INHIBITION OF UNINTENDED MUTATIONS IN GENE EDITING

BACKGROUND

[0001] 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 applications in life sciences researches, biotechnology/agricultural technology development and most importantly pharmaceutical/clinical innovations. For example, genome editing can be used to correct the driver mutations underlying genetic diseases and leading to complete cure of these diseases in living organisms. Genome editing can also be used to engineer the genome of crops, increasing the yield of crops and conferring crops resistance to environmental contamination or pathogen infection. In addition, microbial genome transformation through accurate genome editing is of great significance in the development of renewable bio-energy.

[0002] 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 organisms. Directed by a guide RNA (gRNA), the 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.

[0003] In general, two major DNA repair pathways can 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, leading to open reading frame (ORF) shift and ultimately gene inactivation. In contrast, when HDR is triggered, the genomic DNA sequence at the target site can 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).

[0004] Base editors (BE), which integrate the CRISPR/Cas system with the APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) AID (activation-induced cytidine deaminase) family, was recently developed that greatly enhanced the efficiency of CRISPR/Cas-meditated gene correction. Through the fusion with the Cas9 nickase (nCas9) or a catalytically dead Cpf1 (dCpf1 also known as dCas12a), the cytosine (C) deamination activity of APOBEC/AID family members can be purposely directed to the target bases in the genome and to catalyze C to Thymine (T) substitutions at these bases.

[0005] However, as APOBEC/AID family members can induce C-to-T base substitution mutations in single-stranded DNA (ssDNA) regions, the specificity of the current base editing system is compromised, thereby limiting the applications, for instance, using BEs to restore the T-to-C mutation that result in human diseases for therapeutic purposes. Hence, creating novel BEs that can specifically edit cytosines in target region but not cause C-to-T mutations in other ssDNA regions is desirable. Such novel BEs will enable us to perform more specific base editing in various living organisms. Importantly, the high specificity of such BEs will promote the potential clinical translation, particularly in the gene therapies that involve restoring disease-related T-to-C mutations.

SUMMARY

[0006] The present disclosure, in some embodiments, provides base editors useful for genome editing that cause reduced or even no off-target mutations common to current base editors. In certain embodiments, a nucleobase deaminase inhibitor is cleavably fused to a nucleobase deaminase involved for genome editing. In the presence of the nucleobase deaminase inhibitor, the nucleobase deaminase is unable to (is less able to) react with a nucleotide molecule. At a target editing location, the nucleobase deaminase inhibitor can be cleaved releasing a fully active nucleobase deaminase that can then carry out the editing as desired.

[0007] Accordingly, in one embodiment, provided is a fusion protein comprising: a first fragment comprising a nucleobase deaminase or a catalytic domain thereof, a second

fragment comprising a nucleobase deaminase inhibitor, and a protease cleavage site between the first fragment and the second fragment.

[0008] In some embodiments, the nucleobase deaminase is an adenosine deaminase. In some embodiments, the adenosine deaminase is selected from the group consisting of tRNA-specific adenosine deaminase (TadA), adenosine deaminase tRNA specific 1 (ADAT1), adenosine deaminase tRNA specific 2 (ADAT2), adenosine deaminase tRNA specific 3 (ADAT3), adenosine deaminase RNA specific B1 (ADARB1), adenosine deaminase RNA specific B2 (ADARB2), adenosine monophosphate deaminase 1 (AMPD1), adenosine monophosphate deaminase 2 (AMPD2), adenosine monophosphate deaminase 3 (AMPD3), adenosine deaminase (ADA), adenosine deaminase 2 (ADA2), adenosine deaminase like (ADAL), adenosine deaminase domain containing 1 (ADAD1), adenosine deaminase domain containing 2 (ADAD2), adenosine deaminase RNA specific (ADAR) and adenosine deaminase RNA specific B1 (ADARB1).

[0009] In some embodiments, the nucleobase deaminase is a cytidine deaminase. In some embodiments, the cytidine deaminase is selected from the group consisting of APOBEC3B (A3B), APOBEC3C (A3C), APOBEC3D (A3D), APOBEC3F (A3F), APOBEC3G (A3G), APOBEC3H (A3H), APOBEC1 (A1), APOBEC3 (A3), APOBEC2 (A2), APOBEC4 (A4) and AICDA (AID). In some embodiments, the cytidine deaminase is a human or mouse cytidine deaminase. In some embodiments, the catalytic domain is mouse A3 cytidine deaminase domain 1 (CDA1) or human A3B cytidine deaminase domain 2 (CDA2).

[0010] In some embodiments, the nucleobase deaminase inhibitor is an inhibitory domain of a nucleobase deaminase. In some embodiments, the nucleobase deaminase inhibitor is an inhibitory domain of a is a cytidine deaminase. In some embodiments, the nucleobase deaminase inhibitor is an inhibitory domain of an adenosine deaminase. In some embodiments, the nucleobase deaminase inhibitor comprises an amino acid sequence selected from SEQ ID NO: 1-2 and Tables 1 and 2 (SEQ ID NO: 48-135), or an amino acid sequence having at least 85% sequence identity to any of the amino acid sequence selected from SEQ ID NO: 1-2 and Tables 1 and 2. In some embodiments, the nucleobase deaminase inhibitor comprises the amino acid sequence of SEQ ID NO:1, amino acids residues AA76-AA149 of SEQ ID NO:1 or the amino acid sequence of SEQ ID NO:2.

[0011] In some embodiments, the first fragment further comprises a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein. 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, xSpCas9, SpCas9-NG, 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, and RanCas13b.

[0012] In some embodiments, the protease cleavage site is a protease cleavage site a protease selected from the group consisting of TuMV protease, PPV protease, PVY protease, ZIKV protease and WNV protease.

[0013] In some embodiments, the protease cleavage site is a self-cleavage site. In some embodiments, the protease cleavage site is a TEV protease cleavage site. In some embodiments, the fusion protein further comprises a third fragment comprising a TEV protease or a fragment thereof. In some embodiments, the third fragment comprises a TEV protease fragment which alone is not able to cleave the TEV protease cleavage site.

[0014] Also provided, in another embodiment, is a fusion protein comprising: a first fragment comprising a cytidine deaminase or a catalytic domain thereof, a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, and a first TEV protease fragment, a second fragment comprising a cytidine deaminase inhibitor, and a TEV protease cleavage site between the first fragment and the second fragment, wherein the first TEV protease fragment alone is not able to cleave the TEV protease cleavage site.

[0015] In some embodiments, the fusion protein further comprises a uracil glycosylase inhibitor (UGI). In some embodiments, the cytidine deaminase inhibitor, the TEV protease cleavage site, the cytidine deaminase or a catalytic domain thereof, the Cas protein, and the first TEV protease fragment are arranged from the N-terminus to the C-terminus. In some embodiments, the first TEV protease fragment is the N-terminal domain (SEQ ID NO:3) or the C-terminal domain (SEQ ID NO:4) of the TEV protease. In some embodiments, the TEV protease cleavage site has the amino acid sequence of SEQ ID NO:5.

[0016] Further provided, in one embodiment, is a method for conducting genomic editing in a cell at a target site, comprising introducing to the cell: (a) a fusion protein of the present

disclosure, (b) a guide RNA that targets the target site or a crRNA that targets the target site and a tracrRNA, and further comprises a tag sequence, and (c) a second TEV protease fragment coupled to an RNA recognition peptide that is able to bind the tag sequence.

[0017] In some embodiments, one or more of the molecules is introduced to the cell by a polynucleotide encoding the molecule. In some embodiments, the first TEV protease fragment and the second TEV protease fragment, when in interaction, are able to cleave the TEV protease cleavage site. In some embodiments, the second TEV protease fragment is fused to the RNA recognition peptide.

[0018] In some embodiments, the tag sequence comprises a MS2 sequence (SEQ ID NO:16). In some embodiments, the RNA recognition peptide comprises a MS2 coat protein (MCP, SEQ ID NO:22). In some embodiments, the tag sequence comprises a PP7 sequence (SEQ ID NO:18) and the RNA recognition peptide comprises a PP7 coat protein (PCP, SEQ ID NO: 23), or the tag sequence comprises a boxB sequence (SEQ ID NO:20) and the RNA recognition peptide comprises a boxB coat protein (N22p, SEQ ID NO:24).

[0019] Also provided, in one embodiment, is a kit or package for conducting gene editing, comprising: (a) a fusion protein of the present disclosure, and (b) a second TEV protease fragment coupled to an RNA recognition peptide that is able to bind an RNA sequence.

[0020] Yet another embodiment provides a fusion protein comprising: a first fragment comprising first cytidine deaminase or a catalytic domain thereof, and a second fragment comprising an inhibitory domain of a second cytidine deaminase, wherein the first cytidine deaminase is same as or different from the second cytidine deaminase.

[0021] In another embodiment, a fusion protein is provided comprising a first fragment comprising: a nucleobase deaminase or a catalytic domain thereof, a nucleobase deaminase inhibitor, a first RNA recognition peptide, and a TEV protease cleavage site between the nucleobase deaminase or a catalytic domain thereof and the nucleobase deaminase inhibitor.

[0022] In some embodiments, the fusion protein further comprises a second fragment comprising: a TEV protease fragment which alone is not able to cleave the TEV protease cleavage site, and a second RNA recognition peptide. In some embodiments, the fusion protein further comprises a self-cleavage site between the first fragment and the second fragment.

[0023] In some embodiments, the fusion protein further comprises a third fragment comprising a second TEV protease fragment, wherein the first TEV protease fragment is able to cleave the TEV protease site in the presence of the second TEV protease fragment. In some embodiments, the fusion protein further comprises a second self-cleavage site between the second fragment and the third fragment, ad upon cleavage of the second self-cleavage site, the fusion protein releases the second TEV protease fragment which is not fused to any RNA recognition peptide.

[0024] Also provided, in one embodiment, is a dual guide RNA system, comprising: a target single guide RNA comprising a first spacer having sequence complementarity to a target nucleic acid sequence proximate to a first PAM site, a helper single guide RNA comprising a second spacer having sequence complementarity to a second nucleic acid sequence proximate to a second PAM site, a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, and a nucleobase deaminase, wherein the second PAM site is from 34 to 91 bases from the first PAM site. In some embodiments, the second spacer is 8-15 bases in length. In some embodiments, the second spacer is 9-12 bases in length.

[0025] In one embodiment, provided is guide RNA comprising a scaffold which comprises, from the 5' to 3' direction, a first stem loop portion, a second stem loop portion, a third stem loop portion, and a fourth stem loop portion, wherein the third stem loop comprises five base pairings within. In another embodiment, the present disclosure provides a guide RNA comprising a scaffold derived from SEQ ID NO:31 by introducing a base pairing between the bases at positions 45 and 55. In some embodiments, the scaffold comprises a sequence selected from the group consisting of SEQ ID NO:32-43. In some embodiments, the guide RNA is at least 100, or 120 nucleotides in length.

[0026] Another embodiment provides a method for conducting genetic editing in a cell at a target site, comprising introducing to the cell: a first viral particle enclosing a first construct encoding a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, and a second viral particle enclosing a second construct encoding a reverse-transcriptase fused to an RNA recognition peptide.

[0027] In some embodiments, the second construct further encodes a guide RNA comprising an RNA recognition site that the RNA recognition peptide binds to. In some embodiments, the Cas protein is SpCas9-NG (SEQ ID NO:46) or xSpCas9 (SEQ ID NO:47).

[0028] Polynucleotides encoding the fusion proteins of the present disclosure, constructs containing the polynucleotides, cells containing the polynucleotides or the constructs, and compositions comprising any of the above are also provided, without limitation.

BRIEF DESCRIPTION OF THE DRAWINGS

[0029] FIG. 1A-C: Unintended base substitutions caused by current BEs in Sa-SITE31 ssDNA region. 1A: Schematic diagram illustrating the co-expression of SaD10A nickase and Sa-sgSITE31 to trigger the formation of ssDNA region at Sa-sgSITE31 on-target site. **1B**: Schematic diagram illustrating the co-transfection of the plasmid expressing Sa-sgSITE31 and the plasmid expressing SaD10A nickase with the plasmid expressing BE3, the plasmid expressing hA3A-BE3 or an empty vector. **1C**: The untended base substitutions caused by BE3 and hA3A-BE3. Dashed boxes represent the locations of untended base substitutions at Sa-sgSITE31 target site.

[0030] FIG. 2A-C: Unintended base substitutions caused by current BEs in Sa-SITE42 ssDNA region. 2A: Schematic diagram illustrating the co-expression of SaD10A nickase and Sa-sgSITE42 to trigger the formation of ssDNA region at Sa-sgSITE42 on-target site. **2B**: Schematic diagram illustrating the co-transfection of the plasmid expressing Sa-sgSITE42 and the plasmid expressing SaD10A nickase with the plasmid expressing BE3, the plasmid expressing hA3A-BE3 or an empty vector. **2C**: The untended base substitutions caused by BE3 and hA3A-BE3. Dashed boxes represent the locations of untended base substitutions at Sa-sgSITE42 target site.

[0031] FIG. 3A-C: Unintended base substitutions caused by current BEs in Sa-F1 ssDNA region. 3A: Schematic diagram illustrating the co-expression of SaD10A nickase and Sa-sgF1 to trigger the formation of ssDNA region at Sa-sgF1 on-target site. **3B**: Schematic diagram illustrating the co-transfection of the plasmid expressing Sa-sgF1 and the plasmid expressing SaD10A nickase with the plasmid expressing BE3, the plasmid expressing hA3A-BE3 or an empty vector. **3C**: The untended base substitutions caused by BE3 and hA3A-BE3. Dashed boxes represent the locations of untended base substitutions at Sa-sgF1 target site.

[0032] FIG. 4A-C: mA3CDA2 inhibits C-to-T base editing activity in TET1 region. 4A: Schematic diagram illustrating the regions of CDA domains in mA3, rA1 and hA3A. **4B**: Schematic diagram illustrating the co-transfection of the plasmid expressing sgTET1 with the

plasmid expressing mA3-BE3, the plasmid expressing mA3CDA1-BE3, the plasmid expressing mA3rev-BE3, the plasmid expressing mA3rev-2A-BE3, the plasmid expressing BE3, the plasmid expressing mA3CDA2-BE3, the plasmid expressing mA3CDA2-2A-BE3, the plasmid expressing hA3A-BE3, the plasmid expressing mA3CDA2-hA3A-BE3 or the plasmid expressing mA3CDA2-2A-hA3A-BE3. **4C**: mA3CDA2 inhibits C-to-T editing activity of mA3CDA1-BE3, BE3 and hA3A-BE3. Dashed boxes represent the locations of C-to-T base editing at sgTET1 target site.

[0033] FIG. 5A-C: mA3CDA2 inhibits C-to-T base editing activity in RNF2 region. 5A: Schematic diagram illustrating the regions of CDA domains in mA3, rA1 and hA3A. **5B**: Schematic diagram illustrating the co-transfection of the plasmid expressing sgRNF2 with the plasmid expressing mA3-BE3, the plasmid expressing mA3CDA1-BE3, the plasmid expressing mA3rev-BE3, the plasmid expressing mA3rev-2A-BE3, the plasmid expressing BE3, the plasmid expressing mA3CDA2-BE3, the plasmid expressing mA3CDA2-2A-BE3, the plasmid expressing hA3A-BE3, the plasmid expressing mA3CDA2-hA3A-BE3 or the plasmid expressing mA3CDA2-2A-hA3A-BE3. **5C**: mA3CDA2 inhibits C-to-T editing activity of mA3CDA1-BE3, BE3 and hA3A-BE3. Dashed boxes represent the locations of C-to-T base editing at sgRNF2 target site.

[0034] FIG. 6A-C: mA3CDA2 inhibits C-to-T base editing activity in SITE3 region. 6A: Schematic diagram illustrating the regions of CDA domains in mA3, rA1 and hA3A. **6B**: Schematic diagram illustrating the co-transfection of the plasmid expressing sgSITE3 with the plasmid expressing mA3-BE3, the plasmid expressing mA3CDA1-BE3, the plasmid expressing mA3rev-BE3, the plasmid expressing mA3rev-2A-BE3, the plasmid expressing BE3, the plasmid expressing mA3CDA2-BE3, the plasmid expressing mA3CDA2-2A-BE3, the plasmid expressing hA3A-BE3, the plasmid expressing mA3CDA2-hA3A-BE3 or the plasmid expressing mA3CDA2-2A-hA3A-BE3. **6C**: mA3CDA2 inhibits C-to-T editing activity of mA3CDA1-BE3, BE3 and hA3A-BE3. Dashed boxes represent the locations of C-to-T base editing at sgSITE3 target site.

[0035] FIG. 7A-C: hA3BCDA1 inhibits C-to-T base editing activity in TET1 region. 7A: Schematic diagram illustrating the regions of CDA domains in hA3B. **7B**: Schematic diagram illustrating the co-transfection of the plasmid expressing sgTET1 with the plasmid expressing hA3B-BE3, the plasmid expressing hA3BCDA2-BE3 or the plasmid expressing hA3B-2A-

BE3. **7C**: hA3BCDA1 inhibits C-to-T editing activity of hA3BCDA2-BE3. Dashed boxes represent the locations of C-to-T base editing at sgTET1 target site.

[0036] FIG. 8A-C: hA3BCDA1 inhibits C-to-T base editing activity in RNF2 region. 8A: Schematic diagram illustrating the regions of CDA domains in hA3B. **8B**: Schematic diagram illustrating the co-transfection of the plasmid expressing sgRNF2 with the plasmid expressing hA3B-BE3, the plasmid expressing hA3BCDA2-BE3 or the plasmid expressing hA3B-2A-BE3. **8C**: hA3BCDA1 inhibits C-to-T editing activity of hA3BCDA2-BE3. Dashed boxes represent the locations of C-to-T base editing at sgRNF2 target site.

[0037] FIG. 9A-C: hA3BCDA1 inhibits C-to-T base editing activity in SITE3 region. 9A: Schematic diagram illustrating the regions of CDA domains in hA3B. **9B**: Schematic diagram illustrating the co-transfection of the plasmid expressing sgSITE3 with the plasmid expressing hA3B-BE3, the plasmid expressing hA3BCDA2-BE3 or the plasmid expressing hA3B-2A-BE3. **9C**: hA3BCDA1 inhibits C-to-T editing activity of hA3BCDA2-BE3. Dashed boxes represent the locations of C-to-T base editing at sgSITE3 target site.

[0038] FIG. 10A-C: Mapping the split site of mA3 by examining base editing efficiency in FANCF region. 10A: Schematic diagram illustrating the regions of two CDA domains in mA3 and the sites (AA196/AA197, AA207/AA208, AA215/AA216, AA229/AA230, AA237/AA238) used to split mA3. **10B**: Schematic diagram illustrating the co-transfection of the plasmid expressing sgFANCF with the plasmid expressing mA3rev-BE3-196, the plasmid expressing mA3rev-2A-BE3-196, the plasmid expressing mA3rev-BE3, the plasmid expressing mA3rev-2A-BE3, the plasmid expressing mA3rev-BE3-215, the plasmid expressing mA3rev-2A-BE3-215, the plasmid expressing mA3rev-BE3-229, the plasmid expressing mA3rev-2A-BE3-229, the plasmid expressing mA3rev-BE3-237 or the plasmid expressing mA3rev-2A-BE3-237. **10C**: The splitting sites spanning from AA196/AA197 to AA237/AA238 generally keep C-to-T editing efficiencies. Dashed boxes represent the locations of C-to-T base editing at sgFANCF target site.

[0039] FIG. 11A-C: Mapping the split site of mA3 by examining base editing efficiency in SITE2 region. 11A: Schematic diagram illustrating the regions of two CDA domains in mA3 and the sites (AA196/AA197, AA207/AA208, AA215/AA216, AA229/AA230, AA237/AA238) used to split mA3. **11B**: Schematic diagram illustrating the co-transfection of the plasmid expressing sgSITE2 with the plasmid expressing mA3rev-BE3-196, the plasmid

expressing mA3rev-2A-BE3-196, the plasmid expressing mA3rev-BE3, the plasmid expressing mA3rev-2A-BE3, the plasmid expressing mA3rev-BE3-215, the plasmid expressing mA3rev-2A-BE3-215, the plasmid expressing mA3rev-BE3-229, the plasmid expressing mA3rev-2A-BE3-229, the plasmid expressing mA3rev-BE3-237 or the plasmid expressing mA3rev-2A-BE3-237. **11C**: The splitting sites spanning from AA196/AA197 to AA237/AA238 generally keep C-to-T editing efficiencies. Dashed boxes represent the locations of C-to-T base editing at sgSITE2 target site.

