

METHODS AND COMPOSITIONS FOR GENOMIC INTEGRATION**CROSS-REFERENCE**

[0001] This application claims the benefit of U.S. Provisional Application No. 63/608,616, filed on December 11, 2023, U.S. Provisional Application No. 63/618,090, filed on January 05, 2024, U.S. Provisional Application No. 63/625,640, filed on January 26, 2024, U.S. Provisional Application No. 63/626,486 filed on January 29, 2024, U.S. Provisional Application No. 63/642,047, filed on May 03, 2024, U.S. Provisional Application No. 63/698,812, filed on September 25, 2024, and U.S. Provisional Application No. 63/705,442, filed on October 09, 2024, each of which is incorporated herein by reference in their entireties.

BACKGROUND

[0002] Cell therapy is a rapidly developing field for addressing difficulties in treating diseases, such as cancer, persistent infections and certain diseases that are refractory to other forms of treatment. Cell therapy often utilizes cells that are engineered *ex vivo* and administered to an organism to correct deficiencies within the body. An effective and reliable system for manipulation of a cell's genome is crucial, in the sense that when the engineered cell is administered into an organism, it functions optimally and with prolonged efficacy. Likewise, reliable mechanisms of genetic manipulation form the cornerstone in the success of gene therapy. However, severe deficiencies exist in methods for delivering nucleic acid cargo (e.g., large cargo) in a therapeutically safe and effective manner. Viral delivery mechanisms are frequently used to deliver large nucleic acid cargo in a cell but are tied to safety issues and cannot be used to express the cargo in some cell types. Additionally, subjecting a cell to repeated gene manipulation can affect cell health, induce alterations of cell cycle and render the cell unsuitable for therapeutic use. Advancements are continually sought in the area for efficacious delivery and stabilization of an exogenously introduced genetic material for therapeutic purposes.

SUMMARY

[0003] Provided herein is a composition comprising: (a) an RNA molecule comprising: (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; (ii) a 5' homology arm and a 3' homology arm; and (iii) a sequence encoding a human ORF2p polypeptide, wherein the sequence encoding a human ORF2p polypeptide is

upstream of the 5' homology arm or downstream of the 3' homology arm; (b) an endonuclease or a polynucleic acid sequence encoding the endonuclease; and (c) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.

[0004] Also provided herein is a composition comprising: (a) an RNA molecule comprising: (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; (ii) a 5' homology arm and a 3' homology arm; and (iii) a sequence encoding a polypeptide with reverse transcriptase activity, wherein the sequence encoding a polypeptide with reverse transcriptase activity is upstream of the 5' homology arm or downstream of the 3' homology arm; (b) an endonuclease or a polynucleic acid sequence encoding the endonuclease; and (c) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.

[0005] Also provided herein is a composition comprising: (a) an RNA molecule comprising: (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; (ii) a 5' homology arm that is upstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide and a 3' homology arm that is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; and (iii) a sequence encoding a polypeptide with reverse transcriptase activity, wherein the sequence encoding a polypeptide with reverse transcriptase activity is upstream of

the 5' homology arm or downstream of the 3' homology arm; (b) an endonuclease or a polynucleic acid sequence encoding the endonuclease; and (c) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.

[0006] In some embodiments, the RNA molecule of (a) comprises the polynucleic acid sequence encoding the endonuclease of (b).

[0007] In some embodiments, the sequence of the 5' homology arm or the 3' homology arm is about 50 nucleotides in length.

[0008] In some embodiments, the polynucleic acid sequence encoding the endonuclease is upstream of the 5' homology arm sequence or downstream of the 3' homology arm sequence.

[0009] In some embodiments, the polynucleic acid sequence encoding the endonuclease is present in a polynucleic acid molecule that is different than the RNA molecule of (a).

[0010] In some embodiments, the RNA molecule of (a) comprises the one or more polynucleic acids encoding the one or more guide RNA molecules of (c), and wherein the one or more polynucleic acids encoding the one or more guide RNA molecules are upstream of the 5' homology arm sequence or downstream of the 3' homology arm sequence.

[0011] In some embodiments, the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of a second strand of the genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of a first strand of the genomic DNA of the target cell.

[0012] In some embodiments, the number of nucleobases between first target sequence of the second strand of the genomic DNA of a target cell and the second target sequence of the first strand of the of the genomic DNA of the target cell is about 10-1000 nucleobases, about 50-500 nucleobases, or about 100-300 nucleobases.

[0013] In some embodiments, the 5' homology arm sequence comprises a sequence capable of hybridizing to a target sequence of the second strand of the genomic DNA of the target cell and the 3' homology arm sequence comprises a sequence capable of hybridizing to a target sequence of the first strand of the genomic DNA of the target cell.

[0014] In some embodiments, the sequence of the 3' homology arm comprises a sequence that primes with the 3'-flap on the first strand created by Cas 9 nicking.

[0015] In some embodiments, the 3' end of the target sequence of the second strand of the genomic DNA to which the 5' homology arm hybridizes is at a distance of at most 50, 40, 30, 20, 10, 5, 4, 3, 2, or 1 bases from the 3' end of the target sequence of the first strand to which the 3' homology arm hybridizes.

[0016] In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA overlaps with the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence.

[0017] In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 50, 40, 30, 20, 10, 5, 4, 3, 2, or 1 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence.

[0018] In some embodiments, the exogenous sequence encoding a polypeptide is flanked by the 5' homology arm and the 3' homology arm.

[0019] In some embodiments, the endonuclease is a nickase.

[0020] In some embodiments, the nickase is a Cas9 nickase.

[0021] In some embodiments, the Cas9 nickase comprises an H840A mutation relative to the wild type Cas9.

[0022] In some embodiments, the Cas9 nickase is encoded by a nucleic acid sequence with at least 90% sequence identity to SEQ ID NO: 124.

[0023] In some embodiments, the polynucleic acid sequence encoding the endonuclease is an RNA sequence.

[0024] In some embodiments, the one or more guide RNA molecules comprises the first guide RNA molecule comprising a first guide RNA sequence and the second guide RNA molecule comprising a second guide RNA sequence, and wherein the first guide RNA molecule forms a complex with a first endonuclease molecule and the second guide RNA molecule forms a complex with a second endonuclease molecule.

[0025] In some embodiments, the one or more guide RNA molecules direct the endonuclease to cut (i) a first strand of the genome DNA of the target cell and (ii) a second strand of the genomic DNA of the target cell.

[0026] In some embodiments, the one or more guide RNA molecules do not direct the endonuclease to create a double strand break (DSB) in the genome DNA of the target cell.

[0027] In some embodiments, the first guide RNA molecule directs the endonuclease to cut a first strand of the genome DNA of the target cell and the second guide RNA molecule directs the endonuclease to cut a second strand of the genomic DNA of the target cell.

[0028] In some embodiments, the number of nucleobases between the first cut of the first strand of the genomic DNA and the second cut of the second strand of the genomic DNA is about 10-1000 bases, about 50-500 bases, or about 100-300 bases.

[0029] In some embodiments, the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell are within a same region of the genomic DNA of the target cell.

[0030] In some embodiments, the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell are within a same locus of the genomic DNA of the target cell.

[0031] In some embodiments, the locus is a genomic safe harbor locus.

[0032] In some embodiments, the locus is non-ribosomal DNA.

[0033] In some embodiments, the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell do not comprise a sequence of ribosomal DNA.

[0034] In some embodiments, the locus is a human ortholog of the mouse Rosa26 locus, adeno-associated virus site 1 (AAVS1), CCR5 gene, HEK3, PRNP, or IDS loci.

[0035] In some embodiments, the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell comprise sequences from a single gene within the genomic DNA of the target cell.

[0036] In some embodiments, the single gene is a non-essential gene.

[0037] In some embodiments, the polypeptide with reverse transcriptase activity promotes (i) reverse transcription of the reverse complement sequence of the insert sequence thereby producing the insert sequence, (ii) integration of the insert sequence into the genomic DNA of the target cell, or (iii) integration of the insert sequence into the genomic DNA of the target cell via target primed reverse transcription (TPRT).

[0038] In some embodiments, the polypeptide with reverse transcriptase activity promotes (i) integration of the insert sequence into the genomic DNA of the target cell at the first target sequence of the genomic DNA of the target cell, (ii) integration of the insert sequence into the genomic DNA of the target cell at the second target sequence of the genomic DNA of the target cell, or (iii) both.

[0039] In some embodiments, the sequence encoding a polypeptide with reverse transcriptase activity comprises a mobile genetic element sequence.

[0040] In some embodiments, the sequence encoding a polypeptide with reverse transcriptase activity comprises a human mobile genetic element sequence.

