVDH 151 QGCPFPWDG LDEHSQALSG RLRAILQNQG N	9
Human APOBEC3A isoform b D113Y	1 MEASPASGPR HKTYLCYEVE RLDNGTSVKM DQHRGFLHNQ AKNLLCGFYG 51 RHAELRFLDL VPSLQLDPAQ IYRVTWFISW SPCFSWGCAG EVRAFLQENT	10

	101 HVRLRIFAAR IY<u>Y</u>YDPLYKE ALQMLRDAGA QVSIMTYDEF KHCWDTFVDH 151 QGCPFQPWDG LDEHSQALSG RLRAILQNQG N	
Human APOBEC3A isoform a Y130F-D131E-Y132D	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARI<u>F</u> <u>EDD</u>DPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTEVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	22
Human APOBEC3A isoform a Y130F-D131Y-Y132D	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARI<u>F</u> <u>YDD</u>DPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTEVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	23
Human APOBEC3A isoform a W98Y	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFIS<u>Y</u>SP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARIY DYDPLYKEAL QMLRDAGAQV 150 SIMTYDEFKH CWDTEVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	24
Human APOBEC3A isoform a P134Y	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARIY DYD<u>Y</u>LYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTEVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	25
Human APOBEC3A isoform a W98Y+W104A	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFIS<u>Y</u>SP 101 CFS<u>A</u>GCAGEV RAFLQENTHV RLRIFAARIY DYDPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTEVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	26
Human APOBEC3A isoform a W98Y+P134Y	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFIS<u>Y</u>SP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARIY DYD<u>Y</u>LYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTEVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	27
Human APOBEC3A isoform a W104A+P134Y	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFS<u>A</u>GCAGEV RAFLQENTHV RLRIFAARIY DYD<u>Y</u>LYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTEVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	28
Human APOBEC3A isoform a W98Y+W104A+Y130F	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFIS<u>Y</u>SP 101 CFS<u>A</u>GCAGEV RAFLQENTHV RLRIFAARI<u>F</u> DYDPLYKEAL QMLRDAGAQV	29

	151 SIMTYDEFKH CWDTFVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	
Human APOBEC3A isoform a W98Y+W104A+ Y132D	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFIS Y SP 101 CFS A GCAGEV RAFLQENTHV RLRIFAARIY DD DPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	30
Human APOBEC3A isoform a W104A+Y130F +P134Y	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFS A GCAGEV RAFLQENTHV RLRIFAARI F DD D Y LYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	31
Human APOBEC3A isoform a W104A+Y132D +P134Y	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFS A GCAGEV RAFLQENTHV RLRIFAARIY DD D Y LYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	32
Human APOBEC3A isoform a W104A+Y130F	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFS A GCAGEV RAFLQENTHV RLRIFAARI F DD DPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	33
Human APOBEC3A isoform a W104A+Y132D	1 MEASPASGPR HLMDPHIETS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFS A GCAGEV RAFLQENTHV RLRIFAARIY DD DPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGN	34
Human APOBEC3A isoform b W80Y	1 MEASPASGPR HKTYLCYEVE RLDNGTSVKM DQHRGFLHNQ AKNLLCGFYG 51 RHAELRFLDL VPSLQLDPAQ IYRVTWFIS Y SPCFSWGCAG EVRAFLQENT 101 HVRLRIFAAR IYDYDPLYKE ALQMLRDAGA QVSIMTYDEF KHCWDTFVDH 151 QGCPFPWDG LDEHSQALSG RLRAILQNQG N	35
Human APOBEC3A isoform b P116Y	1 MEASPASGPR HKTYLCYEVE RLDNGTSVKM DQHRGFLHNQ AKNLLCGFYG 51 RHAELRFLDL VPSLQLDPAQ IYRVTWFISW SPCFSWGCAG EVRAFLQENT 101 HVRLRIFAAR IYDYD Y LYKE ALQMLRDAGA QVSIMTYDEF KHCWDTFVDH 151 QGCPFPWDG LDEHSQALSG RLRAILQNQG N	36

In some embodiments, the APOBEC3A in the fusion protein of the present disclosure is human isoform a or a sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, 99% of sequence identity to isoform a. In some embodiments, the APOBEC3A in the fusion protein of the present disclosure is human isoform b or a sequence having at least 70%, 75%, 80%, 85%, 90%, 95%,

98%, 99% of sequence identity to isoform b. In some embodiments, the APOBEC3A in the fusion protein of the present disclosure is rat APOBEC3 or a sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, 99% of sequence identity to the rat APOBEC3. In some embodiments, the APOBEC3A in the fusion protein of the present disclosure is mouse APOBEC3 or a sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, 99% of sequence identity to the mouse APOBEC3. In some embodiments, the sequence retains the cytidine deaminase activity.

In some embodiments, the APOBEC3A includes a Y130F mutation, according to residue numbering in SEQ ID NO:1 (the numbering would be different in human isoform b and rat and mouse sequences, but can readily be converted). In some embodiments, the APOBEC3A includes a Y132D mutation, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes a W104A mutation, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes a D131Y mutation, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes a D131E mutation, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes a W98Y mutation, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes a P134Y mutation, according to residue numbering in SEQ ID NO:1.

In some embodiments, the APOBEC3A includes mutations Y130F, D131E, and Y132D, according to residue numbering in SEQ ID NO:1 (the numbering would be different in human isoform b and rat and mouse sequences, but can readily be converted). In some embodiments, the APOBEC3A includes mutations Y130F, D131Y, and Y132D, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes mutations W98Y and W104A, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes mutations W98Y and P134Y, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes mutations W104A and P134Y, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes mutations W98Y, W104A, and Y130F, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes mutations W98Y, W104A, and Y132D, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes mutations W104A, Y130F, and P134Y, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes mutations W104A, Y132D, and P134Y, according to residue

numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes mutations W104A and Y130F, according to residue numbering in SEQ ID NO:1. In some embodiments, the APOBEC3A includes mutations W104A and Y132D, according to residue numbering in SEQ ID NO:1.

Example APOBEC3A sequences are shown in SEQ ID NO:1-10 and 22-36.

The APOBEC3A protein can allow further modifications, such as addition, deletion and/or substitutions, at other amino acid locations as well. Such modifications can be substitution at one, two or three or more positions. In one embodiment, the modification is substitution at one of the positions. Such substitutions, in some embodiments, are conservative substitutions. In some embodiments, the modified APOBEC3A protein still retains the cytidine deaminase activity. In some embodiments, the modified APOBEC3A protein retains the mutations tested in the experimental examples.

In various embodiments, the APOBEC3A can be substituted with another deaminase such as A3B (APOBEC3B), A3C (APOBEC3C), A3D (APOBEC3D), A3F (APOBEC3F), A3G (APOBEC3G), A3H (APOBEC3H), A3 (APOBEC3), and AID (AICDA).

In some embodiments, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3B (APOBEC3B) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein. In some embodiments, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3C (APOBEC3C) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein. In some embodiments, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3D (APOBEC3D) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein. In some embodiments, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3F (APOBEC3F) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein. In some embodiments, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3G (APOBEC3G) and a second fragment comprising a clustered regularly interspaced short

palindromic repeats (CRISPR)-associated (Cas) protein. In some embodiments, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3H (APOBEC3H) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein. In some embodiments, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3 (APOBEC3) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein. In some embodiments, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit AID (AICDA) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein.

In some embodiments, the APOBEC protein is a human protein. In some embodiments, the APOBEC protein is a mouse or rat protein. Some example APOBEC proteins are listed in the table below.

Deaminase	Example version	NCBI Accession Nos.
A3B (APOBEC3B)	hA3B (human)	NP_001257340, NP_004891
A3C (APOBEC3C)	hA3C (human)	NP_055323
A3D (APOBEC3D)	hA3D (human)	NP_689639, NP_001350710
A3F (APOBEC3F)	hA3F (human)	NP_660341, NP_001006667
A3G (APOBEC3G)	hA3G (human)	NP_068594, NP_001336365, NP_001336366, NP_001336367
A3H (APOBEC3H)	hA3H (human)	NP_001159474, NP_001159475, NP_001159476, and NP_861438
A1 (APOBEC1)	hA1 (human)	NP_001291495, NP_001635, NP_005880
	mA1 (mouse)	NP_001127863, NP_112436
A3 (APOBEC3)	mA3 (mouse)	NP_001153887, NP_001333970, NP_084531
AID (AICDA)	hAID (human)	NP_001317272, NP_065712
	mAID (mouse)	NP_033775
	cAICDA (channel catfish)	NP_001187114

A “conservative amino acid substitution” is one in which the amino acid residue is replaced with an amino acid residue having a similar side chain. Families of amino acid residues having similar side chains have been defined in the art, including basic side chains (*e.g.*, lysine, arginine, histidine), acidic side chains (*e.g.*, aspartic acid, glutamic acid), uncharged polar side chains (*e.g.*, glycine, asparagine, glutamine, serine, threonine, tyrosine, cysteine), nonpolar side chains (*e.g.*, alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan), beta-branched side chains (*e.g.*, threonine, valine, isoleucine) and aromatic side chains (*e.g.*, tyrosine,

phenylalanine, tryptophan, histidine). Thus, a nonessential amino acid residue in an immunoglobulin polypeptide is preferably replaced with another amino acid residue from the same side chain family. In another embodiment, a string of amino acids can be replaced with a structurally similar string that differs in order and/or composition of side chain family members.

Non-limiting examples of conservative amino acid substitutions are provided in the table below, where a similarity score of 0 or higher indicates conservative substitution between the two amino acids.

Table A. Amino Acid Similarity Matrix

	C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W
W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17
Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
K	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
H	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Q	-5	-1	0	-1	0	-1	2	2	1	4										
N	-4	0	-1	1	0	0	2	1	2											
E	-5	0	-1	0	0	0	3	4												
D	-5	1	-1	0	0	0	4													
T	-2	0	0	1	1	3														
A	-2	1	1	1	2															
S	0	1	1	1																
P	-3	-1	6																	
G	-3	5																		
C	12																			

Table B. Conservative Amino Acid Substitutions

For Amino Acid	Substitution With
Alanine	D-Ala, Gly, Aib, β -Ala, L-Cys, D-Cys
Arginine	D-Arg, Lys, D-Lys, Orn D-Orn
Asparagine	D-Asn, Asp, D-Asp, Glu, D-Glu Gln, D-Gln
Aspartic Acid	D-Asp, D-Asn, Asn, Glu, D-Glu, Gln, D-Gln
Cysteine	D-Cys, S-Me-Cys, Met, D-Met, Thr, D-Thr, L-Ser, D-Ser
Glutamine	D-Gln, Asn, D-Asn, Glu, D-Glu, Asp, D-Asp
Glutamic Acid	D-Glu, D-Asp, Asp, Asn, D-Asn, Gln, D-Gln

Glycine	Ala, D-Ala, Pro, D-Pro, Aib, β -Ala
Isoleucine	D-Ile, Val, D-Val, Leu, D-Leu, Met, D-Met
Leucine	Val, D-Val, Met, D-Met, D-Ile, D-Leu, Ile
Lysine	D-Lys, Arg, D-Arg, Orn, D-Orn
Methionine	D-Met, S-Me-Cys, Ile, D-Ile, Leu, D-Leu, Val, D-Val
Phenylalanine	D-Phe, Tyr, D-Tyr, His, D-His, Trp, D-Trp
Proline	D-Pro
Serine	D-Ser, Thr, D-Thr, allo-Thr, L-Cys, D-Cys
Threonine	D-Thr, Ser, D-Ser, allo-Thr, Met, D-Met, Val, D-Val
Tyrosine	D-Tyr, Phe, D-Phe, His, D-His, Trp, D-Trp
Valine	D-Val, Leu, D-Leu, Ile, D-Ile, Met, D-Met

The term “clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein” or simply “Cas protein” refers to RNA-guided DNA endonuclease enzymes associated with the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) adaptive immunity system in *Streptococcus pyogenes*, as well as other bacteria. Non-limiting examples of Cas proteins include *Streptococcus pyogenes* Cas9 (SpCas9), *Staphylococcus aureus* Cas9 (SaCas9), *Acidaminococcus* sp. Cas12a (Cpf1), *Lachnospiraceae bacterium* Cas12a (Cpf1), *Francisella novicida* Cas12a (Cpf1). Additional examples are provided in Komor *et al.*, “CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes,” Cell. 2017 Jan 12;168(1-2):20-36.

Example Cas proteins include SpCas9, FnCas9, St1Cas9, St3Cas9, NmCas9, SaCas9, AsCpf1, LbCpf1, FnCpf1, VQR SpCas9, EQR SpCas9, VRER SpCas9, RHA FnCas9, KKH SaCas9, NmeCas9, StCas9, CjCas9, AsCpf1, FnCpf1, SsCpf1, PcCpf1, BpCpf1, CmtCpf1, LiCpf1, PmCpf1, Pb3310Cpf1, Pb4417Cpf1, BsCpf1, EeCpf1, BhCas12b, AkCas12b, EbCas12b, LsCas12b, RfCas13d, LwaCas13a, PspCas13b, PguCas13b, RanCas13b, CasX, CasY and those provided in **Table C** below.

Table C. Example Cas Proteins

Cas protein types	Cas proteins
Cas9 proteins	Cas 9 from <i>Streptococcus pyogenes</i> (SpCas9) Cas9 from <i>Staphylococcus aureus</i> (SaCas9) Cas9 from <i>Neisseria meningitidis</i> (NmeCas9) Cas9 from <i>Streptococcus thermophilus</i> (StCas9)

	Cas9 from <i>Campylobacter jejuni</i> (CjCas9)
Cas12a (Cpf1) proteins	Cas12a (Cpf1) from <i>Lachnospiraceae bacterium</i> Cas12a (LbCpf1) Cas12a (Cpf1) from <i>Acidaminococcus</i> sp BV3L6 (AsCpf1) Cas12a (Cpf1) from <i>Francisella novicida</i> sp BV3L6 (FnCpf1) Cas12a (Cpf1) from <i>Smithella</i> sp SC K08D17 (SsCpf1) Cas12a (Cpf1) from <i>Porphyromonas crevioricanis</i> (PcCpf1) Cas12a (Cpf1) from <i>Butyrivibrio proteoclasticus</i> (BpCpf1) Cas12a (Cpf1) from <i>Candidatus Methanoplasma termitum</i> (CmtCpf1) Cas12a (Cpf1) from <i>Leptospira inadai</i> (LiCpf1) Cas12a (Cpf1) from <i>Porphyromonas macacae</i> (PmCpf1) Cas12a (Cpf1) from <i>Peregrinibacteria bacterium</i> GW2011 WA2 33 10 (Pb3310Cpf1) Cas12a (Cpf1) from <i>Parcubacteria bacterium</i> GW2011 GWC2 44 17 (Pb4417Cpf1) Cas12a (Cpf1) from <i>Butyrivibrio</i> sp. NC3005 (BsCpf1) Cas12a (Cpf1) from <i>Eubacterium eligens</i> (EcCpf1)
Cas12b (C2c1) proteins	Cas12b (C2c1) <i>Bacillus hisashii</i> (BhCas12b) Cas12b (C2c1) <i>Bacillus hisashii</i> with a gain-of-function mutation (see, e.g., Strecker et al., Nature Communications 10 (article 212) (2019)) Cas12b (C2c1) <i>Alicyclobacillus kakegawensis</i> (AkCas12b) Cas12b (C2c1) <i>Elusimicrobia bacterium</i> (EbCas12b) Cas12b (C2c1) <i>Laceyella sediminis</i> (Ls) (LsCas12b)
Cas13 proteins	Cas13d from <i>Ruminococcus flavefaciens</i> XPD3002 (RfCas13d) Cas13a from <i>Leptotrichia wadei</i> (LwaCas13a) Cas13b from <i>Prevotella</i> sp. P5-125 (PspCas13b) Cas13b from <i>Porphyromonas gulae</i> (PguCas13b) Cas13b from <i>Riemerella anatipestifer</i> (RanCas13b)
Engineered Cas proteins	Nickases (mutation in one nuclease domain) Catalytically inactive mutant (dCas; mutations in both of the nuclease domains) Enhanced variants with improved specificity (see, e.g., Chen <i>et al.</i> , Nature, 550, 407-410 (2017))

In some embodiments, the Cas protein is a mutant of protein selected from the above, wherein the mutant retains the DNA-binding capability but does not introduce double strand DNA breaks.

For example, it is known that in SpCas9, residues Asp10 and His840 are important for Cas9's catalytic (nuclease) activity. When both residues are mutated to Ala, the mutant loses the nuclease activity. In another embodiment, only the Asp10Ala mutation is made, and such a mutant protein cannot generate a double strand break; rather, a nick is generated on one of the strands. Such a mutant is also referred to as a Cas9 nickase. A non-limiting example of a Cas9 nickase is provided is SEQ ID NO: 11. Non-limiting example of a Cas12a nickase are provided is SEQ ID NO:37-39. Cas proteins also encompass mutants of known Cas proteins that have

certain sequence identity (e.g., 70%, 75%, 80%, 85%, 90%, 95%, 98%, 99% or more). In some embodiments, the Cas protein retains the catalytic (nuclease) activity.

In some embodiments, the Cas protein in a fusion protein of the present disclosure is a Cas12a (Cpf1, CRISPR-associated endonuclease in *Prevotella* and *Francisella* 1) protein. In conventional base editors, Cas9 is the commonly used DNA endonuclease. The Cas12a (Cpf1) has the advantage of recognizing A/T rich sequence when used together with APOBEC1 in base editors. In another surprising discovery of the present disclosure, when APOBEC1 was replaced with A3A, the editing efficiency was greatly increased (see, e.g., **Examples 3-5** and **FIG. 7B, 9B and 11B**). Yet, the editing efficiency of such a Cas12a-A3A can be further increased when the A3A includes a few tested mutations (**Examples 3-5** and **FIG. 7B, 9B and 11B**) and the editing window such a Cas12a-A3A can be narrowed to achieve more precise editing when even more tested mutations are included in A3A (**Examples 3-5** and **FIG. 8B, 10B and 12B**).

In some embodiments, therefore, provided is a fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A) and a second fragment comprising a CRISPR-associated endonuclease in *Prevotella* and *Francisella* 1 (Cpf1). Examples of APOBEC3A, as well as its alternatives (e.g., A3B (APOBEC3B), A3C (APOBEC3C), A3D (APOBEC3D), A3F (APOBEC3F), A3G (APOBEC3G), A3H (APOBEC3H), A3 (APOBEC3), or AID (AICDA)) and biological equivalents (homologues) have been disclosed above. Non-limiting example fusion sequences are provided in SEQ ID NO:40-50.

In some embodiments, the fusion protein further comprises a uracil glycosylase inhibitor (UGI). A non-limiting example of UGI is found in *Bacillus phage AR9* (YP_009283008.1). In some embodiments, the UGI comprises the amino acid sequence of SEQ ID NO:12 or has at least at least 90% sequence identity to SEQ ID NO:12 and retains the uracil glycosylase inhibition activity.

In some embodiments, the UGI is not fused to the fusion protein, but rather is provided separately (free UGI, not fused to a Cas protein or a cytosine deaminase) when the fusion protein is used for genomic editing. In some embodiments, the free UGI is provided with the fusion protein which also includes a UGI portion.

Preferably, a peptide linker is provided between each of the fragments in the fusion protein. In some embodiments, the peptide linker has from 1 to 100 amino acid residues (or 3-20, 4-15,

without limitation). In some embodiments, at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% or 90% of the amino acid residues of peptide linker are amino acid residues selected from the group consisting of alanine, glycine, cysteine, and serine. In some embodiments, the peptide linker has an amino acid sequence of SEQ ID NO:13 or 14.

The APOBEC3A, Cas protein, and UGI can be arranged in any manner. However, in a preferred embodiment, APOBEC3A is placed at the N-terminal side of the Cas protein. In one embodiment, the Cas protein is placed at the N-terminal side of the UGI.

In some embodiments, the fusion protein further comprises a nuclear localization sequence such as SEQ ID NO:15.

Non-limiting examples of fusion proteins include those having an amino acid sequence selected from the group consisting of SEQ ID NO:16-20.

Table 2. Additional Sequences

Name	Sequence	SEQ ID NO:
Cas9-Nickase	1 MYPYDVPDYA SPKKKPKVEA SDKKYSIGL ^A IGTVNSVGWAV ITDEYKVPSPK 51 KFKVLGNTDR HSIKKNLIGA LLFDSGETAE ATRLKRTARR RYTRRKNRIC 101 YLQEIFSNEM AKVDDSFHHR LEESFLVEED KKHERHPIFG NIVDEVAYHE 151 KYPTIYHLRK KLVDSTDKAD LRLLIYLALAH MIKFRGHFLI EGDLPDNDSD 201 VDKLFIQLVQ TYNQLFEENP INASGVDAKA ILSARLSKSR RLENLIAQLP 251 GEKKNGLFGN LIALSLGLTP NFKSNFDLAE DAKLQLSKDT YDDDLNLLA 301 QIGDQYADLF LAAKNLSDAI LLSDILRVNT EITKAPLSAS MIKRYDEHHQ 351 DLTLLKALVR QQLPEKYKEI FFDQSKNGYA GYIDGGASQE EFKYFIKPIL 401 EKMDGTEELL VKLNREDLLR KQRTFDNGSI PHQIHLGELH AILRRQEDFY 451 PFLKDNREKI EKILTFRIPY YVGPLARGNS RFAWMTRKSE ETITPWNFEE 501 VVDKGASQS FIERMTNFDR NLPNEKVLPK HSLLYEYFTV YNELTKVKYV 551 TEGMRKPAFL SGEQKKAIVD LLFETNRKVT VKQLKEDYFK KIECFDSVEI 601 SGVEDRFNAS LGTYHDLRI IKDKDFLDNE ENEDILEDIV LTLTLFEDRE 651 MIEERLKTYA HLFDDKVMKQ LKRRRYTGWG RLSRKLINGI RDKQSGKTIL 701 DFLKSDGFAN RNFQMQLIHD SLTFKEDIQK AQVSGQGDSL HEHIANLAGS 751 PAIKKGILQT VKVVDLVKV MGRHKPENIV IEMARENQTT QKGQKNSRER 801 MKRIEEGIKE LGSQILKEHP VENTQLQNEK LYLYYLQNGR DMYVDQELDI 851 NRLSDYDVH IVPQSFLKDD SIDNKVLTRS DKNRGKSDNV PSEEVVKKMK 901 NYWRQLLNAK LITQRKFDNL TKAERGGLSE LDKAGFIKRQ LVETRQITKH 951 VAQILD SRMN TKYDENDKLI REVKVITLKS KLVSDFRKDF QFYKVVREINN 1001 YHHAHDAYLN AVVGTAIIKK YPKLESEFVY GDYKVYDVRK MIAKSEQEIG 1051 KATAKYFFYS NIMNFFKTEI TLANGEIRKR PLIETNGETG EIVWDKGRDF 1101 ATVRKVL SMP QVNIVKKTEV QTGGFSKESI LPKRNSDKLI ARKKDWDPPK	11

	1151 YGGFDSPTVA YSVLVVAKVE KGKSKKLKSV KELLGITIME RSSFEKNPID 1201 FLEAKGYKEV KKDIIKLPK YSLFELENGR KRMLASAGEL QKGNELALPS 1251 KYVNFLYLAS HYEKLKGSPE DNEQKQLFVE QHKHYLDEII EQISEFSKRV 1301 ILADANLDKV LSAYNKHARDK PIREQAENII HLFTLTNLGA PAAFYFDTT 1351 IDRKPRTSTK EVLDATLIHQ SITGLYETRI DLSQLGGDSF KKKRKVEAS	
Uracil-DNA-glycosylase inhibitor (UGI)	1 TNLSDIIEKE TGKQLVIQES ILMLPREEVEE VIGNKPESDI LVHTAYDEST 51 DENVMMLTSD APEYKPWALV IQDSNGENKI KML	12
Linker 1	1 SGSETPGTSE SATPES	13
Linker 2	1 SGGG	14
Nuclear localization sequence	1 PKKKPKV	15
Fusion protein 1	1 MEASPASGPR HMDPHIFTS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARIY DYDPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CPFQFWDGLD EHSQALSGRL RAILQNGGNS 201 GSETPGTSES ATPESDKKYS IGLAIGTNSV GWAVITDEYK VPSKKFKVLG 251 NTDRHSIKKN LIGALLFDSG ETAEATRLKR TARRRYTRRK NRICYLQEIF 301 SNEMAKVDDS FFHRLEESFL VEEDKKHERH PIFGNIVDEV AYHEKYPTIY 351 HLRKKLV DST DKADLRILIY ALAHMIKFRG HFLIEGDLNP DNSDVKLFI 401 QLVQTYNQLF EENPINASGV DAKAILSARL SKSRRLNLI AQLPGEKKNG 451 LFGNLIALLSL GLTFNFKSNF DLAEDAKLQL SKDLYDDDLN NLLAQIGDQY 501 ADLFLAAKNL SDAILLSDIL RVNTEITKAP LSASMIKRYD EHHQDLTLLK 551 ALVRQQLPEK YKEIFFDQSK NGYAGYIDGG ASQEEFYKFI KPILKMDGT 601 EELLVKLNRE DLLRKQRTFD NGSIPHQIHL GELHAILRRQ EDFYPFLKDN 651 REKIEKILTF RIPPYVGPLA RGNRFAWMT RKSEETITPW NFEVVDKGA 701 SAQSFIERMT NFDKNLPNEK VLPKHSLLYE YFTVYNELTK VKYVTEGMRK 751 PAFLSGEQKK AIVDLLFRTN RKVTVKQLKE DYFKKIECFD SVEISGVEDR 801 FNASLGTYHD LLKIIKDKDF LDNEENEDIL EDIVLTTLTF EDREMIEERL 851 KTYAHLFDDK VMKQLKRRRY TGWGRLSRKL INGIRDKQSG KTILDFLKSD 901 GFANPNFMQL IHDDSLTFKE DIQKAQVSGQ GDSLHEHIAN LAGSFAIKKG 951 ILQTVKVVDL LVKVMGRHKP ENIVIEWARE NQTQKGQKN SRERMKRIEE 1001 GIKELGSQIL KEHPVENTQL QNEKLYLYYL QNGRDMYVDQ ELDINRLSDY 1051 DVDHIVFQSF LKDDSIDNKV LTRSDKNRGK SDNVFSEEVV KKMKNYWRQL 1101 LNAKLITQRK FDNLTKAERG GLSELDKAGF IKRQLVETRQ ITKHVAQILD 1151 SRMNTKYDEN DKLIREVKVI TLKSKLVSDF RKDFQFYKVR EINNYHHAHD 1201 AYLNAAVVGTA LIKKYPKLES EFVYGDKVY DVRKMIKASE QEIGKATAKY 1251 FFYSNIMNEF KTEITLANGE IRKRPLIETN GETGEIVWDK GRDFATVRKV 1301 LSMPQVNIVK KTEVQTGGFS RESILPKRNS DKLIARKKDW DPKKYGGFDS 1351 PTVAYSVLVV AKVEKGKSKK LKSVKELLGI TIMERSSEFEK NPIDFLEAKG 1401 YKEVKDLII KLPKYSLEFEL ENGRKRMLAS AGELQKGNEL ALPSKYVNFL 1451 YLASHYEKLLK GSPEDNEQKQ LFVEQHKHYL DEIEEQISEF SKRVILADAN 1501 LDKVLSAYNK HRDKPIREQA ENIIHLFTLT NLGAPAAFKY FDTTIDRKRY 1551 TSTKEVL DAT LIHQ SITGLY ETRIDLSQLG GDSGGSTNLS DIIKETGKQ	16