[0040] FIG. 12A-C: Mapping the split site of mA3 by examining base editing efficiency in SITE4 region. 12A: Schematic diagram illustrating the regions of two CDA domains in mA3 and the sites (AA196/AA197, AA207/AA208, AA215/AA216, AA229/AA230, AA237/AA238) used to split mA3. **12B**: Schematic diagram illustrating the co-transfection of the plasmid expressing sgSITE4 with the plasmid expressing mA3rev-BE3-196, the plasmid expressing mA3rev-2A-BE3-196, the plasmid expressing mA3rev-BE3, the plasmid expressing mA3rev-2A-BE3, the plasmid expressing mA3rev-BE3-215, the plasmid expressing mA3rev-2A-BE3-215, the plasmid expressing mA3rev-BE3-229, the plasmid expressing mA3rev-2A-BE3-229, the plasmid expressing mA3rev-BE3-237 or the plasmid expressing mA3rev-2A-BE3-237. **12C**: The splitting sites spanning from AA196/AA197 to AA237/AA238 generally keep C-to-T editing efficiencies. Dashed boxes represent the locations of C-to-T base editing at sgSITE4 target site.

[0041] FIG. 13A-B: Mapping the minimal region of mA3 that contains the base-editing inhibitory effect in FANCF region. 13A: Schematic diagram illustrating the co-transfection of the plasmid expressing sgFANCF with the plasmid expressing mA3rev-BE3-237, the plasmid expressing mA3rev-BE3-237-Del-255, the plasmid expressing mA3rev-BE3-237-Del-285 or the plasmid expressing mA3rev-BE3-237-Del-333. **13B**: The region spanning from AA334 to AA429 of mA3 contains the base-editing inhibitory effect. Dashed boxes represent the locations of C-to-T base editing at sgFANCF target site.

[0042] FIG. 14A-B: Mapping the minimal region of mA3 that contains the base-editing inhibitory effect in SITE2 region. 14A: Schematic diagram illustrating the co-transfection of the plasmid expressing sgSITE2 with the plasmid expressing mA3rev-BE3-237, the plasmid expressing mA3rev-BE3-237-Del-255, the plasmid expressing mA3rev-BE3-237-Del-285 or the plasmid expressing mA3rev-BE3-237-Del-333. **14B**: The region spanning

from AA334 to AA429 of mA3 contains the base-editing inhibitory effect. Dashed boxes represent the locations of C-to-T base editing at sgSITE2 target site.

[0043] FIG. 15A-B: Mapping the minimal region of mA3 that contains the base-editing inhibitory effect in SITE4 region. 15A: Schematic diagram illustrating the co-transfection of the plasmid expressing sgSITE4 with the plasmid expressing mA3rev-BE3-237, the plasmid expressing mA3rev-BE3-237-Del-255, the plasmid expressing mA3rev-BE3-237-Del-285 or the plasmid expressing mA3rev-BE3-237-Del-333. **15B**: The region spanning from AA334 to AA429 of mA3 contains the base-editing inhibitory effect. Dashed boxes represent the locations of C-to-T base editing at sgSITE4 target site.

[0044] FIG. 16A-B: Schematic diagram illustrating the working process of BEsafe and BE3 or hA3A-BE3. 16A: BEsafe induces C-to-T base editing at on-target site and avoids causing mutations in non-relevant ssDNA regions. **16B**: BE3 or hA3A-BE3 induces C-to-T base editing at on-target site but causes C-to-T mutations in non-relevant ssDNA regions.

[0045] FIG. 17A-D: Comparison of hA3A-BE3 and BEsafe in non-relevant Sa-SITE31 ssDNA region and at TET1 on-target site. 17A: Schematic diagram illustrating the co-expression of SaD10A nickase and Sa-sgSITE31 to trigger the formation of ssDNA region at Sa-sgSITE31 on-target site. **17B**: Schematic diagram illustrating the co-transfection of the plasmid expressing Sa-sgSITE31 and the plasmid expressing SaD10A nickase with the plasmid expressing hA3A-BE3 and the plasmid expressing sgTET1, with the plasmid expressing BEsafe and the plasmid expressing MS2-sgTET1 and MCP-TEVc or with the plasmid expressing MCP-TEVc and the plasmid expressing MS2-sgTET1 and BEsafe. **17C**: Comparing the untended C-to-T mutation frequencies triggered by hA3A-BE3 and BEsafe in the non-relevant Sa-SITE31 ssDNA region. Dashed boxes represent the locations of untended base substitutions at Sa-sgSITE31 target site. **17D**: Comparing the base editing efficiencies of hA3A-BE3 and BEsafe at TET1 site. Dashed boxes represent the locations of C-to-T base editing at sgTET1 target site.

[0046] FIG. 18A-D: Comparison of hA3A-BE3 and BEsafe in non-relevant Sa-SITE32 ssDNA region and at RNF2 on-target site. 18A: Schematic diagram illustrating the co-expression of SaD10A nickase and Sa-sgSITE32 to trigger the formation of ssDNA region at Sa-sgSITE32 on-target site. **18B**: Schematic diagram illustrating the co-transfection of the plasmid expressing Sa-sgSITE32 and the plasmid expressing SaD10A nickase with the

plasmid expressing hA3A-BE3 and the plasmid expressing sgRNF2, with the plasmid expressing BEsafe and the plasmid expressing MS2-sgRNF2 and MCP-TEVc or with the plasmid expressing MCP-TEVc and the plasmid expressing MS2-sgRNF2 and BEsafe. **18C**: Comparing the untended C-to-T mutation frequencies triggered by hA3A-BE3 and BEsafe in the non-relevant Sa-SITE32 ssDNA region. Dashed boxes represent the locations of untended base substitutions at Sa-sgSITE32 target site. **18D**: Comparing the base editing efficiencies of hA3A-BE3 and BEsafe at RNF2 site. Dashed boxes represent the locations of C-to-T base editing at sgRNF2 target site.

[0047] FIG. 19A-D: Comparison of hA3A-BE3 and BEsafe in non-relevant Sa-F1 ssDNA region and at SITE3 on-target site. 19A: Schematic diagram illustrating the co-expression of SaD10A nickase and Sa-sgF1 to trigger the formation of ssDNA region at Sa-sgF1 on-target site. **19B**: Schematic diagram illustrating the co-transfection of the plasmid expressing Sa-sgF1 and the plasmid expressing SaD10A nickase with the plasmid expressing hA3A-BE3 and the plasmid expressing sgSITE3, with the plasmid expressing BEsafe and the plasmid expressing MS2-sgSITE3 and MCP-TEVc or with the plasmid expressing MCP-TEVc and the plasmid expressing MS2-sgSITE3 and BEsafe. **19C**: Comparing the untended C-to-T mutation frequencies triggered by hA3A-BE3 and BEsafe in the non-relevant Sa-F1 ssDNA region. Dashed boxes represent the locations of untended base substitutions at Sa-sgF1 target site. **19D**: Comparing the base editing efficiencies of hA3A-BE3 and BEsafe at SITE3 site. Dashed boxes represent the locations of C-to-T base editing at sgSITE3 target site.

[0048] FIG. 20a-f: Identification of cytidine deaminase inhibitors. 20a: Schematic diagrams illustrate the APOBEC family members that have single- or dual- CDA domains (left) and paired base editors that were constructed with one or two CDAs of dual-domain APOBECs (right). **20b**: Editing frequencies induced by the indicated BEs at one representative genomic locus. **20c**: Statistical analysis of normalized editing frequencies, setting the ones induced by the single-CDA-containing BEs as 100%. n = 78 from three independent experiments at 26 editable cytosine sites shown in (b). **20d**: Schematic diagrams illustrate the conjugation of different cytidine deaminase inhibitors (CDIs) to the N-terminus of mA3CDA1-nSpCas9-BE. **20e**: Editing frequencies induced by the indicated BEs at one representative genomic locus. **20f**: Statistical analysis of normalized editing frequencies, setting the ones induced by the BEs without CDI as 100%. n = 57 from three independent

experiments at 19 editable cytosine sites shown in (e). (b), (e) Means ± s.d. were from three independent experiments. NT, non-transfected control. (c), (f) *P* value, one-tailed Student's *t* test. The median and interquartile range (IQR) are shown.

[0049] FIG. 21a-f: Conjugation of mA3CDI reduced unintended base editing at sgRNA-independent OTss sites. 21a: Schematic diagrams illustrate that BE3 induces C-to-T mutations but CDI-conjugated iBE1 remains dormant at sgRNA-independent OTss sites. **21b**: Comparison of C-to-T editing frequencies induced by BE3 and iBE1 in the ssDNA regions triggered by the nSaCas9-generated SSBs. **21c**: Statistical analysis of normalized accumulative editing frequencies at four ssDNA sites shown in (b), setting the ones induced by BE3 as 100%. n = 12 from three independent experiments. **21d**: Schematic diagrams illustrate that the sgRNA-mediated cleavage of CDI restores the editing activity of iBE at on-target sites. **21e**: Comparison of C-to-T editing frequencies induced by BE3 and iBE1 at on-target sites. **21f**: Statistical analysis of normalized accumulative editing frequencies at four on-target sites shown in (e), setting the ones induced by BE3 as 100%. n = 12 from three independent experiments. (c), (f) Means ± s.d. were from three independent experiments. (d), (g) *P* value, one-tailed Student's *t* test. The median and interquartile range (IQR) are shown.

[0050] FIG. 22a-e: neSpCas9 reduced the unintended editing of iBE1 at OTsg sites. 22a: Schematic diagrams illustrate that iBE1 but not iBE2 induces C-to-T editing at the OTsg sites that are partially complementary to sgRNAs. **22b**: Comparison of C-to-T editing frequencies induced by iBE1 and the targeting-specificity-improved iBEs at indicated OTsg sites. **22c**: Statistical analysis of normalized accumulative editing frequencies at OTsg sites for two sgRNAs used in (b), setting the ones induced by iBE1 as 100%. n = 6 from three independent experiments. **22d**: Comparison of C-to-T editing frequencies induced by iBE1 and the targeting-specificity-improved iBEs at on-target sites. **22e**: Statistical analysis of normalized accumulative editing frequencies at the six on-target sites shown in (d), setting the ones induced by iBE1 as 100%. n = 18 from three independent experiments. (b), (d) Means ± s.d. were from three independent experiments. (c), (e) *P* value, one-tailed Student's *t* test. The median and interquartile range (IQR) are shown.

[0051] FIG. 23a-e: Comparison of the base editing induced by hA3A-BE3 and iBE2. 23a: Comparison of C-to-T editing frequencies induced by hA3A-BE3 and iBE2 at representative OTss, OTsg and on-target sites. **23b-c**: Statistical analysis of normalized accumulative editing frequencies at the OTss, OTsg (b) and on-target (c) sites for three

sgRNAs used in (a), setting the ones induced by hA3A-BE3 as 100%. n = 9 from three independent experiments. **23d**: Statistical analysis of the normalized ratios of on-target editing frequencies to the total editing frequencies at OTss and OTsg sites for three sgRNAs used in (a), setting the ones induced by the hA3A-BE3 as 1. n = 9 from three independent experiments. **23e**: Schematic diagrams illustrate that iBE2 induces specific base editing at on-target sites but not at OTss or OTsg sites, whereas hA3A-BE3 induces base editing at on-target sites and both OTss and OTsg sites. (a) Means ± s.d. were from three independent experiments. (b-d) *P* value, one-tailed Student's *t* test. The median and IQR are shown.

[0052] FIG. 24A-B. Schematic diagram illustrating the working process of isplitBE and regular base editors. 24A: isplitBE induces C-to-T base editing only at on-target site and avoids to cause mutations in non-relevant off-target ssDNA regions (OTss) or at the off-target sites with sequence similarity to the spacer region of sgRNA (OTsg). **24B**: BE3 or hA3A-BE3 induces C-to-T base editing at on-target site but causes C-to-T mutations in OTss and OTsg regions.

[0053] FIG. 25. Schematic diagram illustrating different strategies to remove cytidine deaminase inhibitor (mA3CDA2) at on-target site.

[0054] FIG. 26A-B. C-to-T editing at EMX1-ON, Sa-SITE31-OTss and EMX1-OTsg sites induced by different combinations of nCas9 (D10A), APOBEC cytidine deaminase, cytidine deaminase inhibitor (CDI), uracil DNA glycosylase inhibitor (UGI) and TEV protease. 26A: Schematic diagram illustrating the co-transfection of the plasmid expressing Sa-sgSITE31 and the plasmid expressing SaD10A nickase with the indicated ten pairs of plasmid expressing various base editors. **26B**: Comparison of editing efficiencies at EMX1-ON, Sa-SITE31-OTss and EMX1-OTsg sites. isplitBE-rA1 (Pair 9) induced substantial editing at ON site but no editing at OTss or OTsg sites.

[0055] FIG. 27A-B. C-to-T editing at FANCF-ON, Sa-VEGFA-7-OTss and FANCF-OTsg sites induced by different combinations of nCas9 (D10A), APOBEC cytidine deaminase, cytidine deaminase inhibitor (CDI), uracil DNA glycosylase inhibitor (UGI) and TEV protease. 27A: Schematic diagram illustrating the co-transfection of the plasmid expressing Sa-sgVEGFA-7 and the plasmid expressing SaD10A nickase with the indicated ten pairs of plasmid expressing various base editors. **27B**: Comparison of editing efficiencies

at FANCF-ON, Sa-VEGFA-7-OTss and FANCF-OTsg sites. isplitBE-rA1 (Pair 9) induced substantial editing at ON site but no editing at OTss or OTsg sites.

[0056] FIG. 28A-B. C-to-T editing at V1B-ON, Sa-SITE42-OTss and V1B-OTsg sites induced by different combinations of nCas9 (D10A), APOBEC cytidine deaminase, cytidine deaminase inhibitor (CDI), uracil DNA glycosylase inhibitor (UGI) and TEV protease. 28A: Schematic diagram illustrating the co-transfection of the plasmid expressing Sa-sgSITE42 and the plasmid expressing SaD10A nickase with the indicated ten pairs of plasmid expressing various base editors. **28B**: Comparison of editing efficiencies at V1B-ON, Sa-SITE42-OTss and V1B-OTsg sites. isplitBE-rA1 (Pair 9) induced substantial editing at ON site but no editing at OTss or OTsg sites.

[0057] FIG. 29A-C. Effect of the distance between helper sgRNA (hsgRNA) and sgRNA on base editing efficiency. 29A: Schematic diagram illustrating the distance between hsgRNA and sgRNA at DNTET1, EMX1 and FANCF sites. **29B**: Base editing frequencies induced by the indicated sgRNAs and hsgRNAs. **29C**: Summary of the effect of distance between hsgRNA and sgRNA. The range of distance for best base editing efficiency is -91 to -34 bp from the PAM of hsgRNA to the PAM of sgRNA.

[0058] FIG. 30A-C. Effect of hsgRNA spacer length on base editing efficiency. 30A: Schematic diagram illustrating the co-transfection of sgRNA and the hsgRNAs with different spacer lengths at DNEMX1, FANCF and V1A sites. **30B**: Base editing frequencies induced by the indicated sgRNAs and hsgRNAs at the target sties of hsgRNA and sgRNA. **30C**: Statistic analysis of the effects of hsgRNA spacer length. The use of hsgRNA with 10-bp spacer greatly reduce the editing efficiency at hsgRNA target sites but maintain the editing efficiency at sgRNA target sites.

[0059] FIG. 31. Comparison of editing efficiency of isplitBE-rA1 and BE3. Editing frequencies induced by indicated base editors at different target sties.

[0060] FIG. 32A-C. Comparison of genome-wide C-to-T mutations induced by isplitBE-rA1 and BE3. 32A: mRNA expression levels in wild-type 293FT cells and the APOBEC3 knockout 293FT cells (293FT-A3KO). **32B:** Schematic diagram illustrating the procedures to determine genome-wide C-to-T mutations induced by base editors. **32C:** On-target editing efficiencies (left) and the number of genome-wide C-to-T mutations induced by Cas9, BE3, hA3A-BE3-Y130F (Y130F) and isplitBE-rA1.

[0061] FIG. 33A-C. Comparison of transcriptome-wide C-to-U mutations induced by isplitBE-mA3, BE3 and hA3A-BE3-Y130F (Y130F). 33A: The number of transcriptome-wide C-to-U mutations induced by Cas9, BE3, hA3A-BE3-Y130F (Y130F) and isplitBE-mA3. **33B:** RNA C-to-U editing frequencies induced by Cas9, BE3, hA3A-BE3-Y130F (Y130F) and isplitBE-mA3. **33C:** Distribution of RNA C-to-U editing induced by BE3 replicate 1 and isplitBE-mA3 replicate 1.

[0062] FIG. 34A-D. Stop codon induced by isplitBE-mA3 in human PCSK9 gene. 34A: Schematic diagram illustrating the co-transfection of sgRNA and the hsgRNAs with isplitBE-mA3 and nCas9. **34B-34D:** Editing efficiency induced by isplitBE-mA3 at indicated sites.

[0063] FIG. 35A-B. Inhibiting effect of mA3CDA2 on the editing efficiency of adenine base editor (ABE). 35A: Schematic diagram illustrating the co-transfection of sgRNA and ABE fused with mA3CDA2 or not. **35B:** Editing efficiency induced by indicated ABEs at RNF2 and FANCF sites.

[0064] FIG. 36A-G. Enhanced prime editing by manipulating prime editing guide RNA (pegRNA). 36A: Schematic diagram illustrating the change of RNA base pairs to increase stem stability of enhanced pegRNA (epegRNA). **36B:** Schematic diagram illustrating the co-transfection of PE2, nicking sgRNA with pegRNA or epegRNA-GC. **36C-36D:** Comparison of prime editing efficiencies induced with pegRNA and epegRNA-GC. **36E:** Schematic diagram illustrating the change of RNA base pairs to increase stem stability of enhanced pegRNA (epegRNA). **36F:** Schematic diagram illustrating the co-transfection of PE2, nicking sgRNA with pegRNA or epegRNA-CG. **36G:** Comparison of prime editing efficiencies induced with pegRNA and epegRNA-CG.

[0065] FIG. 37A-B. Prime editing system by using PEs containing different Cas9 proteins. 37A: Schematic diagram illustrating the co-transfection of pegRNA, nicking sgRNA with PE2-NG or xPE2. **37B:** Prime editing efficiencies induced by PE2-NG and xPE2.

[0066] FIG. 38A-C. Split prime editing (split-PE) system. 38A: Schematic diagram illustrating the working process of PE and split-PE systems. **38B:** Schematic diagram illustrating the co-transfection of PE and split-PE systems. **38C:** Editing efficiency induced by PE and split-PE systems at EMX1 site.

[0067] FIG. 39A-C. Alignment of mA3CDA2 core region with other cytidine deaminase domains.

[0068] FIG. 40A-D. Alignment of hA3BCDA1 with other cytidine deaminase domains.

DETAILED DESCRIPTION

Definitions

[0069] 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.

[0070] 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.

[0071] "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.

[0072] 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.

[0073] 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.

[0074] 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.

Use of Nucleobase Deaminase Inhibitor to Reduce Random Insertions and Deletions

[0075] As shown in the experimental examples and **FIG. 1-3**, the currently commonly used base editors BE3 and hA3A-BE3 induced C-to-T mutations in off-target single-stranded DNA regions.

[0076] It was discovered surprisingly, however, that the use of the mouse APOBEC3 (mA3) in mA3-BE3 (**FIG. 4B, 5B, 6B**) generally did not induce C-to-T editing at the tested target sites (**FIG. 4C, 5C, 6C**). mA3 has two cytidine deaminase (CDA) domains, CDA1 and CDA2 (**FIG. 4A, 5A, 6A**). When the CDA2 domain was removed from the full-length mA3, the resulting base editor mA3CDA1-BE3 (**FIG. 4B, 5B, 6B**) induced substantial C-to-T editing (**FIG. 4C, 5C, 6C**). These results indicate that the mA3-CDA2 domain is an inhibitor of base editing.

[0077] Also surprisingly, the mA3-CDA2 domain not only can inhibit the base editing activity of mA3-CDA1, it can also inhibit other nucleobase deaminases. For instance, when mA3-CDA2 was fused to the N-terminus of each of three active BEs, mA3CDA1-BE3, BE3 and hA3A-BE3, the fusion proteins mA3rev-BE3, mA3-CDA2-BE3 and mA3-CDA2-hA3A-BE3 (**FIG. 4B, 5B, 6B**) had clearly reduced base editing efficiencies (**FIG. 4C, 5C, 6C**).

[0078] Moreover, cleavage of mA3-CDA2 from the fusion proteins restored the base editing efficiency (**FIG. 4C, 5C, 6C**), suggesting that the inhibition of mA3-CDA2 is associated with its covalent connection to the BEs.