[0041] In some embodiments, the mobile human mobile genetic element comprises LINE-1 or a fragment thereof.

[0042] In some embodiments, the polypeptide with reverse transcriptase activity comprises a human ORF2p polypeptide or a functional fragment thereof.

[0043] In some embodiments, the human ORF2p polypeptide or functional fragment thereof lacks endonuclease (EN) activity.

[0044] In some embodiments, the human ORF2p polypeptide or functional fragment thereof comprises a mutation in the endonuclease domain that results in a lack of endonuclease activity.

[0045] In some embodiments, the human ORF2p polypeptide or functional fragment thereof comprises a mutation at D205 relative to wild type human ORF2p.

[0046] In some embodiments, the mutation at D205 comprises a D205A mutation relative to wild type human ORF2p.

[0047] In some embodiments, the human ORF2p polypeptide or functional fragment thereof is a functional fragment of a human ORF2p polypeptide that lacks an endonuclease domain.

[0048] In some embodiments, the composition further comprises (d) an RNA molecule comprising a sequence encoding a human ORF2p polypeptide that lacks a functional endonuclease domain, wherein the RNA molecule of (d) lacks sequences encoding a human ORF1p polypeptide or the exogenous sequence encoding a polypeptide.

[0049] In some embodiments, the human ORF2p polypeptide that lacks a functional endonuclease domain comprises an exogenous DNA binding domain.

[0050] In some embodiments, the exogenous DNA binding domain increases DNA binding activity of the human ORF2p polypeptide or functional fragment thereof.

[0051] In some embodiments, the exogenous DNA binding domain increases TPRT processivity of the human ORF2p polypeptide or functional fragment thereof.

[0052] In some embodiments, the exogenous DNA binding domain comprises a Sso7d DNA binding domain.

[0053] In some embodiments, the Sso7d DNA binding domain comprises an amino acid sequence with at least 90% sequence identity to SEQ ID NO: 152.

[0054] In some embodiments, the RNA molecule of (d) comprises a sequence with at least 90% sequence identity to SEQ ID NO: 144.

[0055] In some embodiments, the exogenous DNA binding domain is derived from an HMGD protein.

[0056] In some embodiments, the exogenous DNA binding domain comprises an amino acid sequence with at least 90% sequence identity to SEQ ID NO: 154.

[0057] In some embodiments, the RNA molecule of (d) comprises a sequence with at least 90% sequence identity to SEQ ID NO: 164.

[0058] In some embodiments, the exogenous DNA binding domain increases direct recruitment of the endonuclease to the target sequence in the genomic DNA of the target cell.

[0059] In some embodiments, the exogenous DNA binding domain is derived from an hPARP1 protein.

[0060] In some embodiments, the exogenous DNA binding domain comprises ZnF1 domain and ZnF2 domain of the hPARP1 protein.

[0061] In some embodiments, the exogenous DNA binding domain comprises an amino acid sequence with at least 90% sequence identity to SEQ ID NO: 155.

[0062] In some embodiments, the RNA molecule of (d) comprises a sequence with at least 90% sequence identity to SEQ ID NO: 146.

[0063] In some embodiments, the additional DNA binding domain is operably linked to the functional fragment of a human ORF2p polypeptide via a linker.

[0064] In some embodiments, the linker is a flexible linker.

[0065] In some embodiments, the flexible linker comprises a 33XTEN linker.

[0066] In some embodiments, the human ORF2p polypeptide that lacks a functional endonuclease domain is a human ORF2p polypeptide that lacks an endonuclease domain.

[0067] In some embodiments, a ratio of an amount of the RNA molecule of (a) to that of the RNA molecule of (d) is about 5:1.

[0068] In some embodiments, the RNA molecule of (a) further comprises a sequence encoding a human ORF1p polypeptide, wherein the sequence encoding the human ORF1p polypeptide is upstream of the 5' homology arm or downstream of the 3' homology arm.

[0069] In some embodiments, the sequence encoding a human ORF1p polypeptide and the sequence encoding a human ORF2p polypeptide are separated by an interORF sequence.

[0070] In some embodiments, the sequence encoding a human ORF1p polypeptide and the sequence encoding a human ORF2p polypeptide are separated by (a) a sequence encoding a GSS linker and (b) a sequence encoding a T2A cleavage sequence.

[0071] In some embodiments, the sequence encoding the human ORF1 polypeptide comprises a sequence with at least 90% sequence identity to SEQ ID NO: 10.

[0072] In some embodiments, the polypeptide is a therapeutic polypeptide.

[0073] In some embodiments, the polypeptide is selected from the group consisting of a ligand, an antibody, a receptor, an enzyme, a transport protein, a structural protein, a hormone, a contractile protein, a storage protein and a transcription factor.

[0074] In some embodiments, the polypeptide is a receptor selected from the group consisting of a chimeric antigen receptor (CAR) and a T cell receptor (TCR).

[0075] In some embodiments, the target cell is a mammalian cell, a human cell, a primary cell, an ex vivo cell, or an in vivo cell.

[0076] In some embodiments, the target cell is an immune cell selected from the group consisting of a T cell, a B cell, a myeloid cell, a monocyte, a macrophage, and a dendritic cell.

[0077] In some embodiments, the composition further comprises one or more delivery vehicles for delivery of the RNA molecule, the endonuclease or the polynucleic acid sequence encoding the endonuclease, and the one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules into the target cell.

[0078] In some embodiments, the one or more delivery vehicles is selected from the group consisting of: a nanoparticle delivery vehicle, a plasmid vector, and a viral vector.

[0079] In some embodiments, the viral vector is an adenoviral vector, an adeno-associated viral vector, a lentiviral vector, or a retroviral vector.

[0080] In some embodiments, the nanoparticle delivery vehicle is selected from the group consisting of a lipid nanoparticle and a polymeric nanoparticle.

[0081] In some embodiments, the 3' homology arm comprises a nucleic acid sequence of any of SEQ ID NOs: 49, 51, 53, 55, 57, 59, 61, or 63.

[0082] In some embodiments, the 5' homology arm comprises a nucleic acid sequence of any of SEQ ID NOs: 48, 50, 52, 54, 56, 58, 60, or 62.

[0083] In some embodiments, the RNA molecule of (a) comprises a nucleic acid sequence encoding a human ORF2p with D205A mutation, and wherein the nucleic acid sequence comprises a sequence with at least 90% sequence identity to SEQ ID NO: 129.

[0084] Also provided herein is a pharmaceutical composition comprising (a) the composition of any one of the preceding embodiments; and (b) a pharmaceutically acceptable excipient.

[0085] Also provided herein is a gene editing system for use in incorporating an exogenous sequence encoding an exogenous therapeutic polypeptide in a genome of a target cell, the system comprising a composition comprising: (a) an endonuclease or a polynucleic acid encoding the same; (b) a first guide RNA sequence or a polynucleic acid encoding the same; (c) a second guide RNA sequence or polynucleic acid encoding the same; and (d) an RNA comprising a sequence encoding a retrotransposon machinery, the RNA comprising: (i) a sequence that is a reverse complement of the sequence encoding the exogenous therapeutic polypeptide; and (ii) a 5' homology arm and a 3' homology arm flanking the sequence encoding the exogenous therapeutic polypeptide, and wherein each homology arm is complementary to a sequence of a target site in the genome; and (iii) a sequence encoding a human LINE1 ORF1p or

functional fragment thereof; and (iv) a sequence encoding a human ORF2p comprising a reverse transcriptase (RT), wherein the endonuclease activity of ORF2p is inactivated.

[0086] In some embodiments, the endonuclease is a nickase, and wherein the nickase is (i) a Cas9 nickase, or (ii) a Cas 9 nickase comprising an H840A mutation relative to the wild type Cas9.

[0087] In some embodiments, the nucleic acid sequence flanked by the two homology arms does not comprise or overlap with the sequences encoding the endonuclease, the LINE1 ORF1p, or the ORF2p, or a fragment thereof.

[0088] In some embodiments, the system further comprises (e) an RNA molecule comprising a sequence encoding a human ORF2p polypeptide that lacks a functional endonuclease domain, wherein the RNA molecule of (d) lacks sequences encoding a human ORF1p polypeptide or the exogenous sequence encoding a polypeptide, wherein the human ORF2p polypeptide that lacks a functional endonuclease domain comprises an exogenous DNA binding domain, wherein the exogenous DNA binding domain increases (i) DNA binding activity of the human ORF2p polypeptide or functional fragment thereof, (ii) TPRT processivity of the human ORF2p polypeptide or functional fragment thereof, or (iii) direct recruitment of the endonuclease to the target sequence in the genomic DNA of the target cell.