	1601 LVIQESILML PEEVEEVIGN KPESDILVHT AYDESTDENV MLLTSDAPEY 1651 KPWALVIQDS NGENKIKMLS GGSPKKKRKV	
Fusion protein 2 (Y130F)	1 MEASPASGPR HLMDFPHIFTS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARI F DYDPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGNS 201 GSETPGTSES ATPESDKKYS IGLAIGTNSV GWAVITDEYK VPSKKFKVLG 251 NTDHRSIKKN LIGALLFDSDG ETAEATRLKR TARRRYTRRK NRICYLQEIF 301 SNEMAKVDDS FFHRLEESFL VEEDKKHERH PIFGNIVDEV AYHEKYPTIY 351 HLRKKLVDST DKADLRILIY ALAHMIKFRG HFLIEGDLNP DNSDVKLFI 401 QLVQTYNQLF EENPINASGV DAKAILSARL SKSRRLNLI AQLPGEKKNK 451 LFGNLIALLSL GLTPNFKSNF DLAEDAKLQL SKDTYDDDLN NLLAQIGDQY 501 ADLFLAAKNL SDAILLSDIL RVNTEITKAP LSASMIKRYD EHHQDLTLK 551 ALVRQQLPEK YKEIFFDQSK NGYAGYIDGG ASQEEFYKFI KPILEKMDGT 601 EELLVKLNRE DLLRKQRTFD NGSIPHQIHL GELHAILRRQ EDFYPFLKDN 651 REKIEKILTF RIPPYYVGPLA RGNSRFAWMT RKSEETITPW NFEVVVDKGA 701 SAQSFIERTM NFDKNLPNEK VLPKHSLLYE YFTVYNELTK VKYVTEGMRK 751 PAFLSGEQKK AIVDLLFKTN RKVTVKQLKE DYFKKIECFD SVEISGVEDR 801 FNASLGTYHD LLKIKDKDF LDNEENEDIL EDIVLTITLF EDREMIEERL 851 KTYAHLFDDK VMKQLKRRRY TGWGRLSRKL INGIRPKQSG KTILDFLKSD 901 GFANRNFMQL IHDDSLTFKE DIQKAQVSGQ GDSLHEHIAN LAGSPAIKKG 951 ILQTVKVVDL LVKVMGRHKP ENIVIAMARE NQTTQKGQKN SRERMKRIEE 1001 GIKELGSQIL KEHPVENTQL QNEKLYLYYL QNGRDMYVDQ ELDINRLSDY 1051 DVDHIVPQSF LKDDSIDNKV LTRSDKNRGK SDNVPSEEVV KMKKNYWRQL 1101 LNAKLITQRK FDNLTKAERG GLSELDKAGF IKRQLVETRO ITKHVAQILD 1151 SRMNTKYDEN DKLIREVKVI TLKSKLVSDF RKDFQFYKVR EINNYHHAHD 1201 AYLNAVVGTA LIKKYPKLES EFVYGDYKVY DVRKMIKASE QEIGKATAKY 1251 FFYSNIMNFF KTEITLANGE IRKRPLIETN GETGEIVWDK GRDFATVRKV 1301 LSMPOVNIVK KTEVQTGGFS KESILPKRNS DKLIARKKDW DPKKYGGFDS 1351 PTVAYSVLVW AKVEKGKSKK LKSVKELLGI TIMERSSEFEK NPIDFLEAKG 1401 YKEVKKDLII KLPKYSLFEL ENGRKRMLAS AGELQKGNEL ALPSKYVNF 1451 YLASHYEKLLK GSPEDNEQKQ LFVEQHKHYL DEIEQISEF SKRVILADAN 1501 LDKVLSAYNK HRDKPIREQA ENIIHLFTLT NLGAPAAFKY FDTTIDRKRY 1551 TSTKEVLDAI LIHQSIITGLY ETRIDLSQLG GDSGGSTNLS DIIEKETGKQ 1601 LVIQESILML PEEVEEVIGN KPESDILVHT AYDESTDENV MLLTSDAPEY 1651 KPWALVIQDS NGENKIKMLS GGSPKKKRKV	17
Fusion protein 3 (Y132D)	1 MEASPASGPR HLMDFPHIFTS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARIY DD DPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGNS 201 GSETPGTSES ATPESDKKYS IGLAIGTNSV GWAVITDEYK VPSKKFKVLG 251 NTDHRSIKKN LIGALLFDSDG ETAEATRLKR TARRRYTRRK NRICYLQEIF 301 SNEMAKVDDS FFHRLEESFL VEEDKKHERH PIFGNIVDEV AYHEKYPTIY 351 HLRKKLVDST DKADLRILIY ALAHMIKFRG HFLIEGDLNP DNSDVKLFI	18

	401 QLVQTYNQLF EENPINASGV DAKAILSARL SKSRRLLENLI AQLPGEKKNG 451 LFGNLIALLSL GLTPNFKSNF DLAEDAKLQL SKDTYDDDLN NLLAQIGDQY 501 ADLFLAAKNL SDAILLSDIL RVNTEITKAP LSASMIKRYD EHHQDLTLK 551 ALVRQQLPEK YKEIFFDQSK NGYAGYIDGG ASQEEFYKFI KPILEKMDGT 601 EELLVKLNRE DLLRKQRTFD NGSIPHQIHL GELHAILRRQ EDFYFPLKDN 651 REKIEKILTF RIPPYVGPLA RGNSRFAWMT RKSEETITPW NFEVVVDKGA 701 SAQSFIERTM NFDKNLPNEK VLPKHSLLYE YFTVYNELTK VKYVTEGMRK 751 PAFLSGEQKK AIVDLLFKTN RKVTVKQLKE DYFKKIECFD SVEISGVEDR 801 FNASLGTYHD LLKLIKDKDF LDNEENEDIL EDIVLTTLTF EDREMIEERL 851 KTYAHLFDDK VMKQLKRRRY TGWGRLSRKL INGIRDQSG KTILDFLKSD 901 GFANRNFMQL IHDDSLTFKE DIQKAQVSGQ GDSLHEHIAN LAGSPAIKKG 951 ILQTVKVVDE LVKVMGRHPK ENIVEMARE NQTQKGQKN SRERMKRIEE 1001 GIKELGSQIL KEHPVENTQL QNEKLYLYYL QNGRDMYVDQ ELDINRLSDY 1051 DVDHIVPQSF LKDDSIDNKV LTRSDKNRGK SDNVPSEEVV KMKNYWRQL 1101 LNAKLITQRK FDNLTKAERG GLSELDKAGF IKRQLVETRO ITHKVAQILD 1151 SRMNTKYDEN DKLIREVKVI TLKSKLVSDF RKDFQFYKVR EINNYHHAHD 1201 AYLNNAVVGTA LIKKYPKLES EFVYGDYKVY DVRKMIKASE QEIGKATAY 1251 FFYSNIMNFF KTEITLANGE IRKRPLIETN GETGEIVWDK GRDFATVRKV 1301 LSMQVQNVK KTEVQTGGFS KESILPKRNS DKLIARKKDW DPKKYGGFDS 1351 PTVAYSVLVV AKVEKGSKK LKSVKELLGI TIMERSSEFEK NPIDFLEAKG 1401 YKEVKKDLII KLPKYSLFEL ENGRKRMLAS AGELQKGNEL ALPSKYVNFL 1451 YLASHYEKLLK GSPEDNEQKQ LFVEQHKHYL DEIEEQISEF SKRVILADAN 1501 LDKVLSAYNK HRDKPIREQA ENIIHLFTLT NLGAPAAFKY FDTTIDRKRY 1551 TSTKEVLDT LIHQSIITGLY ETRIDLSQLG GDSGGSTNLS DIIKETGKQ 1601 LVIQESILML PEEVEEVIGN KPESDILVHT AYDESTDENV MLLTSDAPEY 1651 KPWALVIQDS NGENKIKMLS GGSPKKKRKV	
Fusion protein 4 (W104A)	1 MEASPASGPR HLMDFPHIFTS NFNNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSAGCAGEV RAFLQENTHV RLRIFAARIY DYDPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CFFQPWDGLD EHSQALSGRL RAILQNQGNS 201 GSETPGTSSES ATPESDKKYS IGLAIGTNSV GWAVITDEYK VPSKKFKVLG 251 NTDHRSIKKN LIGALLFDSDG ETAEATRLKR TARRRYTRRK NRICYLQEIF 301 SNEMAKVDDS FFHRLSEESFL VEEDKKHERH PIFGNIVDEV AYHEKYPTIY 351 HLRKKLV DST DKADLRLLYL ALAHMIKFRG HFLIEGDLNP DNSDVKLFI 401 QLVQTYNQLF EENPINASGV DAKAILSARL SKSRRLLENLI AQLPGEKKNG 451 LFGNLIALLSL GLTPNFKSNF DLAEDAKLQL SKDTYDDDLN NLLAQIGDQY 501 ADLFLAAKNL SDAILLSDIL RVNTEITKAP LSASMIKRYD EHHQDLTLK 551 ALVRQQLPEK YKEIFFDQSK NGYAGYIDGG ASQEEFYKFI KPILEKMDGT 601 EELLVKLNRE DLLRKQRTFD NGSIPHQIHL GELHAILRRQ EDFYFPLKDN 651 REKIEKILTF RIPPYVGPLA RGNSRFAWMT RKSEETITPW NFEVVVDKGA 701 SAQSFIERTM NFDKNLPNEK VLPKHSLLYE YFTVYNELTK VKYVTEGMRK 751 PAFLSGEQKK AIVDLLFKTN RKVTVKQLKE DYFKKIECFD SVEISGVEDR 801 FNASLGTYHD LLKLIKDKDF LDNEENEDIL EDIVLTTLTF EDREMIEERL 851 KTYAHLFDDK VMKQLKRRRY TGWGRLSRKL INGIRDQSG KTILDFLKSD 901 GFANRNFMQL IHDDSLTFKE DIQKAQVSGQ GDSLHEHIAN LAGSPAIKKG	19

	951 ILQTVKVVDE LVKVMGRHKP ENIVIAMARE NQTTQKGQKN SRERMKRIEE 1001 GIKELGSQIL KEHPVENTQL QNEKLYLYYL QNGRDMYVDQ ELDINRLSDY 1051 DVDHIVPQSF LKDDSIDNKV LTRSDKNRGK SDNVPSEEVV KKMKNYWRQL 1101 LNAKLITQRK FDNLTKAERG GLSELDKAGF IKRQLVETRQ ITKHVAQILD 1151 SRMNTKYDEN DKLIREVKVI TLKSKLVSDF RKDFQFYKVR EINNYYHHAHD 1201 AYLNNAVVGTA LIKKYPKLES EFVYGDYKVY DVRKMIKASE QEIGKATAKY 1251 FFYSNIMNFF KTEITLANGE IRKRPLIETN GETGEIVWDK GRDFATVRKV 1301 LSMPQVNIVK KTEVQTGGFS KESILPKRNS DKLIARKKDW DPKKYGGFDS 1351 PTVAYSVLVV AKVEKGKSKK LKSVKELLGI TIMERSSEFEK NPIDFLEAKG 1401 YKEVKKDLII KLPKYSLEFEL ENGRKRMLAS AGELQKGNEL ALPSKYVNFL 1451 YLASHYEKLEK GSPEDNEQKQ LFVEQHKHYL DEIEQISEF SKRVILADAN 1501 LDKVLSAYNK HRDKPIREQA ENIIHLFTLT NLGAPAAFKY FDTTIDRKRY 1551 TSTKEVL DAT LIHQ SITGLY ETRIDLSQLG GDSGGSTNLS DIIEKETGKQ 1601 LVIQESILML PEEVEEVIGN KPESDILVHT AYDESTDENV MLLTSDAPEY 1651 KPWALVIQDS NGENKIKMLS GGSPPKKRKV	
Fusion protein 5 (D131Y)	1 MEASPASGPR HMDPHIFTS NFMNGIGRHK TYLCYEVERL DNGTSVKMDQ 51 HRGFLHNQAK NLLCGFYGRH AELRFLDLVP SLQLDPAQIY RVTWFISWSP 101 CFSWGCAGEV RAFLQENTHV RLRIFAARIY <u>Y</u> YDPLYKEAL QMLRDAGAQV 151 SIMTYDEFKH CWDTFVDHQG CPFQPWDGLD EHSQALSGRL RAILQNQGNS 201 GSETPGTSES ATPESDKKYS IGLAIGTNSV GWAVITDEYK VPSKKFKVLG 251 NTDHRSIKKN LIGALLFD SG ETAEATRLKR TARRRYTRRK NRICYLQEIF 301 SNEMAKVDDS FFHRLEESFL VEEDKKHERH PIFGNIVDEV AYHEKYPTIY 351 HLRKKLV DST DKADLR LIYL ALAHMIKFRG HFLIEGDLNP DNSDVKLFI 401 QLVQTYNQLF EENPINASGV DAKAILSARL SKSRRLNLI AOLPGEKKNG 451 LFGNLIALLSL GLTFNFKSNF DLAEDAKLQL SKDTYDDDL NLLAQIGDQY 501 ADLFLAAKNL SDAILLSDIL RVNTEITKAP LSASMIKRYD EHHQDLTLLK 551 ALVRQQLPEK YKEIFFDQSK NGYAGYIDGG ASQEEFYKFI KPILEKMDGT 601 EELLVKLNRE DLLRKQRTFD NGSIPHQIHL GELHAILRRQ EDFYPFLKDN 651 REKIEKILTF RIPPYVGPLA RGNSRFAMMT PKSEETITPW NFEENVDKGA 701 SAQSFIERMT NFDKNLPNEK VLPKHSLLYE YFTVYNELTK VKYVTEGMRK 751 PAFLSGEQKK AIVDLLFKTN RKVTVKQLKE DYFKKIECFD SVEISGVEDR 801 FNASLGTYHD LLKIIKDKDF LDNEENEDIL EDIVLTTLTF EDREMIEERL 851 KTYAHLFDDK VMKQLKRRRY TGWGRLSRKL INGIRDKQSG KTILDFLKSD 901 GFANRNFMQL IHDDSLTFKE DIQKAQVSGQ GDSLHEHIAN LAGSPAIKKG 951 ILQTVKVVDE LVKVMGRHKP ENIVIAMARE NQTTQKGQKN SRERMKRIEE 1001 GIKELGSQIL KEHPVENTQL QNEKLYLYYL QNGRDMYVDQ ELDINRLSDY 1051 DVDHIVPQSF LKDDSIDNKV LTRSDKNRGK SDNVPSEEVV KKMKNYWRQL 1101 LNAKLITQRK FDNLTKAERG GLSELDKAGF IKRQLVETRQ ITKHVAQILD 1151 SRMNTKYDEN DKLIREVKVI TLKSKLVSDF RKDFQFYKVR EINNYYHHAHD 1201 AYLNNAVVGTA LIKKYPKLES EFVYGDYKVY DVRKMIKASE QEIGKATAKY 1251 FFYSNIMNFF KTEITLANGE IRKRPLIETN GETGEIVWDK GRDFATVRKV 1301 LSMPQVNIVK KTEVQTGGFS KESILPKRNS DKLIARKKDW DPKKYGGFDS 1351 PTVAYSVLVV AKVEKGKSKK LKSVKELLGI TIMERSSEFEK NPIDFLEAKG 1401 YKEVKKDLII KLPKYSLEFEL ENGRKRMLAS AGELQKGNEL ALPSKYVNFL 1451 YLASHYEKLEK GSPEDNEQKQ LFVEQHKHYL DEIEQISEF SKRVILADAN	20

	1501 LDKVLSAYNK HRDKPIREQA ENIIHLFTLT NLGAPAAFKY FDTTIDRKRY 1551 TSTKEVL DAT LIHQSI TGLY ETRIDLSQLG GDSGGSTNLS DIIEKETGKQ 1601 LVIQESILML PEEVEEVIGN KPESDILVHT AYDESTDENV MLLTSDAPEY 1651 KPWALVIQDS NGENKIKMLS GGSPKKRKV	
DNA construct	1 Atatgccaag tacgccccct attgaagtc aatgacggtaa atggcccgcc 51 tggcattatg ccaggtacat gacottatgg gactttccta cttggcagta 101 catctacgta ttagtcatcg ctattacat ggtgatgcgg ttttggcagt 151 acatcaatgg gcgtggatag cggtttgact caccgggatt tccaagtctc 201 caccgccattg acgtcaatgg gagtttggtt tggcaccaaa atcaacggga 251 ctttccaaaa tgtcgttaaca actccgcccc attgacgcaa atgggcggta 301 ggcggtgtacg gtgggaggtc tatataagca gagctgggtt agtgaaccgt 351 cagatccgct agagatccgc gcccgcta atcgaactcact atagggagag 401 ccgccaccat ggaagccagc ccagcatccg gccccagaca cttgatggat 451 ccacacatat tcacttccaa ctttaacaat ggcattggaa ggcataagac 501 ctacctgtgc tacgaagtgg agcgccctgga caatggcacc tcggtcaaga 551 tggaccagca caggggcttt ctacacaacc aggctaagaa tcttctctgt 601 ggcttttacg gccgccaatgc ggagctgcgc ttcttggacc tggttccttc 651 tttgcagttg gaccgggccc agatctacag ggtcacttgg ttcattctct 701 ggagccctcg cttctcctgg ggtgtgcgc ggaagtgcg tgcgttctct 751 caggagaaca cacacgtgag actgcgtatc ttcgctgccc gcattctatga 801 ttacgacccc ctatataagg aggcactgca aatgctgcgg gatgctgggg 851 cccaagtctc catcatgacc tacgatgaat ttaagcactg ctgggacacc 901 tttgtggacc accagggatg tcccttccag cctgggatg gactagatga 951 gcacagccaa gccctgagtg ggaggctgcg gccattctc cagaatcagg 1001 gaaacagcgg cagcgagact cccgggacct cagagtcgc cacaccggaa 1051 agtgataaaa agtattctat tggtttagcc atcggcacta attccgttgg 1101 atgggctgtc ataaccgatg aatacaaaagt accttcaaag aaatttaagg 1151 tgttggggaa cacagaccgt cattcgatta aaaagaatct tatcggtgcc 1201 ctctattctg atagtggcga aacggcagag gcgactcgcc tgaaacgaac 1251 cgctcggaga aggtatacac gtgcgaagaa ccgaatatgt tacttacaag 1301 aaatttttag caatgagatg gccaaagttg acgattcttt ctttcaccgt 1351 ttggaagagt ccttccttgt cgaagaggac aagaaacatg aacggcacc 1401 catctttgga aacatagtag atgaggtggc atatcatgaa aagtacccaa 1451 cgatttatca cctcagaaaa aagctagttg actcaactga taaagcggac 1501 ctgaggttaa tctacttggc tcttgcccat atgataaagt tccgtgggca 1551 ctttctcatt gaggtgatc taaatccgga caactcgat gtgcacaaac 1601 tgttcattca gttagtacaa acctataatc agttgtttga agagaacct 1651 ataaatgcaa gtggcggtga tgcgaaggct attcttagcg ccgcctctc 1701 taaatccgga cggctagaaa acctgatcgc acaattaccc ggagagaaga 1751 aaaatgggtt gttcggtaac cttatagcgc tctcactagg cctgacacca 1801 aatttttaagt cgaacttcga cttagctgaa gatgccaat tgcagcttag 1851 taaggacacg tacgatgacg atctcgacaa tctactggca caaattggag 1901 atcagtatgc ggacttattt ttggtgcga aaaaccttag cgatgcaatc 1951 ctctatctg acatactgag agttaatact gagattacca aggcgcgtt 2001 atccgcttca atgatcaaaa ggtacgatga acatcaacca gacttgacac	21

2051	ttctcaaggc	cctagtcogt	cagcaactgc	ctgagaaata	taaggaaata
2101	ttctttgatc	agtcgaaaaa	cgggtacgca	ggttatattg	acggcggagc
2151	gagtcaagag	gaattctaca	agtttatcaa	acccatatta	gagaagatgg
2201	atgggaacgga	agagttgctt	gtaaaactca	atcgogaaga	tctactgcga
2251	aagcagcgga	ctttcgacaa	cggtagcatt	ccacatcaaa	tccacttagg
2301	cgaattgcat	gctatactta	gaaggcagga	ggatTTTTat	ccgttctctca
2351	aagacaatcg	tgaaaagatt	gagaaaatcc	taacctttcg	cataccttac
2401	tatgtgggac	ccctggcccg	agggaactct	cggttcgcat	ggatgacaag
2451	aaagtccgaa	gaaacgatta	ctccatggaa	ttttgaggaa	gtgtgcgata
2501	aagggtgcgc	agctcaatcg	ttcatcgaga	ggatgaccaa	ctttgacaag
2551	aattttacga	acgaaaaagt	attgcctaag	cacagtttac	tttacgagta
2601	tttcacagtg	tacaatgaac	tcacgaaagt	taagtatgtc	actgagggca
2651	tgcgtaaacc	cgcctttcta	agcggagaa	agaagaaagc	aatagtagat
2701	ctgttattca	agaccaaccg	caaagtgaca	gttaagcaat	tgaaagagga
2751	ctactttaag	aaaattgaat	gcttcgattc	tgctgagatc	tccggggtag
2801	aagatcgatt	taatgcgtca	cttggtaagt	atcatgacct	cctaaagata
2851	attaaagata	aggacttctt	ggataacgaa	gagaatgaag	atatcttaga
2901	agatatagtg	ttgactctta	ccctctttga	agatcgggaa	atgattgagg
2951	aaagactaaa	aacatacgtc	cacctgttcg	acgataaggt	tatgaaacag
3001	ttaaagaggc	gtcgtctata	gggctgggga	cgattgtcgc	ggaaacttat
3051	caacgggata	agagacaagc	aaagtggtaa	aactattctc	gattttctaa
3101	agagcgacgg	cttcgccaat	aggaaacttta	tgagctgat	ccatgatgac
3151	tctttaacct	tcaaagagga	tatacaaaa	gcacagggtt	ccggacaagg
3201	ggactcattg	cacgaacata	ttgcgaatct	tgctggttcg	ccagccatca
3251	aaaagggcat	actccagaca	gtcaaagtag	tggatgagct	agttaaggtc
3301	atgggaacgc	acaaacggga	aaacattgta	atcgagatgg	cacgcgaaaa
3351	tcaaacgact	cagaaggggc	aaaaaaacag	tcgagagcgg	atgaagagaa
3401	tagaagaggg	tattaaagaa	ctgggcagcc	agatcttaaa	ggagcatcct
3451	gtggaaaaata	cccaattgca	gaacgagaaa	ctttacctct	attacctaca
3501	aaatggaagg	gacatgtatg	ttgatcagga	actggacata	aaccgtttat
3551	ctgattacga	cgtcgatcac	attgtacccc	aatccttttt	gaaggacgat
3601	tcaatcgaca	ataaagtgtc	tacacgctcg	gataagaacc	gagggaaaa
3651	tgacaatggt	ccaagcgagg	aagtcgtaaa	gaaaatgaag	aactattggc
3701	ggcagctcct	aatgcgaaa	ctgataacgc	aaagaaagtt	cgataactta
3751	actaaagctg	agaggggtgg	cttgtctgaa	cttgacaagg	ccggatttat
3801	taaacgtcag	ctcgtggaaa	ccgcgcaaat	cacaaagcat	gttgacacaga
3851	tactagattc	ccgaatgaat	acgaaatcgc	acgagaacga	taagctgatt
3901	cgggaagtca	aagtaatcac	tttaaagtca	aaattgggtg	cggacttcag
3951	aaaggatttt	caattctata	aagttaggga	gataaataac	taccaccatg
4001	cgcacgacgc	ttatcttaat	gccgtcgtag	ggaccgcact	cattaagaaa
4051	taccogaagc	tagaaagtga	gtttgtgtat	ggtgattaca	aagtttatga
4101	cgtccgtaag	atgatcgaga	aaagcgaaca	ggagataggc	aaggctacag
4151	ccaaatactt	cttttattct	aacattatga	atttctttta	gacggaaaac
4201	actctggcaa	acggagagat	acgcaaacga	cctttaattg	aaaccaatgg
4251	ggagacaggt	gaaatcgtat	gggataaggg	ccgggacttc	gcgacggtga