[0079] Like mA3, the human APOBEC3B (hA3B) also has two cytidine deaminase (CDA) domains, CDA1 and CDA2 (**FIG. 7A, 8A, 9A**). Incorporation of the full-length hA3B in hA3B-BE3 (**FIG. 7B, 8B, 9B**) only induced relatively low levels of C-to-T editing at three tested target sites (**FIG. 7C, 8C, 9C**). However, hA3B-CDA2-BE3, which was generated by deleting the hA3B-CDA1 domain (**FIG. 7B, 8B, 9B**) induced higher C-to-T editing (**FIG. 7C, 8C, 9C**). These results indicate that hA3B-CDA1 is another inhibitor of base editing and the inhibition of hA3B-CDA1 is associated with its covalent connection to the BEs.

[0080] Using the sequences of mA3-CDA2 and hA3B-CDA1, the instant inventors were able to identify additional nucleobase deaminase inhibitors/domains in the protein database. Table 1 shows 44 proteins/domains that have significant sequence homology to mA3-CDA2 core sequence (**FIG. 39**) and Table 2 shows 43 proteins/domains that have significant sequence

homology to hA3B-CDA1 (**FIG. 40**). All of these proteins and domains, as well as their variants and equivalents, are contemplated to have nucleobase deaminase inhibition activities.

Fusion Proteins

[0081] Based on these surprising and expected findings, a fusion protein is designed that can be used to generate a base editor with improved base editing specificity and efficiency. In one embodiment, the present disclosure provides a fusion protein that includes a first fragment comprising a nucleobase deaminase or a catalytic domain thereof, a second fragment comprising a nucleobase deaminase inhibitor, and a protease cleavage site between the first fragment and the second fragment.

[0082] A base editor that incorporates such a fusion protein has reduced or even no editing capability and accordingly will generate reduced or no off-target mutations. Upon cleavage of the protease cleavage site and release of the nucleobase deaminase inhibitor from the fusion protein at a target site, the base editor that is at the target site will then be able to edit the target site efficiently.

[0083] The term "nucleobase deaminase" as used herein, refers to a group of enzymes that catalyze the hydrolytic deamination of nucleobases such as cytidine, deoxycytidine, adenosine and deoxyadenosine. Non-limiting examples of nucleobase deaminases include cytidine deaminases and adenosine deaminases.

[0084] "Cytidine deaminase" refers to enzymes that catalyze the irreversible hydrolytic deamination of cytidine and deoxycytidine to uridine and deoxyuridine, respectively. Cytidine deaminases maintain the cellular pyrimidine pool. A family of cytidine deaminases is APOBEC ("apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like"). Members of this family are C-to-U editing enzymes. Some APOBEC family members have two domains, one domain of APOBEC like proteins is the catalytic domain, while the other domain is a pseudocatalytic domain. More specifically, the catalytic domain is a zinc dependent cytidine deaminase domain and is important for cytidine deamination. RNA editing by APOBEC-1 requires homodimerisation and this complex interacts with RNA binding proteins to form the editosome.

[0085] Non-limiting examples of APOBEC proteins include APOBEC1, APOBEC2, APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D, APOBEC3F, APOBEC3G, APOBEC3H, APOBEC4, and activation-induced (cytidine) deaminase (AID).

[0086] Various mutants of the APOBEC proteins are also known that have bring about different editing characteristics for base editors. For instance, for human APOBEC3A, certain mutants (e.g., W98Y, Y130F, Y132D, W104A, D131Y and P134Y) even outperform the wildtype human APOBEC3A in terms of editing efficiency or editing window. Accordingly, the term APOBEC and each of its family member also encompasses variants and mutants that have certain level (e.g., 70%, 75%, 80%, 85%, 90%, 95%, 98%, 99%) of sequence identity to the corresponding wildtype APOBEC protein or the catalytic domain and retain the cytidine deaminating activity. The variants and mutants can be derived with amino acid additions, deletions and/or substitutions. Such substitutions, in some embodiments, are conservative substitutions.

[0087] "Adenosine deaminase", also known as adenosine aminohydrolase, or ADA, is an enzyme (EC 3.5.4.4) involved in purine metabolism. It is needed for the breakdown of adenosine from food and for the turnover of nucleic acids in tissues.

[0088] Non-limiting examples of adenosine deaminases include tRNA-specific adenosine deaminase (TadA), adenosine deaminase tRNA specific 1 (ADAT1), adenosine deaminase tRNA specific 2 (ADAT2), adenosine deaminase tRNA specific 3 (ADAT3), adenosine deaminase RNA specific B1 (ADARB1), adenosine deaminase RNA specific B2 (ADARB2), adenosine monophosphate deaminase 1 (AMPD1), adenosine monophosphate deaminase 2 (AMPD2), adenosine monophosphate deaminase 3 (AMPD3), adenosine deaminase (ADA), adenosine deaminase 2 (ADA2), adenosine deaminase like (ADAL), adenosine deaminase domain containing 1 (ADAD1), adenosine deaminase domain containing 2 (ADAD2), adenosine deaminase RNA specific (ADAR) and adenosine deaminase RNA specific B1 (ADARB1).

[0089] Some of the nucleobase deaminases have a single, catalytic domain, while others also have other domains, such as an inhibitory domain as currently discovered by the instant inventors. In some embodiments, therefore, the first fragment only includes the catalytic domain, such as mA3-CDA1 and hA3B-CDA2. In some embodiments, the first fragment includes at least a catalytic core of the catalytic domain. For instance, as demonstrated in the

experimental examples, when mA3-CDA1 was truncated at residues 196/197 the CDA1 domain still retained substantial editing efficiencies (**FIG. 10C, 11C, 12C**).

[0090] The present disclosure tested two nucleobase deaminase inhibitors, mA3-CDA2 and hA3B-CDA1, which are the inhibitory domains of the corresponding nucleobase deaminase. Additional nucleobase deaminase inhibitors and inhibitory domains were also identified in the protein databases (see Tables 1 and 2). Their biological equivalents (e.g., having at least about 80%, 85%, 90%, 95%, 97%, 98%, 99%, 99.5% sequence identity, or having one, two, or three amino acid addition/deletion/substitution, and having nucleobase deaminase inhibitor activity) can also be prepared with known methods in the art, such as conservative amino acid substitutions. A "nucleobase deaminase inhibitor," accordingly, refers to a protein or a protein domain that inhibits the deaminase activity of a nucleobase deaminase. In some embodiments, the second fragment includes at least an inhibitory core of the inhibitory protein/domain. For instance, as demonstrated in the experimental examples, when mA3-CDA2 retained residues 334-429, the CDA2 still had the inhibitory effect of base editing (**FIG. 13B, 14B, 15B**).

[0091] In some embodiments, the fusion protein further includes a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, optionally in the first fragment, next to the nucleobase deaminase or the catalytic domain thereof.

[0092] The term "Cas protein" or "clustered regularly interspaced short palindromic repeats (CRISPR)-associated (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. Cas proteins include Cas9 proteins, Cas12a (Cpf1) proteins, Cas12b (formerly known as C2c1) proteins, Cas13 proteins and various engineered counterparts. Example Cas proteins include SpCas9, FnCas9, St1Cas9, St3Cas9, NmCas9, SaCas9, AsCpf1, LbCpf1, FnCpf1, VQR SpCas9, EQR SpCas9, VRER SpCas9, SpCas9-NG, xSpCas9, 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 and those provided in **Table A** below.

Table A. Example Cas Proteins

Cas protein types	Cas proteins
Cas9 proteins	Cas9 from *Staphylococcus aureus* (SaCas9) Cas9 from *Neisseria meningitidis* (NmeCas9) Cas9 from *Streptococcus thermophilus* (StCas9) Cas9 from *Campylobacter jejuni* (CjCas9)
Cas12a (Cpf1) proteins	Cas12a (Cpf1) from *Acidaminococcus* sp BV3L6 (AsCpf1) Cas12a (Cpf1) from *Francisella novicida* sp BV3L6 (FnCpf1) Cas12a (Cpf1) from *Smithella* sp SC K08D17 (SsCpf1) Cas12a (Cpf1) from *Porphyromonas crevioricanis* (PcCpf1) Cas12a (Cpf1) from *Butyrivibrio proteoclasticus* (BpCpf1) Cas12a (Cpf1) from *Candidatus Methanoplasma termitum* (CmtCpf1) Cas12a (Cpf1) from *Leptospira inadai* (LiCpf1) Cas12a (Cpf1) from *Porphyromonas macacae* (PmCpf1) Cas12a (Cpf1) from *Peregrinibacteria bacterium GW2011 WA2 33 10* (Pb3310Cpf1) Cas12a (Cpf1) from *Parcubacteria bacterium GW2011 GWC2 44 17* (Pb4417Cpf1) Cas12a (Cpf1) from *Butyrivibrio sp. NC3005* (BsCpf1) Cas12a (Cpf1) from *Eubacterium eligens* (EeCpf1)
Cas12b (C2c1) proteins	Cas12b (C2c1) *Bacillus hisashii* (BhCas12b) Cas12b (C2c1) *Bacillus hisashii* with a gain-of-function mutation (see, e.g., Strecker et al., Nature Communications 10 (article 212) (2019) Cas12b (C2c1) *Alicyclobacillus kakegawensis* (AkCas12b) Cas12b (C2c1) *Elusimicrobia bacterium* (EbCas12b) Cas12b (C2c1) *Laceyella sediminis* (Ls) (LsCas12b)
Cas13 proteins	Cas13d from *Ruminococcus flavefaciens* XPD3002 (RfCas13d) Cas13a from *Leptotrichia wadei* (LwaCas13a) Cas13b from *Prevotella* sp. P5-125 (PspCas13b) Cas13b from *Porphyromonas gulae* (PguCas13b) Cas13b from *Riemerella anatipestifer* (RanCas13b)
Engineered Cas proteins	Nickases (mutation in one nuclease domain) Catalytically inactive mutant (dCas9; mutations in both of the nuclease domains) Enhanced variants with improved specificity (see, e.g., Chen *et al.*, *Nature*, 550, 407-410 (2017)

[0093] The protease cleavage site between the first fragment and the second fragment can be any known protease cleavage site (peptide) for any proteases. Non-limiting examples of proteases include TEV protease, TuMV protease, PPV protease, PVY protease, ZIKV protease and WNV protease. The protein sequences of example proteases and their corresponding cleavage sites are provided in **Table B**.

Table B. Example Sequences

Name	Sequence	SEQ ID NO:
Mouse APOBEC3 cytidine deaminase domain 2	MSSSTLSNICLTKGLPETRFWVEGRRMDPLSEEEFYSQFYNQRVKHLCY YHRMKPYLCYQLEQFNGQAPLKGCLLSEKGKQHAEILFLDKIRSMELSQ VTITCYLTWSPCPNCAWQLAAFKRDRPDLILHIYTSRLYFHWKRPFQKG LCSLWQSGILVDVMDLPQFTDCWTNFVNPKRPFWPWKGLEIISRRTQRR LRRIKESWGLQDLVNDFGNLQLGPPMS	1
Human APOBEC3B cytidine deaminase domain 1	MNPQIRNPMERMYRDTFYDNFENEPILYGRSYTWLCYEVKIKRGRSNLL WDTGVFRGQVYFKPQYHAEMCFLSWFCGNQLPAYKCFQITWFVSWTPCP DCVAKLAEFLSEHPNVTLTISAARLYYYWERDYRRALCRLSQAGARVKI MDYEEFAYCWENFVYNEGQ	2
TEV protease N-terminal domain	MGESLFKGPRDYNPISSTICHLTNESDGHTTSLYGIGFGPFIITNKHLF RRNNGTLLVQSLHGVFKVKNTTTLQQHLIDGRDMIIIRMPKDFPPFPQK LKFREPQREERICLVTTNFQT	3
TEV protease C-terminal domain	MKSMSSMVSDTSCTFPSSDGIFWKHWIQTKDGQCGSPLVSTRDGFIVGI HSASNFTNTNNYFTSVPKNFMELLTNQEAQQWVSGWRLNADSVLWGGHK VFMVKPEEPFQPVKEATQ	4
TEV protease cleavage site	ENLYFQS	5
TuMV protease	MASSNSMFRGLRDYNPISNNICHLTNVSDGASNSLYGVGFGPLILTNRH LFERNNGELVIKSRHGEFVIKNTTQLHLLPIPDRDLLLIRLPKDVPPFP QKLGFRQPEKGERICMVGSNFQTKSITSIVSETSTIMPVENSQFWKHWI STKDGQCGSPMVSTKDGKILGLHSLANFQNSINYFAAFPDDFAEKYLHT IEAHEWVKHWKYNTSAISWGSLNIQASQPSGLFKVSKLISDLDSTAVYA Q	6
TuMV protease cleavage site	GGCSHQS	7
PPV protease	MASSKSLFRGLRDYNPIASSICQLNNSSGARQSEMFGLGFGGLIVTNQH LFKRNDGELTIRSHHGEFVVKDTKTLKLLPCKGRDIVIIRLPKDFPPFP RRLQFRTPTTEDRVCLIGSNFQTKSISSTMSETSATYPVDNSHFWKHWI STKDGHCGLPIVSTRDGSILGLHSLANSTNTQNFYAAFPDNFETTYLSN QDNDNWIKQWRYNPDEVCWGSLQLKRDIPQSPFTICKLLTDLDGEFVYT Q	8
PPV protease cleavage site	QVVVHQSK	9
PVY protease	MASAKSLMRGLRDFNPIAQTVCRLKVSVEYGASEMYGFGFGAYIVANHH LFRSYNGSMEVQSMHGTFRVKNLHSLSVLPIKGRDIILIKMPKDFPVFP QKLHFRAPTQNERICLVGTNFQEKYASSIITETSTTYNIPGSTFWKHWI ETDNGHCGLPVVSTADGCIVGIHSLANNAHTTNYYSAFDEDFESKYLRT NEHNEWVKSWVYNPDTVLWGPLKLKDSTPKGLFKTTKLVQDLIDHDVVV EQ	10
PVY protease cleavage site	YDVRHQSR	11
ZIKV protease	MASDMYIERAGDITWEKDAEVTGNSPRLDVALDESGDFSLVEEDGPPMR EGGGGSGGGGSGALWDVPAPKEVKKGETTDGVYRVMTRRLLGSTQVGVG VMQEGVFHTMWHVTKGAALRSGEGRLDPYWGDVKQDLVSYCGPWKLDAA WDGLSEVQLLAVPPGERARNIQTLPGIFKTKDGDIGAVALDYPAGTSGS PILDKCGRVIGLYGNGVVIKNGSYVSAITQGKREEETPVECFE	12
ZIKV protease cleavage site	KERKRRGA	13
WNV protease	MASSTDMWIERTADISWESDAEITGSSERVDVRLDDDGNFQLMNDPGAP WKGGGGSGGGGGVLWDTPSPKEYKKGDTTTGVYRIMTRGLLGSYQAGAG VMVEGVFHTLWHTTKGAALMSGEGRLDPYWGSVKEDRLCYGGPWKLQHK	14

	WNGQDEVQMIVVEPGKNVKNVQTKPGVFKTPEGEIGAVTLDFPTGTSGS PIVDKNGDVIGLYGNGVIMPNGSYISAIVQGERMDEPIPAGFEPEML	
WNV protease cleavage site	KQKKRGGK	15
MS2	ACAUGAGGAUCACCCAUGU	16
sgRNA scaffold with 2×MS2	GUUUGAGAGCUAGGCCA**ACAUGAGGAUCACCCAUGU**CUGCAGGGCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGCCA**ACAUGAGGAUC ACCCAUGU**CUGCAGGGCCAAGUGGCACCGAGUCGGUGC	17
PP7	GGAGCAGACGAUAUGGCGUCGCUCC	18
sgRNA scaffold with 2×PP7	GUUUGAGAGCUACCGGAGCAGACGAUAUGGCGUCGCUCCGGUAGCAAGU UCAAAUAAGGCUAGUCCGUUAUCAACUUGGAGCAGACGAUAUGGCGUCG CUCCAAGUGGCACCGAGUCGGUGC	19
boxB	GCCCUGAAGAAGGGC	20
sgRNA scaffold with 2×boxB	GUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAGCAAGUUCAAAUAAGG CUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAAGUGGCACCGA GUCGGUGC	21
MS2 coat protein (MCP)	MASNFTQFVLVDNGGTGDVTVAPSNFANGIAEWISSNSRSQAYKVTCSV RQSSAQNRKYTIKVEVPKGAWRSYLNMELTIPIFATNSDCELIVKAMQG LLKDGNPIPSAIAANSGIY	22
PP7 coat protein (PCP)	MGSKTIVLSVGEATRTLTEIQSTADRQIFEEKVGPLVGRLRLTASLRQN GAKTAYRVNLKLDQADVVDSGLPKVRYTQVWSHDVTIVANSTEASRKSL YDLTKSLVATSQVEDLVVNLVPLGR	23
boxB coat protein (N22p)	MGNARTRRRERRAEKQAQWKAAN	24
UGI	TNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDES TDENVMLLTSDAPEYKPWALVIQDSNGENKIKML	25
P2A	GSGATNFSLLKQAGDVEENPGP	26
T2A	GSGEGRGSLLTCGDVEENPGP	27
E2A	GSGQCTNYALLKLAGDVESNPGP	28

[0094] In some embodiments, the protease cleavage site is a self-cleaving peptide, such as the 2A peptides. “2A peptides” are 18-22 amino-acid-long viral oligopeptides that mediate “cleavage” of polypeptides during translation in eukaryotic cells. The designation “2A” refers to a specific region of the viral genome and different viral 2As have generally been named after the virus they were derived from. The first discovered 2A was F2A (foot-and-mouth disease virus), after which E2A (equine rhinitis A virus), P2A (porcine teschovirus-1 2A), and T2A (thosea asigna virus 2A) were also identified. A few non-limiting examples of 2A peptides are provided in SEQ ID NO:26-28.

[0095] In some embodiments, the protease cleavage site is a cleavage site (e.g., SEQ ID NO:5) for the TEV protease. In some embodiments, the fusion protein further includes a third fragment that includes the TEV protease or a fragment thereof. In some embodiments, the

TEV protease fragment in the fusion protein in not active, that, is, not able to cleave the TEV cleavage site on its own. However, in the presence of the remaining portion of the TEV protease, this fragment will be able to execute the cleavage. As further described below, such an arrangement provides additional control and flexible of the base editing capabilities. The TEV fragment may be the TEV N-terminal domain (e.g., SEQ ID NO:3) or the TEV C-terminal domain (e.g., SEQ ID NO:4).

[0096] Various arrangement of the fragments can be made. Non-limiting examples include, from N-terminal side to C-terminal side:

(1) first fragment (e.g., catalytic domain) – protease cleavage site – second fragment (e.g., inhibitory domain);

(2) first fragment (e.g., catalytic domain and Cas protein) – protease cleavage site – second fragment (e.g., inhibitory domain);

(3) first fragment (e.g., catalytic domain, Cas protein and TEV N-terminal domain) – protease cleavage site (e.g., TEV cleavage site) – second fragment (e.g., inhibitory domain);

(4) second fragment (e.g., inhibitory domain) – protease cleavage site (e.g., TEV cleavage site) – first fragment (e.g., catalytic domain, Cas protein and TEV N-terminal domain); and

(5) second fragment (e.g., inhibitory domain) – protease cleavage site (e.g., TEV cleavage site) – first fragment (e.g., Cas protein, catalytic domain, and TEV C-terminal domain).

[0097] In some embodiments, provided are fusion proteins comprising a first fragment comprising first nucleobase deaminase (e.g., cytidine deaminase) or a catalytic domain thereof, and a second fragment comprising an inhibitory domain of a second nucleobase deaminase, wherein the first nucleobase deaminase is different from the second nucleobase deaminase. In some embodiments, each of the first and second nucleobase deaminases is independently selected from the group of human and mouse APOBEC3B (A3B), APOBEC3C (A3C), APOBEC3D (A3D), APOBEC3F (A3F), APOBEC3G (A3G), APOBEC3H (A3H), APOBEC1 (A1), APOBEC3 (A3), APOBEC2 (A2), APOBEC4 (A4) and AICDA (AID).

[0098] The fusion proteins may include other fragments, such as uracil DNA glycosylase inhibitor (UGI) and nuclear localization sequences (NLS).