[0089] Also provided herein is a method of incorporating an exogenous sequence encoding a therapeutic polypeptide at a specific site in a genome of a target mammalian cell, the method comprising contacting the target mammalian cell with the composition of any one of the preceding embodiments, wherein only the exogenous sequence encoding a therapeutic polypeptide is integrated into the genome of the target mammalian cell.

[0090] Also provided herein is a method of incorporating an exogenous sequence encoding a therapeutic polypeptide at a specific site in a genome of a target mammalian cell, the method comprising: (a) contacting the target mammalian cell with a composition comprising: (i) an endonuclease or a polynucleic acid encoding the same; (ii) a first guide RNA sequence or a polynucleic acid encoding the same; (iii) a second guide RNA sequence or a polynucleic acid sequence encoding the same; and (iv) an RNA molecule comprising: (A) an RNA sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide; (B) a 5' homology arm and a 3' homology arm; and (C) a human mobile genetic element comprising a sequence encoding a polypeptide that promotes integration of the exogenous sequence encoding the therapeutic polypeptide into the genome of the target mammalian cell via target primed reverse transcription (TPRT); and (b) expressing the exogenous sequence encoding the therapeutic polypeptide from a genomically incorporated sequence of the target mammalian cell;

[0091] wherein, only the exogenous sequence encoding the therapeutic polypeptide is the genomically incorporated sequence that is incorporated at the specific site in the genome of the target mammalian cell.

[0092] In some embodiments, the nickase is a Cas9 nickase.

[0093] In some embodiments, the 5' homology arm and 3' homology arm flank the RNA sequence that is a reverse complement of an exogenous sequence encoding the therapeutic polypeptide.

[0094] In some embodiments, the nucleic acid sequence flanked by the two homology arms does not comprise or overlap with the sequences encoding the endonuclease or the human mobile genetic element, or a fragment thereof.

[0095] In some embodiments, the human mobile genetic element comprises one or more elements of human LINE1.

[0096] In some embodiments, the one or more elements of human LINE1 comprise ORF1p or a functional fragment thereof.

[0097] In some embodiments, the sequence encoding a human ORF1p polypeptide and the sequence encoding a human ORF2p polypeptide are separated by (a) a sequence encoding a GSS linker and (b) a sequence encoding a T2A cleavage sequence.

[0098] In some embodiments, the one or more elements of human LINE 1 comprise an ORF2p reverse transcriptase.

[0099] In some embodiments, the one of more elements of human LINE 1 comprises an ORF2p endonuclease domain (EN), and wherein the endonuclease activity of ORF2p is inactivated.

[0100] In some embodiments, the endonuclease activity of ORF2p is inactivated by introduction of one or more mutations, selected from the group consisting of S228P, Y1180A and D205A.

[0101] In some embodiments, the endonuclease activity of ORF2p is inactivated by truncation.

[0102] In some embodiments, the composition in step (a) further comprises (v) an RNA molecule comprising a sequence encoding a human ORF2p polypeptide that lacks a functional endonuclease domain, wherein the RNA molecule of (d) lacks sequences encoding a human ORF1p polypeptide or the exogenous sequence encoding a polypeptide.

[0103] In some embodiments, the human ORF2p polypeptide that lacks a functional endonuclease domain lacks an endonuclease domain.

[0104] In some embodiments, the human ORF2p polypeptide comprises an exogenous DNA binding domain.

[0105] In some embodiments, the exogenous DNA binding domain is derived from (i) a Sso7d protein, (ii) an HMGD protein, or (iii) an PARP1 protein.

[0106] In some embodiments, the exogenous DNA binding domain is operably linked to the ORF2p via a linker.

[0107] In some embodiments, the linker is a 33XTEN linker.

[0108] In some embodiments, a ratio of an amount of the RNA molecule of (iv) to that of the RNA molecule of (v) is about 5:1.

[0109] In some embodiments, the Cas9 nickase cuts the genomic DNA at a site that is within 1000, 800, 600, 400, 200, or 100 nucleotides from the specific site in the genome of a target mammalian cell.

[0110] In some embodiments, the method prevents double stranded breaks on the genomic DNA.

[0111] In some embodiments, the first guide RNA sequence comprises a sequence that has Watson-Crick pairing with a plurality of contiguous nucleotides on a first strand of the genomic DNA, and the second guide RNA sequence comprises a sequence that has Watson-Crick pairing with a plurality of contiguous nucleotides on the second and opposite strand of the genomic DNA.

[0112] In some embodiments, the first guide RNA sequence and the second guide RNA sequence do not comprise sequences that have Watson-Crick pairing with a plurality of contiguous nucleotides of the genomic DNA that are themselves complementary to each other.

[0113] In some embodiments, contacting the target mammalian cell with a composition comprises delivering the composition into the target mammalian cell.

[0114] In some embodiments, the composition is delivered to the target mammalian cell using a nanoparticle selected from the group consisting of a lipid nanoparticle and a polymeric nanoparticle.

[0115] In some embodiments, the therapeutic polypeptide is selected from the group consisting of a ligand, an antibody, a receptor, an enzyme, a transport protein, a structural protein, a hormone, a contractile protein, a storage protein and a transcription factor.

[0116] In some embodiments, the therapeutic polypeptide is a receptor selected from the group consisting of a chimeric antigen receptor (CAR) and a T cell receptor (TCR).

[0117] In some embodiments, the target mammalian cell is an immune cell selected from the group consisting of a T cell, a B cell, a myeloid cell, a monocyte, a macrophage, and a dendritic cell.

[0118] In some embodiments, the human mobile genetic element comprising a sequence encoding a polypeptide that promotes integration of the exogenous sequence encoding the therapeutic polypeptide into the genome of the target mammalian cell via target primed reverse

transcription (TPRT) comprises (i) a sequence encoding the ORF1p and (ii) a sequence encoding the ORF2p, wherein both (i) and (ii) are on the same polynucleic acid molecule.

[0119] In some embodiments, the ORF2p endonuclease comprises a sequence having a D205A mutation.

[0120] In some embodiments, the endonuclease is Cas9 nickase that is encoded by a polynucleotide molecule that is separate from the polynucleotide molecule encoding the human mobile genetic element.

[0121] In some embodiments, the first guide RNA sequence and the second guide RNA sequence are encoded by a polynucleotide molecule that is separate from the polynucleotide molecule encoding the human mobile genetic element.

[0122] In some embodiments, contacting the target mammalian cell with (i) a first composition comprising the RNA molecule comprising the mRNA sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide; the 5' homology arm and the 3' homology arm; and the human mobile genetic element is at a time separate from contacting the target mammalian cell with (ii) a second composition comprising the first guide RNA sequence and the second guide RNA sequence or one or more polynucleic acid sequences encoding the same.

[0123] In some embodiments, contacting comprises contacting the target mammalian cell with the first composition and the second composition at an interval of at least 2 hours.

[0124] In some embodiments, the target mammalian cell is contacted with the second composition at 3, 4, 5, 6, 7 or 8 hours after contacting with the first composition.

[0125] In some embodiments, the endonuclease is Cas9 nickase, wherein the Cas 9 nickase is contacted to the target mammalian cell at the same time as contacting with the first composition.

[0126] In some embodiments, the target mammalian cell is contacted with a composition comprising Cas9 nickase or a polynucleic acid encoding the same (i) prior to contacting with a first composition comprising the RNA molecule comprising the mRNA sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide; the 5' homology arm and the 3' homology arm; and the human mobile genetic element; or (ii) prior to contacting with a second composition comprising the first guide RNA sequence and the second guide RNA sequence or one or more polynucleic acid sequences encoding the same; or (iii) after contacting with the second composition.

[0127] In some embodiments, the target mammalian cell is contacted with an RNA encoding Cas9 nickase.

[0128] In some embodiments, the orientation of the 5' homology arm and the 3' homology arm with respect to the reverse complement of the exogenous sequence encoding the therapeutic polypeptide is reversed for insertion of the exogenous sequence in the opposite orientation.

[0129] In some embodiments, the sequence of the 5'-homology arm or the 3' homology arm between 10-80 nucleotides, between 20-70 nucleotides, between 30-60 nucleotide, between 40-60 nucleotides, between 15-20 nucleotides, less than 20 nucleotides, about 18 nucleotides, or about 17 nucleotides in length.

[0130] In some embodiments, the sequence of the 5'-homology arm or the 3' homology arm is about 30 nucleotides in length.

[0131] In some embodiments, the sequence of the 5'-homology arm or the 3' homology arm is about 50 nucleotides in length.