4301	gaaaagtttt	gtccatgccc	caagtcaaca	tagtaaagaa	aactgaggtg
4351	cagaccggag	ggttttcaaa	ggaatcgatt	cttccaaaaa	ggaatagtga
4401	taagctcatc	gctcgtaaaa	aggactggga	cccgaaaaag	tacgggtggct
4451	tcgatagccc	tacagttgcc	tattctgtcc	tagtagtggc	aaaagttgag
4501	aagggaaaat	ccaagaaact	gaagtcagtc	aaagaattat	tggggataac
4551	gattatggag	cgctcgtctt	ttgaaaagaa	ccccatcgac	ttccttgagg
4601	cgaaaggtta	caaggaagta	aaaaaggatc	tcataattaa	actaccaaag
4651	tatagtctgt	ttgagttaga	aatggccga	aaacggatgt	tggctagcgc
4701	cggagagctt	caaaaaggga	acgaactcgc	actacogtct	aaatacgtga
4751	atttcctgta	tttagcgtcc	cattacgaga	agttgaaagg	ttcacctgaa
4801	gataacgaac	agaagcaact	ttttgttgag	cagcaciaac	attatctcga
4851	cgaaatcata	gagcaaatct	cggaattcag	taagagagtc	atcctagctg
4901	atgccaatct	ggacaaaagta	ttaagcgcac	acaacaagca	cagggataaa
4951	cccatcagtg	agcaggcggg	aaatattatc	catttgttta	ctcttaccaa
5001	cctcggcgct	ccagccgcac	tcaagtatct	tgacacaacg	atagatcgca
5051	aacgatacac	ttctaccaag	gaggtgctag	acgcgacact	gattcaccaa
5101	tccatcacgg	gattatatga	aactcggata	gatttgctac	agcttggggg
5151	tgactctggt	ggttctacta	atctgtcaga	tattattgaa	aaggagaccg
5201	gtaagcaact	ggttatccag	gaatccatcc	tcattgctcc	agaggaggtg
5251	gaagaagtca	ttgggaacaa	gccggaaagc	gatatactcg	tgacacccgc
5301	ctacgaacgag	agcaccgacg	agaatgtcat	gcttctgact	agcgacgccc
5351	ctgaatacaa	gccttgggct	ctggtcatac	aggatagcaa	cggtgagaa
5401	aagattaaga	tgctctctgg	tggttctccc	aagaagaaga	ggaaagtcta
5451	accggtcatc	atcaccatca	ccattgagtt	taaacccgct	gatcagcctc
5501	gactgtgctt	tctagttgcc	agccatctgt	tgtttgcccc	tcccccgctc
5551	cttccttgac	cctggaaggt	gccactccca	ctgtcctttc	ctaataaaat
5601	gaggaaattg	catcgcatct	tctgagtagg	tgctattcta	ttctgggggg
5651	tgggtggggg	caggacagca	agggggagga	ttgggaagac	aatagcaggc
5701	atgctgggga	tgcggtgggc	tctatggctt	ctgaggcgga	aagaaccagc
5751	tggggtctga	taccgtcgac	ctctagctag	agcttggcgt	aatcatggct
5801	atagctgttt	cctgtgtgaa	attgttatcc	gctcacaatt	ccacacaaca
5851	tacgagccgg	aagcataaag	tgtaaagcct	agggtgccct	atgagtgagc
5901	taactccatc	taattgcggt	gcgctcactg	cccgcctttc	agtcgggaaa
5951	cctgtcgtgc	cagctgcatt	aatgaatcgg	ccaacgcgcg	gggagaggcg
6001	gtttgcgtat	tgggcgctct	tccgcttctc	cgctcaactga	ctcgtcgcgc
6051	tcggtcgttc	ggctgcggcg	agcggtatca	gctcaactca	aggcggtaat
6101	acggttatcc	acagaatcag	gggataacgc	aggaaagaac	atgtgagcaa
6151	aaggccagca	aaaggccagg	aaccgtaaaa	aggccgcggt	gctggcgctt
6201	ttccataggc	tccgcccccc	tgacgagcat	cacaaaaatc	gacgctcaag
6251	tcagaggtgg	cgaaaaccga	caggactata	aagataccag	gcgtttcccc
6301	ctggaagctc	cctcgtgcgc	tctcctgttc	cgaccctgcc	gcttacggga
6351	tacctgtccg	cctttctccc	ttcgggaagc	gtggcgcttt	ctcatagctc
6401	acgctgtagg	tatctcagtt	cgggtgtagt	cgctcgcctc	aagctgggct
6451	gtgtgcacga	accccccggt	cagcccgacc	gctgcgcctt	atccggtaac
6501	tatcgtcttg	agtccaaccc	ggtaagacac	gacttatcgc	cactggcagc

	6551 agccactggt aacaggatta gcagagcgag gtatgtaggc ggtgctacag 6601 agttcttgaa gtgggtggcct aactacggct acactagaag aacagtatctt 6651 ggtatctgcg ctctgctgaa gccagttacc ttcggaaaaa gagttggtag 6701 ctcttgatcc ggcaaacaaa ccaccgctgg tagcgggtggg ttttttgggtt 6751 gcaagcagca gattacgcgc agaaaaaaag gatctcaaga agatcctttg 6801 atctttttcta cgggggtctga cgctcagtggt aacgaaaact cacgttaagg 6851 gatttttggtc atgagattat caaaaaggat cttcacctag atcctttttaa 6901 attaaaaatg aagtttttaa tcaatctaaa gtatatatga gtaaaacttg 6951 tctgacagtt accaatgctt aatcagtgag gcacctatct cagcgatctg 7001 tctatttctgt tcatccatag ttgcctgact ccccgctgtg tagataacta 7051 cgatacggga gggcttaacca tctggcccca gtgctgcaat gataccgcga 7101 gacccacgct caccggtcc agatttatca gcaataaacc agccagcgg 7151 aagggccgag cgcagaagtg gtccctgcaac tttatccgcc tccatccagt 7201 ctattaattg ttgccgggaa gctagagtaa gtagttcgcc agttaatagt 7251 ttgcgcaacg ttgttgccat tgctacaggc atcgtgggtg caogctcgtc 7301 gtttgggtatg gcttcattca gctccggttc ccaacgatca aggcgagtta 7351 catgatcccc catgtttgtg aaaaaagcgg ttagctcctt cggctcctcg 7401 atcgtttgtca gaagtaagtt ggccgcagtg ttatcaacta tggttatggc 7451 agcaactgcat aattctctta ctgtcatgcc atccgtaaga tgcctttctg 7501 tgactggtga gtactcaacc aagtcattct gagaatagtg tatgcggcga 7551 ccgagttgct cttgcccggc gtcaatacgg gataataccg cgcacatag 7601 cagaacttta aaagtgtca tcattggaaa acgttcttcg gggcgaaaaac 7651 tctcaaggat cttaccgctg ttgagatcca gttcgatgta acccaactgt 7701 gcacccaact gatcttcagc atcttttact ttcaccagcg tttctgggtg 7751 agcaaaaaaca ggaaggcaaa atgccgcaaa aaagggaata agggcgacac 7801 ggaaatgttg aatactcata ctcttccttt ttcaatatta ttgaagcatt 7851 tatcagggtt attgtctcat gagcggatac atatttgaat gtatttagaa 7901 aaataaaca ataggggttc cgcgcacatt tccccgaaaa gtgccacctg 7951 acgtcgacgg atcgggagat cgatctcccg atccctagg gtgcactctc 8001 agtacaatct gctctgatgc cgcatagtta agccagtatc tgcctcctgc 8051 ttgtgtgttg gaggtcgctg agtagtgcgc gagcaaaatt taagctacaa 8101 caaggcaagg cttgaccgac aattgcatga agaactctgt taggggttagg 8151 cgttttgcgc tgcttcgcga tgtacgggcc agatatacgc gttgacattg 8201 attattgact agttattaat agtaatcaat tacggggtca ttagttcata 8251 gccatataat ggagttccgc gttacataac ttacggtaaa tggcccgctt 8301 ggctgaccgc ccaacgaccc ccgccattg acgtcaataa tgacgtatgt 8351 tcccatagta acgccaatag ggactttcca ttgacgtcaa tgggtggagt 8401 atttacggta aactgcccac ttggcagtac atcaagtgtc tc	
Lb-dCas12a	1 MSKLEKFTNC YLSKTLRFK AIPVGKTQEN IDNKRLLED EKRAEDYKGV 51 KKLLDRYLS FINDVLHSIK LKNLNNYISL FRKTRTEKE NKELENLEIN 101 LRKEIAKAFK GNEGYKSLFK KDIIETILPE FLDDKDEIAL VNSFNGFTTA 151 FTGFFDNREN MFSEEAKSTS IAFRCINENL TRYISNMDIF EKVDIAFDKH 201 EVQEIKEKIL NSDYDVEDFF EGEFFNFVLT QEGIDVYNAI IGGFVTSERGE 251 KIKGLNEYIN LYNQKTKQKL PKFKPLYKQV LSDRESLSFY GEGYTSDEEV 301 LEVFRNTLNK NSEIFSSIKK LEKLFKNFDE YSSAGIFVKV GPAISTISKD	37

	351 IFGEWNVIRD KWNAEYDDIH LKKKAVVTEK YEDDRKSFK KIGSFSLEQL 401 QEYADADLSV VEKLKEIIIIQ KVDEIYKVYG SSEKLFDADF VLEKSLKKND 451 AVVAIMKDLL DSVKSFENYI KAFFGEGKET NRDESFGYGF VLAYDILLKV 501 DHIYDAIRNY VTQKPYSKDK FKLYFQNPQF MGGWDKDKET DYRATILRYG 551 SKYYLAIMDK KYAKCLQKID KDDVNGNYEK INYKLLPGPN KMLPKVFFSK 601 KWMAYYNPSE DIQKIYKNGT FKKGDMFNLN DCHKLIDFFK DSISRYPKWS 651 NAYDFNFSET EKYKDIAGFY REVEEQGYKV SFESASKKEV DKLVEEGKLY 701 MFQIYNKDFS DKSHGTPNLH TMYFKLLFDE NNHGQIRLSG GAELFMRRAS 751 LKKEELVVHP ANSPIANKNP DNPCKTTTLS YDVYKDKRFS EDQYELHIPI 801 AINKCPKNIF KINTEVRVLL KHDDNPYVIG IARGERNNLY IVVVDGKGNL 851 VEQYSLNEII NNFNGIRIKT DYHSLLDKKE KERFEARQNW TSINIENIKELK 901 AGYISQVVKH ICELVEKYDA VIALADLNSG FKNSRVKVEK QVYQKFEKML 951 IDKLNVMVDK KSNPCATGGA LKGYQITNKF ESFKSMSTQN GFIFYIPAWL 1001 TSKIDPSTGF VNLLKTKYTS IADSKKFIS FDRIMYVPEE DLFEFALDYK 1051 NFSRTDADYI KKWKLYSYGN RIRIFRNPCK NNVFDWEVC LTSAYKELFN 1101 KYGINYQQGD IRALLCEQSD KAFYSSFMAL MSLMLQMRNS ITGRTDVAFV 1151 ISPVKNSDGI FYDSRNYEAO ENAILPKNAD ANGAYNIARK VLWAIQGFKK 1201 AEDEKLDKVK IAISNKEWLE YAQTSVKHGS	
AsCas12a	1 MTQFEGFTNL YQVSKTLRFE LIPQGKTLKH IQEQGFIEED KARNDHYKEL 51 KPIIDRIYKT YADQCLQLVQ LDWENLSAAI DSYRKEKTEE TRNALIEEQA 101 TYRNAIHDFY IGRDNLTDI INKRHAIEYK GLFKAELFNG KVLKQLGTVT 151 TTEHENALLR SFDKFTTYFS GFYENRKNVF SAEDISTAIP HRIVQDNFVK 201 FKENCHIFTR LITAVPSLRE HFENVKKAIG IFVSTSIEEV FSFPFYNNLL 251 TQTQIDLYNQ LLGGISREAG TEKIKGLNEV LNLAIQKNDI TAHIIASLPH 301 RFIPLFKQIL SDRNTLSFIL EEFKSDEEVI QSFCKYKTLN RNENVLETAE 351 ALFNEINLSD LTHIFISHKK LETISSALCD HWDTLRNALY ERRISELTGK 401 ITKSAKEKVQ RSLKHEDINL QEIIISAAGKE LSEAFKQKTS EILSHAHAA 451 DQPLPTTLKK QEEKEILKSQ LDSLLGLYHL LDWFAVDEN EVDPEFSARL 501 TGIKLEMEPS LSFYNKARNY ATKKPYSVEK FKLNFQMPTL ASGWDVNKEK 551 NNGAILFVKN GLYYLGIMPK QKGRYKALS FEPTEKTSEGF DKMYDYDFPD 601 AAKMIPKCST QLKAVTAHFQ THTPILLSN NFIEPLEITK EIYDLNNPEK 651 EPKKFQTAYA KKTGDQKGYR EALCKWIDFT RDFSLSKYTKT TSIDLSSLRP 701 SSQYKDLGEY YAEINPLLYH ISFQRIAEKE IMDAVETGKL YLFQIYNKDF 751 AKGHHGKPNL HTLYWTGLFS PENLAKTSIK LNGQAELFYR PKSRMKRMAH 801 RLGEKMLNKK LKDQKTPIPD TLYQELYDYV NHRLSHDLSD EARALLPNVI 851 TKEVSHEIIK DRRFTSDKFF FHVPIITLNYQ AANSPSKFNQ RVNAYLKEHP 901 ETPIIGIDRG ERNLIYITVI DSTGKILEQR SLNTIQQFDY QKKLDNREKE 951 RVAARQAWSV VGTIKDLKQG YLSQVIHEIV DLMIHQAVV VLENLNFQFK 1001 SKRTGIAEKA VYQQFEKMLI DKLNLCLVLK YPAEKVGGVL NPYQLTDQFT 1051 SFAKMGTSQ FLYFVPAPYT SKIDPLTGFV DPFVWKTIKN HESRKHFLG 1101 FDFLHYDVKT GDFILHFKMN RNLSFQGRGL GFMPAWDIVF EKNETQFQDAK 1151 GTPFIAGKRI VPVIENHRFT GRYRDLYPAN ELIALLEEKG IVFRDGSNLI 1201 PKLENDSDH AIDTMVALIR SVLQMRNSNA ATGEDYINSP VRDLNGVCDF 1251 SRFQNPPEWPM DADANGAYHI ALKGQQLLNH LKESKDLKLQ NGISNQDWLA 1301 YIQELRN	38

FnCas12a	1 MSIQEFVNK YSLSKTLRFE LIPQGKTLEN IKARGLILDD EKRAKDYKKA 51 KQIIDKYHQF FIEEILSSVC ISEDLLQNY S DVYFKLKKSD DDNLQKDFKS 101 AKDTIKKQIS EYIKDSEKFK NLFNQNLIDA KKGQESDLIL WLKQSKDNGI 151 ELFKANS DIT DIDEALEIIK SFGWTTYFK GFHENRKNVY SSNDIPTSII 201 YRIVDDNLPK FLENKAKYES LKDKAPEAIN YEQIKKDLAE ELTFDIDYKT 251 SEVNQRVFSL DEVFEIANFN NYLNQSGITK FNTIIGGKFV NGENTKRKGI 301 NEYINLYSQQ INDKTLKKYK MSVLFKQILS DTESKSFVID KLEDDSDVVT 351 TMQS FYEQIA AFKTVEEKSI KETLSLLFDD LKAQKLDLSK IYFKNDKSLT 401 DLSQQVFDDY SVIGTAVLEY ITQQIAPKNL DNPSKKEQEL IAKKTEKAKY 451 LSLETIKLAL EEFNKHRDID KQCRFEEILA NFAAIPMIFD EIAQNKDNL A 501 QISIKYQNGG KKDLLQASAE DDVKAIKDLL DQTNNLLHKL KIFHISQSED 551 KANILDKDEH FYLVFEECYF ELANIVPLYN KIRNYITQKP YSDEKFKLNF 601 ENSTLANGWD KNKEPDNTAI LFIKDDKYLL GVMNKKNNKI FDDKAIKENK 651 GEGYKKIVYK LLPGANKMLP KVFFSAKSIK FYNPSEDILR IRNHSTHTKN 701 GSPQKGYEKF EFNIEDCRKF IDFYKQSISK HPEWKDFGFR FSDTQRYNSI 751 DEFYREVENQ GYKLTFENIS ESYIDSVVNQ GKLYLFQIYN KDFSAYSKGR 801 PNLHTLYWKA LFDERNLQDV VYKLNGEAE L FYRKQSIPKK ITHPAKEAIA 851 NKNKDNPKKE SVFEYDLIKD KRFTEDKFFF HCPITINFKS SGANKFNDEI 901 NLLKKEKAND VHILSIDRGE RHLAYYTLVD GKGNIKQDT FNIIGNDRMK 951 TNYHDKLAAI EKDRDSARKD WKKINNIKEM KEGYLSQVVH EIAKLVI EYN 1001 AIVVFEDLNF GFKRGRFKVE KQVYQKLEKM LIEKLNLYLV KDNEDDKTGG 1051 VLRAYQLTAP FETFKKMGKQ TGIYYVPAG FTSKICPVTG FVNQLYPKYE 1101 SVSKSQEFFS KFDKICYNLD KGYFEFSFDY KNFGDKAAKG KWTIASFGSR 1151 LINFNRSDKN HNWDTREVPY TKELEKLLKD YSIEYGHGEC IKAACGESD 1201 KKFFAKLTSV LNTILQMRNS KTGTELDYL I SPVADVNGNF FDSRQAPKNM 1251 PQDADANGAY HIGLKGLMLL GRIKNNQEGK KLNLVIKNEE YFEFVQNRNN	39
dCas12a-hA3A-BE	1 MPKKKRK VME ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVEERLDN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFISWSPCF SWGCAGEVRA FLQENTHVR L RIFAARIYDY DPLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGR LRA 201 ILQNQGN SGT ETPGTSESAT PESMSKLEKF TNCYSLSKTL RFKAI PVGKT 251 QENIDNKRL L VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLN NY 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGN EGYKS LFKKDIIETI 351 LPEFLDDKDE IALVNSFNGF TTAFTGFFDN RENMFSEEAK STSIAFR CIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQTK QKLPKFKPLY 501 KQVLS DRESL SFYGEGYTSD EEVLEVFRNT LKNKSEIFSS IKKLEKLFKN 551 FDEYSSAGIF VKNGPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRRK SFKKIGSFSL EQLOEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLFD ADFVLEKSLK KNDAVVAIMK DLLDSVKSFE NYIKAFFGEG 701 KETNRDES FY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDKD KETD YRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF 851 NLNDCHKLID FFKDSIS RYP KWSNAYDFNF SETEKYKDIA GFYREVVEEQG	40

	901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHGQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNPKKTT 1001 TLSYDVYKDK RFSEDQYELH IPIAINKCPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVVDGK GNIVEQYSLN EIINNFNIGIR IKTDYHSLLD 1101 KKEKERFEAR QNWTISIENIK ELKAGYISQV VHKICELVEK YDAVIALADL 1151 NSGFKNSRVK VEKQVYQKFE KMLIDKLNYM VDKKSNPCAT GGALKGYQIT 1201 NKFESFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNNVFDWE EVCLTSAYKE LFNKYGINYO QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGRTDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIGQ FKKADEKLD KVKIAISNKE WLEYAQTSVK 1451 HGSPKKKRKV SGGSTNLSDI IEKETGKQLV IQESILMLPE EVEEVIGNKP 1501 ESDILVHTAY DESTDENVML LTSDAPEYKP WALVIQDSNG ENKIKMLSGG 1551 SPKKKRKV	
dCas12a-hA3A-BE-W98Y	1 MPKKKRKVM ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVEERLDN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFISYSPCF SWGCAGEVRA FLQENTHURL RIFAARIYDY DPLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGRRLRA 201 ILQNGGNSGS ETPGTSESAT PESMSKLEKF TNCYSLSKTL RFKAIPVGKT 251 QENIDNKRL L VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLNNY 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGNEGYKS LFKKDIIETI 351 LPEFLDDKDE IALVNSFNGF TTAFTGFFDN RENMFSEEAK STSIAFRCIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY 501 KQVLSRESL SFYGEYTS D EEVLEVFRT LNKNSEIFSS IKKLEKLFKN 551 FDEYSSAGIF VKNGPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRR SFKKIGSFSL EQLQEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLF D ADFVLEKSLK KNDAVVAIMK DLLDSVKSEF NYIKAFFGEG 701 KETNRDESFY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDK D KETDYRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF 851 NLNDCHKLID FFKDSISRYP KWSNAYDFNF SETEKYKDIA GFYREVVEEQG 901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHGQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNPKKTT 1001 TLSYDVYKDK RFSEDQYELH IPIAINKCPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVVDGK GNIVEQYSLN EIINNFNIGIR IKTDYHSLLD 1101 KKEKERFEAR QNWTISIENIK ELKAGYISQV VHKICELVEK YDAVIALADL 1151 NSGFKNSRVK VEKQVYQKFE KMLIDKLNYM VDKKSNPCAT GGALKGYQIT 1201 NKFESFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNNVFDWE EVCLTSAYKE LFNKYGINYO QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGRTDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIGQ FKKADEKLD KVKIAISNKE WLEYAQTSVK 1451 HGSPKKKRKV SGGSTNLSDI IEKETGKQLV IQESILMLPE EVEEVIGNKP 1501 ESDILVHTAY DESTDENVML LTSDAPEYKP WALVIQDSNG ENKIKMLSGG	41

	1551 SPKKKRKV	
dCas12a-hA3A- BE-W104A	1 MPKKKRKVME ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVEERLDN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFIWSWPCF SAGCAGEVRA FLQENTHVR L RIFAARIYDY DPLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGRLRA 201 ILQNQGNSGS ETPGTSESAT PESMSKLEKF TNCYLSKTL RFAKIPVGKT 251 QENIDNKRL L VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLNYY 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGN EGYKS LFKKDIIETI 351 LPEFLDDKDE IALVNSFN GF TTAFTGFFDN RENMFSEEAK STSIAFR CIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY 501 KQVLSDRESL SFYGE GYTSD EEVLEVFRNT LKNKSEIFSS IKKLEKLFKN 551 FDEYSSAGIF VKNGPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRRK SFKKIGSFSL EQLQEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLF D ADFVLEKSLK KNDAVVAIMK DLLDSVKSFE NYIKAFFGEG 701 KETNRDESFY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDKD KETD YRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF 851 NLNDCHKLID FFKDSIS RYP KWSNAYDFNF SETEKYKDIA GFYREV EEQG 901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNP KKT 1001 TLSYDVYKDK RFSEDQYELH IPIAINKCPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVVDGK GNIVEQYSLN EIINN FN GIR IKTDYHSLLD 1101 KKEKERFEAR QNWT SIENIK ELKAGYISQV VHKICELVEK YDAVIALADL 1151 NSGFKNSRVK VEKQVYQKFE KMLIDKLN YM VDKKSNPCAT GGALKGYQIT 1201 NKFESFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNNVFDWE EVCLTSAYKE LFNKYGIN YQ QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGR TDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIQ FKKAEDEKLD KVKIAISNKE WLEYAQTSVK 1451 HGSPKKKRKV SGGSTNLS DI IEKETGKQLV IQESILMLPE EVEEVIGNKP 1501 ESDILVHTAY DESTDENV ML LTSDAPEYKP WALVIQDSNG ENKIKMLSGG 1551 SPKKKRKV	42
dCas12a-hA3A- BE-P134Y	1 MPKKKRKVME ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVEERLDN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFIWSWPCF SWGCAGEVRA FLQENTHVR L RIFAARIYDY DYLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGRLRA 201 ILQNQGNSGS ETPGTSESAT PESMSKLEKF TNCYLSKTL RFAKIPVGKT 251 QENIDNKRL L VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLNYY 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGN EGYKS LFKKDIIETI 351 LPEFLDDKDE IALVNSFN GF TTAFTGFFDN RENMFSEEAK STSIAFR CIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY 501 KQVLSDRESL SFYGE GYTSD EEVLEVFRNT LKNKSEIFSS IKKLEKLFKN	43

	551 FDEYSSAGIF VKNGPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRRK SFKKIGSFSL EQLQEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLFD ADFVLEKSLK KNDVAVAIMK DLLDSVKSEF NYIKAFFGEG 701 KETNRDESFY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDKD KETDYRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF 851 NLNDCHKLID FFKDSISRYP KWSNAYDFNF SETEKYKDIA GFYREVEEQG 901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHGQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNPCKTT 1001 TLSYDVYKDK RFSEDQYELH IPIAINKCPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVVDGK GNIVEQYSLN EIINNFNIGIR IKTDYHSLLD 1101 KKEKERFEAR QNWTISIENIK ELKAGYISQV VHKEICELVEK YDAVIALADL 1151 NSGFKNSRVK VEKQVYQKFE KMLIDKLNLM VDKKSNPCAT GGALKGYQIT 1201 NKFESFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNNVDFWE EVCLTSAYKE LFNKYGINYO QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGRTDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIGQ FKKAEDEKLD KVKIAISNKE WLEYAQTSVK 1451 HGSPKKRKV SGGSTNLSDI IEKETGKQLV IQESILMLPE EVEEVIGNKP 1501 ESDILVHTAY DESTDENVM L TS DAPEYKP WALVIQDSNG ENKIKMLSGG 1551 SPKKRKV	
dCas12a-hA3A- BE-W98Y- W104A	1 MPKKRKVM ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVEERLDN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFISSYSPCF SAGCAGEVRA FLQENTHVR LIFAARIYDY DPLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGLRLA 201 ILQNGNSGS ETPGTSESAT PESMSKLEKF TNCYSLSKTL RFKAIPVGKT 251 QENIDNKRL VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLNNY 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGNEGYKS LFKKDIETI 351 LPEFLDDKDE IALVNSFNGF TTAFTGFFDN RENMFSEEAK STSIAFRGIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY 501 KQVLSDRESL SFYGEYTS D EEVLEVFRT LNKNSEIFSS IKKLEKLFKN 551 FDEYSSAGIF VKNGPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRRK SFKKIGSFSL EQLQEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLFD ADFVLEKSLK KNDVAVAIMK DLLDSVKSEF NYIKAFFGEG 701 KETNRDESFY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDKD KETDYRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF 851 NLNDCHKLID FFKDSISRYP KWSNAYDFNF SETEKYKDIA GFYREVEEQG 901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHGQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNPCKTT 1001 TLSYDVYKDK RFSEDQYELH IPIAINKCPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVVDGK GNIVEQYSLN EIINNFNIGIR IKTDYHSLLD 1101 KKEKERFEAR QNWTISIENIK ELKAGYISQV VHKEICELVEK YDAVIALADL 1151 NSGFKNSRVK VEKQVYQKFE KMLIDKLNLM VDKKSNPCAT GGALKGYQIT	44

	1201 NKFESFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNNVFDWE EVCLTSAYKE LFNKYGINYO QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGRTDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIGQ FKKADEKLD KVKIAISNKE WLEYAQTSVK 1451 HGSPKKRKV SGGSTNLSDI IEKETGKQLV IQESILMLPE EVEEVIGNKP 1501 ESDILVHTAY DESTDENVML LTSDAPEYKP WALVIQDSNG ENKIKMLSGG 1551 SPKKRKV	
dCas12a-hA3A- BE-W98Y- P134Y	1 MPKKRKVME ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVRDLN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFISYSPCF SWGCAGEVRA FLQENTHVR LIFAARIYDY DYLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGLRLA 201 ILQNGNSGS ETPGTSESAT PESMSKLEKF TNCYLSKTL RFKAIPVGKT 251 QENIDNKRL VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLNNY 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGNEGYKS LFKKDIETI 351 LPEFLDDKDE IALVNSFNGF TTAFTGFFDN RENMFSEEAK STSIAFRGIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFEFEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY 501 KQVLSDRESL SFYGEYTS D EEVLEVFRT LKNSEIFSS IKKLEKLFKN 551 FDEYSSAGIF VKNGPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRK SFKKIGSFS EQLOEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLF D ADFVLEKSLK KNDAVVAIMK DLLDSVKSFE NYIKAFFGEG 701 KETNRDESFY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDK KETDYRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF 851 NLNDCHKLID FFKDSISRYP KWSNAYDFNF SETEKYKDIA GFYREVVEEQG 901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNPKKT 1001 TLSYDVYKDK RFSEDQYELH IPIAINCKPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVDGK GNIVEQYSLN EIINNFGIR IKTDYHSLD 1101 KKEKERFEAR QNWTSIENIK ELKAGYISQV VHKICELVEK YDAVIALADL 1151 NSGFKNSRVK VEKQVYQKFE KMLIDKLN Y VDKKSNPCAT GGALKGYQIT 1201 NKFESFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNNVFDWE EVCLTSAYKE LFNKYGINYO QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGRTDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIGQ FKKADEKLD KVKIAISNKE WLEYAQTSVK 1451 HGSPKKRKV SGGSTNLSDI IEKETGKQLV IQESILMLPE EVEEVIGNKP 1501 ESDILVHTAY DESTDENVML LTSDAPEYKP WALVIQDSNG ENKIKMLSGG 1551 SPKKRKV	45
dCas12a-hA3A- BE-W104A- P134Y	1 MPKKRKVME ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVRDLN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFISWSPCF SAGCAGEVRA FLQENTHVR LIFAARIYDY DYLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGLRLA	46