[0099] The “Uracil Glycosylase Inhibitor” (UGI), which can be prepared from *Bacillus subtilis* bacteriophage PBS1, is a small protein (9.5 kDa) which inhibits *E. coli* uracil-DNA glycosylase (UDG) as well as UDG from other species. Inhibition of UDG occurs by reversible protein binding with a 1:1 UDG:UGI stoichiometry. UGI is capable of dissociating UDG-DNA complexes. 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:25 or has at least at least 70%, 75%, 80%, 85%, 90% or 95% sequence identity to SEQ ID NO:25 and retains the uracil glycosylase inhibition activity.

[0100] The fusion protein, in some embodiments, may include one or more nuclear localization sequences (NLS).

[0101] A “nuclear localization signal or sequence” (NLS) is an amino acid sequence that tags a protein for import into the cell nucleus by nuclear transport. Typically, this signal consists of one or more short sequences of positively charged lysines or arginines exposed on the protein surface. Different nuclear localized proteins may share the same NLS. An NLS has the opposite function of a nuclear export signal (NES), which targets proteins out of the nucleus. A non-limiting example of NLS is the internal SV40 nuclear localization sequence (iNLS).

[0102] In some embodiments, a peptide linker is optionally 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.

[0103] For any fusion protein of the present disclosure, biological equivalents thereof are also provided. In some embodiments, the biological equivalents have at least about 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity with the reference fusion protein. Preferably, the biological equivalents retained the desired activity of the reference fusion protein. In some embodiments, the biological equivalents are derived by including one, two, three, four, five or more amino acid additions, deletions, substitutions, of the combinations thereof. In some embodiments, the substitution is a conservative amino acid substitution.

[0104] 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.

[0105] 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 C. 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 D. 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

On-Target Activation of Fusion Proteins

[0106] The present disclosure also provides compositions and methods in which a fusion protein of the present disclosure, which includes both a nucleobase deaminase or the catalytic domain thereof and an inhibitor, is activated where its activity is desired. The technology is illustrated in **FIG. 16**.

[0107] In an illustrative configuration, the fusion protein (A) includes (a) a first fragment comprising a nucleobase deaminase (e.g., cytidine deaminase) or a catalytic domain thereof, optionally with a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, and a first TEV protease fragment, (b) a second fragment comprising a nucleobase deaminase inhibitor, and (c) a TEV protease cleavage site between the first fragment and the second fragment. In some embodiments, wherein the first TEV protease fragment alone is not able to cleave the TEV protease cleavage site.

[0108] When the fusion protein is used, *in vitro* or *in vivo*, to conduct gene editing in a cell, two additional molecules can be introduced. In one example, one molecule (B) is a single guide RNA (sgRNA) that further incorporates a tag sequence that can be recognized by an RNA recognition peptide. The sgRNA, alternatively, can be replaced by a crRNA that targets the target site and a CRISPR RNA (crRNA) alone, or in combination with a trans-activating CRISPR RNA (tracrRNA). Examples of tag sequences and corresponding RNA recognition

peptides include MS2/MS2 coat protein (MCP), PP7/PP7 coat protein (PCP), and boxB/boxB coat protein (N22p), the sequences of which are provided in **Table B**. The molecule (B) may be provided as a DNA sequence encoding the RNA molecule.

[0109] The other additional molecule (C), in some embodiments, includes a second TEV protease fragment coupled to the RNA recognition peptide (e.g., MCP, PCP, N22p). The first TEV fragment and the second TEV fragment, in some embodiments, when present together, are able to cleave a TEV protease site.

[0110] Such co-presence can be triggered by the molecule (C) binding to the molecule (B) by virtue of the tag sequence-RNA recognition protein interaction. Meanwhile, the fusion protein (A) and the molecule (B) will be both present at the target genome locus for gene editing. Therefore, the molecule (B) brings both of the TEV protease fragments from the fusion protein (A) and molecule (C) together, which will activate the TEV protease, leading to removal of the nucleobase deaminase inhibitor from the fusion protein and activation of the base editor. It can be readily appreciated that such activation only occurs at the target genome site, not at off-target single-stranded DNA regions. As such, base editing does not occur at the single-stranded DNA regions that sgRNA does not bind to (as demonstrated in **FIG. 17-19**).

[0111] "Guide RNAs" are non coding short RNA sequences which bind to the complementary target DNA sequences. A guide RNA first binds to the Cas enzyme and the gRNA sequence guides the complex via pairing to a specific location on the DNA, where Cas performs its endonuclease activity by cutting the target DNA strand. A "single guide RNA," frequently simply referred to as "guide RNA", refers to synthetic or expressed single guide RNA (sgRNA) that consists of both the crRNA and tracrRNA as a single construct. The tracrRNA portion is responsible for Cas endonuclease activity and the crRNA portion binds to the target specific DNA region. Therefore, the trans activating RNA (tracrRNA, or scaffold region) and crRNA are two key components and are joined by tetraloop which results in formation of sgRNA.

[0112] The scaffold of the guide RNA has a stem-loop structure in itself and attaches to the endonuclease enzyme. A typical scaffold has a structure as illustrated in FIG. 36A (upper), which includes, from the 5' to the 3' end, (a) a repeat region, (b) a tetraloop, (c) an anti-repeat that is at least partially complementary to the repeat region, (d) stem loop 1, (e) a linker, (f)

stem loop 2, and (g) stem loop 3. The scaffold sequence is generally conserved, but the loops in stem loop 1 and stem loop 3 can have different sequences. More importantly, the loops of tetraloop and stem loop 2 can be entirely replaced with even much longer sequences. Sequences such as RNA tags (e.g., MS2, PP7, boxB) can be inserted here, enabling recognition by the corresponding recognition peptides. Example scaffold sequences are shown below.

Table E. Example sgRNA Scaffold Sequences

SEQ ID NO:	Sequence
29	GUUUUAGAGCUA**GAAA**UAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUU**GAAA**AAGUGGCACCGAGUCGGUGC
30	GUUUUAGAGCUA**GAAA**UAGCAAGUUAAAAUAAGGCAUGUCCGUUAUCAACUU**GAAA**AAGUGGCACCGAUUCGGUGC
196	GUUUGAGAGCUA**GAAA**UAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUU**GAAA**AAGUGGCACCGAGUCGGUGC

[0113] With reference to these example scaffold sequences, the fragment of positions 1-12 (e.g., GUUUUAGAGCUA, SEQ ID NO:197; GUUUGAGAGCUA, SEQ ID NO:198) represents the repeat region, which forms about 8-12 base pairings with the anti-repeat, which includes positions 17-30 (e.g., UAGCAAGUUAAAAU, SEQ ID NO:199). The GAAA loop (SEQ ID NO:200) between them is the tetraloop. As shown in SEQ ID NO:17, this entire loop can be replaced with a MS2 sequence. Stem loop 1 roughly includes positions 31-39 and includes a small loop (e.g., UA, AU, AA, or UU, without limitation). Stem loop 1 generally has 3-4 base parings in the stem. Stem loop 2, including positions 48-61 (e.g., AACUUGAAAAAGUG, SEQ ID NO:201), generally includes 4 base parings in the stem, and a GAAA (SEQ ID NO:200) loop which can be totally replaced. The remaining, positions 62-76 (e.g., GCACCGAGUCGGUGC, SEQ ID NO:202; GCACCGAUUCGGUGC; SEQ ID NO:203) constitute stem loop 3, which generally includes 4 base pairings in the stem. The small loop (U and G here in the example) can be any nucleotide.

[0114] Accordingly, the sequence of the scaffold can be expressed as: GUUUNAGAGCUA**X**$_1$UAGCAAGUUNAAAUAAGGCNNGUCCGUUAUCAACUU**X**$_2$ AAGUGGCACCGANUCGGUGC (SEQ ID NO:31), where N represents any base, and X1 and X2 denotes any nucleotide sequence of a length of 2-50 bases. The terms "guide RNA" and "single guide RNA" encompasses those that include additional sequences, such as MS2, PP7 and boxB, inserted into one or more loops in the RNA.

[0115] Various embodiments and examples of nucleobase deaminases, catalytic domains, nucleobase deaminase inhibitors, and Cas proteins are provided in the disclosures. For instance, a nucleobase deaminase can be a cytidine deaminases and adenosine deaminases.

Non-limiting examples of cytidine deaminases include APOBEC1, APOBEC2, APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D, APOBEC3F, APOBEC3G, APOBEC3H, APOBEC4, and activation-induced (cytidine) deaminase.

[0116] Non-limiting examples of adenosine deaminases include tRNA-specific adenosine deaminase (TadA), adenosine deaminase tRNA specific 1 (ADAT1), adenosine deaminase tRNA specific 2 (ADAT2), adenosine deaminase tRNA specific 3 (ADAT3), adenosine deaminase RNA specific B1 (ADARB1), adenosine deaminase RNA specific B2 (ADARB2), adenosine monophosphate deaminase 1 (AMPD1), adenosine monophosphate deaminase 2 (AMPD2), adenosine monophosphate deaminase 3 (AMPD3), adenosine deaminase (ADA), adenosine deaminase 2 (ADA2), adenosine deaminase like (ADAL), adenosine deaminase domain containing 1 (ADAD1), adenosine deaminase domain containing 2 (ADAD2), adenosine deaminase RNA specific (ADAR) and adenosine deaminase RNA specific B1 (ADARB1).

[0117] Example Cas proteins include SpCas9, FnCas9, St1Cas9, St3Cas9, NmCas9, SaCas9, AsCpf1, LbCpf1, FnCpf1, VQR SpCas9, EQR SpCas9, VRER SpCas9, RHA FnCas9, and KKH SaCas9 and those provided in **Table A**.

[0118] The fusion proteins may include other fragments, such as uracil DNA glycosylase inhibitor (UGI) and nuclear localization sequences (NLS), each of which is discussed herein.

[0119] The base editors and base editing methods described in this disclosure can be applied to perform high-specificity and high-efficiency base editing in the genome of various eukaryotes.

[0120] The present disclosure 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.

[0121] 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.

[0122] In one embodiment, provided is a method of editing a cytosine on a nucleic acid sequence in a sample. In some embodiments, the method entails contacting the sample a fusion protein of the present disclosure, or a polynucleotide encoding the fusion protein. In some embodiments, further added is a suitable guide RNA. Design of the guide RNA is readily available to the skilled artisan.

[0123] 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. The in vivo contacting may be administration to a live subject, such as a human, an animal, a yeast, a plant, a bacterium, a virus, without limitation.

Configurations of Induced and Split Base Editors

[0124] Various configurations of constructs have been tested to implement the induced and split base editor (isplitBE) design (FIG. 24). Among the configuration tested (FIG. 25), Pair 9 of Example 3 exhibited superior editing efficiency and minimized off-target editing (greatly improved specificity). Pair 9 employs a dual sgRNA system, in which a helper sgRNA (hsgRNA) is used to target a site proximate the main target site. Such dual targeting improves specificity (FIG. 32-33).

[0125] In configuration Pair 9 (FIG. 25-28), the nucleobase deaminase inhibitor is only released when both sgRNA are bound to the target sequences, ensuring that the nucleobase deaminase does not edit at off-target sites. Configuration Pair 9 includes 6 different molecules, which can be produced from two separate constructs, for instance (FIG. 26A and 34A).

[0126] The first molecule can include just a Cas protein, which has a suitable size for packaging in a common vehicle, AAV. The second molecule includes, among others, a nucleobase deaminase (e.g., APOBEC), a nucleobase deaminase inhibitor (e.g., mA3-CDA2), and an RNA recognition peptide (e.g., MCP). A protease cleavage site (e.g., TEV site) is inserted between the nucleobase deaminase and the nucleobase deaminase inhibitor, which

enables removal of the nucleobase deaminase inhibitor at proper timing/location. Optionally, the second molecule further includes a UGI.

[0127] The third molecule is a fusion between an inactive portion of the protease (e.g., TEVc) fused to different RNA recognition peptide (e.g., N22p). The fourth molecule is a standalone TEVn which, in combination with the first portion, can carry out the protease activity to remove the nucleobase deaminase inhibitor from the second molecule.

[0128] The fifth molecule is a helper sgRNA containing an RNA recognition site (e.g., MS2) recognizable by the RNA recognition peptide in the 2nd molecule. The sixth molecule is a regular sgRNA that contains an RNA recognition site (e.g., boxB) recognizable by the RNA recognition peptide in the 3rd molecule.

[0129] At the correct target site in the genome (or RNA), both the hsgRNA and the sgRNA will bind, and each recruits a Cas protein to the binding site. The hsgRNA will also recruit the 2nd molecule by virtue of the MS2-MCP binding, and the sgRNA will recruit the 3rd molecule by virtue of the boxB-N22p binding. Therefore, the TEVc of the 3rd molecule is in contact with the TEV site. Since the standalone TEVn is present in the entire cell, it can also be present here, which ensures that the TEVc is active and cleaves the nucleobase deaminase inhibitor from the nucleobase deaminase in molecule 2, thereby activating the nucleobase deaminase.

[0130] It is further discovered that an optimal distance between the hsgRNA binding site and the regular sgRNA binding site is from 34-91 bp (from PAM to PAM), with the hsgRNA on the upstream.

[0131] Moreover, even though the proper binding of both hsgRNA and regular sgRNA is required for the intended editing in the target site for regular sgRNA, the editing in the target site for hsgRNA is not desirable. It is discovered herein that when the spacer length of the hsgRNA (spacer is the target complementary region) is from 8-15 bases, such a hsgRNA is still sufficient to provide dual recognition to ensure binding specificity, but greatly reduce the editing in the hsgRNA target site.

[0132] In accordance with one embodiment of the disclosure, therefore, provided is a fusion protein comprising a first fragment comprising: a nucleobase deaminase or a catalytic domain thereof, a nucleobase deaminase inhibitor, a first RNA recognition peptide, and a TEV

protease cleavage site between the nucleobase deaminase or a catalytic domain thereof and the nucleobase deaminase inhibitor.

[0133] In some embodiments, the fusion protein further comprises a second fragment comprising: a TEV protease fragment which alone is not able to cleave the TEV protease cleavage site, and a second RNA recognition peptide. In some embodiments, the fusion protein further comprises a self-cleavage site between the first fragment and the second fragment.

[0134] In some embodiments, the fusion protein further comprises a third fragment comprising a second TEV protease fragment, wherein the first TEV protease fragment is able to cleave the TEV protease site in the presence of the second TEV protease fragment. In some embodiments, the fusion protein further comprises a second self-cleavage site between the second fragment and the third fragment, ad upon cleavage of the second self-cleavage site, the fusion protein releases the second TEV protease fragment which is not fused to any RNA recognition peptide.

[0135] Also provided, in one embodiment, is a dual guide RNA system, comprising: a target single guide RNA comprising a first spacer having sequence complementarity to a target nucleic acid sequence proximate to a first PAM site, a helper single guide RNA comprising a second spacer having sequence complementarity to a second nucleic acid sequence proximate to a second PAM site, a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, and a nucleobase deaminase.

[0136] In some embodiments, the second PAM site is located within 150 bases, or alternatively within 140, 130, 120, 110, 100, 95, 94, 93, 92, 91, 90, 89, 88, 87, 86, 85, 84, 83, 82, 81, 80, 75 or 70 bases from the second PAM site. In some embodiments, the second PAM site is located at least 10 bases, or alternatively at least 15, 20, 25, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 45, 50, 55, or 60 bases from the first PAM. In some embodiments, the second PAM site is upstream from the first PAM site. In some embodiments, the second PAM site is downstream from the first PAM site. In some embodiments, the distance is from 20-100, 25-95, 30-95, 34-95, 34-91, 34-90, 35-90, 40-90, 40-84, 45-85, or 50-80 bases, without limitation.

[0137] In some embodiments, the second (helper) spacer is 8-15 bases in length. In some embodiments, the second spacer is 8-14, 8-13, 8-12, 8-11, 8-10, 9-15, 9-14, 9-13, 9-12, 9-11,

9-10, 10-15, 10-14, 10-13, 10-12, 10-11, 11-15, 11-14, 11-13, 11-12, 12-15, 12-14, 12-13, or 13-15 bases in length. The first spacer, by contrast, is at least 16, 17, 18, or 19 bases in length.

[0138] Various "split" base editing systems are also described here, which allow the Cas protein and the nucleobase deaminase to be packaged into separate delivery vehicles (e.g., AAV).

[0139] In some embodiments, pairs of regular sgRNA and hsgRNA are provided that can mediate efficient editing to generate early stop codons in PCSK9 genes, which can have clinic benefits. Based on the discoveries here, suitable target sites for sgRNA and hsgRNA have been selected for converting a non-stop codon to a stop codon. Take C-to-T/U editing as an example, a non-stop codon can be CAG, CAA or CGA.

[0140] Examples of such target sites are illustrated in Table 4. It is readily understood that the sequences in Table 4 are used to show the location of the target. The actual sgRNA and hsgRNA, however, does not need to bind to the entire sequence. In fact, for hsgRNA for instance, a binding of 8-15 nucleotides could well be enough as explained above. Accordingly, the spacer sequence on the hsgRNA can be complementary to a sub-sequence of any shown in Table 4, or even overlap with any of them. The same is true for sgRNA as well, with a preferred spacer length of 18-24 nucleotides, without limitation.

[0141] In one embodiment, provided is a pair of helper guide RNA/guide RNA for editing a human PCSK9 nucleic acid sequence, wherein the guide RNA specifically targets a first site on the PCSK9 nucleic acid to enable base editing to convert a non-stop codon to a stop codon, and wherein the helper guide RNA specifically targets a second site on the PCSK9 nucleic acid that is 20 to 100 bases from the first site. In some embodiments, the second site is about 20-100, 25-95, 30-95, 34-95, 34-91, 34-90, 35-90, 40-90, 40-84, 45-85, or 50-80 bases away from the first site.

[0142] In some embodiments, the hsgRNA has a spacer that is 8-15 bases in length. In some embodiments, the spacer is 8-14, 8-13, 8-12, 8-11, 8-10, 9-15, 9-14, 9-13, 9-12, 9-11, 9-10, 10-15, 10-14, 10-13, 10-12, 10-11, 11-15, 11-14, 11-13, 11-12, 12-15, 12-14, 12-13, or 13-15 bases in length. In some embodiments, the sgRNA has a spacer that is at least 16, 17, 18, or 19 bases in length.

[0143] Spacer sequences for the sgRNA/hsgRNA can be readily designed. For instance, for each target site shown in Table 4, a spacer may be the complementary sequence of the desired length (i.e., complementary to a sub-sequence of any of SEQ ID NO:166-180 or 181-195). Specific examples of pairs of binding sites include, without limitation, SEQ ID NO:166 and 181; SEQ ID NO:167 and 182; SEQ ID NO:168 and 183; SEQ ID NO:169 and 184; SEQ ID NO:170 and 185; SEQ ID NO:171 and 186; SEQ ID NO:172 and 187; SEQ ID NO:173 and 188; SEQ ID NO:174 and 189; SEQ ID NO:175 and 190; SEQ ID NO:176 and 191; SEQ ID NO:177 and 192; SEQ ID NO:178 and 193; SEQ ID NO:179 and 194; and SEQ ID NO:180 and 195.

[0144] Example sgRNA/hsgRNA sequences have also been designed and tested. See Table 3. Moreover, polynucleotide sequences encoding the helper guide RNA and guide RNA are also provided.

[0145] With such pairs of sgRNA/hsgRNA sequences, methods of inactivating a PSCK9 gene in a cell can be carried out. In some embodiments, the method entails contacting the cell with a pair of helper guide RNA and guide RNA of the present disclosure, a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, and a nucleobase deaminase. Each of these elements have been further described in the instant disclosure.

Enhanced Prime Editing

[0146] Improved prime editing systems are also provided, in some embodiments. In particular, certain prime editing guide RNA (pegRNA) molecules provided herein have improved stability. These pegRNA contain a scaffold that, compared to the conventional guide RNA, has one additional base pairing (see, FIG. 36A and 36E). Using the standard scaffold (SEQ ID NO:31) at a template, the improved scaffold may have a sequence of any of SEQ ID NO:32-43.

[0147] As discussed above, a typical guide RNA scaffold has a structure which includes, from the 5’ to the 3’ end, (a) a repeat region, (b) a tetraloop, (c) an anti-repeat that is at least partially complementary to the repeat region, (d) stem loop 1, (e) a linker, (f) stem loop 2, and (g) stem loop 3. In other words, the scaffold includes 4 stem loops. The third stem loop

(counted from 5’ to 3’), also referred to as “Stem loop 2”, includes 4 base pairings in the conventional design. In the new design, this stem loop has 5 base pairing.

[0148] In one embodiment, provided is a guide RNA comprising a scaffold which comprises, from the 5’ to 3’ direction, a first stem loop portion, a second stem loop portion, a third stem loop portion, and a fourth stem loop portion, wherein the third stem loop comprises five base pairings within.