[0132] Also provided herein is a single mRNA molecule comprising a genome editing system for incorporating an exogenous sequence encoding a therapeutic polypeptide in a genome of a target cell, the genomic editing system comprising: (a) a Cas endonuclease or a functional fragment thereof; (b) a human LINE1 retrotransposon reverse transcriptase or a fragment thereof, and (c) the exogenous sequence encoding the therapeutic polypeptide flanked by one or more homology arms comprising non-ribosomal genomic sequence; or (d) a sequence comprising reverse complement of (c).

[0133] In some embodiments, the Cas endonuclease is a Cas 9 endonuclease, and the Cas endonuclease is a Cas 9 nickase.

[0134] In some embodiments, the exogenous sequence encoding the therapeutic polypeptide is greater than 1kb, 1.2kb, 1.5kb, 1.7kb, 1.8 kb, 1.9kb, 2 kb, 2.1kb, 2.5kb, 2.7kb, 2.8 kb, 2.9kb, 3 kb, 3.1kb, 4kb, or 5 kb in length.

[0135] Also provided herein is a composition comprising the single mRNA molecule of any one of the preceding embodiments and a nucleic acid delivery vehicle, and wherein the nucleic acid delivery vehicle comprises one or more lipids.

[0136] In some embodiments, the composition further comprises one or more guide RNAs.

[0137] In some embodiments, the composition further comprises one or more siRNAs.

[0138] In some embodiments, composition further comprises one or more nuclear localization sequences (NLS), and wherein the NLS helps transport of one or more ORF polypeptides into the nucleus.

[0139] In some embodiments, the NLS is SV40 NLS.

[0140] Also provided herein is a pharmaceutical composition for treating a disease in a subject, comprising: (a) an RNA comprising a sequence encoding Cas9 nickase; (b) an RNA comprising a sequence that is a reverse complement of the exogenous sequence encoding the therapeutic

polypeptide; a 5' homology arm and a 3' homology arm, wherein the 5' homology arm or the 3' homology arm is in reverse complement orientation with respect to a sequence on one strand of the genomic DNA; and a sequence encoding a human mobile genetic element; (c) a composition comprising a first guide RNA sequence and a second guide RNA sequence or one or more polynucleic acid sequences encoding the same; and a therapeutically acceptable excipient, wherein each of the compositions of (a), (b) and (c) are comprised in a separate delivery vehicle.

[0141] Also provided herein is a composition comprising: (a) an RNA molecule comprising: (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; (ii) a 5' homology arm and a 3' homology arm; and (iii) a sequence encoding a modified human LINE1 sequence or fragment thereof, wherein the sequence encoding the modified human LINE1 sequence or fragment thereof is upstream of the 5' homology arm or downstream of the 3' homology arm; (b) an RNA molecule encoding an endonuclease, wherein the endonuclease is a nickase; (c) an RNA molecule comprising a sequence encoding a human ORF2p polypeptide comprising an endonuclease wherein the endonuclease is functionally deficient; and (d) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.

[0142] In some embodiments, the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF2p polypeptide, wherein the human ORF2p polypeptide comprises an endonuclease and a reverse transcriptase, wherein the endonuclease comprises a mutation.

[0143] In some embodiments, the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF1p polypeptide upstream of the sequence encoding the human ORF2p polypeptide, and is separated by a LINE1 interORF sequence.

[0144] In some embodiments, the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF1p polypeptide upstream of the

sequence encoding the human ORF2p polypeptide, and is separated by an IRES sequence or a cleavage site.

[0145] In some embodiments, the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF1p polypeptide upstream of the sequence encoding the human ORF2p polypeptide, and is separated by (a) a sequence encoding a GSS linker and (b) a T2A cleavage sequence.

[0146] In some embodiments, the sequence encoding the modified human LINE1 sequence as well as the sequence encoding the human ORF2p polypeptide comprise a mutation in ORF2p endonuclease domains, wherein the mutation in the ORF2p endonuclease domains render the endonucleases functionally deficient.

[0147] In some embodiments, none of the ORF2p endonucleases have endonuclease activity.

[0148] In some embodiments, the human ORF2p polypeptide lacks an endonuclease domain, and wherein the human ORF2p polypeptide comprises an exogenous DNA binding domain.

[0149] In some embodiments, the exogenous DNA binding domain is derived from (i) a Sso7d protein, (ii) an HMGD protein, or (iii) an PARP1 protein.

[0150] In some embodiments, the RNA molecule comprising a sequence encoding a human ORF2p polypeptide comprises a 5' - methylated guanosyl cap structure (m7G cap).

[0151] In some embodiments, the RNA molecule comprising a sequence encoding a human ORF2p polypeptide is translated via a canonical cap-dependent mechanism.

[0152] In some embodiments, the sequence encoding a human ORF2p polypeptide in the RNA molecule (c) is located at the 5' end of the RNA molecule (c).

[0153] In some embodiments, the sequence encoding the human ORF2p polypeptide in the modified human LINE1 sequence comprises an N-terminal NLS and a C-terminal NLS.

[0154] In some embodiments, the nickase is a functional nickase, and wherein the nickase comprises an N-terminal NLS, a C-terminal NLS or both N-terminal and C-terminal NLS.

[0155] In some embodiments, the sequence encoding the modified human LINE1 sequence or fragment thereof comprises one or more STOP codons at the end of the sequence encoding the human ORF2p polypeptide, the sequence encoding the human ORF1p polypeptide or both.

[0156] In some embodiments, any of the RNA molecules comprise one or more ribosomal entry sites, one or more cleavage sites, or one or more oligomerization domains.

[0157] In some embodiments, each of the RNA molecules comprise a poly A sequence at the 3' end.

[0158] In some embodiments, the RNA molecule of (c) further comprises a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide

operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide.

[0159] In some embodiments, the RNA molecule of (c) further comprises a 5' homology arm and a 3' homology arm.

[0160] In some embodiments, the RNA molecule of (c) does not comprise a sequence encoding ORF1p.

[0161] In some embodiments, the RNA molecule of (c) does not comprise a reverse complement sequence of an insert sequence.

[0162] In some embodiments, the RNA molecule of (c) consists of the sequence encoding the human ORF2p polypeptide with at least one NLS.

[0163] In some embodiments, the promoter comprises a EF1alpha promoter or a functional variant.

[0164] In some embodiments, the EF1alpha promoter is a full-length EF1alpha promoter.

[0165] In some embodiments, the EF1alpha promoter is a short EF1alpha promoter.

[0166] Also provided herein is a composition comprising: (a) an RNA molecule comprising: (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; (ii) a 5' homology arm and a 3' homology arm; and (iii) a sequence encoding a human ORF2p polypeptide comprising an endonuclease wherein the endonuclease is functionally deficient; (b) an RNA molecule encoding an endonuclease, wherein the endonuclease is a nickase; (c) an RNA molecule comprising a sequence encoding a human ORF1p, wherein the RNA molecule does not comprise a sequence encoding a human ORF2p; (d) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.

[0167] In some embodiments, the RNA molecule of (c) consists of a sequence encoding the human ORF1p.

[0168] Also provided herein is a method for incorporating an exogenous sequence encoding a therapeutic polypeptide at a specific site in a genome of a target mammalian cell, the method comprising contacting the mammalian target cell with the composition of any one of the preceding embodiments, wherein only the exogenous sequence encoding a therapeutic polypeptide is integrated into the genome of the target mammalian cell.

[0169] In some embodiments, the integration efficiency is increased at least by 0.1%, 0.2%, or 0.5% compared to a system where the sequence that is a reverse complement of an exogenous sequence encoding the polypeptide is not operably linked to the sequence that is a reverse complement of the promoter sequence or does not comprise a promoter.

[0170] Also provided herein is a composition comprising: (a) an RNA molecule comprising: (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; (ii) a 5' homology arm and a 3' homology arm; and (iii) a sequence encoding a human ORF1p polypeptide and a human ORF2p polypeptide, the human ORF2p polypeptide comprising an endonuclease wherein the endonuclease is functionally deficient; (b) an RNA molecule encoding an endonuclease, wherein the endonuclease is a nickase; (c) an RNA molecule comprising a sequence encoding a human ORF2p, wherein the RNA molecule does not comprise a sequence encoding a human ORF1p or the reverse complement sequence of an insert sequence; (d) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell, wherein the integration efficiency of the insert sequence is increased by at least 0.05% compared to a composition that lacks (c).

[0171] Also provided herein is a method of treating a disease in a subject, comprising administering to the subject (a) the composition of any one of the preceding embodiments or (b) the pharmaceutical composition of any one of the preceding embodiments.

[0172] In some embodiments, the disease is cancer.