	201 ILQNQGNSGS ETPGTSESAT PESMSKLEKF TNCYLSKTL RFKAIPVGKT 251 QENIDNKRL L VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLN 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGNEGYKS LFKKDIIETI 351 LPEFLDDKDE IALVNSFNGF TTAFTGFFDN RENMFSEEAK STSIAFRGIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY 501 KQVLSDRESL SFYGEGYTS D EEVLEVFRTN LKNSEIFSS IKKLEKLFKN 551 FDEYSSAGIF VKN GPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRRK SFKKIGSFSL EQLQEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLF D ADFVLEKSLK KND AVVAIMK DLLDSVK SFE NYIKAFFGEG 701 KETNRDES FY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDKD KETDYRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF 851 NLNDCHKLID FFKDSIS RYP KWSNAYDFNF SETEKYKDIA GFYREVVEEQG 901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHGQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNP KKT 1001 TLSYDVYKDK RFSEDQYELH IPIAINKCPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVVDGK GNIVEQYSLN EIINN FN GIR IKTDYHSLLD 1101 KKEKERFEAR QNWT SIENIK ELKAGYISQV VHKICELVEK YDAVIALADL 1151 NSGFKNSRVK VEKQVYQKFE KMLIDKLN YM VDKKSNPCAT GGALKGYQIT 1201 NKFESFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNVFDWE EVCLTSAYKE LFNKYGIN YQ QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGR TDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIQ FKKAEDEKLD KVKIAISNKE WLEYAQT SVK 1451 HGSPKKRKV SGGSTNLS DI IEKETGKQLV IQESILMLPE EVEEVIGNKP 1501 ESDILVHTAY DESTDEN VML LTSDAPEYKP WALVIQDSNG ENKIKMLSGG 1551 SPKKKRKV	
dCas12a-hA3A- BE-W98Y- W104A-Y130F	1 MPKKKRKVME ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVE RL DN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFISYSPCF SAGCAGEVRA FLQENTHVR L RIFAARIFDY DPLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGR LRA 201 ILQNQGNSGS ETPGTSESAT PESMSKLEKF TNCYLSKTL RFKAIPVGKT 251 QENIDNKRL L VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLN 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGNEGYKS LFKKDIIETI 351 LPEFLDDKDE IALVNSFNGF TTAFTGFFDN RENMFSEEAK STSIAFRGIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY 501 KQVLSDRESL SFYGEGYTS D EEVLEVFRTN LKNSEIFSS IKKLEKLFKN 551 FDEYSSAGIF VKN GPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRRK SFKKIGSFSL EQLQEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLF D ADFVLEKSLK KND AVVAIMK DLLDSVK SFE NYIKAFFGEG 701 KETNRDES FY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDKD KETDYRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF	47

	851 NLNDCHKLID FFKDSISRYP KWSNAYDFNF SETEKYKDIA GFYREVVEEQG 901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHGQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNPKKTT 1001 TLSYDVYKDK RFSEDQYELH IPIAINKCPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVVDGK GNIVEQYSLN EIINNFNIGIR IKTDYHSLLD 1101 KKEKERFEAR QNWTISIENIK ELKAGYISQV VHKICELVEK YDAVIALADL 1151 NSGFKNSRVK VEKQVYQKFE KMLIDKLNLM VDKKSNPCAT GGALKGYQIT 1201 NKFESFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNNVFDWE EVCLTSAYKE LFNKYGINYO QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGRTDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIQ FKKAEDEKLD KVKIAISNKE WLEYAQTSVK 1451 HGSPKKRKV SGGSTNLSDI IEKETGKQLV IQESILMLPE EVEEVIGNKP 1501 ESDILVHTAY DESTDENVML LTSDAPEYKP WALVIQDSNG ENKIKMLSGG 1551 SPKKKRKV	
dCas12a-hA3A- BE-W98Y- W104A-Y132D	1 MPKKKRKVM ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVEERLDN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFISYSPCF SAGCAGEVRA FLQENTHVR LIFAARIYDD DPLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGRLRA 201 ILQNQGNSSG ETPGTSESAT PESMSKLEKF TNCYSLSKTL RFKAIPVGKT 251 QENIDNKRL VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLNNY 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGNEGYKS LFKKDIETI 351 LPEFLDDKDE IALVNSFNGF TTAFTGFFDN RENMFSEEAK STSIAFRGIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY 501 KQVLSDRESL SFYGEGYTSD EEVLEVFRTN LKNSEIFSS IKKLEKLFKN 551 FDEYSSAGIF VKNGPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRK SFKKIGSFSL EQLOEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLF DAFVLEKSLK KNDVAVAIMK DLLDSVKSFE NYIKAFFGEG 701 KETNRDESFY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDK KETDYRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF 851 NLNDCHKLID FFKDSISRYP KWSNAYDFNF SETEKYKDIA GFYREVVEEQG 901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHGQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNPKKTT 1001 TLSYDVYKDK RFSEDQYELH IPIAINKCPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVVDGK GNIVEQYSLN EIINNFNIGIR IKTDYHSLLD 1101 KKEKERFEAR QNWTISIENIK ELKAGYISQV VHKICELVEK YDAVIALADL 1151 NSGFKNSRVK VEKQVYQKFE KMLIDKLNLM VDKKSNPCAT GGALKGYQIT 1201 NKFESFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNNVFDWE EVCLTSAYKE LFNKYGINYO QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGRTDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIQ FKKAEDEKLD KVKIAISNKE WLEYAQTSVK 1451 HGSPKKRKV SGGSTNLSDI IEKETGKQLV IQESILMLPE EVEEVIGNKP	48

	1501 ESDILVHTAY DESTDENVML LTSDAPEYKP WALVIQDSNG ENKIKMLSGG 1551 SPKKKRKV	
dCas12a-hA3A- BE-W104A- Y130F-P134Y	1 MPKKKRKVME ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVEERLDN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFIWSWPCF SAGCAGEVRA FLQENTHVRL RIFAARIFDY DYLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGLRLA 201 ILQNQGNSSGS ETPGTSESAT PESMSKLEKF TNCYSLSKTL RFKAIPVGKT 251 QENIDNKRL L VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLNYY 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGNENGYKS LFKKDIETI 351 LPEFLDDKDE IALVNSFNGF TTAFTGFFDN RENMFSEEAK STSIAFRCIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFEFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY 501 KQVLSDRESL SFYGEYTS D EEVLEVFRTN LKNSEIFSS IKKLEKLFKN 551 FDEYSSAGIF VKNGPAISTI SKDIFGEWNV IRDKWNAEYD DIHLKKKAVV 601 TEKYEDDRRK SFKKIGSFSL EQLOEYADAD LSVVEKLKEI IIQKVDEIYK 651 VYGSSEKLFD ADFVLEKSLK KNDAVVAIMK DLLDSVKSFE NYIKAFFGEG 701 KETNRDESFY GDFVLAYDIL LKVDHIYDAI RNYVTQKPYS KDKFKLYFQN 751 PQFMGGWDKD KETDYRATIL RYGSKYLLAI MDKKYAKCLQ KIDKDDVNGN 801 YEKINYKLLP GPNKMLPKVF FSKKWMAYYN PSEDIQKIYK NGTFKKGDMF 851 NLNDCHKLID FFKDSISRYK KWSNAYDFNF SETEKYKDIA GFYREVEEQG 901 YKVSFESASK KEVDKLVEEG KLYMFQIYNK DFSDKSHGTP NLHTMYFKLL 951 FDENNHGQIR LSGGAELFMR RASLKKEELV VHPANSPIAN KNPDNPKKTT 1001 TLSYDVYKDK RFSEDQYELH IPIAINKCPK NIFKINTEVR VLLKHDDNPY 1051 VIGIARGERN LLYIVVVDGK GNIVEQYSLN EIINNENFNGIR IKTDYHSLLD 1101 KKEKERFEAR QNWTISIENIK ELKAGYISQV VHKICELVEK YDAVIALADL 1151 NSGFKNRSVK VEKQVYQKFE KMLIDKLYNM VDKKSNPCAT GGALKGYQIT 1201 NKFESEFKSMS TQNGFIFYIP AWLTSKIDPS TGFVNLLKTK YTSIADSKKF 1251 ISSFDRIMYV PEEDLFEFAL DYKNFSRTDA DYIKKWKLYS YGNRIRIFRN 1301 PKKNNVDFWE EVCLTSAYKE LFNKYGINYO QGDIRALLCE QSDKAFYSSF 1351 MALMSLMLQM RNSITGRTDV AFLISPVKNS DGIFYDSRNY EAQENAILPK 1401 NADANGAYNI ARKVLWAIQ FKKAEDEKLD KVKAISNKE WLEYAQTSVK 1451 HGSPPKKRKV SGGSTNLSDI IEKETGKQLV IQESILMLPE EVEEVIGNKP 1501 ESDILVHTAY DESTDENVML LTSDAPEYKP WALVIQDSNG ENKIKMLSGG 1551 SPKKKRKV	49
dCas12a-hA3A- BE-W104A- Y132D-P134Y	1 MPKKKRKVME ASPASGPRHL MDPHIFTSNF NNGIGRHKTY LCYEVEERLDN 51 GTSVKMDQHR GFLHNQAKNL LCGFYGRHAE LRFLDLVPSL QLDPAQIYRV 101 TWFIWSWPCF SAGCAGEVRA FLQENTHVRL RIFAARIYDD DYLYKEALQM 151 LRDAGAQVSI MTYDEFKHCW DTFVDHQGCP FQPWDGLDEH SQALSGLRLA 201 ILQNQGNSSGS ETPGTSESAT PESMSKLEKF TNCYSLSKTL RFKAIPVGKT 251 QENIDNKRL L VEDEKRAEDY KGVKKLLDRY YLSFINDVLH SIKLKNLNYY 301 ISLFRKKTRT EKENKELENL EINLRKEIAK AFKGNENGYKS LFKKDIETI 351 LPEFLDDKDE IALVNSFNGF TTAFTGFFDN RENMFSEEAK STSIAFRCIN 401 ENLTRYISNM DIFEKVDAIF DKHEVQEIKE KILNSDYDVE DFEFEGEFFNF 451 VLTQEGIDVY NAIIGGFVTE SGEKIKGLNE YINLYNQKTK QKLPKFKPLY	50

501	KQVLSDRESL	SFYGEGYTS	EEVLEVFRNT	LNKNSEIFSS	IKKLEKLFKN
551	FDEYSSAGIF	VKNGPAISTI	SKDIFGEWNV	IRDKWNAEYD	DIHLKKKAVV
601	TEKYEDDRRK	SFKKIGSFSL	EQLQEYADAD	LSVVEKLKEI	IIQKVDEIYK
651	VYGSSEKLF	ADFLVLEKSLK	KNDVAVAIMK	DLSDSVKSFE	NYIKAFFGEG
701	KETNRDESFY	GDFVLAYDIL	LKVDHIYDAI	RNYVTQKPYS	KDKFKLYFQN
751	PQFMGGWDKD	KETDYPATIL	RYGSKYYLAI	MDKKYAKCLQ	KIDKDDVNGN
801	YEKINYKLLP	GPNKMLPKVF	FSKKWMAYYN	PSEDIQKIYK	NGTFKKGDMF
851	NLNDCHKLID	FFKDSISRYP	KWSNAYDFNF	SETEKYKDIA	GFYREVVEEQG
901	YKVSFESASK	KEVDKLVEEG	KLYMFQIYNK	DFSDKSHGTP	NLHTMYFKLL
951	FDENNHGQIR	LSGGAELFMR	RASLKKEELV	VHPANSPIAN	KNPDNPCKTT
1001	TLSYDVYKDK	RFSEDQYELH	IPIAINKCPK	NIFKINTEVR	VLLKHDDNPY
1051	VIGIARGERN	LLYIVVVDGK	GNIVEQYSLN	EIINNFGNIR	IKTDYHSLLD
1101	KKEKERFEAR	QNWTSIENIK	ELKAGYISQV	VHKICELVEK	YDAVIALADL
1151	NSGFKNSRVK	VEKQVYQKFE	KMLIDKLNLM	VDKKSNPCAT	GGALKGYQIT
1201	NKFESFKSMS	TQNGFIFYIP	AWLTSKIDPS	TGFVNLLKTK	YTSIADSKKF
1251	ISSFDRIMYV	PEEDLFEFAL	DYKNFSRTDA	DYIKKWKLYS	YGNRIRIFRN
1301	PKKNNVFDWE	EVCLTSAYKE	LFNKYGINYQ	QGDIRALLCE	QSDKAFYSSF
1351	MALMSLMLQM	RNSITGRTDV	AFLISPVKNS	DGIFYDSRNY	EAQENAILPK
1401	NADANGAYNI	ARKVLWAIQ	FKKADEKLD	KVKIAISNKE	WLEYAQTSVK
1451	HGSPKKRKRV	SGGSTNLSDI	IEKETGKQLV	IQESILMLPE	EVEEVIGNKP
1501	ESDILVHTAY	DESTDENVML	LTSDAPEYKP	WALVIQDSNG	ENKIKMLSGG
1551	SPKKKRKV				

The present disclosure also provides isolated polynucleotides or nucleic acid molecules (e.g., SEQ ID NO:21) encoding the fusion proteins, variants or derivatives thereof of the disclosure. Methods of making fusion proteins are well known in the art and described herein.

Compositions and Methods

The present disclosure also provides compositions and methods. Such compositions comprise an effective amount of a fusion protein, and an acceptable carrier. In some embodiments, the composition further includes a guide RNA that has a desired complementarity to a target DNA. Such a composition can be used for base editing in a sample.

The fusion proteins and the compositions can be used for base editing. In one embodiment, a method for editing a target polynucleotide is provided, comprising contacting to the target polynucleotide a fusion protein of the present disclosure and a guide RNA having at least partial sequence complementarity to the target polynucleotide, wherein the editing comprises deamination of a cytosine (C) in the target polynucleotide.

It is shown that the presently disclosed fusion proteins can edit cytosine at any location and in any context, such as in CpC, ApC, GpC, TpC, CpA, CpG, CpC, CpT. It is surprising and unexpected, however, that these fusion proteins can edit C in a GpC dinucleotide context, and even when the C is methylated.

The contacting between the fusion protein (and the guide RNA) and the target polynucleotide can be *in vitro*, in particular in a cell culture. When the contacting is *ex vivo*, or *in vivo*, the fusion proteins can exhibit clinical/therapeutic significance.

EXAMPLES

Example 1: Base Editors

Human apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A, hA3A; SEQ ID NO:1) was included in an expression vector that further included a Cas9 nickase (SEQ ID NO:11) and a uracil-DNA-glycosylase inhibitor [Bacillus phage AR9] (SEQ ID NO:12). The Cas9 nickase contained a Asp10Ala mutation that inactivated its double strand nuclease activity, while allowing it to introduce a nick on one of the strands.

The fusion vector, hA3A-nCas9-UGI (hA3A-BE, SEQ ID NO:21), and a sgRNA expression vector were co-transfected into eukaryotic cells (**FIG. 1A**) to perform C-to-T base editing at sgRNA target site in the genome. After PCR amplification of the target genomic DNA, the C-to-T base editing efficiency at targeted site in genome were determined through Sanger DNA sequencing. As illustrated in two sgRNA target sites (sgFANCF-M-L6 and sgSITE4), efficient C-to-T base editing was executed on C of GpC through co-expressing hA3A-BE and sgRNA, as compared to co-expressing BE3 (APOBEC1-nCas9-UGI) and sgRNA (**FIG 1B**, dashed box).

Next, mutations Y130F (SEQ ID NO:2) and Y132D (SEQ ID NO:3) were individually introduced into the hA3A gene in the construct, thereby generating the base editor hA3A-BE-Y130F or hA3A-BE-Y132D (**FIG. 2A**). The Y130F and Y132D mutations in hA3A-BE narrowed the window of base editing, and further improved the editing precision of hA3A-BE (**FIG. 2B**).

Furthermore, the mutations W104A (SEQ ID NO:4) and D131Y (SEQ ID NO:5) were individually introduced into the hA3A gene of hA3A-BE, thereby generating the base editor hA3A-BE-W104A or hA3A-BE-D131Y (**FIG. 3A**). Both hA3A-BE-W104A and hA3A-BE-

D131Y increased the efficiency of desired C to T base substitutions (**FIG. 3B**), achieving even higher efficiency of base editing as compared to hA3A-BE.

In a further experiment, three amino acid changes (Y130F-D131E-Y132D, SEQ ID NO:22 or Y130F-D131Y-Y132D, SEQ ID NO:23) of human APOBEC3A (hA3A) in hA3A-BE3 (**FIG. 4A**) were tested and it was found that these two base editors (hA3A-BE-Y130F-D131E-Y132D and hA3A-BE-Y130F-D131Y-Y132D) have more narrowed editing windows (position 4-6 in target region) and therefore higher editing precision (**FIG. 4B**).

Example 2: Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion

Base editors (BEs) enable the generation of targeted single-nucleotide mutations, but currently used rat APOBEC1-based BEs are relatively inefficient in editing cytosines in highly-methylated regions or in GpC contexts. By screening a variety of APOBEC/AID deaminases, this example shows that human APOBEC3A-conjugated BEs and versions engineered to have narrower editing windows can mediate efficient C-to-T base editing in regions with high methylation levels and GpC dinucleotide content.

Base editors (BEs), which combine a cytidine deaminase with Cas9 or Cpf1, have been successfully applied to perform targeted base editing, including C-to-T. Numerous human diseases have been reported to be driven by point mutations in genomic DNAs. With recently developed BEs, these disease-related point mutations can be potentially corrected, providing new therapeutic options. By analyzing disease-related T-to-C mutations that can be theoretically reverted to thymines by BEs, the example found that ~ 43% of them are on cytosines in the context of CpG dinucleotides (**FIG. 5a**). It is well known that C of CpG is usually methylated in mammalian cells, and methylation of C strongly suppresses cytidine deamination catalyzed by some APOBEC/AID deaminases. This example shows that CpG dinucleotide methylation hinders the C-to-T base editing by current BEs and has successfully developed BEs for efficient C-to-T base editing in highly methylated regions.

Methods and Materials

Plasmid construction

Primer sets (hA3A_PCR_F/hA3A_PCR_R) were used to amplify the fragment Human_APOBEC3A with template pUC57-Human_APOBEC3A (synthesized by Genscript). Then the fragment Human_APOBEC3A was cloned into the SacI and SmaI linearized pCMV-BE3 (addgene, 73021) with plasmid recombination kit Clone Express® (Vazyme, C112-02) to generate the hA3A-BE3 expression vector pCMV-hAPOBEC3A-XTEN-D10A-SGGS-UGI-SGGS-NLS. hA3B-BE3, hA3C-BE3, hA3D-BE3, hA3F-BE3, hA3G-BE3, hA3H-BE3, hAID-BE3, hA1-BE3, mA3-BE3, mAID-BE3, mA1-BE3, cAICDA-BE3, expression vectors were constructed with the same strategy. The pmCDA1 expression vector pcDNA3.1_pCMV-nCas-PmCDA1-ugi pH1-gRNA(HPRT) was purchased from Addgene (79620).

Primer sets (SupF_PCR_F/SupF_PCR_R) were used to amplify the fragment SupF with template shuttle vector pSP189. Then the fragment SupF was cloned into pEASY-ZERO-BLUNT (TransGen Biotech, CB501) to generate the vector pEASY-SupF-ZERO-BLUNT.

Oligonucleotides SupF_sg1_FOR/SupF_sg1_REV and SupF_sg2_FOR/SupF_sg2_REV were annealed and ligated into BsaI linearized pGL3-U6-sgRNA-PGK-puromycin (addgene, 51133) to generate the sgRNA expression vectors psgSupF-1 and psgSupF-2 that target the *SupF* gene in pEASY-SupF-ZERO-BLUNT.

Two primer sets (hA3A_PCR_F/hA3A_Y130F_PCR_R) (hA3A_Y130F_PCR_F/hA3A_PCR_R) were used to amplify the Y130F-containing fragment hA3A-Y130F. Then the fragment was cloned into the ApaI and SmaI linearized hA3A-BE3 expression vector to generate the hA3A-BE3-Y130F expression vector pCMV-hAPOBEC3A_Y130F-XTEN-D10A-SGGS-UGI-SGGS-NLS. hA3A-BE3-D131Y, hA3A-BE3-Y132D, hA3A-BE3-C101S and hA3A-BE3-C106S expression vectors were constructed with the same strategy.

Primer sets (hA3A_PCR_F/hA3A_PCR_R) were used to amplify the fragment Human_APOBEC3A_Y130F with template hA3A-BE3-Y130F. Then the fragment Human_APOBEC3A_Y130F was cloned into the SacI and SmaI linearized pCMV-eBE-S3¹⁹ to generate the hA3A-eBE-Y130F expression vector pCMV-hAPOBEC3A_Y130F-XTEN-D10A-SGGS-UGI-SGGS-NLS-T2A-UGI-NLS-P2A-UGI-NLS-T2A-UGI-NLS. hA3A-eBE-Y132D expression vector was constructed by the similar way.

Oligonucleotides hEMX1_FOR/hEMX1_REV were annealed and ligated into BsaI linearized pGL3-U6-sgRNA-PGK-puromycin to generate sgEMX1 expression vector psgEMX1. Other sgRNA expression vectors were constructed with the same strategy.

Antibodies

Antibodies were purchased from the following sources: against alpha-tubulin (T6199) - Sigma; against Cas9 (ab204448) - Abcam.

Immunoblotting analysis

Protein samples were incubated at 95 °C for 20 min, separated by SDS-PAGE in sample loading buffer and proteins were transferred to nitrocellulose membranes (Thermo Fisher Scientific). After blocking with TBST (25 mM Tris pH 8.0, 150 mM NaCl, and 0.1% Tween 20) containing 5% (w/v) nonfat dry milk for 2h, the membrane was reacted overnight with indicated primary antibody. After extensive washing, the membranes were reacted with HRP-conjugated secondary antibodies for 1h. Reactive bands were developed in ECL (Thermo Fisher Scientific) and detected with Amersham Imager 600.

Cell culture and transfection

HEK293T cells from ATCC were maintained in DMEM (10566, Gibco/Thermo Fisher Scientific) + 10% FBS (16000-044, Gibco/Thermo Fisher Scientific) and regularly tested to exclude mycoplasma contamination.

The dCas9-Suntag-TetCD system was used to induce targeted demethylation of the genomic regions with natively high levels of methylation, *e.g.*, FANCF, MAGEA1 and MSSK1 regions. The dCas9-DNMT3a-DNMT3l system was used to induce targeted methylation of the genomic regions with natively low levels of methylation, *e.g.*, VEGFA and PDL1 regions. HEK293T cells were transfected by using LIPOFECTAMINE 2000 (Life, Invitrogen) with 3 µg pCAG-scFvGCN4sfGFPTET1CD (synthesized by Genscript) and 1 µg sgRNA expression vector or with 3µg dCas9-DNMT3a-DNMT3l (synthesized by Genscript) and 1 µg sgRNA expression vector. Blasticidin (10 µg/ml, Sigma, 15205) and puromycin (1 µg/ml, Merck, 540411) were added 24 h after transfection. One week later, a portion of cells were collected to determine DNA methylation level and others were stored in liquid nitrogen for base editing. The sgRNAs used to induce genomic DNA methylation/demethylation are the ones used to induce base editing.

For base editing in genomic DNA, HEK293T cells were seeded in a 24-well plate at a density of 1.6×10^5 per well and transfected with 200 μ l serum-free Opti-MEM that contained 5.04 μ l LIPOFECTAMINE LTX (Life, Invitrogen), 1.68 μ l LIPOFECTAMINE plus (Life, Invitrogen), 1 μ g BE3 expression vector (or hA3A-BE3, hA3A-BE3-Y130F, hA3A-BE3-D131Y, hA3A-BE3-Y132D, hA3A-BE3-C101S, hA3A-BE3-C106S, hA3A-eBE-Y130F, hA3A-eBE-Y132D expression vector) and 0.68 μ g sgRNA expression vector. After 72 hr, the genomic DNA was extracted from the cells with QuickExtractTM DNA Extraction Solution (QE09050, Epicentre) or the cells were lysed in 2 $\times$ SDS loading buffer for western blot.

For base editing in plasmid vector, 293T cells were seeded in a 6-well plate at a density of 3×10^5 per well and transfected with 500 μ l serum-free Opti-MEM that contained 4 μ l LIPOFECTAMINE LTX (Life, Invitrogen), 2 μ l LIPOFECTAMINE plus (Life, Invitrogen), 1 μ g BE3 expression vector (or hA3A-BE3, hA3B-BE3, hA3C-BE3, hA3D-BE3, hA3F-BE3, hA3G-BE3, hA3H-BE3, hAID-BE3, hA1-BE3, mA3-BE3, mAID-BE3, mA1-BE3, cAICDA-BE3 or pmCDA1 expression vector) and 0.5 μ g sgRNA expression vector. After 24 hr, these cells were transfected with 500 μ l serum-free Opti-MEM that contained 4 μ l LIPOFECTAMINE LTX, 2 μ l LIPOFECTAMINE plus and 1.5 μ g un-methylated (or methylated) pEASY-SupF-ZERO-BLUNT. After 48 hr, the plasmids were extracted from the cells with TIANprep Mini Plasmid Kit (DP103-A, TIANGEN) or the cells were lysed in 2 $\times$ SDS loading buffer for western blot.

Bisulfite sequencing analysis

Genomic DNA was isolated and treated with bisulfite according to the instruction of EZ DNA methylation-direct Kit (Zymo Research, D5021). The bisulfite-treated DNA was PCR-amplified with TaqTM Hot Start Version (Takara, R007B). The PCR products were ligated into T-Vector pMDTM19 (Takara, 3271). Eight clones were picked out and sequenced by Sanger sequencing (Genewiz). The primers used for bisulfite PCR were listed in Supplementary Table 2.