[0149] The sequence of the scaffold can be expressed as: GUUUNAGAGCUA**X**$_1$UAGCAAGUUNAAAUAAGGCNNGUCCGUUAUCAACUU**X**$_2$ AAGUGGCACCGANUCGGUGC (SEQ ID NO:31), where N represents any base, and X1 and X2 denotes any nucleotide sequence of a length of 2-50 bases (or 2-40, 3-40, 4-40, 4-30, 2-30, 4-20 bases). Accordingly, in some embodiments, the base pairings comprise one between positions 45 and 55, according to the positions in SEQ ID NO:31. In some embodiments, the scaffold has at least 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence identity to SEQ ID NO:31 and includes give base pairings in the third stem loop.

[0150] In one embodiment, therefore, provided is a guide RNA comprising a scaffold derived from SEQ ID NO:31 by introducing a base pairing between the bases at position 45 and position 55, and optionally allowing one, two, three, four, or five base additions, deletions, substitutions, or the combination thereof, as long as it maintains the stem loops structure or the scaffold/guide RNA functionality. In some embodiments, the scaffold comprises a sequence selected from the group consisting of SEQ ID NO:32-43. In some embodiments, the guide RNA is at least 100 nucleotides, or 105, 110, 115, 120, 125, 130, 140 or 150 nucleotides in length. In some embodiments, the guide RNA further includes a spacer (e.g., 8-25 nucleotides), a reverse-transcriptase template, and/or a primer-binding site.

[0151] Improved prime editor proteins are also provided, in some embodiments. In one embodiment, the prime editor includes a Cas protein and a reverse-transcriptase linked through a linker tested to optimized for the performance of the prime editor. In one embodiment, the prime editor comprises the amino acid sequence of SEQ ID NO:44. In one embodiment, the prime editor comprises the amino acid sequence of SEQ ID NO:45. Both these prime editors have been tested and shown to exhibit superior editing efficiency and specificity.

[0152] Various "split" prime editing systems are also described here, which allow the Cas protein and the reverse transcriptase to be packaged into separate delivery vehicles (e.g., AAV).

[0153] With the split prime editing systems, methods for conducting genetic editing in a cell at a target site are also provided. In some embodiments, the method entails introducing to the cell a first viral particle enclosing a first construct encoding a clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein, and a second viral particle enclosing a second construct encoding a reverse-transcriptase fused to an RNA recognition peptide. In some embodiments, the second construct further encodes a guide RNA comprising an RNA recognition site that the RNA recognition peptide binds to.

[0154] In some embodiments, the second construct further encodes a guide RNA comprising an RNA recognition site that the RNA recognition peptide binds to. 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, SpCas9-NG, xSpCas9, 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, and RanCas13b. In some embodiments, the Cas protein is SpCas9-NG or xSpCas9.

[0155] Non-limiting examples of reverse-transcriptases include human immunodeficiency virus (HIV) reverse-transcriptase, moloney murine leukemia virus (MMLV) reverse-transcriptase and avian myeloblastosis virus (AMV) reverse-transcriptase

EXAMPLES

Example 1. Fusion Base Editors with Reduced Off-Target Editing Activity

[0156] Single-guide RNAs (sgRNAs) and base editors (BEs) mentioned in the examples are for SpCas9, unless specifically pointed out, e.g., the sgRNA for SaCas9 (Sa-sgRNAs). To test whether the current base editing system can induce C-to-T mutations in ssDNA regions, we used SaD10A nickase and Sa-sgRNA to make a DNA single-stranded break (SSB), which can trigger end recession to generate a ssDNA region. (Fig. 1A, 2A and 3A). We co-transfected SaD10A, Sa-sgRNA (Sa-sgSITE31, Sa-sgSITE42 and Sa-sgF1) with two

published BEs, i.e., BE3 and hA3A-BE3 or an empty vector (Fig. 1B, 2B and 3B) and determine the mutagenesis around the ssDNA regions triggered by SaD10A. At three tested sites (Sa-SITE31, Sa-SITE42 and Sa-F1) the expression of BE3 or hA3A-BE3 induced C-to-T mutations, whereas the expression of an empty vector did not (Fig. 1C, 2C and 3C). These results indicated that the current base editors, which contain catalytically active cytidine deaminases, indeed cause unintended mutations in non-relevant ssDNA regions (Fig. 1, 2 and 3).

[0157] To inhibit the activity of cytidine deaminase at non-relevant sites, e.g., ssDNA regions, we proposed to fuse base editors with a base editing inhibitor. Mouse APOBEC3 (mA3) comprises two cytidine deaminase (CDA) domains (CDA1 and CDA2, Fig. 4A, 5A, 6A) and the use of full-length mA3 in mA3-BE3 (Fig. 4B, 5B, 6B) did not induce C-to-T editing at three tested target sites (Fig. 4C, 5C, 6C). However, mA3CDA1-BE3, which is generated by deleting mA3CDA2 from mA3-BE3 (Fig. 4B, 5B, 6B), induced substantial C-to-T editing (Fig. 4C, 5C, 6C). These results suggest that mA3CDA2 is a natural inhibitor of base editing. Thus, we added mA3CDA2 to the N-terminus of three active BEs, i.e., mA3CDA1-BE3, BE3 and hA3A-BE3, to generate mA3rev-BE3, mA3CDA2-BE3 and mA3CDA2-hA3A-BE3 (Fig. 4B, 5B, 6B). As we expected, the adding of mA3CDA2 to the N-terminus clearly reduced the base editing efficiencies (Fig. 4C, 5C, 6C).

[0158] Next, we considered whether the cleavage of mA3CDA2 can restore the base editing efficiency. 2A self-cleavage peptides were inserted between mA3CDA2 and the rest part of BE in mA3rev-BE3, mA3CDA2-BE3 and mA3CDA2-hA3A-BE3 to generate mA3rev-2A-BE3, mA3CDA2-2A-BE3 and mA3CDA2-2A-hA3A-BE3 (Fig. 4B, 5B, 6B). Correspondingly, the base editing efficiencies restored in mA3rev-2A-BE3, mA3CDA2-2A-BE3 and mA3CDA2-2A-hA3A-BE3 (Fig. 4C, 5C, 6C), indicating that the inhibition of mA3CDA2 depends on its covalent connection to BEs. We also searched protein database for the domains similar to mA3CDA2 core sequence and found at least 44 proteins have the similar domains (Table 1).

[0159] Human APOBEC3B (hA3B) also comprises two cytidine deaminase (CDA) domains (CDA1 and CDA2, Fig. 7A, 8A, 9A) and the use of full-length hA3B in hA3B-BE3 (Fig. 7B, 8B, 9B) only induced relatively low levels of C-to-T editing at three tested target sites (Fig. 7C, 8C, 9C). However, hA3BCDA2-BE3, which is generated by deleting hA3BCDA1 from hA3B-BE3 (Fig. 7B, 8B, 9B), induced higher C-to-T editing (Fig. 7C, 8C, 9C). In addition,

2A self-cleavage peptides were inserted between hA3BCDA1 and hA3BCDA2 to generate hA3B-2A-BE3 (Fig. 7B, 8B, 9B), which induces higher C-to-T editing efficiencies than hA3B-BE3 (Fig. 7C, 8C, 9C). These results indicate that hA3BCDA1 is another inhibitor of base editing and the inhibition of hA3BCDA1 depends on its covalent connection to BEs. We also searched protein database for the domains similar to hA3BCDA1 and found at least 43 proteins have the similar domains (Table 2).

[0160] Next, we planned to use mA3 to develop novel BEs. The two BEs, mA3rev-BE3 and mA3rev-2A-BE3, were made by splitting mA3 between amino acid (AA)207 and AA208 and then we determined where to split mA3CDA2 can keep the highest editing efficiency (Fig. 10A, 11A, 12A). As mA3CDA1 ends at amino acid (AA)154 and mA3CDA2 starts from AA238, we split mA3CDA2 at AA196/AA197, AA215/AA216, AA229/AA230 and AA237/AA238 to generate mA3rev-BE3-196, mA3rev-2A-BE3-196, mA3rev-BE3-215, mA3rev-2A-BE3-215, mA3rev-BE3-229, mA3rev-2A-BE3-229, mA3rev-BE3-237, and mA3rev-2A-BE3-237 (Fig. 10B, 11B, 12B). Although the splitting of mA3 at AA207/AA208 and AA215/AA216 keeps the highest editing efficiencies, the results also showed that the splitting sites spanning from AA196/AA197 to AA237/AA238 generally keep substantial editing efficiencies (Fig. 10C, 11C, 12C).

[0161] Furthermore, we tried to determine the minimal region of mA3 that has a base editing inhibitory effect. We deleted various N-terminal parts of mA3CDA2 in mA3rev-BE-237 to develop mA3rev-BE-237-Del-255, mA3rev-BE-237-Del-285 and mA3rev-BE-237-Del-333, which contains the AA256-AA429, AA286-AA429 and AA334-AA429 parts of mA3 as the base editing inhibitor respectively (Fig. 13A, 14A, 15A). By comparing with mA3rev-BE-237, which contains the AA238-AA429 part of mA3, mA3rev-BE-237-Del-255, mA3rev-BE-237-Del-285 and mA3rev-BE-237-Del-333 showed similar editing efficiencies (Fig. 13B, 14B, 15B). These results indicated that the AA334-AA429 part of mA3 still has the inhibitory effect of base editing.

[0162] In order to develop base editors that do not cause C-to-T mutation in non-relevant ssDNA regions, we replaced the 2A self-cleavage site with a cleavage site of TEV protease in mA3rev-2A-BE3 and then fused the N-terminal part of TEV protease (TEVn) [Gray et al., 2010, Cell, doi: 10.1016/j.cell.2010.07.014] to the C-terminus of mA3rev-2A-BE3 with another TEV cleavage site. The newly developed BE is named BEsafe. In addition, we put one MS2 loop into the sgRNA to generate MS2-sgRNA [Ma et al., 2016, Nature

Biotechnology, doi: 10.1038/nbt.3526] and then fused the C-terminal part of TEV protease (TEVc) with MS2 coat protein (MCP), which can bind to MS2 loop (Fig. 16A). When BEsafe, MS2-sgRNA and MCP-TEVc were co-expressed, the TEVn fused in BEsafe and the TEVc of MCP-TEVc that can be recruited by MS2-sgRNA would associate and restore the protease activity at on-target site. The subsequent cleavages at TEV sites would remove mA3CDA2 and TEVn from the N- and C-terminus of BEsafe and the resulted mA3CDA1-BE3 can induce efficient base editing at on-target site (Fig. 16A). On contrary, the BEsafe would not induce C-to-T mutations in non-relevant ssDNA regions because the cytidine deaminase activity of mA3CDA1 is inhibited by mA3CDA2 (Fig. 16B).

[0163] We then compared the performance of BEsafe and hA3A-BE3 at on-target sites and non-relevant ssDNA regions (Fig. 17, 18, 19). We co-transfected the plasmid expressing Sa-sgRNA and SaD10A, which can trigger ssDNA formation at Sa-sgRNA target sites (Fig. 17A, 18A, 19A), with the hA3A-BE3 expression plasmid and the sgRNA expression plasmid, with the BEsafe expression plasmid and the plasmid expressing MS2-sgRNA and MCP-TEVc or with the MCP-TEVc expression plasmid and the plasmid expressing MS2-sgRNA and BEsafe (Fig. 17B, 18B, 19B). We examined the C-to-T mutation frequencies in non-relevant ssDNA regions (Sa-sgRNA on-target sites, orthogonal to those of SpCas9) (Fig. 17C, 18C, 19C) and the base editing efficiencies at sgRNA on-target sites of hA3A-BE3 and BEsafe, both of which are SpCas9-derived (Fig. 17D, 18D, 19D). We found that BEsafe did not cause any C-to-T mutation in the non-relevant ssDNA regions (Sa-sgRNA on-target sites) but hA3A-BE3 caused obvious mutations (Fig. 17C, 18C, 19C). At sgRNA on-target sites, BEsafe induced base editing comparable to hA3A-BE3, while the expression of both MS2-sgRNA and BEsafe from one single plasmid yielded higher base editing efficiencies than the expression of only BEsafe from one plasmid did (Fig. 17D, 18D, 19D).

[0164] The base editors and base editing method described in this invention could be applied to perform high-specificity and high-efficiency base editing in the genome of various eukaryotes.

[0165] For the first time, a base editing system was established to avoid causing C-to-T mutations in non-relevant ssDNA regions and to induce efficient base editing at on-target sites. The BEsafe base editing system and the accompanying methods disclosed in this invention could be utilized to perform highly specific base editing that cannot be implemented by the currently existing BEs as the cytidine deaminases in current BEs can

cause unintended mutations in non-relevant ssDNA regions. Importantly, the high specificity and efficiency of this BEsafe base editing system will promote the potential clinical translation, especially in the gene therapies that involve restoring disease-related mutations.

Table 1: mA3CDA2 Core Sequence Related Domains

Name	Sequence	SEQ ID NO:
Mouse APOBEC3 cytidine deaminase domain 2 core (AA282-AA355)	SEKGKQHAEILFLDKIRSMELSQVTITCYLTWSPCPNCAWQLAAFKRDRP DLILHIYTSRLYFHWKRPFQKGLC	48
Mus spicilegus A3 (AA248-AA321)	SEKGKQHAEILFLDKIRSMELSQVTITCYLTWSPCPNCAWQLAAFKRDRP DLIPHIYTSRLYFHWKRPFQKGLC	49
Cricetulus longicaudatus A3 (AA249-AA322)	SEKGKQHAEILFLDKIRSMELSQVTITCYLTWSPCPNCAWRLAAFKRDRP DLILHIYTSRLYFHWKRPFQKGLC	50
Mus terricolor A3 (AA248-AA321)	SEKGKQHAEILFLNKIRSMELSQVTITCYLTWSPCPNCAWQLAAFKKDRP DLILHIYTSRLYFHWKRPFQKGLC	51
Mus caroli A3 (AA260-AA333)	SKKGKQHAEILFLDKIRSMELSQVTITCYLTWSPCPNCAWQLAAFKRDHP DLILHIYTSRLYFHWKRPFQKGLC	52
Mus pahari A3 (AA263-AA336)	SKKGKQHAEILFLEKIRSMELSQMRITCYLTWSPCPNCAWQLAAFQKDRP DLILHIYTSRLYFHWRRIFQKGLC	53
Mus shortridgei A3 (AA233-AA306)	SKKGKQHAEILFLEKIRSMELSQMRITCYLTWSPCPNCAWQLAAFQKDRP DLILHIYTSRLYFHWRRIFQKGLC	54
Mus setulosus A3 (AA29-AA302)	SKKGKQHAEILFLDKIRSMELSQVRITCYLTWSPCPNCAWQLETFKKDRP DLILHIYTSRLYFHWKRAFQEGLC	55
Grammomys surdaster A3 (AA270-AA344)	SKKGKPHAEILFLDKMWSMEELSQVRITCYLTWSPCPNCARQLAAFKKDH PGLILRIYTSRLYFYWRRKFQKGLC	56
Rattus norvegicus A3 (AA256-AA328)	KKGEQHVEILFLEKMRSMELSQVRITCYLTWSPCPNCARQLAAFKKDHPD LILRIYTSRLYFYWRKKFQKGLC	57
Mastomys coucha A3 (AA258-AA331)	SKKGRQHAEILFLEKVRSMQLSQVRITCYLTWSPCPNCAWQLAAFKMDHP DLILRIYASRLYFHWRRAFQKGLC	58
Cricetulus griseus A3B (AA235-AA307)	NKKGKHAEILFIDEMRSLELGQVQITCYLTWSPCPNCAQELAAFKSDHPD LVLRIYTSRLYFHWRRKYQEGLC	59
Peromyscus leucopus A3 (AA266-AA338)	NKKGKHAEILFIDEMRSLELGQARITCYLTWSPCPNCAQKLAAFKKDHPD LVLRVYTSRLYFHWRRKYQEGLC	60
Mesocricetus auratus A3 (AA268-AA340)	NKKDKHAEILFIDKMRSLELCQVRITCYLTWSPCPNCAQELAAFKKDHPD LVLRIYTSRLYFHWRRKYQEGLC	61
Microtus ochrogaster A3B (AA266-AA338)	NKKGKHAEILFIDEMRSLKLSQERITCYLTWSPCPNCAQELAAFKRDHPG LVLRIYASRLYFHWRRKYQEGLC	62
Nannospalax galili A3 (AA231-AA302)	NKRAKHAEILLIDMMRSMELGQVQITCYITWSPCPTCAQELAAFKQDHPD LVLRIYASRLYFHWKRKFQKGL	63
Meriones unguiculatus A3 (AA233-AA305)	NKKGRHAEICLIDEMRSLGLGKAQITCYLTWSPCRKCAQELATFKKDHPD LVLRVYASRLYFHWSRKYQQGLC	64
Dipodomys ordii A3 (AA256-AA330)	NKKGHHAEIRFIERIRSMGLDPSQDYQITCYLTWSPCLDCAFKLAKLKKD FPRLTLRIFTSRLYFHWIRKFQKGL	65
Jaculus jaculus A3 (AA303-AA374)	NKKGKHAEARFVDKMRSMQLDHALITCYLTWSPCLDCSQKLAALKRDHPG LTLRIFTSRLYFHWVKKFQEGL	66
Chinchilla lanigera A3H (AA86-AA161)	SPQKGHHAESRFIKRISSMDLDRSRSYQITCFLTWSPCPSCAQELASFKR AHPHLRFQIFVSRLYFHWKRSYQAGL	67

Heterocephalus glaber A3 (AA277-AA350)	KKGYHAESRFIKRICSMDLGQDQSYQVTCFLTWSPCPHCAQELVSFKRAH PHLRLQIFTARLFFHWKRSYQEGL	68
Octodon degus A3 (AA256-AA329)	KKGQHAEIRFIERIHSMALDQARSYQITCFLTWSPCPFCAQELASFKSTH PRVHLQIFVSRLYFHWKRSYQEGL	69
Urocitellus parryii A3 (AA256-AA330)	NKKGHHAEIRFIKKIRSLDLDQSQNYEVTCYLTWSPCPDCAQELVALTRS HPHVRLRLFTSRLYFHWFWSFQEGL	70
Aotus nancymaae A3H (AA75-AA146)	NRHAEICFIDEIESMGLDKTQCYEVTCYLTWSPCPSCAQKLAAFTKAQVH LNLRIFASRLYYHWRSSYQKGL	71
Cebus capucinus imitator A3H (AA55-AA126)	NRHAEICFIDEIESMGLDKTQCYEVTCYLTWSPCPSCAQKLVAFAKAQDH LNLRIFASRLYYHWRRRYKEGL	72
Saimiri boliviensis boliviensis A3H (AA56-AA125)	HVEICFIDKIASMELDKTQCYDVTCYLTWSPCPSCAQKLAAFAKAQDHLN LRIFASRLYYHWRRSYQKGL	73
Homo sapiens A3H (AA49-AA123)	NKKKCHAEICFINEIKSMGLDETQCYQVTCYLTWSPCSSCAWELVDFIKA HDHLNLGIFASRLYYHWCKPQQKGL	74
Homo sapiens ARP10 (AA48-AA123)	ENKKKCHAEICFINEIKSMGLDETQCYQVTCYLTWSPCSSCAWELVDFIK AHDHLNLGIFASRLYYHWCKPQQKGL	75
Pan paniscus A3H (AA49-AA123)	NKKKCHAEICFINEIKSMGLDETQCYQVTCYLTWSPCSSCAWKLVDFIQA HDHLNLRIFASRLYYHWCKPQQEGL	76
Symphalangus syndactylus A3H (AA49-AA123)	NKKKRHAEIRFINKIKSMGLDETQCYQVTCYLTWSPCPSCAWELVDFIKA HDHLNLGIFASRLYYHWCRHQQEGL	77
Macaca mulatta A3H (AA49-AA123)	NKKKDHAEIRFINKIKSMGLDETQCYQVTCYLTWSPCPSCAGELVDFIKA HRHLNLRIFASRLYYHWRPNYQEGL	78
Theropithecus gelada A3H (AA54-AA128)	NKKKEHAEIRFINKIKSMGLDETQCYQVTCYLTWSPCPSCAGKLVDFIKA HHHLNLRIFASRLYYHWRPNYQEGL	79
Mandrillus leucophaeus A3H (AA49-AA123)	NKKKHHAEIHFINKIKSMGLDETQCYQVTCYLTWSPCPSCARELVDFIKA HRHLNLRIFASRLYYHWRPHYQEGL	80
Bos grunniens A3 (AA74-AA148)	NKKQRHAEIRFIDKINSLDLNPSQSYKIICYITWSPCPNCANELVNFITR NNHLKLEIFASRLYFHWIKPFKMGL	81
Bubalus bubalis A3 (AA74-AA148)	NKKQRHAEIRFIDKINSLDLNPSQSYKIICYITWSPCPNCASELVDFITR NDHLDLQIFASRLYFHWIKPFKRGL	82
Odocoileus virginianus texanus A3H (AA209-AA283)	NKKQRHAEIRFIDKINSLNLDRRQSYKIICYITWSPCPRCASELVDFITG NDHLNLQIFASRLYFHWKKPFQRGL	83
Sus scrofa A3 (AA51-AA125)	NKKKRHAEIRFIDKINSLNLDQNQCYRIICYVTWSPCHNCAKELVDFISN RHHLSLQLFASRLYFHWVRCYQRGL	84
Ceratotherium simum simum A3B (AA232-AA306)	NKKKRHAEIRFIDKIKSLGLDRVQSYEITCYITWSPCPTCALELVAFTRD YPRLSLQIFASRLYFHWRRRSIQGL	85
Equus caballus A3H (AA79-AA153)	NKKKRHAEIRFIDKINSLGLDQDQSYEITCYVTWSPCATCACKLIKFTRK FPNLSLRIFVSRLYYHWFRQNQQGL	86
Enhydra lutris kenyoni A3B (AA243-AA316)	KKKRHAEIRFIDSIRALQLDQSQRFEITCYLTWSPCPTCAKELAMFVQDH PHISLRLFASRLYFHWRWKYQEGL	87
Leptonychotes weddellii A3H (AA50-AA123)	KKKRHAEIRFIDNIKALRLDTSQRFEITCYVTWSPCPTCAKELVAFVRDH RHISLRLFASRLYFHWLRENKKGL	88
Ursus arctos horribilis A3F (AA552-AA626)	NKKKRHAEIRFIDKIRSLQRDSSQTFEITCYVTWSPCFTCAEELVAFVRD HPHVRLRLFASRLYFHWLRKYQEGL	89
Panthera leo bleyenberghi A3H (AA50-AA124)	NKKKRHAEICFIDKIKSLTRDTSQRFEIICYITWSPCPFCAEELVAFVKD NPHLSLRIFASRLYVHWRWKYQQGL	90