INCORPORATION BY REFERENCE

[0173] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference. To the extent publications and patents or patent applications incorporated by reference contradict the disclosure contained in the specification, the specification is intended to supersede and/or take precedence over any such contradictory material.

BRIEF DESCRIPTION OF THE DRAWINGS

[0174] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also "FIG." herein), of which:

[0175] **FIG. 1** depicts exemplary gene editing components for CRISPR-Enabled Autonomous Transposable Element (CREATE) including a polynucleic acid construct encoding ORF1, endonuclease deactivated ORF2, a 5' homology arm, a GFP payload under the control of a CMV promoter, and a 3' homology arm. Also included are a Cas9 nickase and two guide RNAs (gRNAs).

[0176] **FIG. 2** shows a possible mechanism of specific genomic integration of the payload (GFP) initiated by dual nicking by Cas9 nickase guided by dual guide RNAs. First and second strand synthesis is primed by the 3' and 5' homology arms, respectively, resulting in specific integration of only the payload and its promoter.

[0177] **FIG. 3** is a schematic of an experimental process to determine the integration efficiency in HEK239 cells expressing Cas9 nickase. Flow cytometry, PCR, and sequencing are used to confirm integration of the payload.

[0178] **FIGs. 4A-4B** depict results from flow cytometric analysis detecting GFP payload expression. Integration efficiency was measured before (**FIG. 4A**) and after sorting for GFP positive cells (**FIG. 4B**).

[0179] **FIG. 5** depicts gel electrophoresis results from a PCR experiment detecting GFP payload expression in cells contacted with the gene editing system (right) and control cells (left).

[0180] **FIGs. 6A-6B** depict sequencing results obtained using nanopore sequencing confirming integration of the payload at the AAVS1 locus.

[0181] **FIG. 7A** is a schematic showing the design of a CREATE mRNA construct for targeting a payload sequence (encoding GFP, as a surrogate for the exogenous payload sequence) into the safe harbor site AAVS1. The reverse complement sequence of that encoding the GFP is flanked by a 3' homology arm (PBS1) and a 5' homology arm (RC-PBS2) (reverse complement of PBS2). Both sequences encoding ORF1p and ORF2p are on the same mRNA construct. The CREATE system comprises one or more polynucleic acids encoding two guide RNAs, for targeting at PBS1 and PBS2

regions, and a Cas9 nickase (nCAs9) or a polynucleotide encoding the Cas9 nickase, wherein the polynucleotide encoding the Cas9 nickase is mRNA. EF1a, elongation factor 1-alpha subunit, arrow indicating directionality of the mRNA.

[0182] FIG. 7B is a diagrammatic overview of the CREATE complex.

[0183] FIG. 7C is a schematic depiction of the mechanism of action of the CREATE complex in incorporating an exogenous payload polynucleotide sequence in the genomic DNA of a target mammalian cell.

[0184] FIG. 8A shows a schematic of the CREATE mRNA design 1 used the studies shown in FIG. 8B.

[0185] FIG. 8B shows data indicating the specificity of the CREATE design. Only the combination of nCas9, CREATE mRNA design 1 shown in FIG. 8A, and matched dual sgRNA resulted in stable GFP expression.

[0186] FIG. 9A is a diagrammatic view of CREATE mRNA design 2 used in the studies that generated the data shown in FIG. 9B. In this design, ORF2p and ORF1p are encoded by sequences in separate mRNA molecules as shown, known as the split ORF1p-ORF2p design.

[0187] FIG. 9B shows flow cytometry data indicating that split ORF1p-ORF2p designs show an improvement in the initial GFP expression over the CREATE design 1.

[0188] FIG. 9C is a graphical representation of the data shown in FIG. 9B.

[0189] FIG. 9D shows flow cytometry data indicating that split ORF1p-ORF2p design did not show increase in GFP+ cells on day 8 after incorporation.

[0190] FIG. 10A shows a diagrammatic view of the CREATE mRNA design 1, which was used in the studies shown generating data shown in FIG. 10B.

[0191] FIG. 10B shows the results from an assay designed to further enhance the efficiency of CREATE editing by exploring its performance upon varying the conditions of transfection methods. The data indicate that using protocol #2, in which the target mammalian cells were transfected with CREATE mRNA followed by transfection with polynucleotide(s) encoding the guide RNAs at an interval of 4 hours resulted in improvement in genome editing efficiency as evidenced by % GFP+ cells.

[0192] FIG. 11A shows a diagrammatic view of the CREATE mRNA design 1, which was used in the studies shown generating data shown in FIG. 11B.

[0193] FIG. 11B shows the results from an assay where the time interval between mRNA transfection and sgRNA transfection was varied as indicated in the X-axis, and the corresponding % GFP positive cells at 3 days and 8 days are shown. Protocol#2 (mRNA first) with an incubation of 4 or 6 hours between two steps results in the highest GFP expression.

[0194] FIG. 12A shows a diagrammatic view of the CREATE mRNA design 1, which was used in the studies shown generating data shown in FIG. 12B.

[0195] **FIG. 12B** (left graph), shows results of an assay in which the amounts of the CREATE mRNA was varied as indicated on the X-axis, keeping the amounts of dual sgRNA constant to determine an effect on the genome editing efficiency. The editing efficiency plateaus when CREATE mRNA was above 3 μg per 8×10^4 cells, and mRNA above 10 μg started to cause significant cytotoxicity. **FIG. 12B** (right graph), shows results of an assay in which the amounts of each sgRNA was varied as indicated on the X-axis, and the corresponding % GFP positive cells at 3 days and 8 days are shown, keeping the amounts of CREATE mRNA constant at 3 micrograms to determine an effect on the genome editing efficiency. The highest GFP positive cells were obtained when 0.242 micrograms of each sgRNA (0.484 microgram total) was used for transfecting 8×10^4 cells.

[0196] **FIG. 13A** shows a diagrammatic view of the CREATE mRNA design 1, along with an mRNA encoding Cas 9 nickase (comprising the mutation H840A) (nCas9), used for transfecting cells for the assay generating data shown in **FIG. 13B**.

[0197] **FIG. 13B** shows results of an assay demonstrating that the CREATE mRNA could be cotransfected with an mRNA encoding Cas9 H840A nickase to achieve optimal CREATE editing, without the need for preparing a nCas9 expressing cell line.

[0198] **FIG. 14A** shows six different modifications (I-VI) of the CREATE mRNA system to achieve the minimal and optimal machinery for CREATE editing.

[0199] **FIG. 14B** shows results for each modification with respect to the editing performance of the CREATE mRNA design 1, which was used as a benchmark control. The data are statistically significant.

[0200] **FIG. 15** shows experimental results demonstrating CREATE editing in the liver cell line (Huh 7). CREATE mRNA is designed as described above, comprising the reverse complement of sequence encoding GFP payload, to be expressed under the influence of a EF1 α promoter. A first guide RNA and a second guide RNA, as described previously targets the Cas 9 nickase to AAVS1 site followed by retrotransposition at the nicked site by the CREATE mRNA translated products. The ORF2p is endonuclease dead D205A mutant. Huh7 cell line expresses Cas9 nickase. Flow cytometry results show expression of GFP at day 4 following incorporating the CREATE editing complex into the cell.

[0201] **FIGs. 16A-16F** show CREATE editing design and proof-of-concept in mammalian cells. **FIG. 16A** and **FIG. 16B**, Schematics views of the CREATE mRNA design. Payload cassette (EF1 α core promoter driven GFP with SV40 polyA signal) is encoded in the antisense direction. **FIG. 16C** and **FIG. 16D**, Views explaining the CREATE editing mechanism. L1 components ORF1p and ORF2p are produced from CREATE mRNA and co-assemble into ribonuclear complex with CREATE mRNA. Following nicking by Cas9^{H840A} and sgRNAs, the DNA 3'-flap hybridizes with PBS1 to prime 1st strand cDNA synthesis mediated by ORF2p RT domain. 2nd strand synthesis is primed by PBS2 that hybridizes with the reverse transcribed RC-PBS2 sequence in the newly synthesized cDNA strand.

Outcome, of the editing is the targeted insertion of the payload between PBS sites into the genome without the sequences encoding L1 components. **FIG. 16E**, CREATE editing in mammalian cells requires Cas9 nickase activity and dual sgRNAs. GFP+ cells indicate integration and re-expression of GFP payload. Statistical analysis was performed using two-way ANOVA with Dunnett's multiple comparisons test comparing against the first sample. **** indicates $p < 0.0001$. Statistical significance was only labeled for samples with $p < 0.05$. **FIG. 16F**, Improvements of the transfection protocol improved editing efficiency.