Plasmid DNA methylation

For *in vitro* methylation, 1 μ l CpG methyltransferase (M.SssI, Life, EM0821) was used to methylate 2 μ l plasmid DNA in a 20- μ l reaction. After *in vitro* methylation, pEASY-SupF-ZERO-BLUNT was restricted with BstUI (NEB, R0518S) to determine the methylation level.

Blue/white colony screening

The plasmids extracted from transfected cells were transformed into *E. coli* strain MBM7070 (*lacZ*^{uag-amber}), which were grown on LB plates containing 50 µg/ml kanamycin, 1 mM IPTG and 0.03% Blue-gal (Life, Invitrogen) at 37°C overnight and then at room temperature for another day (for maximal color development). The cumulative base editing frequency is calculated by dividing the number of white colonies with the number of total colonies.

DNA library preparation and sequencing

Target genomic sites were PCR amplified by high-fidelity DNA polymerase PrimeSTAR HS (Clontech) with primers flanking each examined sgRNA target site. The PCR primers used to amplify target genomic sequences were listed in Supplementary Table 2. Indexed DNA libraries were prepared by using the TruSeq ChIP Sample Preparation Kit (Illumina) with some minor modifications. Briefly, the PCR products were fragmented by Covaris S220 and then amplified by using the TruSeq ChIP Sample Preparation Kit (Illumina). After being quantitated with Qubit High-Sensitivity DNA kit (Life, Invitrogen), PCR products with different tags were pooled together for deep sequencing by using the Illumina NextSeq 500 (2×150) or HiSeq X Ten (2×150) at CAS-MPG Partner Institute for Computational Biology Omics Core, Shanghai, China. Raw read qualities were evaluated by FastQC (available at www.bioinformatics.babraham.ac.uk/projects/fastqc/). For paired-end sequencing, only R1 reads were used. Adaptor sequences and read sequences on both ends with Phred quality score lower than 28 were trimmed. Trimmed reads were then mapped with the BWA-MEM algorithm (BWA v0.7.9a) to target sequences. After being piled up with samtools (v0.1.18), indels and base substitutions were further calculated.

Indel frequency calculation

Indels were estimated in the aligned regions spanning from upstream eight nucleotides of the target site to downstream 19 nucleotides of PAM sites (50 bp). Indel frequencies were subsequently calculated by dividing reads containing at least one inserted and/or deleted nucleotide by all the mapped reads at the same region.

Base substitution calculation

Base substitutions were selected at each position of the examined sgRNA target sites that mapped with at least 1,000 independent reads, and obvious base substitutions were only observed at the

targeted base editing sites. Base substitution frequencies were calculated by dividing base substitution reads by total reads.

Calculation of BE-targetable genetic variants

The single nucleotide variants (SNVs) from NCBI ClinVar database were overlapped with the pathogenic human allele sequence from NCBI dbSNP database to calculate the pathogenic T-to-C and A-to-G mutations. In 3,089 pathogenic T-to-C or A-to-G mutations, 2,499 are potentially editable by SpCas9-BE3, SaCas9-BE3, dLbCpf1-BE or xCas9-BE3 with nearby PAM sequences. These 2,499 BE-targetable SNVs are further sub-classified according to their 3' adjacent base preferences, *i.e.*, CpA, CpC, CpG and CpT (**FIG. 5a**).

Statistical analysis

P values were calculated from one-tailed Student's *t* test in this study.

Data Availability

The deep-sequencing data from this study are deposited in the NCBI Gene Expression Omnibus (accession no. GSE114999) and the National Omics Data Encyclopedia (accession no. OEP000030).

Results

This example first examined the base editing efficiency of a commonly used BE, the rat APOBEC1 (rA1)-based BE3, in human cells having either increased or decreased levels of methylation. When DNA methylation was promoted by DNMT3 in regions with native low methylation levels, editing frequencies by BE3 decreased. In addition, when DNA methylation was reduced by TET1 in regions with native high methylation levels, BE3-induced editing frequencies increased accordingly. These results suggest that the canonical rA1-based BE3 is less efficient in editing cytosines embedded in highly methylated genomic regions. Notably, C-to-T editing was suppressed by DNA methylation at both CpG and flanking non-CpG sites (median decrement ~28%, $P = 2 \times 10^{-8}$ for CpG sites and ~51%, $P = 7 \times 10^{-10}$ for flanking non-CpG sites). APOBECs deaminate cytidines on single-stranded DNA in a processive manner. CpG methylation may affect the sliding of APOBEC and therefore impairs its binding on the flanking non-CpG sites for deamination.

To screen for efficient base editing in high-methylation background, a series of BEs was obtained by fusing Cas9 nickase with fifteen different APOBEC/AID deaminases (**FIG. 5b**). This example tested these BEs then in an *E. coli*-derived vector system (**FIG. 5b**), which has been previously used to probe mutations. In unmethylated vectors, these BEs showed varied levels of base editing. The BEs containing human APOBEC3A (hA3A-BE3, mean editing frequency ~39%), human APOBEC3B (hA3B-BE3, mean editing frequency ~33%) or human AID (hAID-BE3, mean editing frequency ~28%) mediated base editing at levels that are comparable to BE3 (mean editing frequency ~31%) (**FIG. 5c**). Whereas in methylated vectors, only hA3A-BE3 induced efficient base editing (mean editing frequency ~35%), compared to relatively low editing efficiencies induced by BE3 (mean editing frequency ~12%) or other examined BEs (mean editing frequencies ~1% - 20%) (**FIG. 5c**). Of note, protein products of hA3A-BE3, BE3 and other examined BEs are comparable (**FIG. 5d**).

Similar to the observation in *E. coli*-derived vectors, hA3A-BE3 exhibited significantly higher base editing frequencies than rA1-based BE3 in all tested genomic regions, either those with a native high-methylation background (median ~1.7-fold, $P = 2 \times 10^{-10}$, **FIG. 5e,f**) or those with an induced high-methylation condition (median ~1.8-fold, $P = 5 \times 10^{-4}$). Thus, using hA3A as the deaminase module in BE could generally achieve high base editing efficiency in genomic regions with high methylation levels.

The base editing on cytosines in a GpC context was observed to be generally inefficient by rA1-based BEs. While, this example found that hA3A-BE3 could induce efficient base editing on most of cytosines at GpC sites in both endogenously and induced high-methylation backgrounds (**FIG. 5e**). This example further compared their editing efficiencies under both endogenously and induced low-methylation backgrounds and observed a similar superiority of hA3A-BE3 over BE3 on editing cytosines in the GpC context (**FIG. 5g,h**). Statistical analysis confirmed that the base editing efficiency induced by hA3A-BE3 was significantly higher than that induced by BE3 on cytosines in the GpC context in either high- (median ~2.3-fold, $P = 1 \times 10^{-5}$) or low- (median ~1.8-fold, $P = 6 \times 10^{-9}$) methylation conditions. Notably, hA3A-BE3-mediated base editing was as efficient as BE3 at cytosines in non-GpC contexts in all tested low-methylation regions (median ~1.1-fold, $P = 0.045$). This example also found that hA3A-BE3 yielded less non-C-to-T conversion than BE3 in both high- (median ~97% by hA3A-BE3 comparing to ~94% by BE3, $P = 3 \times 10^{-4}$) and low-methylation regions (median ~92% by hA3A-BE3 comparing to ~90% by BE3, $P = 4 \times 10^{-6}$). Both BE3 and hA3A-BE3 induced less non-C-to-T conversion at CpG sites

with high methylation status than at CpG sites with low methylation status (median ~95% vs ~90%, $P = 3 \times 10^{-5}$ for BE3 and median ~95% vs ~92%, $P = 5 \times 10^{-4}$ for hA3A-BE3). This example also found that hA3A-BE3 induced higher indel frequencies than BE3 (median ~2 in both high- and low- methylation regions). Such an increase may be caused by the high deaminase activity of hA3A, which can trigger downstream DNA repair pathways to generate DNA double strand breaks.

The results suggest that hA3A-BE3 can efficiently induce base editing in a broader scope (**FIG. 5**). However, the editing window of hA3A-BE3 is wider (~12 nt, position 2-13 in the sgRNA target site) than that of BE3 (~5 nt, position 4-8). As the wide editing window of hA3A-BE3 may result from the high deaminase activity of hA3A, mutations in hA3A that can reduce deaminase activity might correspondingly narrow the editing window of hA3A-BE3. Designated mutations (Y130F, D131Y or Y132D) successfully narrowed the editing window with little effect on the base editing efficiency, whereas mutations in the zinc-coordination motif almost completely eliminated the deaminase activity (C101S and C106S).

This example then focused on two engineered hA3A-BE3s (hA3A-BE3-Y130F and hA3A-BE3-Y132D), which have similar editing windows (position 3-8 for hA3A-BE3-Y130F and position 3-7 for hA3A-BE3-Y132D) as BE3 (position 4-8). In highly-methylated regions, hA3A-BE3-Y130F and hA3A-BE3-Y132D induced higher editing efficiencies than BE3 at all editable sites in overlapping editing windows (position 4-7) (**FIG. 6a**, cytosines in pink and **FIG. 6b**, median ~2.3 fold, $P = 0.002$ for hA3A-BE3-Y130F and median ~1.2 fold, $P = 0.03$ for hA3A-BE3-Y132D). For cytosines outside of overlapping editing windows, hA3A-BE3-Y132D induced C-to-T editing frequencies similar to BE3 while hA3A-BE3-Y130F induced higher editing frequencies (**FIG. 6a**, cytosines in black). Similar to the original hA3A-BE3, both engineered hA3A-BE3-Y130F and hA3A-BE3-Y132D edited cytosines in GpC contexts more efficiently than BE3 in overlapping editing windows (**FIG. 6c,d**, median ~2.3 fold, $P = 3 \times 10^{-5}$ for hA3A-BE3-Y130F and median ~1.9 fold, $P = 0.002$ for hA3A-BE3-Y132D). Protein expression levels of hA3A-BE3-Y130F and hA3A-BE3-Y132D were very similar to that of BE3 (**FIG. 6e**), though the two engineered hA3A-BEs induced higher C-to-T editing efficiencies (**FIG. 6b,d**). In terms of product purity, we found that hA3A-BE3-Y130F yielded less non-C-to-T conversion (median ~96.3% by hA3A-BE3-Y130F comparing to ~95.6% by BE3, $P = 0.03$ in high-methylation regions, median ~92% by hA3A-BE3-Y130F comparing to ~90% by BE3, $P = 0.002$ in low-methylation regions) but more indels (median ~2.1 fold, $P = 0.0002$ in high-methylation regions,

median ~1.3 fold in low-methylation regions, $P = 0.12$) than BE3. The product purity induced by hA3A-BE3-Y132D was higher than BE3 in native low-methylation regions (median ~93% by hA3A-BE3-Y132D comparing to ~90% by BE3, $P = 0.001$), but lower in native high-methylation regions (median ~94.9% by hA3A-BE3-Y132D comparing to ~95.6% by BE3, $P = 0.03$). Nevertheless, indel frequencies induced by hA3A-BE3-Y132D were comparable to those induced by BE3 at all tested sites (median ~1.2 fold in both high- and low- methylation regions).

To further enhance C-to-T base editing system, three copies of the 2A- uracil DNA glycosylase inhibitor (UGI) sequence were fused to the C-terminus of hA3A-BE3-Y130F and hA3A-BE3-Y132D to develop hA3A-eBE-Y130F and hA3A-eBE-Y132D. In low-methylation regions, hA3A-eBE-Y130F and hA3A-eBE-Y132D induced significantly higher editing efficiencies (**FIG. 6f,g**, median ~1.2 fold, $P = 0.0004$ for hA3A-eBE-Y130F and median ~1.2 fold, $P = 0.004$ for hA3A-eBE-Y132D), higher product purity (**FIG. 6h**, median ~96% by hA3A-eBE-Y130F comparing to ~94% by hA3A-BE3-Y130F, $P = 0.006$ and median ~96% by hA3A-eBE-Y132D comparing to ~92% by hA3A-BE3-Y132D, $P = 0.004$) and lower indel frequencies (**FIG. 6i**, median decrement ~21%, $P = 4 \times 10^{-5}$ for hA3A-eBE-Y130F and median decrement ~9%, $P = 0.03$ for hA3A-eBE-Y132D) than hA3A-BE3-Y130F and hA3A-BE3-Y132D, respectively. In high-methylation regions, hA3A-eBE-Y130F and hA3A-eBE-Y132D induced significantly higher product purity (median ~97% by hA3A-eBE-Y130F comparing to ~95% by hA3A-BE3-Y130F, $P = 0.003$ and median ~97% by hA3A-eBE-Y132D comparing to ~95% by hA3A-BE3-Y132D, $P = 0.003$) and lower indel frequencies (median decrement ~23%, $P = 2 \times 10^{-7}$ for hA3A-eBE-Y130F and median decrement ~21%, $P = 4 \times 10^{-5}$ for hA3A-eBE-Y132D) than hA3A-BE3-Y130F and hA3A-BE3-Y132D, respectively, though editing efficiencies remained the same (median ~1 fold for hA3A-eBE-Y130F and hA3A-eBE-Y132D). Together, these results indicated that hA3A-BE3-Y130F, hA3A-BE3-Y132D, hA3A-eBE-Y130F and hA3A-eBE-Y132D can mediate highly efficient base editing in narrowed editing windows compared to the original hA3A-BE3 in all examined contexts.

Here, this example demonstrates that hA3A-BE3 and its engineered forms, can comprehensively induce efficient base editing in all examined contexts, including both methylated DNA regions and GpC dinucleotides. It is contemplated that hA3A can also be conjugated with other Cas proteins to further expand the scope of base editing.

Example 3. Gene Editing of Human DYRK1A with dCas12a-hA3A Base Editors

This example tested base editors that combined a Cas12a (Cpf1) and various mutant human A3A proteins.

Methods*Construction of dCas12a-hA3A-BE expression vector*

pUC57-hA3A (synthesized by Genscript Biotechnology Co., Ltd.) was used as a template, using suitable primers. PCR was carried out to obtain the coding sequence of hA3A, and a fragment homologous to the linearized vector at both ends was subjected to gel electrophoresis purification. After purification by gel electrophoresis, the fragment was recombined into the linearized dCas12a-BE vector produced by SacI and SmaI by plasmid recombinant kit Clone Express® to obtain expression vector dCas12a-hA3A-BE.

Construction of dCas12a-hA3A-BE-W98Y expression vector

Using dCas12a-hA3A-BE as a template, two PCR products with a W98Y mutation and a homology arm, and a homologous segment with a linearized vector. After purification by gel electrophoresis, the two fragments were simultaneously recombined into the linearized dCas12a-hA3A-BE vector generated by ApaI and SmaI using plasmid recombinant kit Clone Express® to obtain expression vector dCas12a-hA3A-BE-W98Y.

Likewise, expression vectors dCas12a-hA3A-BE-W104A, dCas12a-hA3A-BE-P134Y, dCas12a-hA3A-BE-W98Y-W104A, dCas12a-hA3A-BE-W98Y-P134Y, dCas12a-hA3A-BE-W104A-P134Y, dCas12a-hA3A-BE-W98Y-W104A-Y130F, dCas12a-hA3A-BE-W98Y-W104A-Y132D, dCas12a-hA3A-BE-W104A-Y130F-P134Y, and dCas12a-hA3A-BE-W104A-Y132D-P134Y. Relevant sequences are shown in Tables 1 and 2.

Construction of gRNA expression plasmid

The nucleotide sequence was annealed to primers and the annealed product was ligated into the gRNA expression vector pLb-Cas12a-pGL3-U6-sgRNA digested with restriction endonuclease BsaI using T4 DNA ligase. gRNA expression plasmid sgDYRK1A targeting human DYRK1A site was obtained.

Eukaryotic cell transfection

The sgDYRK1A and each of dCas12a-hA3A-BE, dCas12a-hA3A-BE-W98Y, dCas12a-hA3A-BE-W104A, dCas12a-hA3A-BE-P134Y, dCas12a-hA3A-BE-W98Y-W104A, dCas12a-hA3A-BE-W98Y-P134Y, dCas12a-hA3A-BE-W104A-P134Y, dCas12a-hA3A-BE-W98Y-W104A-Y130F, dCas12a-hA3A-BE-W98Y-W104A-Y132D, dCas12a-hA3A-BE-W104A-Y130F-P134Y, dCas12a-hA3A-BE-W104A-Y132D-P134Y were mixed into 200 μ l Opti-MEM at a ratio of 0.68 μ g : 1 μ g, added with 1.68 μ l of LIPOFECTAMINE plus, and 5.04 μ l of LIPOFECTAMINE LTX was added, and allowed to stand at room temperature for 5 minutes. 500 μ l DMEM (+ 10% FBS) medium was added for 24-well plates and transfected HEK293T cells 160,000. After 12 h, replaced with fresh medium containing 1% double antibody (cyanin). The cells were harvested after 60 hours of incubation.

EditR analysis of sanger sequencing results

DNA sanger sequencing results were analyzed using EditR software (moriaritylab.shinyapps.io/editr_v10/). EditR is a web version of the sanger sequencing result analysis software developed in 2018 (Kluesner MG, Nedveck DA, Lahr WS, et al. EditR : A Method to Quantify Base Editing from Sanger Sequencing [J]. The CRISPR Journal, 2018, 1 (3): 239-250.). EditR is a simple, accurate and efficient analytical tool for processing the sequencing results of DNA samples based on the sgRNA sequence by using the sanger sequencing signal, and finally outputting the base editing efficiency at the sgRNA target site.

The sequencing results are shown **FIG. 11 and 12**. The EditR analysis results are presented in **FIG. 7 and 8**. When fused to the conventional cytosine deaminase, A1 (APOBEC1), Cas12a (cpf1) exhibited poor efficiency (see, e.g., **FIG. 7B**, the first column in each group). The combination with the hA3A wild-type protein greatly increased the editing efficiency (see, e.g., the second column). Interestingly, the A3A mutation W98Y, W104A, P134Y or the combination of each two further increased the editing efficiency (**FIG. 7**). Also, the editing window such a Cas12a-A3A can be narrowed to achieve more precise editing when the mutation Y130F or Y132D is further included in A3A (**FIG. 8**).

Example 4. Gene Editing of Human SITE6 with dCas12a-hA3A Base Editors

This example tested various indicated base editors with the human gene SITE6.

The experimental procedure is similar to Example 3. The sequencing results are shown in detail in **FIG. 15 and 16** (two replicates of experimental data). The EditR analysis results are shown in **FIG. 9 and 10**. Like in Example 3, the Cas12a-A3A editor had greater editing efficiency than the Cas12a-A1 and the A3A mutation W98Y, W104A, P134Y or the combination of each two further increased the editing efficiency (**FIG. 9**). Also, the editing window such a Cas12a-A3A can be narrowed to achieve more precise editing when the mutation Y130F or Y132D is further included in A3A (**FIG. 10**).

Example 5. Gene Editing of Human RUNX1 with dCas12a-hA3A Base Editors

This example tested various indicated base editors with the human gene RUNX1.

The experimental procedure is similar to Example 3. The sequencing results are shown in detail in **FIG. 17 and 18** (two replicates of experimental data). The EditR analysis results are shown in **FIG. 11 and 12**. Like in Example 3, the Cas12a-A3A editor had greater editing efficiency than the Cas12a-rA1, and the A3A mutation W98Y, W104A, P134Y or the combination of each two further increased the editing efficiency (**FIG. 11**). Also, the editing window such a Cas12a-A3A can be narrowed to achieve more precise editing when the mutation Y130F or Y132D is further included in A3A (**FIG. 12**).

* * *

The present disclosure is not to be limited in scope by the specific embodiments described which are intended as single illustrations of individual aspects of the disclosure, and any compositions or methods which are functionally equivalent are within the scope of this disclosure. It will be apparent to those skilled in the art that various modifications and variations can be made in the methods and compositions of the present disclosure without departing from the spirit or scope of the disclosure. Thus, it is intended that the present disclosure cover the modifications and variations of this disclosure provided they come within the scope of the appended claims and their equivalents.

All publications and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference

CLAIMS

What is claimed is:

1. A fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A) and a second fragment comprising a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein.
2. The fusion protein of claim 1, further comprising a uracil glycosylase inhibitor (UGI).
3. The fusion protein of claim 1, which includes fewer than 2500 amino acid residues in total.
4. The fusion protein of claim 1, wherein the APOBEC3A is a wildtype human APOBEC3A or a mutant of human APOBEC3A having a mutation selected from the group consisting of Y130F, D131Y, D131E, Y132D, W104A, W98Y, P134Y and combinations thereof, according to residue numbering in SEQ ID NO:1, wherein the mutant retains cytidine deaminase activity.
5. The fusion protein of claim 4, wherein the APOBEC3A is mutant human APOBEC3A having mutations selected from the group consisting of Y130F+D131E+Y132D, Y130F+D131Y+Y132D, W98Y+W104A, W98Y+P134Y, W104A+P134Y, W104A+Y130F, W104A+Y132D, W98Y+W104A+Y130F, W98Y+W104A+Y132D, W104A+Y130F+P134Y, and W104A+Y132D+P134Y, according to residue numbering in SEQ ID NO:1.
6. The fusion protein of claim 4, wherein the human APOBEC3A is human APOBEC3A isoform a or isoform b.
7. The fusion protein of claim 4, wherein the APOBEC3A comprises the amino acid sequence of SEQ ID NO:1 or 6, or has at least 90% sequence identity to amino acid residues 29-199 of SEQ ID NO:1 and retains cytidine deaminase activity.
8. The fusion protein of claim 7, wherein the APOBEC3A comprises an amino acid sequence selected from the group consisting of SEQ ID NO:1-10 and 22-36.
9. The fusion protein of claim 1, wherein the Cas protein is selected from the group consisting of SpCas9, FnCas9, St1Cas9, St3Cas9, NmCas9, SaCas9, AsCpf1, LbCpf1, FnCpf1, VQR SpCas9, EQR SpCas9, VRER SpCas9, RHA FnCas9, KKH SaCas9, NmeCas9, StCas9, CjCas9, AsCpf1, FnCpf1, SsCpf1, PcCpf1, BpCpf1, CmtCpf1, LiCpf1, PmCpf1, Pb3310Cpf1,

Pb4417Cpf1, BsCpf1, EeCpf1, BhCas12b, AkCas12b, EbCas12b, LsCas12b, RfCas13d, LwaCas13a, PspCas13b, PguCas13b, RanCas13b, CasX, and CasY.

10. The fusion protein of claim 1, wherein the Cas protein is a mutant of protein selected from the group consisting of SpCas9, FnCas9, St1Cas9, St3Cas9, NmCas9, SaCas9, AsCpf1, LbCpf1, FnCpf1, VQR SpCas9, EQR SpCas9, VRER SpCas9, RHA FnCas9, KKH SaCas9, NmeCas9, StCas9, CjCas9, AsCpf1, FnCpf1, SsCpf1, PcCpf1, BpCpf1, CmtCpf1, LiCpf1, PmCpf1, Pb3310Cpf1, Pb4417Cpf1, BsCpf1, EeCpf1, BhCas12b, AkCas12b, EbCas12b, LsCas12b, RfCas13d, LwaCas13a, PspCas13b, PguCas13b, RanCas13b, CasX, and CasY, wherein the mutant retains the DNA-binding capability but does not introduce double strand DNA breaks.

11. The fusion protein of claim 10, wherein the mutant is capable of introducing a nick to one of the strands of a double stranded DNA bound by the mutant.

12. The fusion protein of claim 10, wherein the Cas protein comprises an amino acid sequence selected from the group consisting of SEQ ID NO:11, and 37-39.

13. The fusion protein of claim 1, wherein the first fragment is at the N-terminal side of the second fragment.

14. The fusion protein of claim 2, wherein the UGI comprises the amino acid sequence of SEQ ID NO:12 or has at least at least 90% sequence identity to SEQ ID NO:12 and retains the uracil glycosylase inhibition activity.

15. The fusion protein of claim 14, wherein the first fragment is at the N-terminal side of the second fragment which is at the N-terminal side of the UGI.

16. The fusion protein of claim 1, further comprising a peptide linker between the first fragment and the second fragment.

17. The fusion protein of claim 16, wherein the peptide linker has from 1 to 100 amino acid residues.

18. The fusion protein of claim 17, wherein at least 40% of the amino acid residues of peptide linker are amino acid residues selected from the group consisting of alanine, glycine, cysteine, and serine.

19. The fusion protein of claim 17, wherein the peptide linker has an amino acid sequence of SEQ ID NO:13 or 14.
20. The fusion protein of claim 1, further comprising a nuclear localization sequence.
21. The fusion protein of claim 1, comprising an amino acid sequence selected from the group consisting of SEQ ID NO:16-20 and 40-50.
22. A fusion protein comprising a first fragment comprising an apolipoprotein B mRNA editing enzyme catalytic subunit 3A (APOBEC3A) and a second fragment comprising a CRISPR-associated endonuclease in *Prevotella* and *Francisella* 1 (Cpf1).
23. The fusion protein of claim 22, wherein the Cpf1 is catalytically inactive.
24. The fusion protein of claim 22 or 21, wherein the Cpf1 is selected from the group consisting of AsCpf1, LbCpf1, and FnCpf1.
25. The fusion protein of claim 22, wherein the Cpf1 is a catalytically inactive *Lachnospiraceae* bacterium Cpf1 (dLbCpf1).
26. The fusion protein of claim 22, wherein the APOBEC3A is a wildtype human APOBEC3A or a mutant of human APOBEC3A having a mutation selected from the group consisting of Y130F, D131Y, D131E, Y132D, W98Y, W104A, P134Y and combinations thereof, according to residue numbering in SEQ ID NO:1, wherein the mutant retains cytidine deaminase activity.
27. A polynucleotide that encodes the fusion protein of any one of claims 1-26.
28. A composition comprising the fusion protein of any one of claims 1-26 and a pharmaceutically acceptable carrier.
29. The composition of claim 28, further comprising a guide RNA.
30. A method of editing a target polynucleotide, comprising contacting to the target polynucleotide a fusion protein of any one of claims 1-26 and a guide RNA having at least partial sequence complementarity to the target polynucleotide, wherein the editing comprises deamination of a cytosine (C) in the target polynucleotide.
31. The method of claim 30, wherein the C is in a GpC context.

32. The method of claim 30, wherein the C is methylated.
33. The method of any one of claims 30-32, wherein the contacting is *in vivo*.
34. The method of any one of claims 30-32, further comprising contacting to the target polynucleotide with a uracil glycosylase inhibitor (UGI) not fused to a Cas protein.