Panthera tigris sumatrae A3H (AA50-AA124)	NKKKRHAEICFIDKIKSLTRDTSQRFEIICYITWSPCPFCAEELVAFVKD NPHLSLRIFASRLYVHWRWKYQQGL	91
Tupaia belangeri A3 (AA46-AA120)	NKKHRHAEVRFIAKIRSMSLDLDQKHQLTCYLTWSPCPSCAQELVTFMAE SRHLNLQVFVSRLYFHWQRDFQQGL	92

Table 2: hA3BCDA1-Related Domains

Name	Sequence	SEQ ID NO:
Gorilla A3B (AA29-AA138)	GRSYNWLCYEVKIKRGRSNLLWNTGVFRGQMYSQPEHHAEMCFLSWFCGN QLPAYKCFQITWFVSWTPCPDCVAKLAEFLAEYPNVTLTISTARLYYYWE RDYRRALCRL	93
Pan paniscus A3B (AA29-AA138)	GRSYTWLCYEVKIRRGHSNLLWDTGVFRGQMYSQPEHHAEMYFLSWFCGN QLPAYKCFQITWFVSWTPCPDCVAKLAEFLAEHPNVTLTISAARLYYYWE RDYRRALCRL	94
Pan troglodytes A3B (AA29-AA138)	GRSYTWLCYEVKIRRGHSNLLWDTGVFRGQMYSQPEHHAEMCFLSWFCGN QLSAYKCFQITWFVSWTPCPDCVAKLAKFLAEHPNVTLTISAARLYYYWE RDYRRALCRL	95
Gorilla A3F (AA30-AA137)	RNTVWLCYEVKTKGPSRPPLDAKIFRGQVYFEPQYHAEMCFLSWFCGNQL PAYKCFQITWFVSWTPCPDCVAKLAEFLAEHPNVTLTISAARLYYYWE	96
Pan troglodytes A3F (AA30-AA137)	RNTVWLCYEVKTKGPSRPRLDTKIFRGQVYFEPQYHAEMCFLSWFCGNQL PAYKCFQITWFVSWTPCPDCVAKLAEFLAEHPNVTLTISAARLYYYWERD YRRALCRL	97
Human sapiens A3F (AA30-AA137)	RNTVWLCYEVKTKGPSRPRLDAKIFRGQVYSQPEHHAEMCFLSWFCGNQL PAYKCFQITWFVSWTPCPDCVAKLAEFLAEHPNVTLTISAARLYYYWERD YRRALCRL	98
Macaca leonine A3F (AA30-AA137)	RNTVWLCYEVKTRGPSMPTWGTKIFRGQVCFEPQYHAEMCFLSRFCGNQL PAYKRFQITWFVSWTPCPDCVAKVAEFLAEHPNVTLTISAARLYYYWETD YRRALCRL	99
Macaca nemestrina A3F (AA30-AA137)	RNTVWLCYEVKTRGPSMPTWGTKIFRGQVCFEPQYHAEMCFLSRFCGNQL PAYKRFQITWFVSWTPCPDCVAKVAEFLAEHPNVTLTISAARLYYYWETD YRRALCRL	100
Rhinopithecus roxellana A3F (AA30-AA137)	RNTVWLCYEVKTRGPSMPTWGAKIFRGQVYFEPQYHAEMCFLSWFCGNQL PAYKRFQITWFVSWTPCPDCVAKVAEFLAEHPNVTLTISAARLYYYWETD YRRALCRL	101
Mandrillus leucophaeus A3F (AA30-AA130)	RNTVWLCYKVKTRGPSMPTWGTKIFRGQVYFQPQYHAEMCFLSWFCGNQL PAYKRFQITWFVSWTPCPDCVVKVAEFLAEHPNVTLTISAARLYYYWETD Y	102
Macaca mulatta A3F (AA30-AA137)	RNTVWLCYEVKTRGPSMPTWDTKIFRGQVYSKPEHHAEMCFLSRFCGNQL PAYKRFQITWFVSWTPCPDCVAKVAEFLAEHPNVTLTISAARLYYYWETD YRRALCRL	103
Theropithecus gelada A3F (AA30-AA137)	RNTVWLCYEVKTRGPSMPTWGTKIFRGQVYFQPQYHAEMCFLSRFCGNQL PAYKRFQITWFVSWNPCPDCVAKVIEFLAEHPNVTLTISAARLYYYWGRD WRRALRRL	104
Cercocebus atys A3B (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWYTGVFRGQVYSKPEHHAEMCFLSRFCGN QLPAYKRFQITWFVSWNPCPDCVAKVIEFLAEHPNVTLTISAARLYYYWS RDWQRALCRL	105
Macaca fascicularis A3B (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVYSKPEHHAEMCFLSRFCGN QLPAYKRFQITWFVSWNPCPDCVAKVIEFLAEHPNVTLTISTARLYYYWG RDWQRALCRL	106
Macaca mulatta A3B (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVYSKPEHHAEMCFLSRFCGN QLPAYKRFQITWFVSWNPCPDCVAKVIEFLAEHPNVTLTISTARLYYYWG RDWQRALCRL	107
Macaca leonina A3B (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVYSKPEHHAEMCFLSRFCGN QLPAYKRFQITWFVSWNPCPDCVVKVIEFLAEHPNVTLTISTARLYYYWG RDWQRALCRL	108
Mandrillus leucophaeus A3B (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWYTGVFRGQVYSKPEHHAEMCFLSRFCGN QLPAYKRFQITWFVSWNPCPDCVAKVIEFLAEHPNVTLTIFTARLYYYWG RDWQRALCRL	109
Macaca nemestrina A3B (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVYSKPEHHAEMCFLSRFCGN QLPAYKRFQITWFVSWNPCPDCVAKVTEFLAEHPNVTLTISTARLYYYWG RDWQRALCRL	110

Rhinopithecus bieti A3F (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVYSEPEHHAEMYFLSWFCGN QLPAYKRFQITWFVSWTPCPDCVAKVAEFLTEHPNVTLTISAARLYYYRG RDWRRALCRL	111
Rhinopithecus roxellana A3B (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVYSEPEHHAEMYFLSWFCGN QLPAYKRFQITWFVSWTPCPDCVAKVAEFLTEHPNVTLTISAARLYYYRG RDWRRALCRL	112
Chlorocebus sabaeus A3B (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQMYSKPEHHAEMCFLSWFCGN QLPAHKRFQITWFVSWTPCPDCVAKVAEFLAEYPNVTLTISAARLYYYWE TDYRRALCRL	113
Nomascus leucogenys A3B (AA30-AA138)	RSYTWLCYEVKIRKDPSKLPWDTGVFRGQMYFQPEYHAEMCFLSWFCGNQ LPAYKRFQITWFVSWTPCPDCVAKVAVFLAEHPNVTLTISAARLYYYWEK DWQRALCRL	114
Cercocebus atys A3F (AA29-AA138)	GRSYTWLCYEVKIKKYPSKLLWDTGVFQGQVYFQPQYHAEMCFLSRFCGN QLPAYKRFQITWFVSWNPCPDCVAKVTEFLAEHPNVTLTISAARLYYYWE KDXRRALRRL	115
Papio anubis A3F (AA29-AA138)	GRSYTWLCYEVKIKEDPSKLLWDTGVFQGQVYFQPQYHAEMCFLSRFCGN QLPAYKRFQITWFVSWNPCPDCVAKVTEFLAEHPNVTLTISAARLYYYWG RDWRRALRRL	116
Chlorocebus aethiops A3D (AA29-AA150)	GRRYTWLCYEVKIKKDPSKLPWDTGVFPGQVRPKFQSNRRYEVYFQPQYH AEMYFLSWFCGNQLPAYKHFQITWFVSWNPCPDCVAKVTEFLAEHRNVTL TISAARLYYYWGKDWRRALCRL	117
Chlorocebus sabaeus A3D (AA29-AA134)	GRRYTWLCYEVKIKKDPSKLPWDTGVFPGQPQYHAEMYFLSWFCGNQLPA YKHFQITWFVSWNPCPDCVAKVTEFLAEHRNVTLTISAARLYYYWGKDWR RALCRL	118
Chlorocebus sabaeus A3F (AA29-AA150)	GRRYTWLCYEVKIKKDPSKLPWDTGVFPGQVRPKFQSNRRQKVYFQPQYH AEMYFLSWFCGNQLPAYKHFQITWFVSWNPCPDCVAKVTEFLAEHRNVTL TISAARLYYYWGKDWRRALCRL	119
Erythrocebus patas A3D (AA29-AA150)	GRRYTWLCYEVKIKKDPSKLPWDTGVFQGQVRPKFQSNRRYEVYFQPQYH AEMCFLSWFCGNQLPAYKHFQITWFVSWNPCPDCVAKVTEFLAEHPNVTL TISAARLYYYWGKDWRRALCRL	120
Macaca fascicularis A3D (AA29-AA159)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVRPKLQSNRRYELSNWECRK RVYFQPQYHAEMYFLSWFCGNQLPANKRFQITWFASWNPCPDCVAKVTEF LAEHPNVTLTISVARLYYYRGKDWRRALRRL	121
Macaca fascicularis A3F (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVYFQPQYHAEMYFLSWFCGN QLPANKRFQITWFASWNPCPDCVAKVTEFLAEHPNVTLTISVARLYYYRG KDWRRALRRL	122
Macaca nemestrina A3D (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRDQVYFQPQYHAEMCFLSWFCGN QLPANKRFQITWFVSWNPCPDCVTKVTEFLAEHPNVTLTISVARLYYYRG KDWRRALRRL	123
Macaca leonina A3D (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWYTGVFRGQVYFQPQYHAEMCFLSWFCGN QLPANKRFQITWFVSWNPCPDCVAKVTEFLAEHPNVTLTISVARLYYYRG KDWRRALRRL	124
Macaca mulatta A3D (AA29-AA138)	GRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVYFQPQYHAEMCFLSWFCGN QLPAYKRFQITWFVSWNPCPDCVAKVTEFLAEHPNVTLTISVARLYYYRG KDWRRALCRL	125
Gorilla A3D (AA29-AA150)	GRSYTWLCYEVKIRRGSSNLLWNTGVFRGPVPPKLQSNHRQEVYFQFENH AEMCFLSWFCGNRLPANRRFQITWFVSWNPCLPCVVKVTKFLAEHPNVTL TISAARLYYYRDREWRRVLRRL	126
Pan paniscus A3D (AA29-AA150)	GRSYTWLCYEVKIKRGCSNLIWDTGVFRGPVLPKLQSNHRQEVYFQFENH AEMCFFSWFCGNRLPANRRFQITWFVSWNPCLPCVVKVTKFLAEHPNVTL TISAARLYYYQDREWRRVLRRL	127
Pan troglodytes A3D (AA29-AA150)	GRSYTWLCYEVKIKRGCSNLIWDTGVFRGPVLPKLQSNHRQEVYFQFENH AEMCFFSWFCGNRLPANRRFQITWFVSWNPCLPCVVKVTKFLAEHPNVTL TISAARLYYYQDREWRRVLRRL	128
Homo sapiens A3D (AA29-AA150)	GRSYTWLCYEVKIKRGRSNLLWDTGVFRGPVLPKRQSNHRQEVYFRFENH AEMCFLSWFCGNRLPANRRFQITWFVSWNPCLPCVVKVTKFLAEHPNVTL TISAARLYYYRDRDWRWVLLRL	129
Nomascus leucogenys A3D (AA29-AA150)	GRSYTWLCYEVKIRKDPSKLPWDKGVFRGQVLPKFQSNHRQEVYFQLENH AEMCFLSWFCGNQLPANRRFQITWFVSWNPCLPCVAKVTEFLAEHPNVTL TISAARLYYYRGRDWRRALRRL	130
Saimiri boliviensis A3C (AA29-AA138)	GKKYTWLCYEVKIKKDTSKLPWNTGVFRGQVNFNPEHHAEMYFLSWFRGK LLPACKRSQITWFVSWNPCLYCVAKVAEFLAEHPNVTLTVSTARLYCYWK KDWRRALRKL	131
Saimiri boliviensis A3F (AA29-AA138)	GKKYTWLCYEVKIKKDTSKLPWNTGVFRGQVNFNPEHHAEMYFLSWFRGK LLPACKRSQITWFVSWNPCLYCVAKVAEFLAEHPNVTLTVSTARLYCYWK KDWRRALRKL	132

Piliocolobus tephrosceles A3F (AA36-AA145)	GRRYTWLCYEVKIMKDHSKLPWYTGVFRGQVYFEPQNHAEMCFLSWFCGN QLPAYECCQITWFVSWTPCPDCVAKVTEFLAEHPNVTLTISAARLYYYRG RDWRRALRRL	133
Colobus angolensis palliatus A3F (AA29-AA138)	GRRYTWLCYEVKISKDPSKLPWDTGIFRGQVYFEPQYHAEMCFLSWYCGN QLPAYKCFQITWFVSWTPCPDCVGKVAEFLAEHPNVTLTISAARLYYYWE TDYRRALCRL	134
Pongo abelii A3F (AA30-AA150)	RNYTWLCYEVKIRKDPSKLAWDTGVFRGQVLPKLQSNHRREVYFEPQYHA EMCFLSWFCGNQLSAYERFQITWFVSWTPCPDCVAMLAEFLAEHPNVTLT VSAARLYYYWERDYRGALRRL	135

Example 2. Further Assessment of Inhibitor-Conjugated Base Editors

[0166] This example developed an efficient method to demonstrate that the APOBEC moiety of base editors (BEs) directly induced mutations at off-target single-stranded DNA (OTss) sites in an sgRNA-independent manner. By testing a series of APOBEC proteins with two cytidine deaminase (CDA) domains, we identified that the catalytically-inactive CDA domains of certain dual-domain APOBECs function as cytidine deaminase inhibitors (CDIs). By taking advantage of this finding and the concept of split-TEV protease, an induced base editor (iBE) by sgRNA-guided cleavage of CDI was developed, which links a nSpCas9-BE and a CDI with a TEV cleavage site. At sgRNA-independent OTss sites, iBE1 remained dormant owing to the covalently linked CDI. Whereas, at on-target sites, iBE1 was activated by sgRNA-guided TEV cleavage of CDI, resulting in efficient base editing. By using 'enhanced specificity' SpCas9 nickase, iBE2 was further developed to reduce unintended OTsg mutations. Due to its minimum off-target effects and uncompromised on-target editing efficiency, the editing specificity of iBEs was significantly higher than that of the previously-reported BEs. Thus, the iBE system described in this example provides a new layer of regulation for the specificity of current base editing system and secures its application against off-target mutations.

METHODS

Cell culture and transfection

[0167] HEK293FT 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.

[0168] For base editing in genomic DNA, HEK293FT cells were seeded in a 24-well plate at a density of 1.1×10^5 per well and transfected with 250 μl serum-free Opti-MEM that contained 5.35 μl LIPOFECTAMINE LTX (Life, Invitrogen), 2.14 μl LIPOFECTAMINE

plus (Life, Invitrogen), 1 μg pCMV-BE3 (or hA3B-BE3, hA3BCDA2-nSpCas9-BE, hA3D-BE3, hA3DCDA2-nSpCas9-BE, hA3F-BE3, hA3FCDA2-nSpCas9-BE, hA3G-BE3, hA3GCDA2-nSpCas9-BE, mA3-BE3, mA3CDA1-nSpCas9-BE, mA3CDA2-mA3CDA1-nSpCas9-BE, hA3FCDA1-mA3CDA1-nSpCas9-BE, hA3BCDA1-mA3CDA1-nSpCas9-BE, mA3CDA2-rA1-nSpCas9-BE, hA3FCDA1-rA1-nSpCas9-BE, hA3BCDA1-rA1-nSpCas9-BE, hA3A-BE3, mA3CDA2-hA3A-nSpCas9-BE, hA3FCDA1-hA3A-nSpCas9-BE, hA3BCDA1-hA3A-nSpCas9-BE, mA3CDA2F1-mA3CDA1-nSpCas9-BE, mA3CDA2F2-mA3CDA1-nSpCas9-BE, mA3CDA2F3-mA3CDA1-nSpCas9-BE, mA3CDAI-T2A-mA3CDA1-nSpCas9-BE, EGFP-mA3CDA1-nSpCas9-BE, EGFP-T2A-mA3CDA1-nSpCas9-BE, mA3CDAI-T2A-rA1-nSpCas9-BE, EGFP-rA1-nSpCas9-BE, EGFP-T2A-rA1-nSpCas9-BE, mA3CDAI-T2A-hA3A-nSpCas9-BE, EGFP-hA3A-nSpCas9-BE, EGFP-T2A-hA3A-nSpCas9-BE, pCMV-dSpCas9, iBE1, iBE2, mA3CDAI-TS-mA3CDA1-nSpCas9HF1-BE-NTEV or mA3CDAI-TS-mA3CDA1-nHypaSpCas9-BE-NTEV) expression vector, 0.64 μg sgRNA expression vector without or with 0.5μg Sa-sg-SaD10A expression vector. After 24 hr, puromycin (ant-pr-1, InvivoGen) was added to the medium at the final concentration of 4 μg/ml. After another 48 hr, the genomic DNA was extracted from the cells with QuickExtract™ DNA Extraction Solution (QE09050, Epicentre) for subsequent sequencing analysis.

DNA library preparation and sequencing

[0169] Target genomic sequences were PCR amplified by high-fidelity DNA polymerase PrimeSTAR HS (Clonetech) with primer sets flanking examined sgRNA target sites. Indexed DNA libraries were prepared by using the TruSeq ChIP Sample Preparation Kit (Illumina) with minor modifications. Briefly, the PCR products amplified from genomic DNA regions were fragmented by Covaris S220. The fragmented DNAs were then PCR amplified by using the TruSeq ChIP Sample Preparation Kit (Illumina). After quantitated with Qubit High-Sensitivity DNA kit (Invitrogen), PCR products with different tags were pooled together for deep sequencing by using the Illumina Hiseq X10 (2×150) or NextSeq 500 (2×150) at CAS-MPG Partner Institute for Computational Biology Omics Core, Shanghai, China. Raw read qualities were evaluated by FastQC. For paired-end sequencing, only R1 reads were used. Adaptor sequences and read sequences on both ends with Phred quality score lower than 30 were trimmed. Trimmed reads were then mapped with the BWA-MEM algorithm (BWA

v0.7.17) to target sequences. After piled up with samtools (v1.9), base substitutions were further calculated.