[0202] **FIGs. 17A-17D** demonstrate CREATE editing confirmation using NGS sequencing. **FIG. 17A**, AAVS1 edited cells before and after sorting for GFP+ cells. **FIG. 17B**, PCR primers were designed to amplify region that covers the target sites for integration. The unedited allele showed expected 570 bp band (red arrow) while the edited allele showed an expected increase in size due to payload integration (green arrow). Lower panel shows a diagrammatic view of the payload. **FIG. 17C**, Amplicon sequencing result shows the percentage of indels at the integration junction. Each plot shows percentage sequencing reads with or without indels. Data from two independent experiments are shown. **FIG. 17D**, Distribution of indels and substitutions at the junction sites. Black arrow indicates sgRNA cut sites at PBS1 and PBS2 junctions.

[0203] **FIG. 17E** shows results of insertion site analysis across the whole genome, shown by chromosome.

[0204] **FIG. 17F** shows amplicon sequencing result shows the percentage of indels at the integration junction. Each plot shows percentage sequencing reads with or without indels at two different PBS region lengths.

[0205] **FIG. 17G** shows amplicon sequencing results using cells with targeted insertion at HEK3, PRNP, or IDS.

[0206] **FIGs. 18A-18D** show mechanistic analysis of CREATE editing. **FIG. 18A**, Mutational analysis of the key enzymatic domains involved in editing. **FIG. 18B**, CREATE editing requires two PBS sequences that matches the two sgRNA target sites. Data for **FIG. 18A** and **FIG. 18B** are from the same experiments and separately presented for clarity. **FIG. 18C**, Influence of the length of PBS sequence in CREATE mRNA on editing efficiency. **FIG. 18D**, Influence of the distance between sgRNA cut sites on editing efficiency. Data for **FIG. 18C** and **FIG. 18D** are from the same experiments and separately presented for clarity. **FIGs. 18E** and **18F** represent data similar to **FIG. 18A** and **18B** respectively. Statistical analysis was performed using two-way ANOVA with Dunnett's multiple comparisons test comparing against the first sample. **** indicates $p < 0.0001$. Statistical significance was only labeled for samples with $p < 0.05$.

[0207] **FIGs. 19A-19D** demonstrate CREATE can be used in other cell type and programmed to other safe harbor sites. **FIG. 19A**, Editing of three additional genomic loci by designing unique sgRNAs and matching PBSs. **FIG. 19B**, FACS sorting of PRNP edited cells confirmed stable

integration. **FIG. 19C**, Sanger sequencing of PCR amplicon covering the junction region confirmed correct integration of payload at PRNP locus. **FIG. 19D**, Co-delivery of CREATE editing components as mRNA/sgRNA achieves efficient editing in 293T and Huh7 cells. **FIG. 19E**, CREATE as a fully RNA-based gene delivery system to multiple cell types (Co-delivery of CREATE editing components as mRNA/sgRNA. RNA-based GFP payload insertion at AAVS1 locus in Huh7 and HEK293T cells. Statistical analysis was performed using two-way ANOVA with Dunnett's multiple comparisons test comparing against the first sample. **** $p < 0.0001$. *** $p < 0.001$. Statistical significance is labeled for samples with $p < 0.01$. Data shown are representative of at least two independent experiments.

[0208] **FIGs. 20A-20C** demonstrate results of PCR detection of potential off-target editing in AAVS1 loci edited cells. Top 5 predicted off-target sites for sgRNA1 and sgRNA2 were examined. Hollow box indicate the expected PCR product if off-target editing occurred at the site. * indicate on-target genomic amplicon as a positive control for PCR reaction.

[0209] **FIG. 21** shows graphical representation of the design of avian R2 based CREATE system. In this 2 guide RNA based system, wherein the R2 endonuclease is functionally deficient (ENdead), is fused with two NLS sequences each on either side, PBS1 and PBS2 are homology arms, directed to the AAVS1 site. GFP encoding sequence is the payload for proof of principle, driven by an EF1 promoter. Both designs shown here include a separate polynucleotide encoding a Cas (nickase) polypeptide. The design shown in the bottom section has separate polynucleic acid encoding a polypeptide comprising Cas nickase fused to the R2 endonuclease.

[0210] **FIG. 22** shows three exemplary designs used in the feasibility studies. Various gene editing constructs described herein are often designated by the term "CREATE" followed by a number ("CREATE #") i.e., CREATE 24, CREATE 27 and so on. A representative schematic diagram of each polypeptide is provided in the relevant set of drawings. The gene editing composition is delivered to cells as mRNA. In general, each construct represents a single mRNA. One or more constructs may be co-delivered, or co-administered into a cell population in vitro for test purposes described herein, and is mentioned accordingly on a case by case basis. A composition for gene editing may comprise CREATE mRNA constructs, guide RNAs as DNA or RNA, separate RNPs or DNAs expressing additional helper proteins or peptides as described in some of the examples below.

[0211] **FIG. 23A** shows GFP expression data at day 3 post transduction that codelivery of CREATE 24 and CREATE 27 constructs lead to an improvement in editing efficiency as compared to delivering CREATE 24 system alone. The CREATE 24 and CREATE 27 mRNA sequences are at the indicated molar ratios when transfected.

[0212] **FIG. 23B** shows GFP expression data at day 7 of the assay depicted in FIG. 23A.

[0213] **FIG. 24A** shows GFP expression data for codelivery of CREATE 24 and CREATE 27 constructs at the indicated ratios at day 3 post transduction. Codelivery improves editing efficiency over single delivery

[0214] **FIG. 24B** shows GFP expression data at day 7 of the assay depicted in **FIG. 24A**.

[0215] **FIG. 25A** shows GFP expression data for codelivery of CREATE 24 and CREATE 27 constructs at the indicated ratios at day 3 post transduction.

[0216] **FIG. 25B** shows GFP expression data at day 7 of the assay depicted in **FIG. 24A**.

[0217] **FIGs. 25C-25E** represent diagrams for additional designs for CREATE constructs. **FIG. 25C** shows a CREATE design where the interORF region is replaced with T2A, P2A, tandem T2A-P2A or an IRES sequences. In this scenario there will not be a need to co-deliver CREATE 27. **FIG. 25D** shows diagram of a construct where CREATE 24 is co-delivered with an mRNA that expresses ORF2p only (CREATE 80) without the RC-PBS2-GFP-EFS-PBS1 payload. **FIG. 25E** shows diagram of a construct where CREATE 24 is co-delivered with an mRNA that expresses ORF1p only at a certain ratio with CREATE 24 or alternatively ORF2p only (e.g., CREATE 80, not shown) added at a certain ratio with CREATE 24, or with payload alone (not shown).

[0218] **FIG. 26A** shows a retrotransposon + Cas (CREATE) editing system design for improvement of retrotransposition mediated editing efficiency. Stop codons denoted by solid diamonds were introduced at the end of each ORF and a Cas 9 encoded by a separate polynucleic acid from the one that comprises the sequences encoding the retrotransposon ORFs and payload (GFP). **FIG. 26B** additionally shows an enlargement of the ORF2p domains in the construct.

[0219] **FIG. 27** shows additional CREATE designs for single mRNA encoding retrotransposon + Cas 9 editing systems.

[0220] **FIG. 28** shows a unique CREATE construct designed to contain a flexible zipper linker attaching the ORF1p and ORF2p polypeptide sequences, representing affinity heterodimer approach.

[0221] **FIG. 29** shows yet another design for the CREATE system comprising a DNA transactivator coupled system, for example designing protein binding sequence connecting the ORF1p and ORF2p polypeptide sequences. In this variation, an affinity heteromer approach is taken, for example to couple separately expressed components. In one example, a tandem repeat of multiple copies of GCN 4 peptide separated by linkers used, for example the SunTag system can be used. SunTag is a repeat of multiple copies of the 19-mer GCN4 peptide, separated by linkers 5 aa long.

[0222] **FIG. 30** shows additional CREATE construct designs to test parameters to improve on the different functions of the CREATE system to effect enhancement of retrotransposition. In this design, addition of a MutL- homolog protein MLH1 protein or a part thereof. The normal function of MLH1 involves DNA mismatch repair during replication and is recruited at the mismatch site. Expressing a dominant negative form of MLH1, efficiency of the CREATE system may be improved, for example, by preventing indel formations by action of natural mismatch proteins of the MMR system.

[0223] **FIG. 31** shows additional CREATE designs to test additional parameters to improve on the different functions of the CREATE system to effect enhancement of retrotransposition.

[0224] **FIG. 32** shows additional CREATE designs to test additional parameters to improve on the different functions of the CREATE system to effect enhancement of retrotransposition.

[0225] **FIG. 33** shows additional CREATE designs to test additional parameters to improve on the different functions of the CREATE system to effect enhancement of retrotransposition.