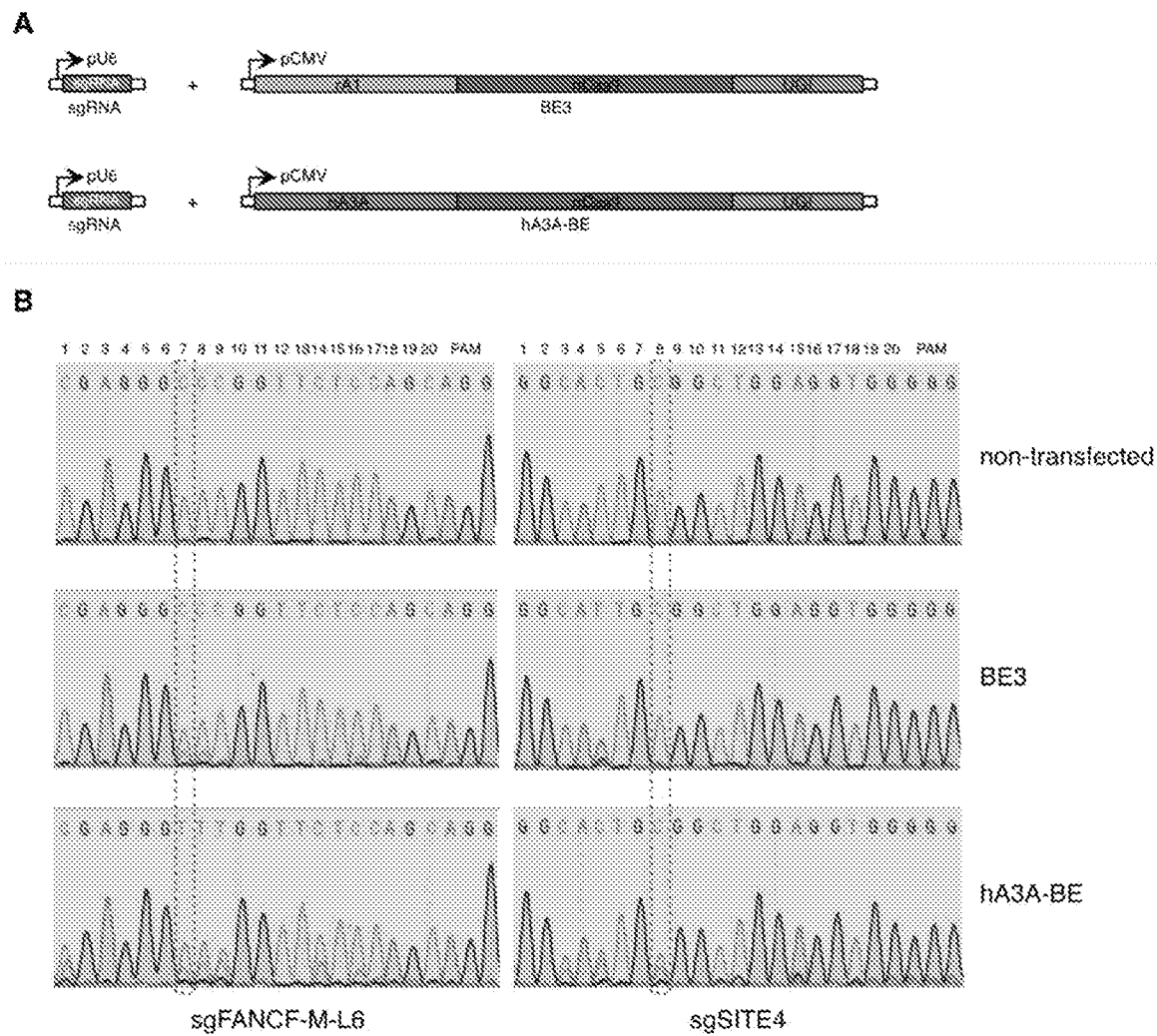


Fig. 1

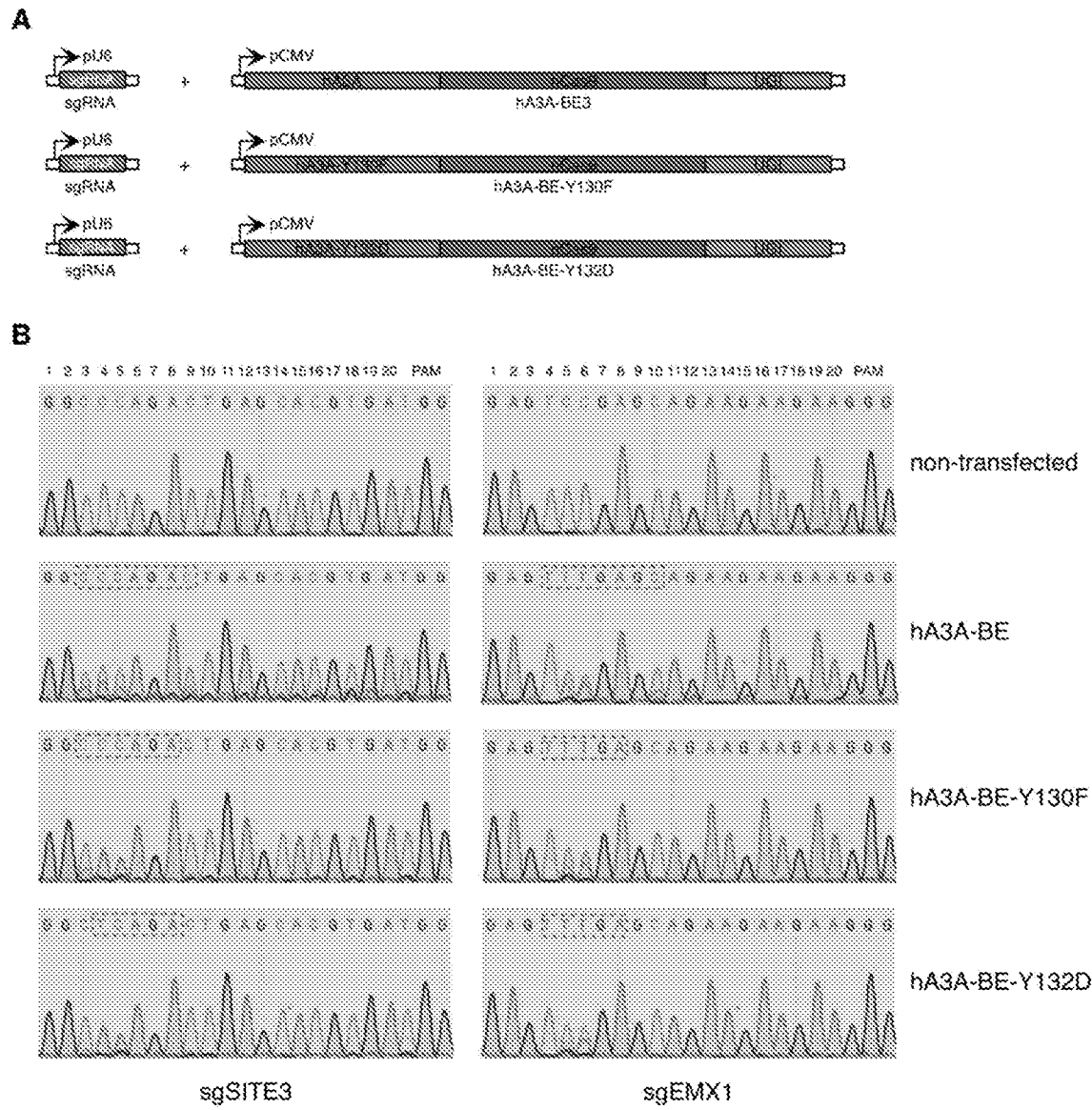


FIG. 2

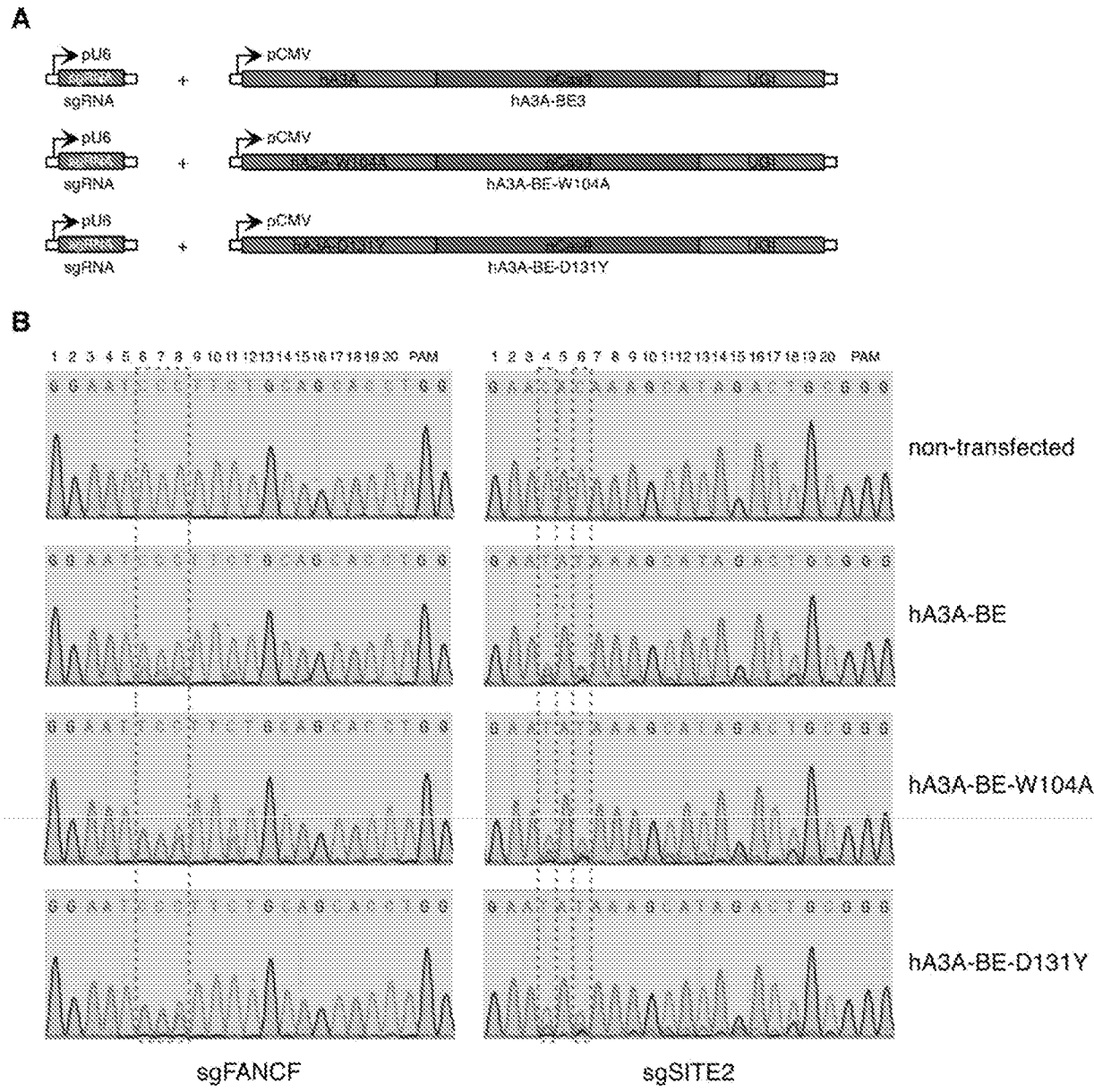


FIG. 3

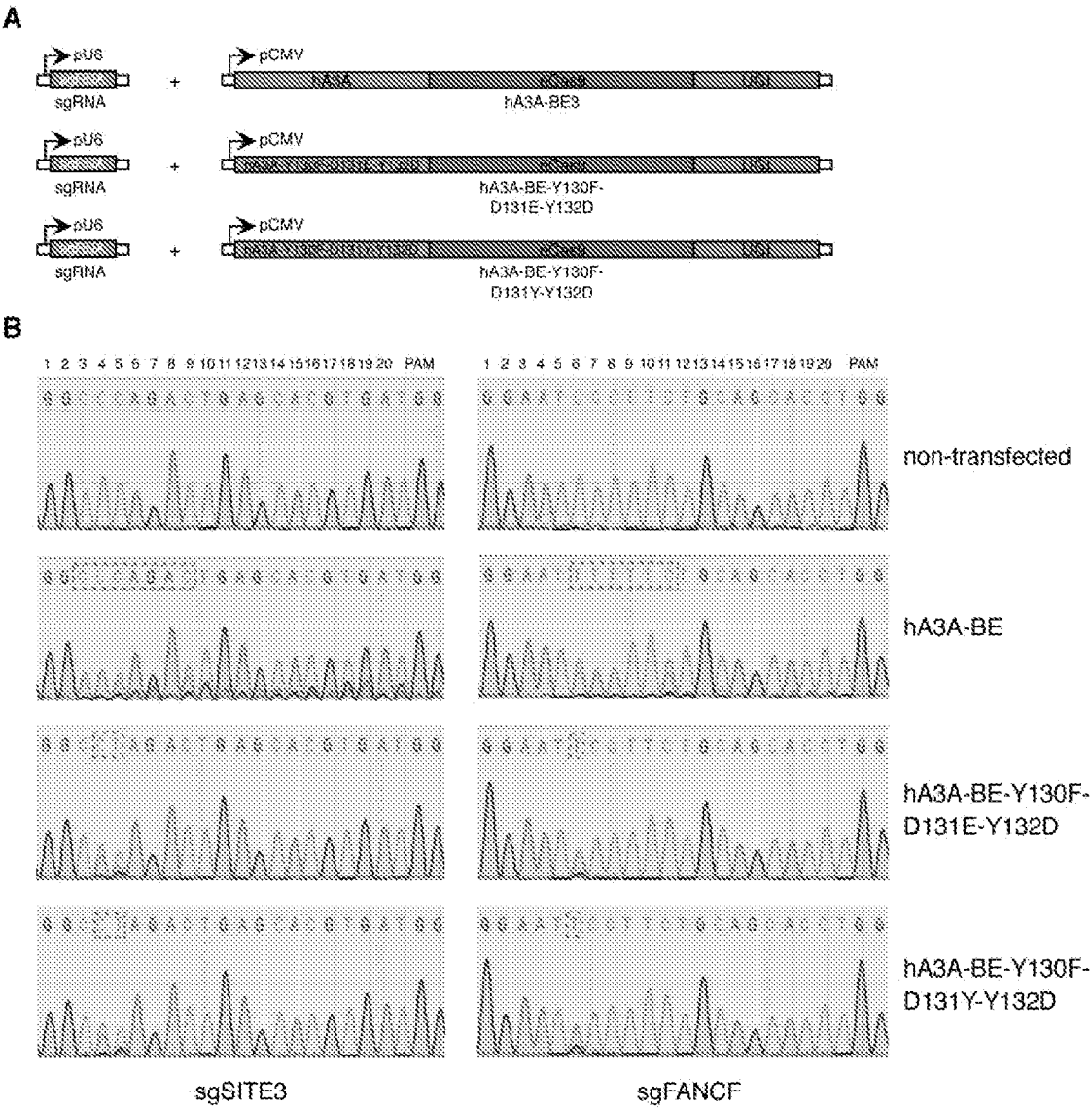


FIG. 4

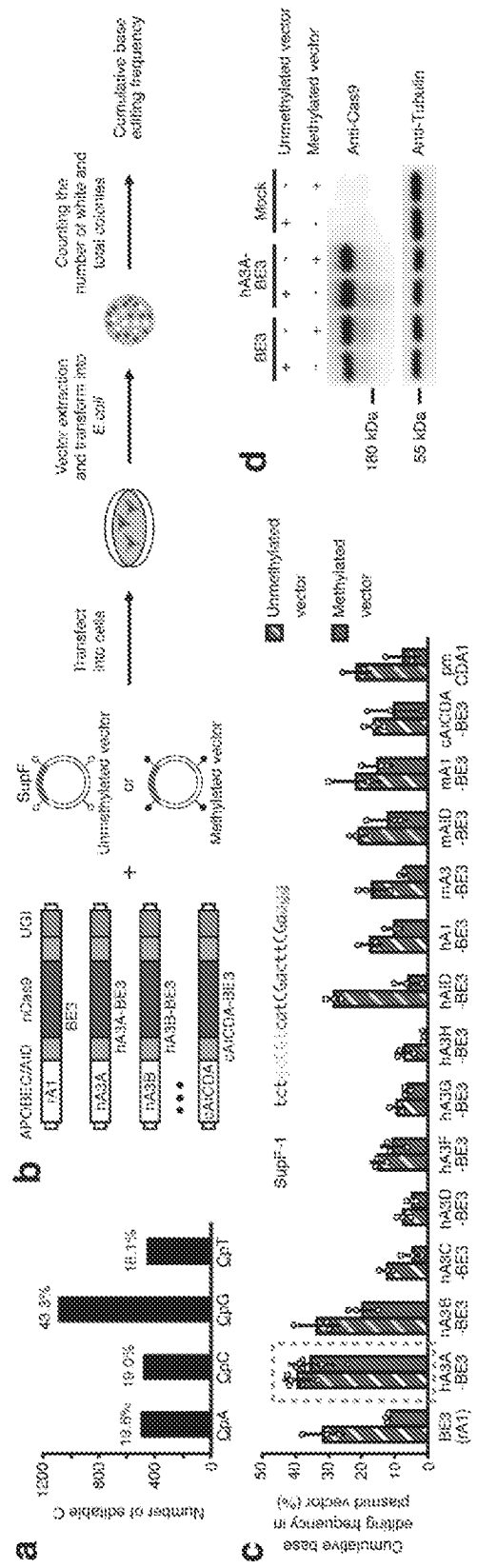
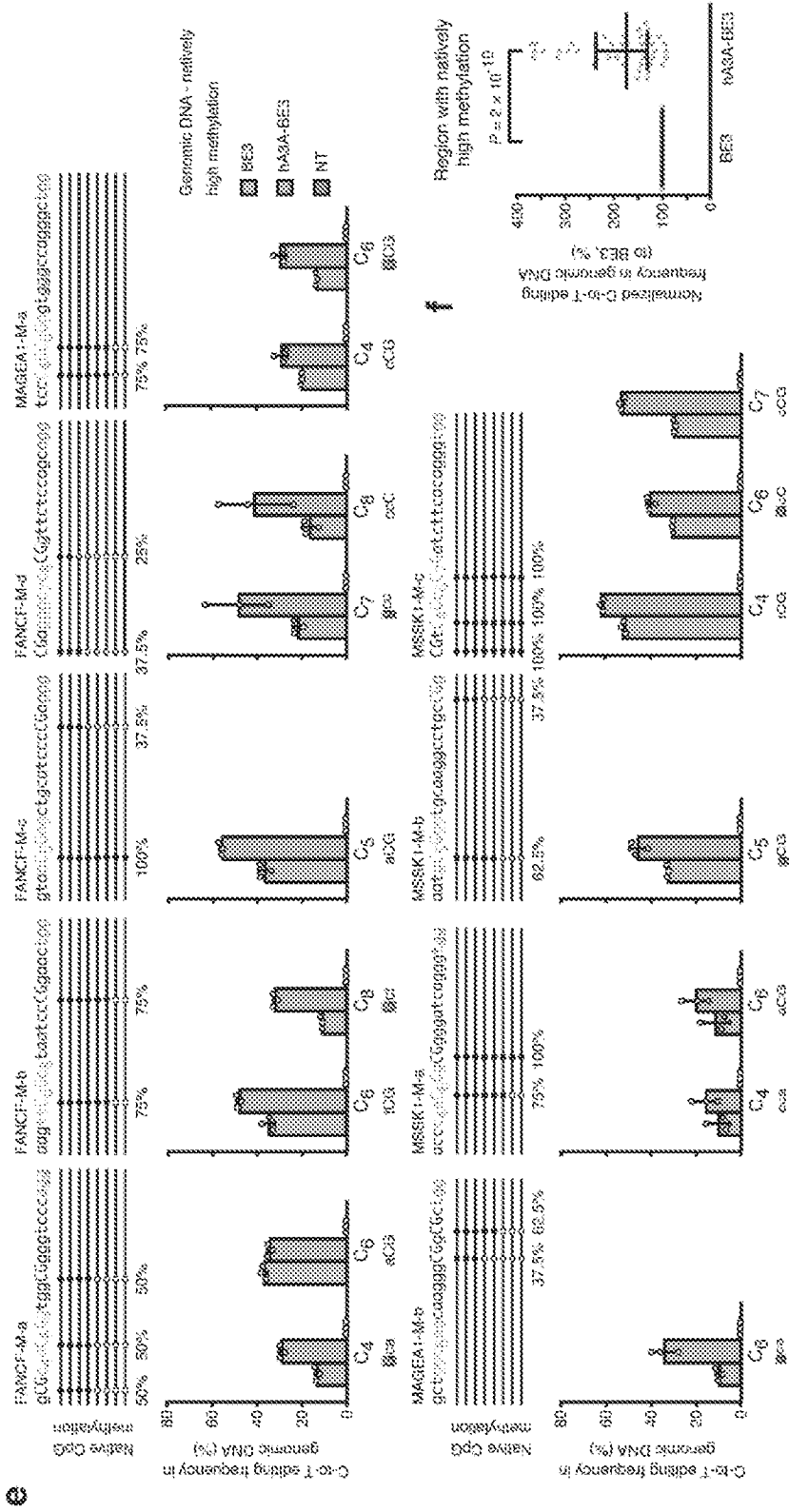