Base substitution calculation

[0170] Base substitutions were selected at each position of the examined sgRNA target sites that were mapped with at least 1000 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. For each sgRNA, the ratio of C-to-T base substitution over indel was calculated by dividing the sum of C-to-T base substitution frequencies at all editing sites by the indel frequency of 50-bp region around sgRNA target site (from upstream eight nucleotides to the target site to downstream 19 nucleotides to PAM sites).

Results

[0171] Cytosine or adenine base editors (CBEs/BEs or ABEs) that fuse native cytidine deaminases or *in vitro* evolved adenosine deaminases with CRISPR-Cas9 have been developed to induce targeted C-to-T or adenine to guanine (A-to-G) conversions with high efficiencies. Because BEs use catalytically dead Cas9 (dCas9) protein or Cas9 nickase (nCas9) to direct their binding to genomic DNA, unintended base substitutions were expected to be induced at OTsg sites that are partially complementary to sgRNA. In this scenario, the use of high-fidelity Cas9 in BEs can reduce these OTsg mutations. Meanwhile, since free APOBECs can induce unexpected C-to-T mutations in single-stranded DNA (ssDNA) regions, the APOBEC moiety of BEs may directly trigger unexpected mutations at OTss sites. In other words, the off-target mutations induced by BEs may also occur at OTss sites independent of the guidance of sgRNA; however, OTss mutations were not revealed due to the lack of a quantitative and reproducible way for detection.

[0172] This example set up an efficient method to quantitatively evaluate BE-induced OTss mutations by co-expressing *S. aureus* and *S. pyogenes* Cas9 orthologs (CESSCO). In CESSCO, the expression of nSaCas9/Sa-sgRNA pairs generated DNA single-strand breaks (SSBs) at specific genomic loci and led to the formation of a genomic ssDNA region in a programmable way. At the same time, co-expressed BE3 in the absence of sgRNA (sgRNA means Sp-sgRNA hereafter) was used to examine whether sgRNA-independent C-to-T base substitutions can be induced by BE3 alone in the ssDNA regions generated around

nSaCas9/Sa-sgRNA-introduced SSBs. After deep-sequencing the genomic regions targeted by nSaCas9/Sa-sgRNA, it was clearly shown that C-to-T mutations at OTss sites were induced by the rat APOBEC1 (rA1) -containing BE3 but not by dSpCas9 in the absence of sgRNA, confirming that OTss mutations are caused by the APOBEC moiety of BEs in an sgRNA-independent manner.

[0173] This example then sought to reduce OTss mutations by exploiting members of APOBEC family suitable for highly-specific BE construction. Most of commonly used BEs were constructed with single domain APOBECs, such as rA1 in BE3, but not with dual-domain APOBECs. Usually, in APOBECs with two CDA domains, one is catalytically active, while the other one is catalytically inactive, which plays a regulatory role on the cytidine deamination activity and thus may be suited for constructing highly-specific BEs with reduced OTss effects. To attempt this possibility, we constructed and compared the C-to-T editing efficiencies of ten paired BEs, which have either one catalytically active CDA domain or two CDA domains of five dual-domain APOBECs (Fig. 20a), *i.e.*, human APOBEC3B (hA3B), human APOBEC3D (hA3D), human APOBEC3F (hA3F), human APOBEC3G (hA3G) and mouse APOBEC3 (mA3).

[0174] As revealed in Fig. 20b,c, the BEs constructed with certain APOBECs (hA3B, hA3F and mA3) containing two CDA domains induced significantly lower editing efficiencies than their paired BEs having only the active CDA domain. This result shows that the catalytically inactive CDA domains from these dual-domain APOBECs, *i.e.*, hA3B, hA3F and mA3, exhibit an inhibitory function on their corresponding active CDA domains.

[0175] To examine whether the inhibitory function is general, we covalently linked the catalytically inactive CDA domains of mA3, hA3F or hA3B individually to the N-terminus of mA3CDA1-nSpCas9-BE (Fig. 20d) and two other commonly used BEs, *i.e.*, BE3 and hA3A-BE3. All these catalytically inactive CDA domains showed broad-spectrum inhibitory effects on all tested BEs, and among them, the CDA2 of mA3 (mA3CDA2) manifested the strongest inhibitory effect (Fig. 20e,f). Detailed mapping analysis further revealed that residues 282-355 of mA3CDA2 exhibited an inhibitory effect similar to that of full-length mA3CDA2. Collectively, these results showed that the catalytically-inactive domains of certain dual-domain APOBECs indeed exhibit general inhibitory effects on cytidine deaminase activity, and we thus defined them as cytidine deaminase inhibitors (CDI).

[0176] Next, we sought to test whether the cleavage of mA3CDI (mA3CDA2) from its covalently-linked BEs can restore their base editing capacity. We used self-cleaving peptides (T2A) to link mA3CDI and mA3CDA1-nSpCas9-BE for examination. After self-cleavage of mA3CDI, the editing efficiency of mA3CDI-T2A-mA3CDA1-nSpCas9-BE was regained to the levels similar to EGFP-mA3CDA1-nSpCas9-BE or EGFP-T2A-mA3CDA1-nSpCas9-BE, ~10-fold higher than that of the non-cleavable-mA3CDI fused BE. The self-cleavage of mA3CDI from BE3 and hA3A-BE3 also enhanced their editing efficiencies, albeit to different extents.

[0177] These results served as a critical proof-of-concept for developing an iBE system for precise base editing with low OTss mutations. iBE1 was constructed by using TEV protease cleavage site (TS) to link three critical modules, *i.e.*, mA3CDI, mA3CDA1-nSpCas9-BE and the N-terminal half of TEV protease (NTEV) (Fig. 21a). In theory, due to the covalent linkage of a CDI, iBE1 remains dormant when it binds to OTss sites by its APOBEC moiety (Fig. 21a). Notably, NTEV itself is inactive but forms a functional TEV protease only when the C-terminal half (CTEV) is recruited. Thus, iBE1 can be guided by its CRISPR-Cas moiety to perform efficient base editing at on-target sites where the CDI is cleaved by the sgRNA-induced assembly of functional TEV protease (Fig. 21d).

[0178] After being expressed in cells, iBE1 remained dormant in the sgRNA-independent OTss regions as expected (Fig. 21b) and induced much lower (~20%) level of C-to-T mutations compared to BE3 (Fig. 21c). At on-target sites, the RNA binding protein (MCP)-fused CTEV can be recruited by the MS2-fused sgRNA (Fig. 21d), which leads to the removal of mA3CDI from iBE1 and therefore enables efficient base editing. The comparison of on-target editing efficiency induced by BE3 and iBE1 across multiple genomic loci (Fig. 21e) demonstrated that iBE1 induced on-target base editing at a similar level as BE3 did (Fig. 21f, ~80% of BE3). Together, this example shows that we have developed an iBE system, which catalyzes efficient base editing at on-target sites with suppressed OTss mutations, through the manipulation of CDI.

[0179] As Cas9 has been known to induce unintended editing at OTsg sites that have partial sequence complementarity to the sgRNA, we also aimed to further reduce OTsg mutations by replacing the unmodified nSpCas9 in iBE1 with its engineered versions that have improved targeting specificity (Fig. 22a). We tested three engineered versions of nSpCas9, *i.e.*, neSpCas9, nSpCas9HF1 and nHypaSpCas9, and found that using either of these targeting-

specificity-improved Cas9 proteins greatly reduced OTsg mutations (Fig. 22b,c). Meanwhile, the use of neSpCas9 did not compromise the on-target editing efficiencies, whereas the use of the other two decreased the on-target editing efficiencies (Fig. 22d,e). In this scenario, we set to replace nSpCas9 with neSpCas9 to construct iBE2.

[0180] As an early developed BE, the editing efficiency of BE3 is restricted under certain conditions and additional BEs with improved editing efficiencies were developed later, *e.g.*, AncBE4max or hA3A-BE3. hA3A-BE3 is a highly active BE in various contexts and we thus compared the performance of iBE2 to that of hA3A-BE3, in terms of editing efficiency and specificity (Fig. 23a). Although the average on-target editing frequency of iBE2 was ~50% of hA3A-BE3 (Fig. 23a,c), the C-to-T mutations induced by iBE2 at OTss and OTsg sites were close to the background level, while hA3A-BE3 induced substantial mutations at these off-target sites (Fig. 23a,b). Taken together, the average editing specificity of iBE2 was ~40-fold higher than that of hA3A-BE3 (Fig. 23d).

[0181] In this example, we first developed an efficient method (CESSCO) to quantitatively evaluate sgRNA-independent OTss mutations and confirmed that the BEs with a regular APOBEC-nCas9 backbone indeed induced OTss mutations in an sgRNA-independent manner (Fig. 21a, 21b, 23a, 23b). Consistent with our findings, recent whole genome sequencing studies also showed that BE3 induced substantial off-target mutations in mice and rice plants, assumedly also in an sgRNA-independent manner. Importantly, we took advantage of our discovery of CDIs to develop iBEs, which remain dormant at OTss sites due to the covalent linkage of CDI but can be activated by sgRNA-mediated cleavage of CDI at on-target sites (Fig. 21a,d). iBEs induced a significantly low level of unintended mutations in sgRNA-independent ssDNA regions, while it performed on-target editing efficiently (Fig. 21b,c,e,f).

[0182] By substituting nSpCas9 with the specificity-improved enSpCas9, highly-specific iBE2 was developed to further reduce unintended editing at OTsg sites (Fig. 22 and 23e). The iBE system is compatible with the BEs having different Cas moieties and the engineered BEs with improved performance, and does not change the characteristics of constructed BEs, such as editing window. In addition, since there are abundant members within APOBEC family, other CDIs may be identified in the future, which would further enrich the repertoire of CDI-conjugated iBE system. As both editing precision and efficiency is essential for base editors, especially in their therapeutic applications, the iBE system developed here will bring a new

layer of regulation for the specificity of current base editing system and secure its application against off-target mutations.

Example 3. Testing of Different Configurations of Induced and Split Base Editors

[0183] This example tested a number of different configurations of molecules for implementing the induced and split base editor (isplitBE) system.

[0184] The working process of isplitBE is illustrated in FIG. 24A, as comparison to the conventional BEs as shown in FIG. 24B. In the illustrated isplitBE system, a nCas9-D10A construction is packaged in an AAV vehicle. A typical AAV vehicle has a 4.7 kb capacity, and the nCas9 construct is about 4.7 kb in length. Another AVV vehicle can package the nucleic acids (about 4.4 kb in total length) for encoding: (a) a fusion protein that includes MCP, UGI, APOBEC, a TEV recognition site (TEV site), and mA3CDA2; (b) a fusion protein with TEVc and N22p; (c) a standalone TEVn, (d) a helper sgRNA (hsgRNA) with a MS2 tag, and (e) another sgRNA with a boxB tag.

[0185] At a target site (ON, left lower branch), each of the hsgRNA and sgRNA binds to two adjacent sites on a target DNA, and the MCP- and N22p-containing fusion proteins bind to the MS2 tag and boxB tag of the hsgRNA and sgRNA, respectively. Due to proximity of TEVc (in the presence of free TEVn) and the TEV site, the TEVc/TEVn cleaves the TEV site, removing the mA3CDA2 from the APOBEC. Without the attached mA3CDA2, the APOBEC can highly efficiently carry out the desired editing.

[0186] At off-target sites, which may be a non-specific binding site (OTss, lower middle branch) or a site that only binds one of the guide RNAs, the TEVc/TEVn complex is not recruited to the TEV site-containing fusion protein, and thus the APOBEC cannot be activated. By contrast, in the conventional BE system (FIG. 24B), the APOBEC is already active and can cause C-to-T editing whenever it is recruited to a single-stranded nucleotide sequence.

[0187] Ten different configurations (Pairs 1-10), as illustrated in FIG. 25, were prepared and tested. For instance, as shown in FIG. 26A, Pair 1 included two constructs, the first of which contained rA1 fused to nCas9-D10A (spD10A), along with a UGI and an NLS, and the second of which contained a sgRNA targeting EMX1. Pair 2 is similar to Pair 1 but rA1 was replaced by hA3A. Pair 3 is also similar and used a mutant hA3A instead (Y130F).

[0188] In Pair 4, the rA1 and the nCas9 proteins were placed on different constructs. rA1 was further fused to an MCP protein which recognizes a MS2 tag on the helper sgRNA. In Pair 5, a mA3CDA2 was further fused to the rA1, via a TEV recognition site (black solid box). In Pair 6, a TEV protein was further fused, through a self-cleavage site 2A, to the rA1-mA3CDA2 fusion. Self cleavage of 2A would release the TEV from the fusion protein.

[0189] Pair 7 is different from Pair 6 by fusing the TEV to a N22p protein, which would recognize the boxB tag on the sgRNA. In Pair 8, the TEV protein was divided into TEVn and TEVc, separated by the 2A self cleavage site. In Pair 9, only the TEVc was fused to a N22p while the TEVn was free of any RNA tag-binding proteins. In Pair 10, the helper sgRNA targeted GFP, rather than a nearby site.

[0190] The constructs in FIG. 26A were designed for C-to-T editing at target site EMX1-ON, and the off-target editing at Sa-SITE31-OTss and EMX1-OTsg sites is examined as well. The testing results are shown in FIG. 26B. isplitBE-rA1 (Pair 9) induced substantial editing at ON site but no editing at OTss or OTsg sites.

[0191] Likewise, all of these configurations were tested with the FANCF-ON, Sa-VEGFA-7-OTss and FANCF-OTsg sites (see schematic diagram in FIG. 27A). FIG. 27B shows the comparison of editing efficiencies for different base editors at FANCF-ON, Sa-VEGFA-7-OTss and FANCF-OTsg sites. Again, isplitBE-rA1 (Pair 9) induced substantial editing at ON site but no editing at OTss or OTsg sites.

[0192] A further testing was done with the V1B-ON, Sa-SITE42-OTss and V1B-OTsg sites (see schematic diagram in FIG. 28A). Again, as shown in FIG. 28B, isplitBE-rA1 (Pair 9) induced substantial editing at ON site but no editing at OTss or OTsg sites.

Example 4. Tuning of Parameters in the isplitBE Systems

[0193] Out of the 10 tested configurations, Pair 9 exhibited the best performance in terms of editing specificity. Pair 9 employs two sgRNA, a helper sgRNA (hsgRNA) and a regular sgRNA. The dual use of sgRNA further enhances specificity as it requires that both target sites are in proximity from each other.

[0194] In a first assay of this example, the optimal distance between the two target sites was assessed. A schematic diagram is presented in FIG. 29A illustrating the distance between

hsgRNA and sgRNA at DNTET1, EMX1 and FANCF sites. FIG. 29B shows the base editing frequencies induced by the indicated sgRNAs and hsgRNAs. The summary of FIG. 29C shows the effect of distance between hsgRNA and sgRNA. Based on the summary, the optimal range of distance for best base editing efficiency is -91 to -34 bp from the PAM of hsgRNA to the PAM of sgRNA.

[0195] The second assay tested the effect of hsgRNA spacer length on base editing efficiency and precision. FIG. 30A presents a schematic diagram illustrating the co-transfection of sgRNA and the hsgRNAs with different spacer lengths at DNEMX1, FANCF and V1A sites. FIG. 30B shows the base editing frequencies induced by the indicated sgRNAs and hsgRNAs at the target sties of hsgRNA and sgRNA. The statistic analysis in FIG. 30C shows the effects of hsgRNA spacer length. As shown, the use of hsgRNA with 10-nt spacer greatly reduced the editing efficiency at hsgRNA target sites but maintained the editing efficiency at sgRNA target sites. Accordingly, a spacer of 9-15 nt in the helper sgRNA sequence can be a good range to ensure efficient editing at the sgRNA target site, while minimizing the editing at the hsgRNA target site.

Example 5. Genome- and Transcriptome-wide Evaluation

[0196] The overall efficiency of the isplitBE system was compared to the conventional BE3. The results are shown in FIG. 31 (editing frequencies induced by indicated base editors at different target sties). There is no apparent sacrifice of efficiency even when the isplitBE had greatly improved specificity.

[0197] Normal cells have a background level of C-to-T mutations attributable to their endogenous APOBEC3 activities. To obtain a more accurate measurement of off-target C-to-T mutations, an APOBEC3 knockout 293FT cell line (293FT-A3KO) was used. FIG. 32A shows the mRNA expression levels in wild-type 293FT cells and the APOBEC3 knockout 293FT cells. FIG. 32B presents a schematic diagram illustrating the procedures to determine genome-wide C-to-T mutations induced by base editors, and the testing results are shown in FIG. 32C (on-target editing efficiencies (left) and the number of genome-wide C-to-T mutations induced by Cas9, BE3, hA3A-BE3-Y130F (Y130F) and isplitBE-rA1). While BE3 and Y130F both had fairly high off-target edits, isplitBE-rA1's off-target editing rate is close to the background (Cas9 only).

[0198] This example then compared transcriptome-wide C-to-U mutations induced by isplitBE-mA3, BE3 and hA3A-BE3-Y130F (Y130F). The numbers of transcriptome-wide C-to-T(U) mutations induced by Cas9, BE3, hA3A-BE3-Y130F (Y130F) and isplitBE-mA3 are shown in FIG. 33A. FIG. 33B shows RNA C-to-U editing frequencies induced by Cas9, BE3, hA3A-BE3-Y130F (Y130F) and isplitBE-mA3. FIG. 33C shows distribution of RNA C-to-U editing induced by BE3 replicate 1 and isplitBE-mA3 replicate 1. Again, isplitBE induced much lower C-to-U editing than BE3.

Example 6. PCSK9 Knockouts

[0199] Proprotein convertase subtilisin/kexin type 9 (PCSK9) is an enzyme encoded by the PCSK9 gene in humans on chromosome 1. It is the 9th member of the proprotein convertase family of proteins that activate other proteins. PCSK9 is inactive when first synthesized, because a section of peptide chains blocks their activity; proprotein convertases remove that section to activate the enzyme. The PCSK9 gene contains one of 27 loci associated with increased risk of coronary artery disease.

[0200] PCSK9 is ubiquitously expressed in many tissues and cell types. PCSK9 binds to the receptor for low-density lipoprotein particles (LDL), which typically transport 3,000 to 6,000 fat molecules (including cholesterol) per particle, within extracellular fluid. The LDL receptor (LDLR), on liver and other cell membranes, binds and initiates ingestion of LDL-particles from extracellular fluid into cells, thus reducing LDL particle concentrations. If PCSK9 is blocked, more LDLRs are recycled and are present on the surface of cells to remove LDL-particles from the extracellular fluid. Therefore, blocking PCSK9 can lower blood LDL-particle concentrations.

[0201] This example tested an approach to inactivate PCSK9 by introducing stop codons through base editing using the present technology. The sequences of the sgRNA/hsgRNA used are shown in Table 3, and the target sites on PCSK9 are shown in Table 4.

[0202] The numbers of stop codons generated by base editing were measured for the human PCSK9 gene. FIG. 34A presents a schematic diagram illustrating the co-transfection of sgRNA and the hsgRNAs with isplitBE-mA3 and nCas9. Editing efficiencies induced by isplitBE-mA3 at indicated sites are shown in FIG. 34B-D. These results demonstrate the high efficiency and specificity of the method.