[0226] **FIG. 34** shows a test design including cotransfection of mRNA encoding ORF2p (CREATE 80) in addition to CREATE 24 mRNA, along with Cas 9 nickase (harboring H840A mutation) to enhance ORF2p expression and increase editing efficiency. Primary T cells (2×10^6) were derived from a human donor; activated using CD3/CD28 antibody complex activator for 68 hours and electroporated with CREATE 24 and CREATE 80 RNAs at a molar ratio of 5:1. Total 7.5 uM of CREATE mRNA-cargo constructs were used. This was followed up with detection of GFP by flow cytometry and imaging on Day 3 and Day 7 post electroporation.

[0227] **FIG. 35A** shows representative flow cytometry test results from experiment detailed above for FIG. 34 at day 3 post electroporation.

[0228] **FIG. 35B** shows a representative flow cytometry test from experiment detailed above for FIG. 34 at day 7 post electroporation.

[0229] **FIG. 36** shows summary graphs from the studies described in FIG. 35A and FIG. 35B, results showing Co-transfection of CREATE 24 and CREATE 80 at 5:1 ratio improved integration efficiency in primary T cells.

[0230] **FIG. 37** is a schematic depicting the expression of ORF2p from a coding sequence comprising an inter-ORF sequence.

[0231] **FIG. 38** depicts the replacement of the interORF sequence with a sequence encoding a GSG linker and a sequence encoding a T2A sequence.

[0232] **FIG. 39** depicts exemplary CREATE mRNA constructs.

[0233] **FIGs. 40A-40C** depict the integration efficiency results of the constructs described in **FIG. 39** by flow cytometry. **FIG. 40A** depicts integration efficiency of control groups. **FIG. 40B** depicts integration efficiency of CREATE mRNA constructs comprising the interORF sequence (left panel) and the integration efficiency of CREATE mRNA constructs comprising the GSG-T2A sequence (right panel). **FIG. 40C** shows a graph summarizing these results.

[0234] **FIG. 41** depicts an ORF2 fusion construct including a high mobility group protein D (HMGC) protein.

[0235] **FIG. 42** depicts exemplary CREATE constructs comprising modified ORF2 having an exogenous DNA binding domain, including Sso7d DNA binding domain, High Mobility Group Protein-D(HMGD), or PARP-1 zinc finger domain (hPARP1).

[0236] **FIGs. 43A-43C** depict the integration efficiency results of the constructs described in **FIG. 42** by flow cytometry. **FIG. 43A** depicts results from control groups (Mock, CREATE-24 only) and results obtained using ORF2 constructs with the interORF2 sequence (CREATE-24 + CREATE-29 + sgRNA) on day 3 (upper panel) and day 8 (lower panel). **FIG. 43B** depicts results of CREATE systems using constructs with modified ORF2 domains in **FIG. 42** on day 3 (upper panel) and day 8 (lower panel). **FIG. 43C** shows a graph summarizing these results in **FIGs. 43A-43B**.

[0237] **FIG. 44** depicts exemplary CREATE mRNA constructs with varied PBS sequence length.

[0238] **FIG. 45** depicts integration efficiency results of the constructs with varied PBS sequence length in **FIG. 44**.

[0239] **FIG. 46** depicts exemplary CREATE systems comprising Cas9 nickase mRNA (CREATE-29), CREATE mRNA constructs with varied PBS sequence length (CREATE-24 or CREATE-37), co-delivered with ORF2 domain (CREATE-80) or modified ORF2 domain (CREATE-82), where CREATE-80 or CREATE-82 is co-delivered with CREATE-37 at 5:1 ratio.

[0240] **FIG. 47** depicts integration efficiency results of the CREATE systems described in **FIG. 46**.

DETAILED DESCRIPTION

[0241] All terms are intended to be understood as they would be understood by a person skilled in the art. Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which the disclosure pertains.

[0242] The section headings used herein are for organizational purposes only and are not to be construed as limiting the subject matter described.

[0243] As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise.

[0244] In this application, the use of "or" means "and/or" unless stated otherwise. The terms "and/or" and "any combination thereof" and their grammatical equivalents as used herein, may be used interchangeably. These terms may convey that any combination is specifically contemplated. Solely for illustrative purposes, the following phrases "A, B, and/or C" or "A, B, C, or any combination thereof" may mean "A individually; B individually; C individually; A and B; B and C; A and C; and A, B, and C." The term "or" may be used conjunctively or disjunctively unless the context specifically refers to a disjunctive use.

[0245] The term "about" or "approximately" may mean within an acceptable error range for the particular value as determined by one of ordinary skill in the art, which will depend in part on how the value is measured or determined, *i.e.*, the limitations of the measurement system. For example, "about" may mean within 1 or more than 1 standard deviation, per the practice in the art. Alternatively, "about" may mean a range of up to 20%, up to 10%, up to 5%, or up to 1% of a given value. Alternatively, particularly with respect to biological systems or processes, the term may mean within an order of magnitude, within 5-fold, and more preferably within 2-fold, of a value. Where particular values are

described in the application and claims, unless otherwise stated the term "about" meaning within an acceptable error range for the particular value should be assumed.

[0246] As used in this specification and claim(s), the words "comprising" (and any form of comprising, such as "comprise" and "comprises"), "having" (and any form of having, such as "have" and "has"), "including" (and any form of including, such as "includes" and "include") or "containing" (and any form of containing, such as "contains" and "contain") are inclusive or open-ended and do not exclude additional, unrecited elements or method steps. It is contemplated that any embodiment discussed in this specification may be implemented with respect to any method or composition of the present disclosure, and vice versa. Furthermore, compositions of the present disclosure may be used to achieve methods of the present disclosure.

[0247] Reference in the specification to "some embodiments," "an embodiment," "one embodiment" or "other embodiments" means that a particular feature, structure, or characteristic described in connection with the embodiments is included in at least some embodiments, but not necessarily all embodiments, of the present disclosures. To facilitate an understanding of the present disclosure, a number of terms and phrases are defined below.

[0248] Although various features of the present disclosure can be described in the context of a single embodiment, the features can also be provided separately or in any suitable combination. Conversely, although the present disclosure can be described herein in the context of separate embodiments for clarity, the disclosure can also be implemented in a single embodiment.

[0249] Provided herein is a composition comprising an RNA molecule comprising a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide, a 5' homology arm and a 3' homology arm, and a sequence encoding a human ORF2p polypeptide, wherein the sequence encoding a human ORF2p polypeptide is upstream of the 5' homology arm or downstream of the 3' homology arm, an endonuclease or polynucleic acid sequence encoding the endonuclease, and one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA sequence and a second guide RNA sequence, or a polynucleic acid encoding the one or more guide RNA molecules, wherein the first guide RNA sequence comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA sequence comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.

[0250] Also provided herein is a composition comprising an RNA molecule comprising a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises a

sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide, a 5' homology arm and a 3' homology arm, a sequence encoding a polypeptide with reverse transcriptase activity, wherein the sequence encoding a polypeptide with reverse transcriptase activity is upstream of the 5' homology arm or downstream of the 3' homology arm, an endonuclease or a polynucleic acid sequence encoding the endonuclease, and one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA sequence and a second guide RNA sequence, or a polynucleic acid encoding the one or more guide RNA molecules, wherein the first guide RNA sequence comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA sequence comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell. Provided herein, as described above is a composition comprising a recombinant polynucleic acid, e.g., an RNA molecule, that can edit genomic DNA of a cell, and comprises an effective working system that combines retroelement enabled CRISPR precision tool for site-specific genomic editing of large DNA segments. In some embodiments, the endonuclease is a Cas endonuclease. In some embodiments, the Cas endonuclease is a nickase. In some embodiments the endonuclease is a Cas9 nickase. In some embodiments, the composition encompasses one or more guide RNA sequences, e.g., two guide RNA sequences, a first guide RNA sequence and a second guide RNA sequence. In some embodiments, the cell may be an ex vivo cell. In some embodiments, the cell may be in vivo.