FIG. 5a-d



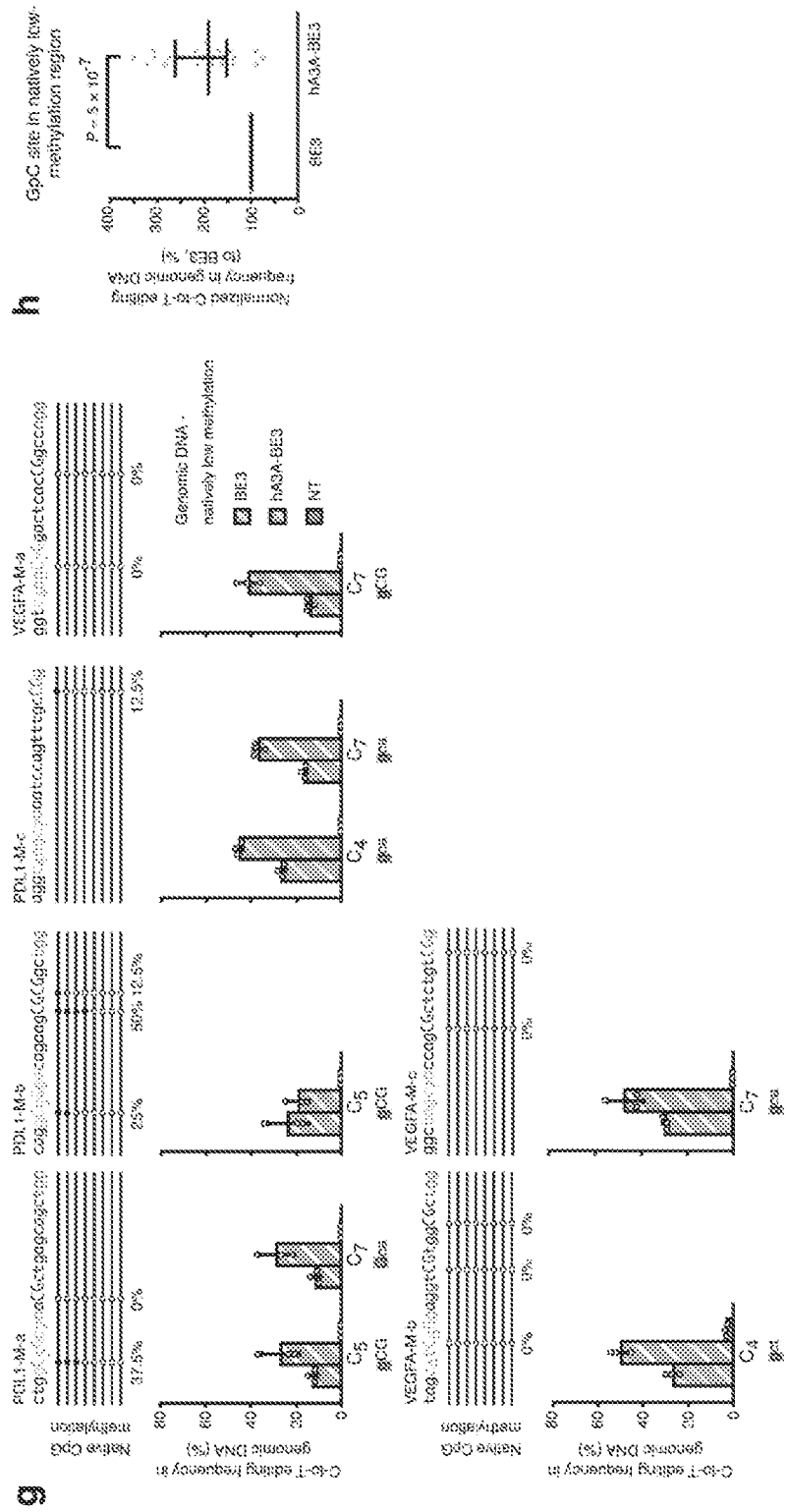


FIG. 5g-h

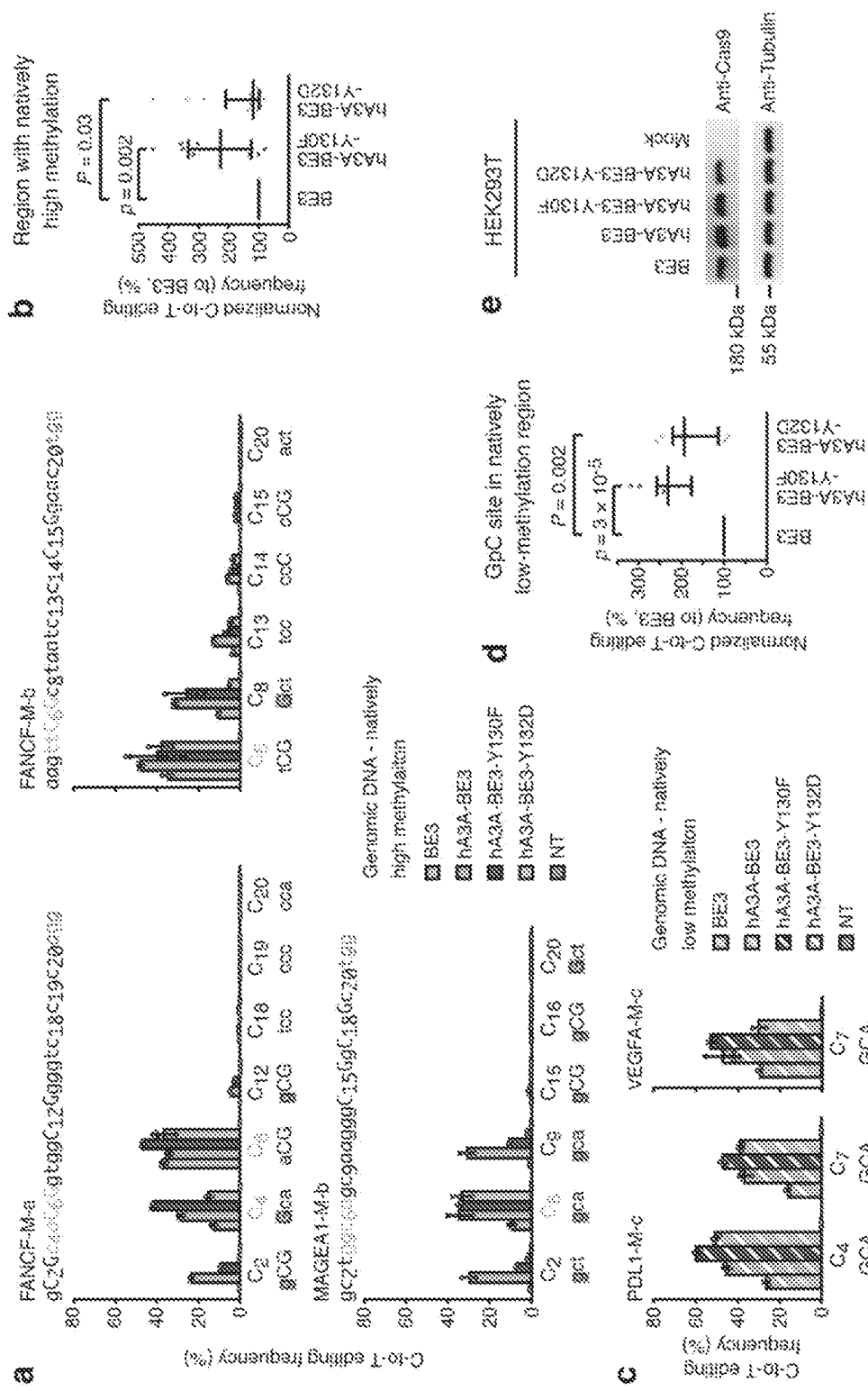


FIG. 6 a-e

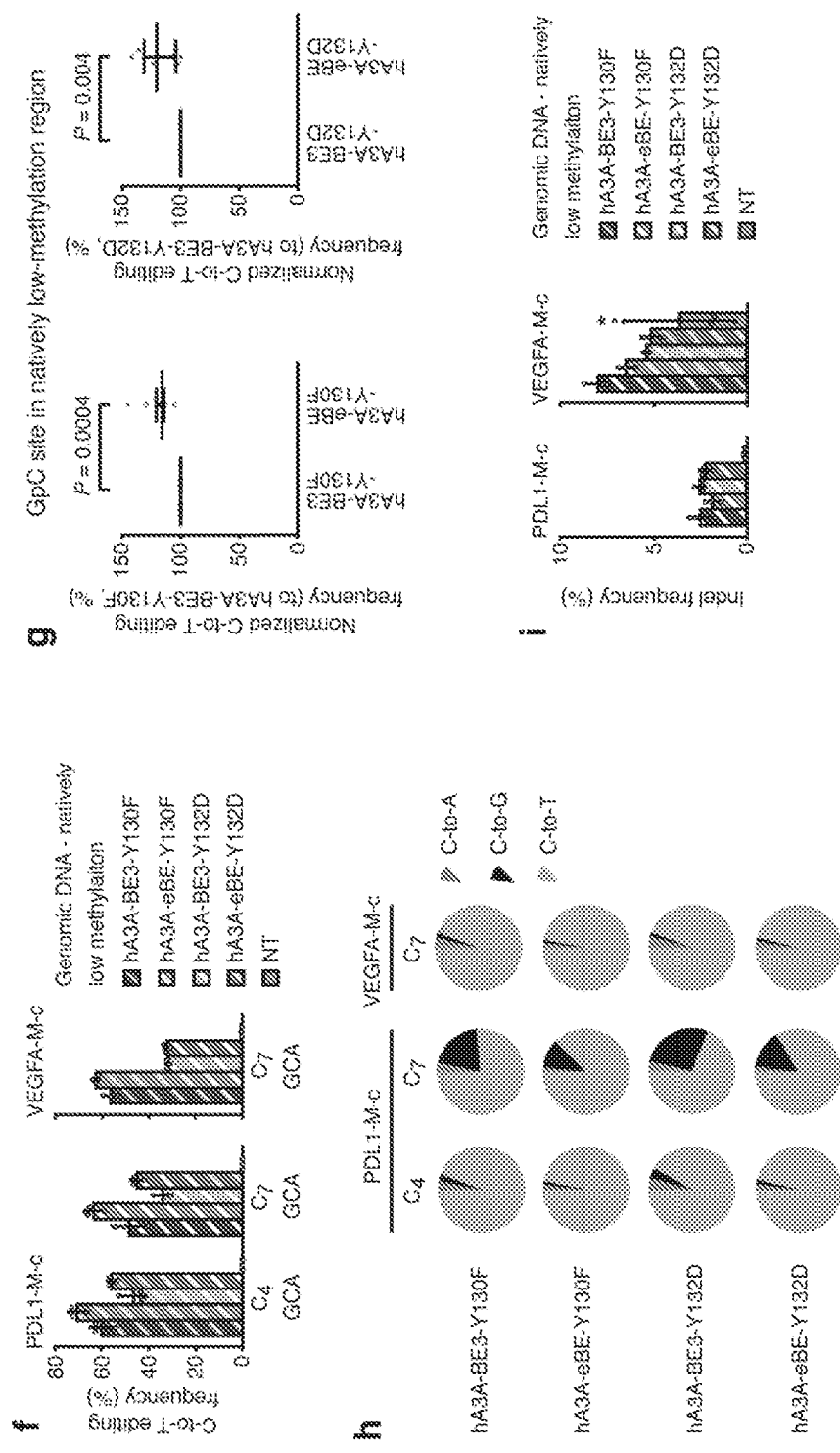


FIG. 6 f-i

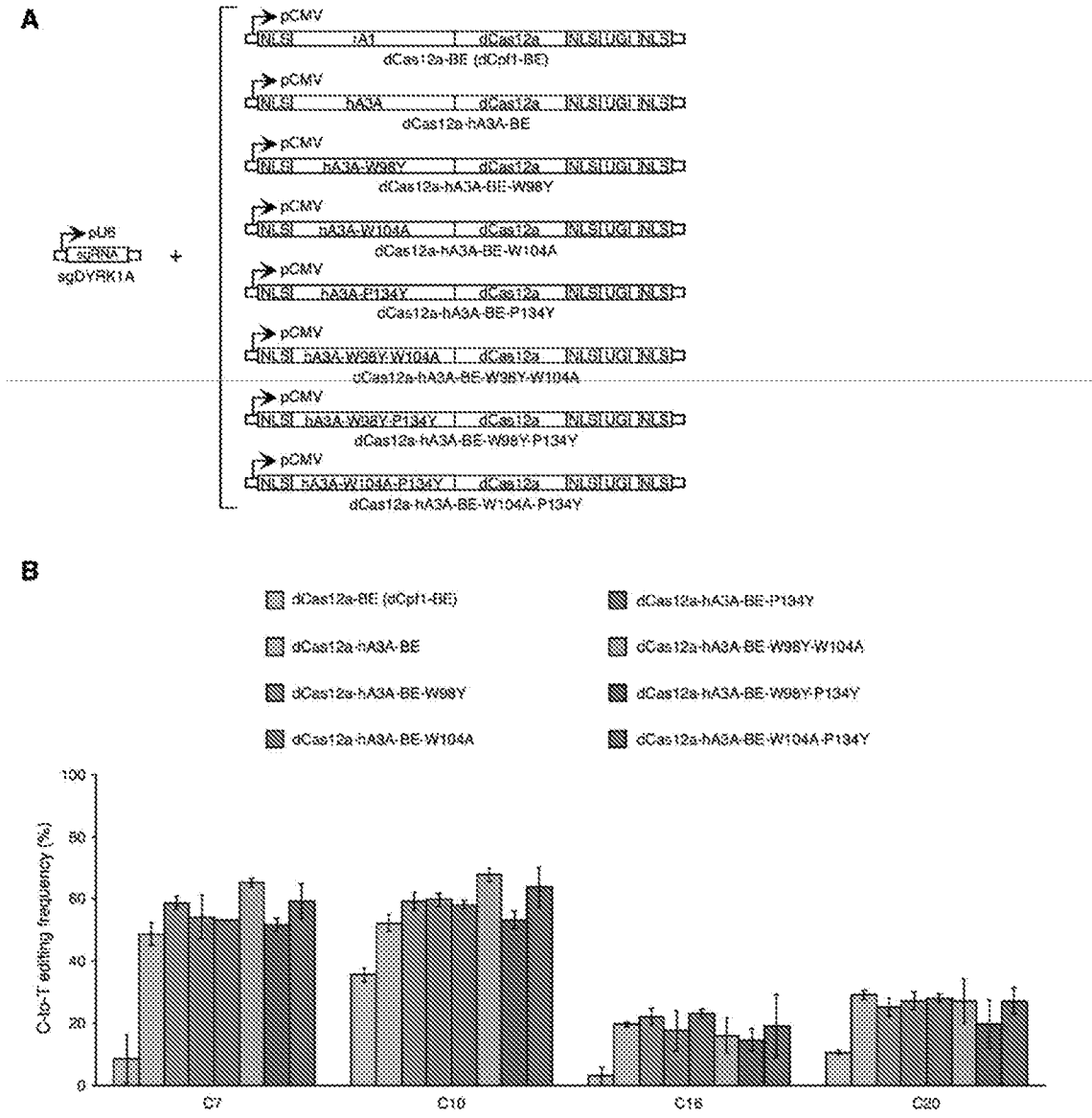


FIG. 7A-B

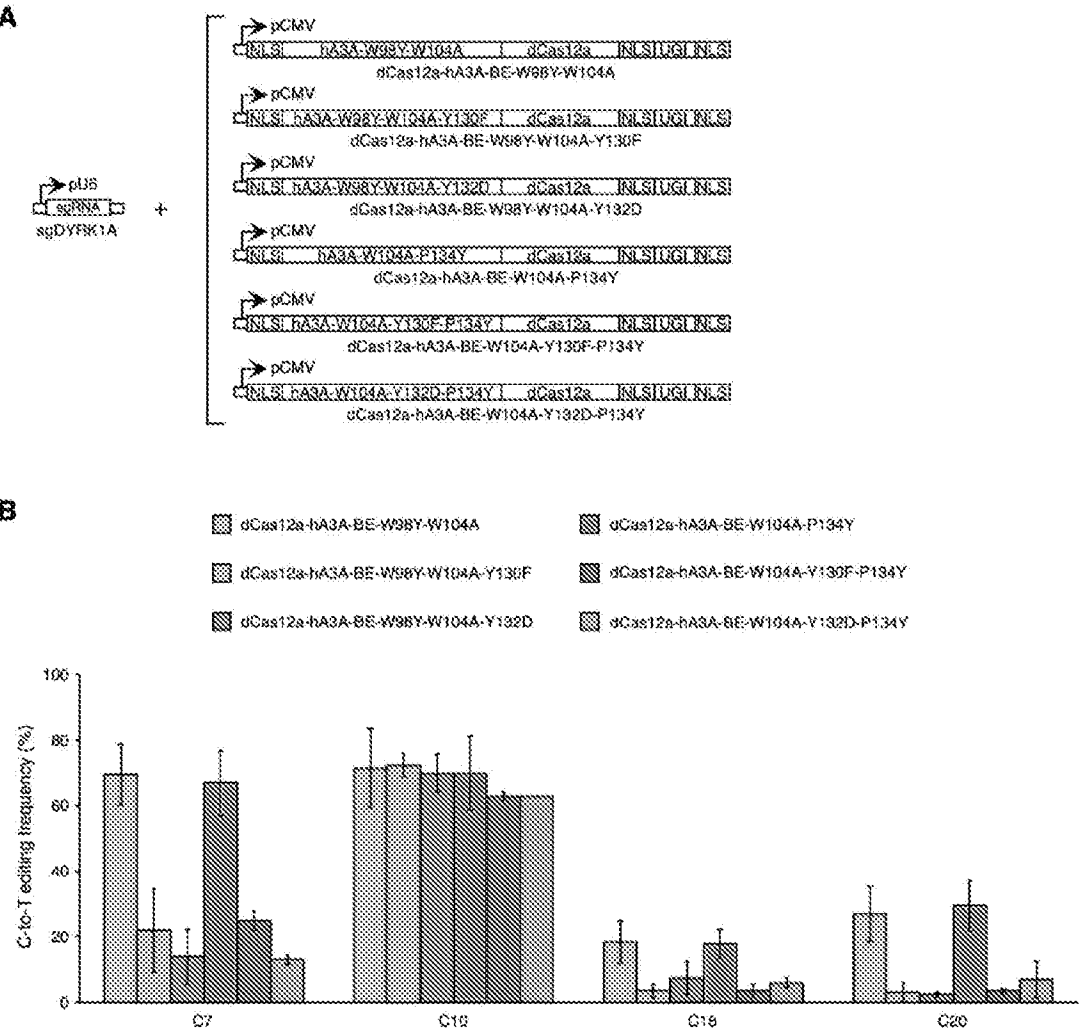


FIG. 8A-B

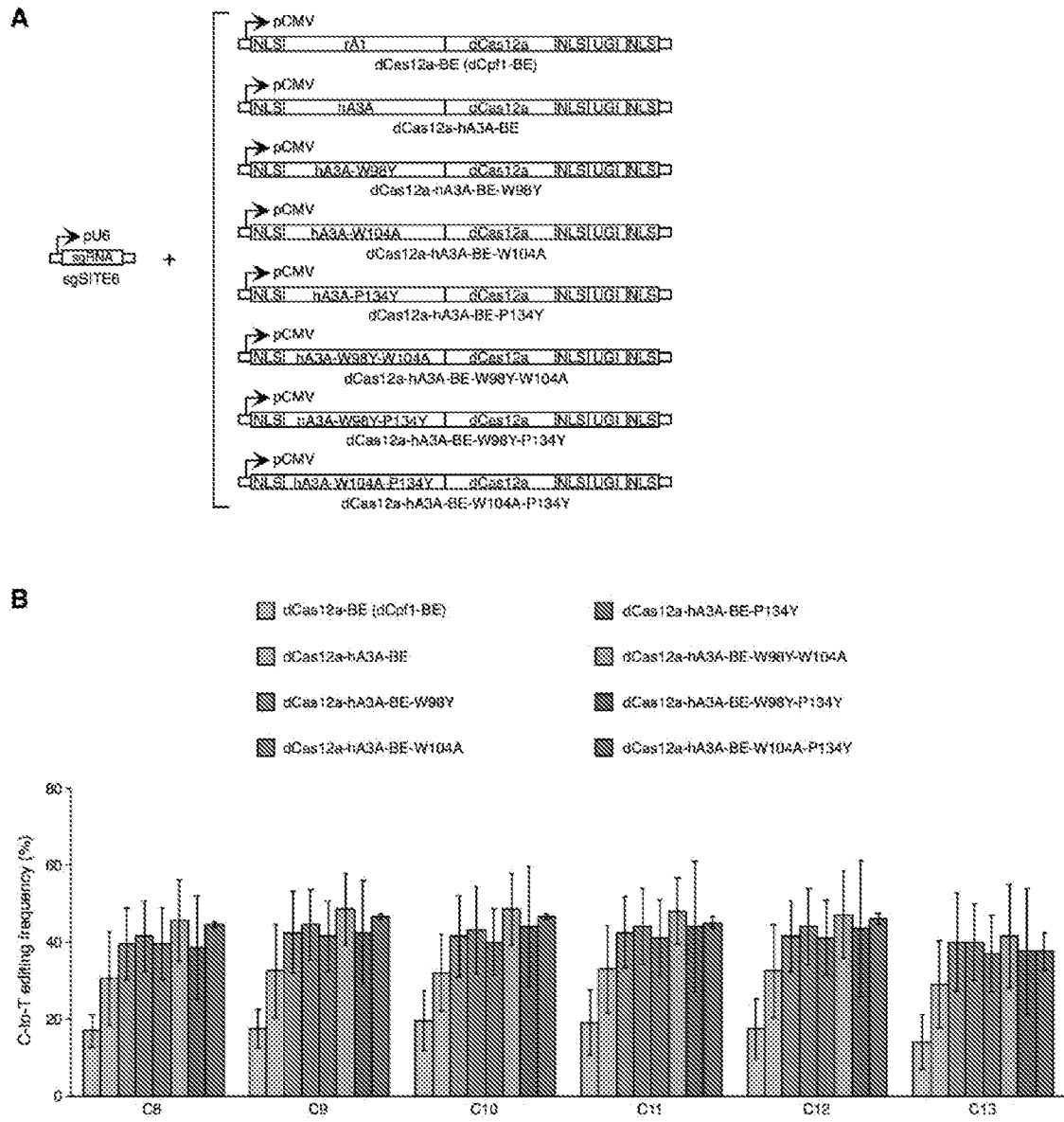


FIG. 9A-B

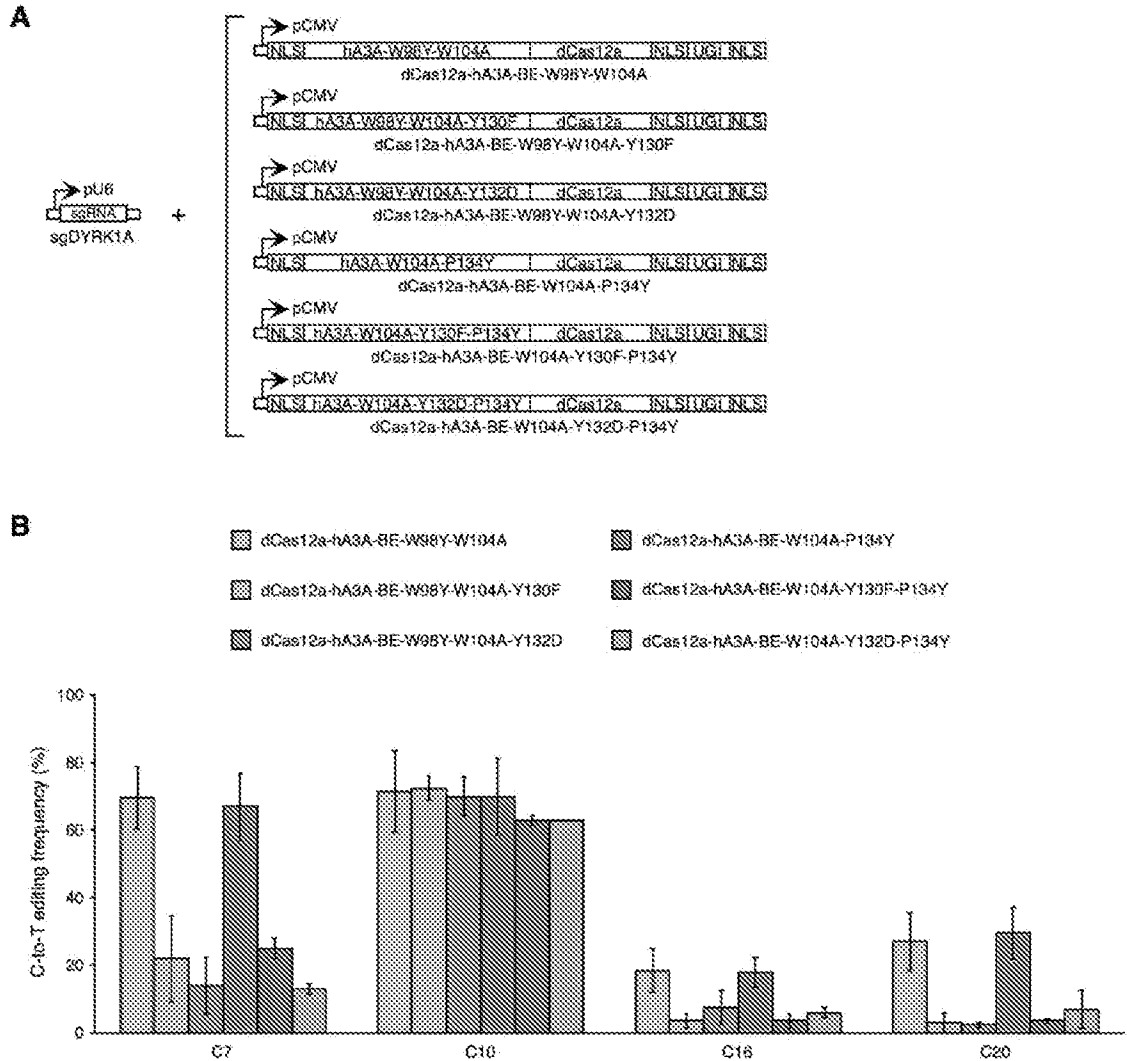


FIG. 10A-B

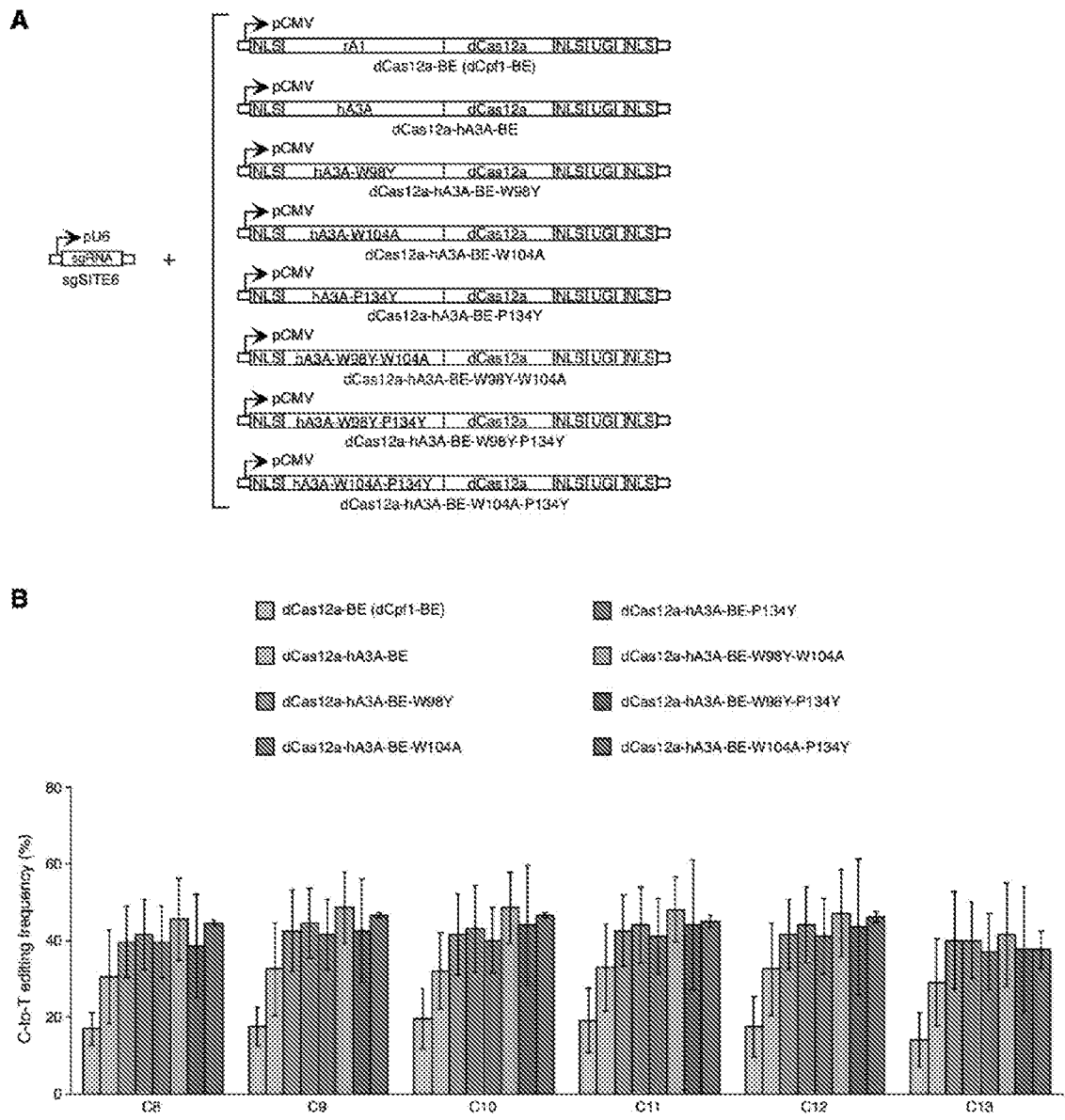


FIG. 11A-B

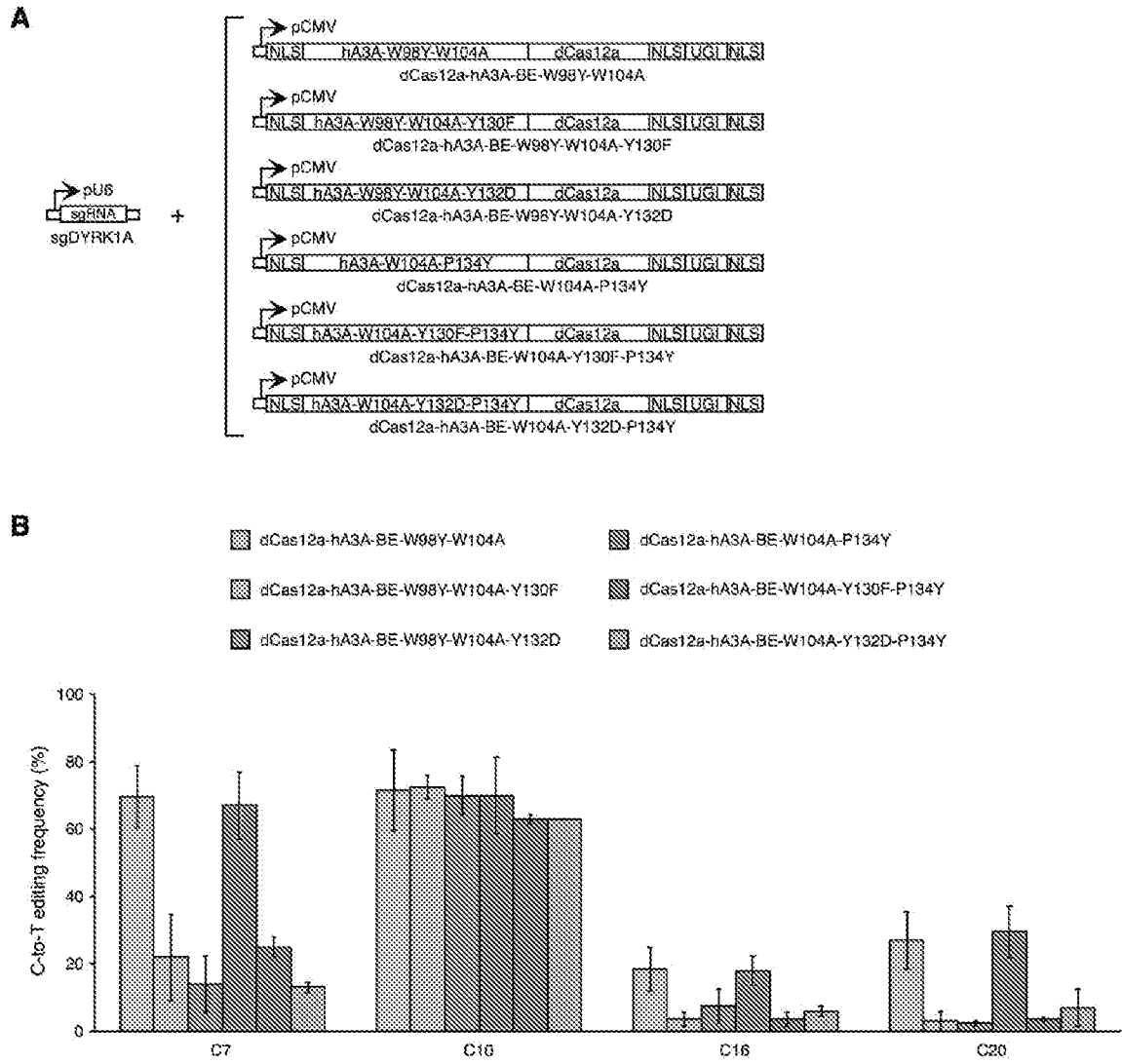


FIG. 12A-B

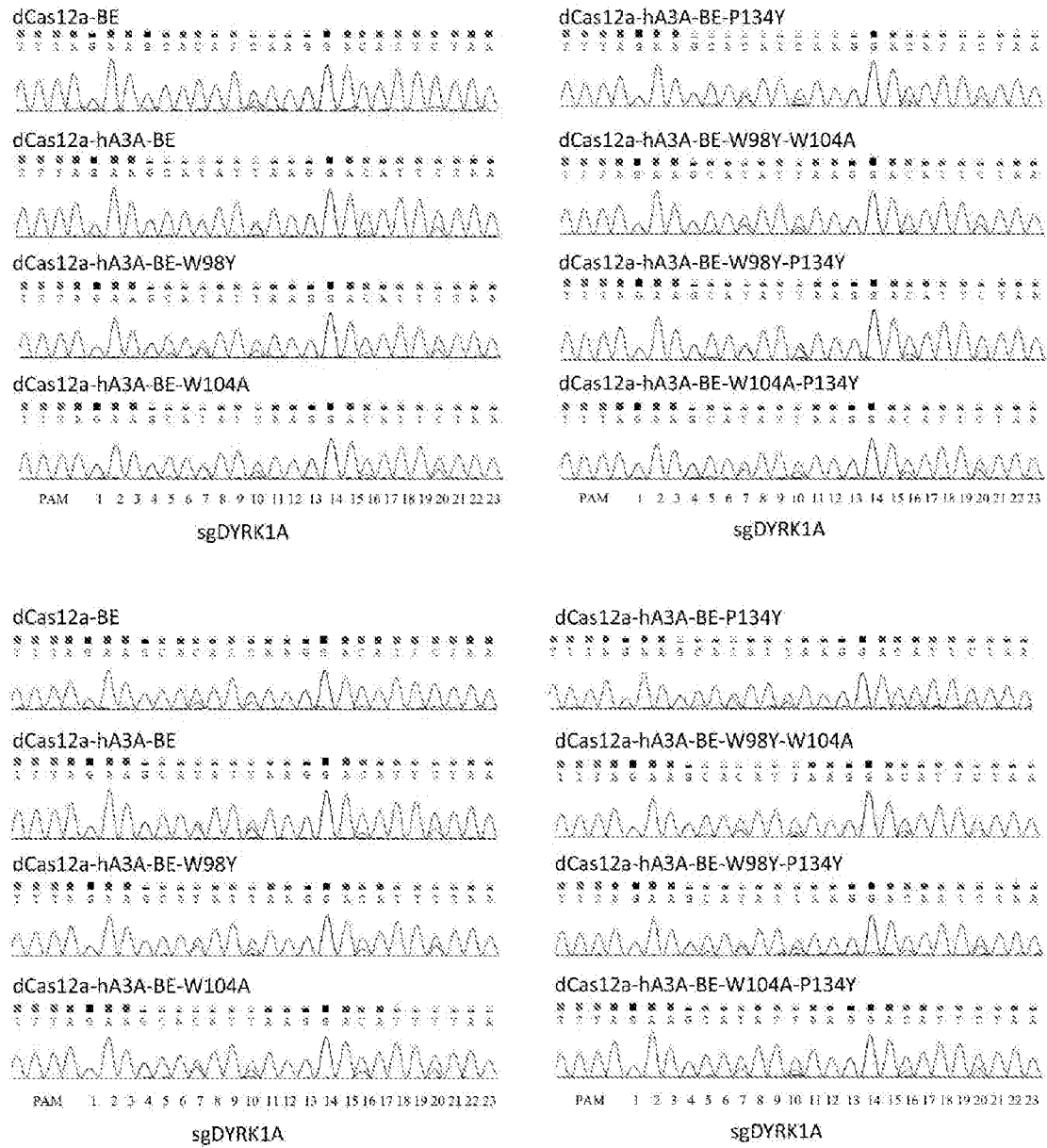


FIG. 13

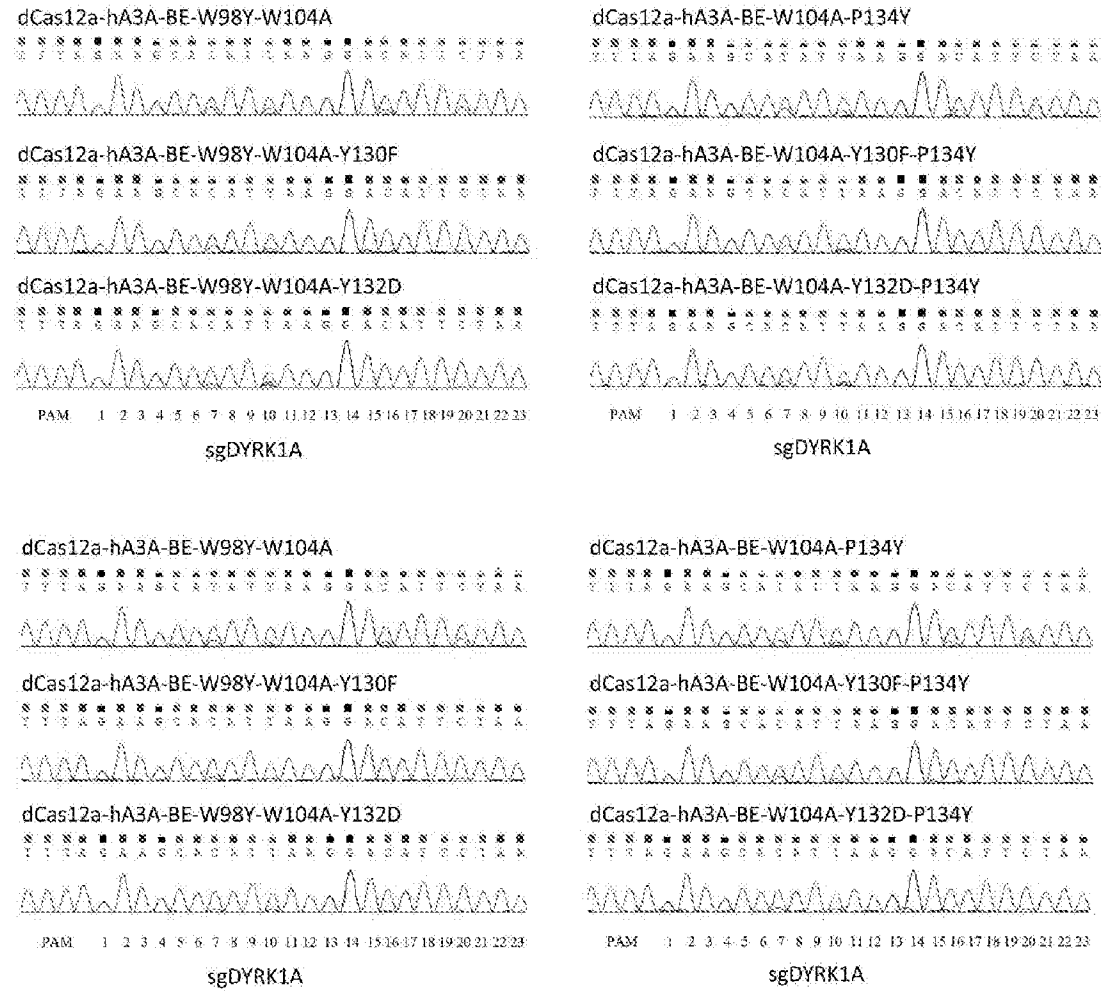


FIG. 14

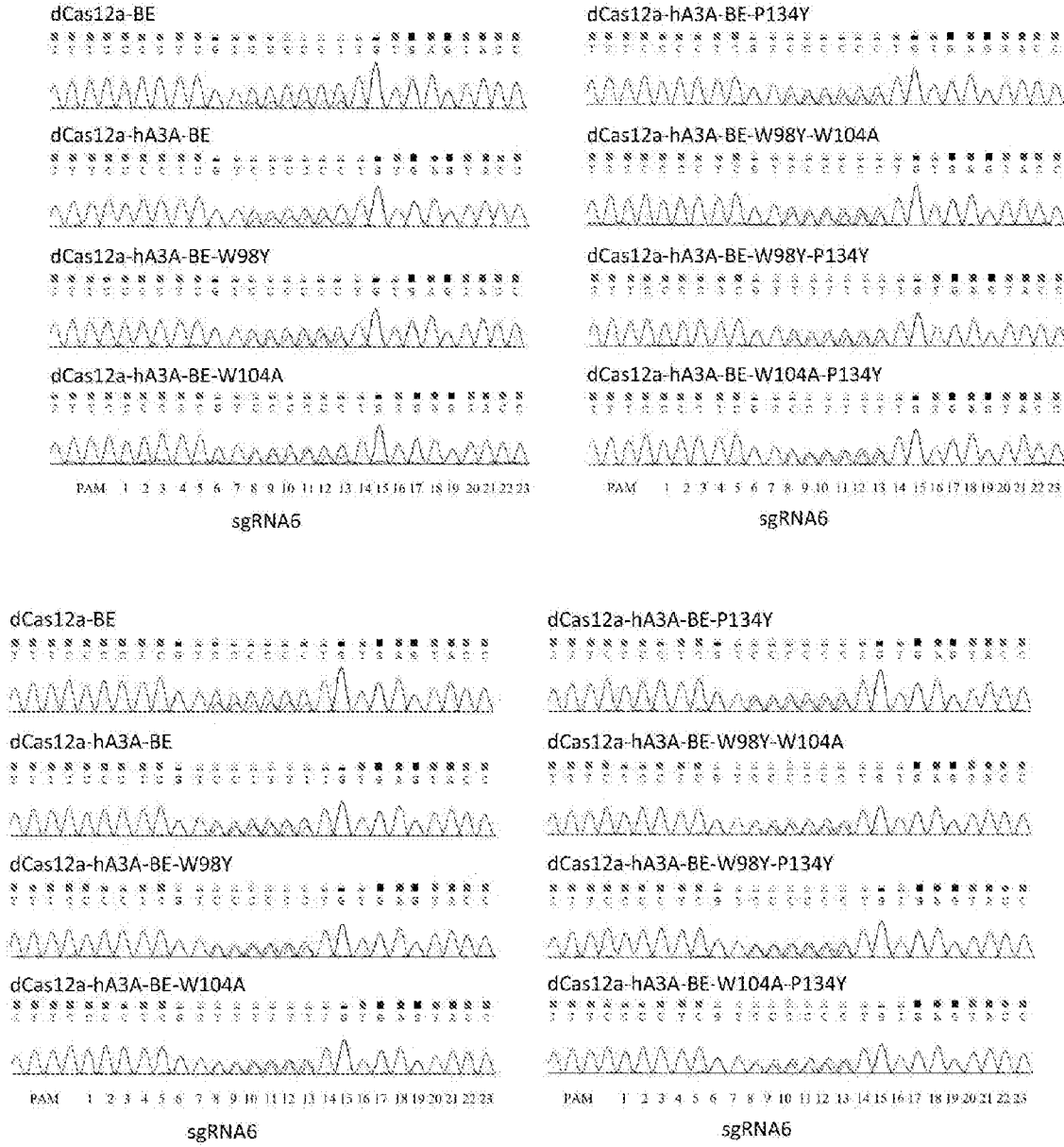


FIG. 15

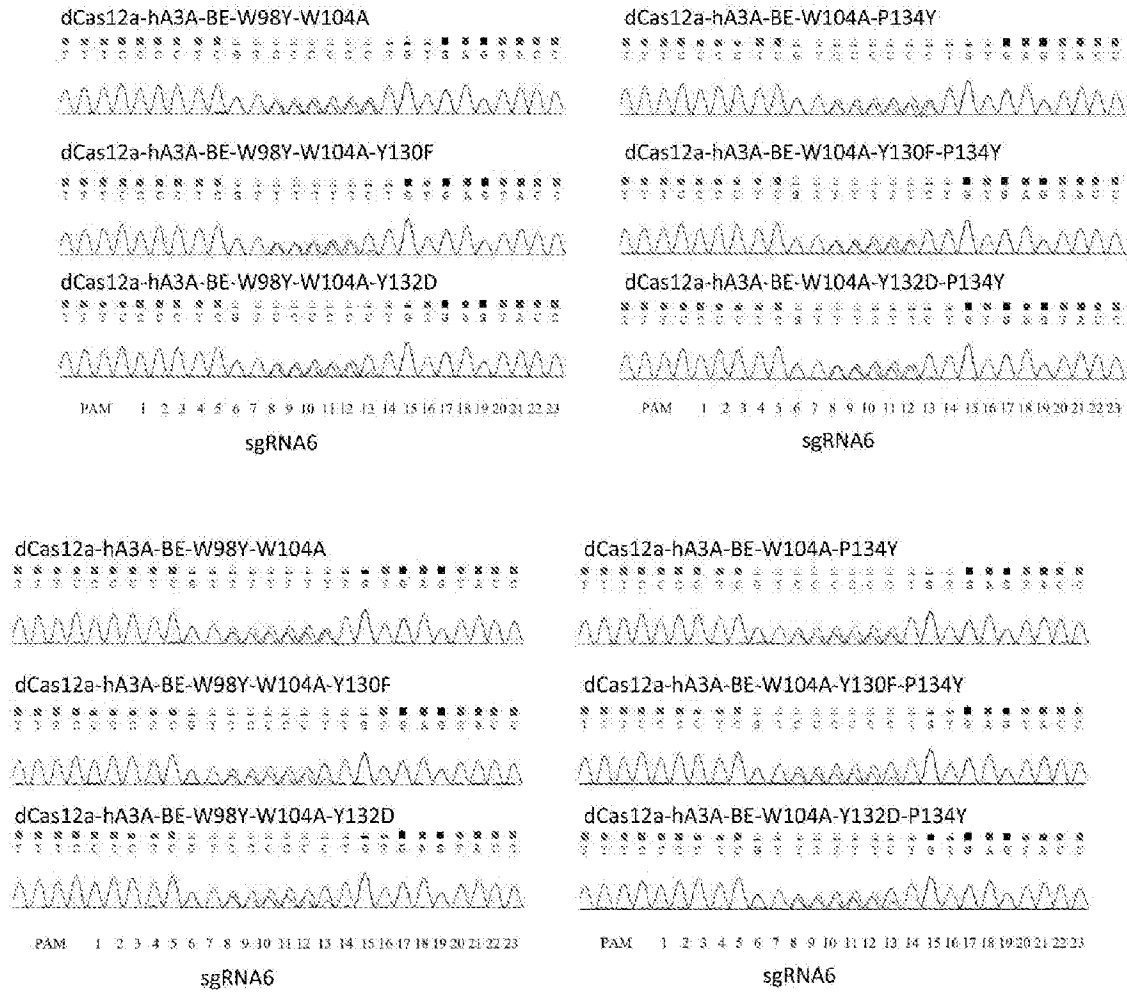


FIG. 16

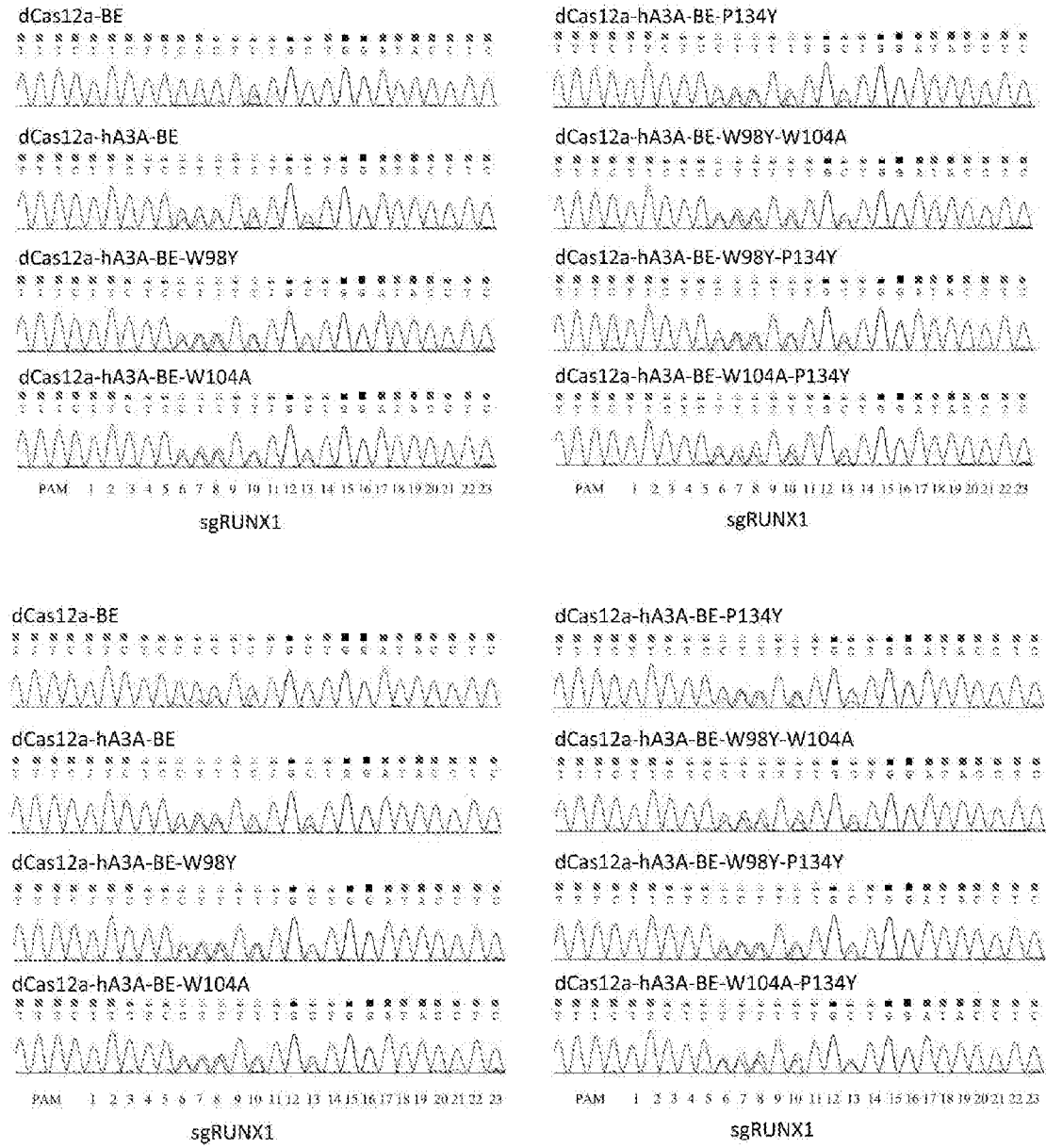


FIG. 17

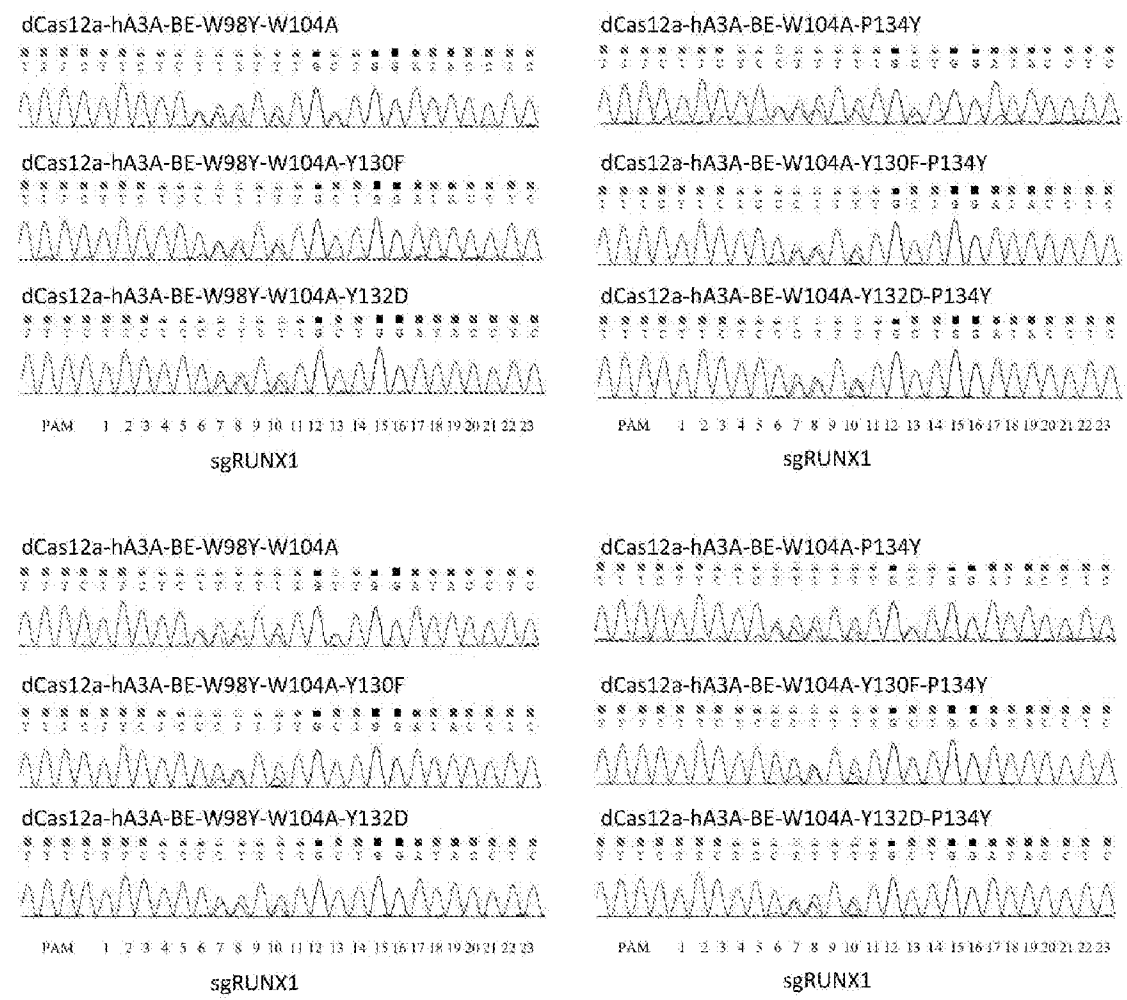


FIG. 18

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/075897

A. CLASSIFICATION OF SUBJECT MATTER

C07K 19/00(2006.01)i; C12N 15/62(2006.01)i; C12N 9/22(2006.01)i; C12N 15/11(2006.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C07K, C12N

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

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

DWPI, SIPOABS, CPRSABS, CPEA, TWABS, CNABS, JPABS, CNTXT, USTXT, WOTXT, EPTXT, CNKI, WANFANG, ISI Web of Science, Elsevier Science, GenBank:Cas, APOBEC3A, CRISPR, cdf1, UGI, mutant, base editing, guide RNA, gRNA, NLS, linker, search on SEQ ID NOs: 1-14, 16-20, 22-36, 40-50

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2018010516 A1 (CHEN QIHAN) 18 January 2018 (2018-01-18) see the description, page 2, lines 5-18 and 23-25, page 18, APOBEC3A, SEQ ID NOs: 202-206	1, 3, 4, 6-13, 16-34
Y	WO 2018010516 A1 (CHEN QIHAN) 18 January 2018 (2018-01-18) see the description, page 2, lines 5-18 and 23-25, page 18, APOBEC3A, SEQ ID NOs: 202-206	2, 14, 15
Y	Lei LQ et al. "APOBEC3 induces mutations during repair of CRISPR-Cas9-generated DNA breaks" <i>Nature Structural & Molecular Biology</i> , Vol. 25, No. 1, 11 December 2017 (2017-12-11), pages 45-52, see page 50, the Discussion	2, 14, 15