Table 3: Regular sgRNA and hsgRNA scaffolds and target sites in PCSK9 genes

Name	Sequence	SEQ ID NO:
hsgCAG2-MS2	GAGGUUGCCUGGCACCUACGGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	136
hsgCAG3-MS2	GAGACCCACCUCUCGCAGUCGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	137
hsgCAG4-MS2	GCCCCAUGUCGACUACAUCGGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	138
hsgCAG7-MS2	AUGGUCACCGACUUCGAGAAGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	139
hsgCAG8-MS2	ACCUUGGCUUUGUUCCUCCCGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	140
hsgCAG9-MS2	GGCUUUGUUCCUCCCAGGCCGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	141
hsgCAG10-MS2	GUGGUGCUGCUGCCCCUGGCGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	142
hsgCAG11-MS2	UGCUGCUGCCCCUGGCGGGUGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	143
hsgCAG12-MS2	ACCCACCUCCUCACCUUUCCGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	144
hsgCAG14-MS2	AGCGACUGCAGCACCUGCUUGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	145
hsgCAG15-MS2	AACGCUUUUGGGGGUGAGGGGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	146
hsgCAG16-MS2	CCACACAGCUCCACCAGCUGGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	147
hsgCAG17-MS2	CACUGGGAGGUGGAGGACCUGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	148
hsgCAG18-MS2	CCCACAAGCCGCCUGUGCUGGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	149
hsgCAG19-MS2	AGGUCUGGAAUGCAAAGUCAGUUUGAGAGCUAGGCCAACAUGAGGAUCACCCAU GUCUGCAGGGCCUAGCAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAA GUGGCACCGAGUCGGUGC	150
sgCAG2-boxB	CUCUCGCAGUCAGAGCGCACGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	151
sgCAG3-boxB	CAGGCCCAGGCUGCCCGCCGGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	152
sgCAG4-boxB	UCUUUGCCCAGAGCAUCCCGGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	153
sgCAG7-boxB	CACAGACAGGUAAGCACGGCGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	154
sgCAG8-boxB	AAGCCAGCUGGUCCAGCCUGGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	155

sgCAG9-boxB	GGUCCAGCCUGUGGGGCCACGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	156
sgCAG10-boxB	CGCCUGCCAGCGCCUGGCGAGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	157
sgCAG11-boxB	UGCCAGCGCCUGGCGAGGGCGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	158
sgCAG12-boxB	AAGACCAGCCGGUGACCCUGGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	159
sgCAG14-boxB	AUCACAGGCUGCUGCCCACGGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	160
sgCAG15-boxB	CUACCCCAGGCCAACUGCAGGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	161
sgCAG16-boxB	CAACAGGGCCACGUCCUCACGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	162
sgCAG17-boxB	AGGUCAGCCCAACCAGUGCGGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	163
sgCAG18-boxB	CCAACCAGUGCGUGGGCCACGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	164
sgCAG19-boxB	CCCCUCAGGAGCAGGUGAAGGUUUGAGAGCUAGGGCCCUGAAGAAGGGCCCUAG CAAGUUCAAAUAAGGCUAGUCCGUUAUCAACUUGGGCCCUGAAGAAGGGCCCAA GUGGCACCGAGUCGGUGC	165

Table 4: Target sites in PCSK9 genes

Name	**Sequence**	**SEQ ID NO:**
hsgCAG2	GAGGTTGCCTGGCACCTACGTGG	166
hsgCAG3	GAGACCCACCTCTCGCAGTCAGA	167
hsgCAG4	GCCCCATGTCGACTACATCGAGG	168
hsgCAG7	ATGGTCACCGACTTCGAGAATGT	169
hsgCAG8	ACCTTGGCTTTGTTCCTCCCAGG	170
hsgCAG9	GGCTTTGTTCCTCCCAGGCCTGG	171
hsgCAG10	GTGGTGCTGCTGCCCCTGGCGGG	172
hsgCAG11	TGCTGCTGCCCCTGGCGGGTGGG	173
hsgCAG12	ACCCACCTCCTCACCTTTCCAGG	174
hsgCAG14	AGCGACTGCAGCACCTGCTTTGT	175
hsgCAG15	AACGCTTTTGGGGGTGAGGGTGT	176
hsgCAG16	CCACACAGCTCCACCAGCTGAGG	177
hsgCAG17	CACTGGGAGGTGGAGGACCTTGG	178
hsgCAG18	CCCACAAGCCGCCTGTGCTGAGG	179
hsgCAG19	AGGTCTGGAATGCAAAGTCAAGG	180
sgCAG2	CTCTCGCAGTCAGAGCGCACTGC	181
sgCAG3	CAGGCCCAGGCTGCCCGCCGGGG	182
sgCAG4	TCTTTGCCCAGAGCATCCCGTGG	183
sgCAG7	CACAGACAGGTAAGCACGGCCGT	184
sgCAG8	AAGCCAGCTGGTCCAGCCTGTGG	185
sgCAG9	GGTCCAGCCTGTGGGGCCACTGG	186
sgCAG10	CGCCTGCCAGCGCCTGGCGAGGG	187
sgCAG11	TGCCAGCGCCTGGCGAGGGCTGG	188
sgCAG12	AAGACCAGCCGGTGACCCTGGGG	189
sgCAG14	ATCACAGGCTGCTGCCCACGTGG	190

sgCAG15	CTACCCCAGGCCAACTGCAGCGT	191
sgCAG16	CAACAGGGCCACGTCCTCACAGG	192
sgCAG17	AGGTCAGCCCAACCAGTGCGTGG	193
sgCAG18	CCAACCAGTGCGTGGGCCACAGG	194
sgCAG19	CCCCTCAGGAGCAGGTGAAGAGG	195

Example 7. Applicability of the isplitBE Design in Adenine Base Editors

[0203] This example confirms the applicability of the induced and split base editor (isplitBE) design in other types of base editors. The inhibitor used was mA3CDA2 and the editor was adenine base editor (ABE).

[0204] A schematic diagram illustrating the co-transfection of sgRNA and ABE fused with mA3CDA2 (or not as control) is shown in FIG. 35A. The editing efficiencies induced by indicated ABEs at RNF2 and FANCF sites are shown in FIG. 35B. With mA3CDA2 attached to the ABE, the editing efficiency was reduced as compared to ABE alone. When the mA3CDA2 was cleaved by 2A, the editing efficiency of ABE was restored, validating the isplitBE approach for ABE.

Example 8. Enhanced Prime Editing

[0205] The conventional base editors are limited to base transitions, not base transversions, insertions or deletions. Recently, a primer editing system was proposed, which employs a primer editor (PE) by conjugating a Cas9 nickase with a reverse transcriptase (RTase). The PE system can write genomes with almost any intended changes, including all types of base substitutions, small indels, and their combinations. The overall efficiency and specificity of the PE systems, however, are still limited.

[0206] In a first assay, this example tested a new design for the primer editing guide RNA (pegRNA). Conventionally, each guide RNA includes a scaffold. A commonly used scaffold sequence is GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUU GAAAAAGUGGCACCGAGUCGGUGC (SEQ ID NO:29). Another example is GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCAUGUCCGUUAUCAACUU GAAAAAGUGGCACCGAUUCGGUGC (SEQ ID NO:30). A more generic consensus sequence is GUUUNAGAGCUA**X**$_1$UAGCAAGUUNAAAUAAGGCNNGUCCGUUAUCAACUU**X**$_2$

AAGUGGCACCGANUCGGUGC (SEQ ID NO:31), where N represents any base, and X1 and X2 denotes any nucleotide sequence of a length of 2-50 bases.

[0207] The scaffold is expected to form a secondary structure (illustrated in FIG. 36A, SEQ ID NO:30) due to its internal complementary sequences. A typical sgRNA used in base editors is about 96 nt in length which includes a spacer that is about 20 nt in length and binds to the target site. In a pegRNA, a reverse-transcription template and a primer-binding site are further added to the 3' end of the scaffold. Surprisingly, it is discovered herein that the original scaffold is not stable enough in the context of the pegRNA.

[0208] A new scaffold was therefore prepared, which forms a new pairing between positions 48 (e.g., A in SEQ ID NO:30) and 61 (e.g., G in SEQ ID NO:30). In the examples shown in FIG. 36A and 36E, the new scaffold has G and C or C and G instead (SEQ ID NO:36, 37). This and additional example mutant scaffolds are shown in Table 5 below.

Table 5. Sequences of Guide RNA Scaffolds

Description	Guide RNA Scaffold	SEQ ID NO:
Original	GUUUNAGAGCUA-X1-UAGCAAGUUNAAAUAAGGCNNGUCCGUUAUC**A**AC UU-X2-AAGU**G**GCACCGANUCGGUGC	31
New 1	GUUUNAGAGCUA-X1-UAGCAAGUUNAAAUAAGGCNNGUCCGUUAUC**G**AC UU-X2-AAGU**C**GCACCGANUCGGUGC	32
New 2	GUUUNAGAGCUA-X1-UAGCAAGUUNAAAUAAGGCNNGUCCGUUAUC**C**AC UU-X2-AAGU**G**GCACCGANUCGGUGC	33
New 3	GUUUNAGAGCUA-X1-UAGCAAGUUNAAAUAAGGCNNGUCCGUUAUC**A**AC UU-X2-AAGU**U**GCACCGANUCGGUGC	34
New 4	GUUUNAGAGCUA-X1-UAGCAAGUUNAAAUAAGGCNNGUCCGUUAUC**U**AC UU-X2-AAGU**A**GCACCGANUCGGUGC	35
Original 2	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCAUGUCCGUUAUC**A**AC UUGAAAAAGU**G**GCACCGAUUCGGUGC	30
New 5	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCAUGUCCGUUAUC**G**AC UUGAAAAAGU**C**GCACCGAUUCGGUGC	36
New 6	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCAUGUCCGUUAUC**C**AC UUGAAAAAGU**G**GCACCGAUUCGGUGC	37
New 7	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCAUGUCCGUUAUC**A**AC UUGAAAAAGU**U**GCACCGAUUCGGUGC	38
New 8	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCAUGUCCGUUAUC**U**AC UUGAAAAAGU**A**GCACCGAUUCGGUGC	39
Original 3	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUC**A**AC UUGAAAAAGU**G**GCACCGAGUCGGUGC	29
New 9	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUC**G**AC UUGAAAAAGU**C**GCACCGAGUCGGUGC	40
New 10	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUC**C**AC UUGAAAAAGU**G**GCACCGAGUCGGUGC	41
New 11	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUC**A**AC UUGAAAAAGU**U**GCACCGAGUCGGUGC	42

New 12	GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUC**U**AC UUGAAAAAGU**A**GCACCGAGUCGGUGC	43

[0209] Constructs for testing the conventional pegRNA and the newly designed, enhanced pegRNA (epegRNA) were prepare as shown in FIG. 36B and 36F for PE2, and the testing results are shown in FIG. 36C-36D and 36G. Comparison of prime editing efficiencies induced with pegRNA and epegRNA. The epegRNA, with greatly improved stem stability, exhibited much higher editing efficiency than pegRNA across the board.

[0210] Likewise, according to the schematic diagram in FIG. 37A, co-transfection of pegRNA, nicking sgRNA with PE2-NG (SEQ ID NO:132) or xPE2 (SEQ ID NO:133) was made to test the editing efficiency for TGATG deletion. The results are shown in FIG. 37B. PE2-NG has an engineered Cas9 that can recognize relaxed NG PAMs (see, e.g., Nishimasu et al., Science 361, 1259-62 (2018)). xPE2 has an engineered Cas9 that can recognize relaxed NG, GAA and GAT PAMs (see, e.g., Hu et al., Nature 556, 57-63 (2018)). The sequences of PE2-NG (SEQ ID NO:44), xPE2 (SEQ ID NO:45), SpCas9-NG (SEQ ID NO:46), and xSpCas9 (SEQ ID NO:47) are shown in Table 6 below.

Table 6. Cas and PE sequences

Name	**Sequence**	**SEQ ID NO:**
PE2-NG	MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVD EVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFI QLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGL TPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNT EITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEF YKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLK DNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMT NFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRK VTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIV LTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDF LKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVV DELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQL QNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDAIVPQSFLKDDSIDNKVLTRSDKNRGKSD NVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKH VAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAV VGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESIRPKRNS DKLIARKKDWDPKKYGGFVSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNP IDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASARFLQKGNELALPSKYVNFLYLAS HYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPI REQAENIIHLFTLTNLGAPRAFKYFDTTIDRKVYRSTKEVLDATLIHQSITGLYETRIDLSQ LGGDGTSGGSSGGSSGSETPGTSESATPESSGGSSGGSSTLNIEDEYRLHETSKEPDVSLGS TWLSDFPQAWAETGGMGLAVRQAPLIIPLKATSTPVSIKQYPMSQEARLGIKPHIQRLLDQG ILVPCQSPWNTPLLPVKKPGTNDYRPVQDLREVNKRVEDIHPTVPNPYNLLSGLPPSHQWYT VLDLKDAFFCLRLHPTSQPLFAFEWRDPEMGISGQLTWTRLPQGFKNSPTLFNEALHRDLAD FRIQHPDLILLQYVDDLLLAATSELDCQQGTRALLQTLGNLGYRASAKKAQICQKQVKYLGY LLKEGQRWLTEARKETVMGQPTPKTPRQLREFLGKAGFCRLFIPGFAEMAAPLYPLTKPGTL FNWGPDQQKAYQEIKQALLTAPALGLPDLTKPFELFVDEKQGYAKGVLTQKLGPWRRPVAYL SKKLDPVAAGWPPCLRMVAAIAVLTKDAGKLTMGQPLVILAPHAVEALVKQPPDRWLSNARM THYQALLLDTDRVQFGPVVALNPATLLPLPEEGLQHNCLDILAEAHGTRPDLTDQPLPDADH	44

	TWYTDGSSLLQEGQRKAGAAVTTETEVIWAKALPAGTSAQRAELIALTQALKMAEGKKLNVY TDSRYAFATAHIHGEIYRRRGWLTSEGKEIKNKDEILALLKALFLPKRLSIIHCPGHQKGHS AEARGNRMADQAARKAAITETPDTSTLLIENSSP	
xPE2	MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVD EVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFI QLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGL TPNFKSNFDLAEDTKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNT EITKAPLSASMIKLYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEF YKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGIIPHQIHLGELHAILRRQEDFYPFLK DNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEKVVDKGASAQSFIERMT NFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGDQKKAIVDLLFKTNRK VTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIV LTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDF LKSDGFANRNFIQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVV DELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQL QNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDAIVPQSFLKDDSIDNKVLTRSDKNRGKSD NVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKH VAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAV VGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNS DKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNP IDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGVLQKGNELALPSKYVNFLYLAS HYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPI REQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQ LGGDGTSGGSSGGSSGSETPGTSESATPESSGGSSGGSSTLNIEDEYRLHETSKEPDVSLGS TWLSDFPQAWAETGGMGLAVRQAPLIIPLKATSTPVSIKQYPMSQEARLGIKPHIQRLLDQG ILVPCQSPWNTPLLPVKKPGTNDYRPVQDLREVNKRVEDIHPTVPNPYNLLSGLPPSHQWYT VLDLKDAFFCLRLHPTSQPLFAFEWRDPEMGISGQLTWTRLPQGFKNSPTLFNEALHRDLAD FRIQHPDLILLQYVDDLLLAATSELDCQQGTRALLQTLGNLGYRASAKKAQICQKQVKYLGY LLKEGQRWLTEARKETVMGQPTPKTPRQLREFLGKAGFCRLFIPGFAEMAAPLYPLTKPGTL FNWGPDQQKAYQEIKQALLTAPALGLPDLTKPFELFVDEKQGYAKGVLTQKLGPWRRPVAYL SKKLDPVAAGWPPCLRMVAAIAVLTKDAGKLTMGQPLVILAPHAVEALVKQPPDRWLSNARM THYQALLLDTDRVQFGPVVALNPATLLPLPEEGLQHNCLDILAEAHGTRPDLTDQPLPDADH TWYTDGSSLLQEGQRKAGAAVTTETEVIWAKALPAGTSAQRAELIALTQALKMAEGKKLNVY TDSRYAFATAHIHGEIYRRRGWLTSEGKEIKNKDEILALLKALFLPKRLSIIHCPGHQKGHS AEARGNRMADQAARKAAITETPDTSTLLIENSSP	45
SpCas9-NG	MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVD EVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFI QLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGL TPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNT EITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEF YKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLK DNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMT NFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRK VTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIV LTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDF LKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVV DELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQL QNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSD NVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKH VAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAV VGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESIRPKRNS DKLIARKKDWDPKKYGGFVSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNP IDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASARFLQKGNELALPSKYVNFLYLAS HYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPI REQAENIIHLFTLTNLGAPRAFKYFDTTIDRKVYRSTKEVLDATLIHQSITGLYETRIDLSQ LGGD	46
xSpCas9	MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVD EVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFI QLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGL TPNFKSNFDLAEDTKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNT EITKAPLSASMIKLYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEF YKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGIIPHQIHLGELHAILRRQEDFYPFLK DNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEKVVDKGASAQSFIERMT	47

	NFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGDQKKAIVDLLFKTNRK VTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIV LTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDF LKSDGFANRNFIQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVV DELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQL QNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSD NVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKH VAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAV VGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNS DKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNP IDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGVLQKGNELALPSKYVNFLYLAS HYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPI REQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQ LGGD	

[0211] A complete prime editor requires a construct (about 11 kb) that is much larger than what an AAV vehicle can accommodate. Accordingly, a Split PE system was designed and tested. The original PE system is illustrated on the left panel of FIG. 38A, and the newly designed Split PE system is illustrated on the right panel, in which the nickase and the RTase are packaged into different AAV particles. The RTase is fused to an RNA binding protein MCP, and the pegRNA includes a binding site MS2. When taken up into a cell, the RTase can be recruited by the pegRNA, through the MS2-MCP binding, and come in contact with the nickase.

[0212] An example co-transfection system is illustrated in FIG. 38B, and the testing results are shown in FIG. 38C, at the EMX1 site.

* * *

[0213] 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.

[0214] 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 a cytidine or adenosine deaminase or a catalytic domain thereof,

a second fragment inhibiting the first fragment, comprising:

i) an amino acid sequence selected from SEQ ID NO: 1-2 and 48-135, or

ii) amino acids residues AA128-AA223 of SEQ ID NO:1, and

a protease cleavage site between the first fragment and the second fragment, wherein protease cleavage site is a protease cleavage site of a protease selected from the group consisting of TEV protease, TuMV protease, PPV protease, PVY protease, ZIKV protease and WNV protease, or is a self-cleavage site.

2. The fusion protein of claim 1, wherein the adenosine deaminase is selected from the group consisting of tRNA-specific adenosine deaminase (TadA), adenosine deaminase tRNA specific 1 (ADAT1), adenosine deaminase tRNA specific 2 (ADAT2), adenosine deaminase tRNA specific 3 (ADAT3), adenosine deaminase RNA specific B1 (ADARB1), adenosine deaminase RNA specific B2 (ADARB2), adenosine monophosphate deaminase 1 (AMPD1), adenosine monophosphate deaminase 2 (AMPD2), adenosine monophosphate deaminase 3 (AMPD3), adenosine deaminase (ADA), adenosine deaminase 2 (ADA2), adenosine deaminase like (ADAL), adenosine deaminase domain containing 1 (ADAD1), adenosine deaminase domain containing 2 (ADAD2), adenosine deaminase RNA specific (ADAR) and adenosine deaminase RNA specific B1 (ADARB1).

3. The fusion protein of claim 1, wherein the cytidine deaminase is selected from the group consisting of APOBEC3B (A3B), APOBEC3C (A3C), APOBEC3D (A3D), APOBEC3F (A3F), APOBEC3G (A3G), APOBEC3H (A3H), APOBEC1 (A1), APOBEC3 (A3), APOBEC2 (A2), APOBEC4 (A4) and AICDA (AID).

4. The fusion protein of claim 3, wherein the catalytic domain is mouse A3 cytidine deaminase domain 1 (CDA1) or human A3B cytidine deaminase domain 2 (CDA2).

5. A fusion protein comprising:

a first fragment comprising a cytidine deaminase or a catalytic domain thereof,

a second fragment inhibiting the first fragment, comprising:

i) an amino acid sequence selected from SEQ ID NO: 1-2 and 48-135, or

ii) amino acids residues AA128-AA223 of SEQ ID NO:1,

a first RNA recognition peptide, and

a TEV protease cleavage site between the nucleobase deaminase or a catalytic domain thereof and the nucleobase deaminase inhibitor.

6. An *in vitro* method for conducting gene editing in a cell, comprising introducing to the cell:

(a) a first molecule comprising a Cas protein;

(b) a second molecule comprising the fusion protein of any one of claims 1-5 and a first RNA recognition peptide;

(c) a third molecule comprising an inactive portion of the TEV protease and a second RNA recognition peptide;

(d) a fourth molecule comprising another portion of the TEV protease which, in combination with the inactive portion of the TEV protease, can carry out the TEV protease activity to cleave the cleavage site in (b);

(e) a fifth molecule comprising a helper sgRNA that comprises a first RNA recognition site recognizable by the first RNA recognition peptide of (b); and

(f) a sixth molecule comprising a sgRNA that comprises a second RNA recognition site recognizable by the second RNA recognition peptide of (c).

7. One or more polynucleotides for conducting gene editing in a subject, wherein the one or more polynucleotides encode:

(a) a first molecule comprising a Cas protein;

(b) a second molecule comprising the fusion protein of any one of claims 1-5 and a first RNA recognition peptide;

(c) a third molecule comprising an inactive portion of the TEV protease and a second RNA recognition peptide;

(d) a fourth molecule comprising another portion of the TEV protease which, in combination with the inactive portion of the TEV protease, can carry out the TEV protease activity to cleave the cleavage site in (b);

(e) a fifth molecule comprising a helper sgRNA that comprises a first RNA recognition site recognizable by the first RNA recognition peptide of (b); and

(f) a sixth molecule comprising a sgRNA that comprises a second RNA recognition site recognizable by the second RNA recognition peptide of (c).