X	Jason MG et al. "An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities" <i>Nature Biotechnology</i> , Vol. 36, No. 10, 30 July 2018 (2018-07-30), pages 977-986, see page 977, left column, lines 6-10 and line 30 to right column, line 8, Figure 1a and 1b	4, 6-8, 27-34

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

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

15 May 2019

Date of mailing of the international search report

03 June 2019

Name and mailing address of the ISA/CN

National Intellectual Property Administration, PRC
6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088
China

Authorized officer

LI, Yang

Facsimile No. (86-10)62019451

Telephone No. 62411031

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/075897

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	Jason MG et al. "An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities" <i>Nature Biotechnology</i> , Vol. 36, No. 10, 30 July 2018 (2018-07-30), pages 977-986, see page 977, left column, lines 6-10 and line 30 to right column, line 8, Figure 1a and 1b	1-3, 9-21
PX	Wang X et al. "Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion" <i>Nature Biotechnology</i> , Vol. 36, No. 10, 20 August 2018 (2018-08-20), pages 946-949, see page 946, left column, line 21 to right column, line 6, Figure 1	1-4, 6-34
PX	Zong Y et al. "Efficient C-to-T base editing in plants using a fusion of nCas9 and human APOBEC3A" <i>Nature Biotechnology</i> , Vol. 36, No. 10, 01 October 2018 (2018-10-01), pages 950-958, see page 950, left column, line 42 to right column, line 10, Figure 1	1-4, 6-21, 27-34
A	WO 2016164889 A1 (HEALTH RESEARCH INC) 13 October 2016 (2016-10-13) see the whole document	1-34

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/075897

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
WO	2018010516	A1	18 January 2018	CN	109477086	A	15 March 2019
WO	2016164889	A1	13 October 2016	US	2018092889	A1	05 April 2018