[0251] In one aspect, provided herein is a CREATE system, wherein a CREATE system may comprise one or more polynucleic acids encoding one or more polypeptides, or polynucleotide sequences for guide RNA molecules; and may further include one or more polypeptides, delivery vehicles and related excipients for a gene editing system combining the retrotransposon's ability to deliver large segments of nucleic acid into a genomic DNA with the targeting specificity of CRISPR-Cas enzymes. 'CREATE' may represent an acronym for CRISPR-Enabled Autonomous Transposable Element. A CREATE mRNA may represent the mRNA described above, or design variants of it, comprising a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide, optionally, operably linked to a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide, a 5' homology arm and a 3' homology arm, a sequence encoding a polypeptide with reverse transcriptase activity, wherein the sequence encoding a polypeptide with reverse transcriptase

activity is upstream of the 5' homology arm or downstream of the 3' homology arm, an endonuclease or a polynucleic acid sequence encoding the endonuclease. In some embodiments, the CREATE system comprises the CREATE mRNA and may further comprise one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA sequence and a second guide RNA sequence, or a polynucleic acid encoding the one or more guide RNA molecules, wherein the first guide RNA sequence comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA sequence comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell. In some embodiments, the CREATE system may further comprise a polynucleic acid encoding the Cas endonuclease as a separate nucleic acid molecule than the mRNA encoding the polypeptide with reverse transcriptase activity. In some embodiments, a CREATE system may comprise a DNA molecule, e.g., a DNA encoding a Cas endonuclease or a guide RNA; a riboprotein; a lipid or a polypeptide. In some cases, a CREATE system may be referred to a CREATE complex. In one embodiment, the one or more variants or design variants may comprise a sequence encoding a mutation, or a rearrangement or a reordering of a sequence of elements arranged on the CREATE polynucleic acid, e.g., the CREATE mRNA; or an inclusion of one or more sequences on a separate RNA molecule than the CREATE mRNA or on a DNA molecule; or inclusion of one or more sequences in the CREATE polynucleic acid, e.g., CREATE mRNA.

[0252] In one aspect, provided herein is a system that is safe and effective for introducing large fragments of genetic material (e.g., greater than 100 bps, greater than 1000 bps etc.) into the genome of a mammalian cell, using a mammalian retrotransposon system, together with precision editing of the genome using the CRISPR-Cas system. This surpasses previous attempts in the field for successful genomic editing. The system described herein is superior in a number of ways considering existing systems, first, it avoids using double stranded breaks that are used in CRISPR-Cas-mediated editing. While base editing methods introduce single stranded breaks, it is not capable of editing large genomic sequences, in other words, a base editing method cannot introduce large fragments of genetic materials, but is only capable of introducing single nucleotide changes. Prime editing capability also ranges over a few nucleotides, and does not extend to introducing nucleotide stretches that are greater than 100 bps, let alone greater than 1000 bp. The combination of retrotransposon system with the specificity of the CRISPR system has now been proven instrumental in incorporating large fragments of genetic material in the genome of a mammalian cell in a safe and effective way. Second, use of mammalian retrotransposon systems, particularly human LINE1 ORF1 and ORF2 retrotransposon systems, ensures tolerance in human cells. Third, integration into the genome has been successful in non-mitochondrial genomic sequences, which is an advantage in terms of both safety and efficacy. Successful site-specific integration into the genome is proven herein by sequencing. Additionally, low

levels of integration is a hallmark of safety features in genome editing of human cells. Finally, introducing the programmable editing system is adapted to mRNA delivery alone, thereby avoiding any viral delivery mechanisms, adding to the safety aspect of genome editing of a human cell. Summarizing, the instant application embodies, to our understanding, the first successful development of human genome editing of large genetic material, by successful use of human retrotransposon system in conjunction with CRISPR-Cas specificity effectively for targeting a non-mitochondrial genomic sequence.

[0253] In some embodiments, the RNA molecule comprises the polynucleic acid sequence encoding the endonuclease. In some embodiments, the polynucleic acid sequence encoding the endonuclease is upstream of the 5' homology arm sequence or downstream of the 3' homology arm sequence. In some embodiments, the polynucleic acid sequence encoding the endonuclease is present in a polynucleic acid molecule that is different from the RNA molecule. In some embodiments, the RNA molecule comprises a sequence encoding the one or more guide RNA molecules. In some embodiments, the sequence encoding the one or more guide RNA molecules is upstream of the 5' homology arm sequence or downstream of the 3' homology arm sequence. In some embodiments, the one or more polynucleic acids encoding the one or more guide RNA molecules comprise one or more polynucleic acid molecules encoding the one or more guide RNA molecules.

[0254] In some embodiments, the first guide RNA sequence comprises a sequence capable of hybridizing to a first target sequence of a second strand of the genomic DNA of a target cell and the second guide RNA sequence comprises a sequence capable of hybridizing to a second target sequence of a first strand of the genomic DNA of the target cell. In some embodiments, the number of nucleobases between the target sequence of the first strand of the genomic DNA of the target cell and the target sequence of the second strand of the genomic DNA of the target cell is about 10-1000 bases. In some embodiments, the number of nucleobases between the target sequence of the first strand of the genomic DNA of the target cell and the target sequence of the second strand of the genomic DNA of the target cell is about 50-500 bases. In some embodiments, the number of nucleobases between the target sequence of the first strand of the genomic DNA of the target cell and the target sequence of the second strand of the genomic DNA of the target cell is 100-300 bases. In some embodiments, the 5' homology arm sequence comprises a sequence capable of hybridizing to a target sequence of a second strand of the genomic DNA of the target cell and the 3' homology arm sequence comprises a sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA of the target cell. In some embodiments, the sequence of the 5' homology arm sequence capable of hybridizing to a target sequence of a second strand of the genomic DNA overlaps with the first target sequence of a second strand of the genomic DNA of the target cell capable of hybridizing to the first guide RNA sequence. In some embodiments, the sequence of the 5' homology arm sequence capable of hybridizing to a target sequence of a second strand of the genomic DNA is at most 50 bases downstream (3') of the

DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 40 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 30 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 20 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 10 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 5 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 4 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 3 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 2 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 1 base downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence. In some embodiments, the exogenous sequence encoding a polypeptide is flanked by the 5' homology arm and the 3' homology arm.

[0255] In some embodiments, the endonuclease is a nickase. In some embodiments, the nickase is a Cas9 nickase. In some embodiments, the Cas9 nickase is a *S. pyogenes* Cas9 endonuclease. In some embodiments, the Cas9 nickase comprises an H840A mutation relative to the wild type Cas9. In some

embodiments, the first guide RNA sequence and the second guide RNA sequence are encoded on a single polynucleic acid molecule, wherein the first guide RNA sequence and the second guide RNA sequence are separated by a cleavable sequence.

[0256] In some embodiments, the one or more guide RNA molecules comprises a guide RNA molecule comprising the first guide RNA sequence and the second guide RNA sequence. In some embodiments, the guide RNA molecule comprising the first guide RNA sequence and the second guide RNA sequence is capable of forming a complex with the endonuclease. In some embodiments, the endonuclease is precomplexed with the guide RNA molecule comprising the first guide RNA sequence and the second guide RNA sequence. In some embodiments, the one or more guide RNA molecules comprises a first guide RNA molecule comprising the first guide RNA sequence and a second guide RNA molecule comprising the second guide RNA sequence. In some embodiments, the first guide RNA molecule forms a complex with a first endonuclease molecule and the second guide RNA molecule forms a complex with a second endonuclease molecule. In some embodiments, the one or more guide RNA molecules directs the endonuclease to cut (i) a first strand of the genome DNA of the target cell and (ii) a second strand of the genomic DNA of the target cell. In some embodiments, the one or more guide RNA molecules do not direct the endonuclease to create a double strand break (DSB) in the genome DNA of the target cell. In some embodiments, the first guide RNA directs the endonuclease to cut a first strand of the genome DNA of the target cell and the second guide RNA directs the endonuclease to cut a second strand of the genomic DNA of the target cell. In some embodiments, the number of nucleobases between the first cut of the first strand of the genomic DNA and the second cut of the second strand of the genomic DNA is about 10-1000 bases. In some embodiments, the number of nucleobases between the first cut of the first strand of the genomic DNA and the second cut of the second strand of the genomic DNA is about 50-500 bases. In some embodiments, the number of nucleobases between the first cut of the first strand of the genomic DNA and the second cut of the second strand of the genomic DNA is about 100-300 bases. In some embodiments, the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell are within a same region of the genomic DNA of the target cell. In some embodiments, the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell are within a same locus of the genomic DNA of the target cell. In some embodiments, the locus is a genomic safe harbor locus. In some embodiments, the locus is non-ribosomal DNA. In some embodiments, the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell do not comprise a sequence of ribosomal DNA. In some embodiments, the locus is a human ortholog of the mouse Rosa26 locus. In some embodiments, the locus is the adeno-associated virus site 1 (AAVS1). In some embodiments, the locus is the CCR5 gene. In some embodiments, the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic

DNA of the target cell comprise sequences from a single gene within the genomic DNA of the target cell. In some embodiments, the single gene is a non-essential gene.

[0257] In some embodiments, the polypeptide with reverse transcriptase activity promotes reverse transcription of the reverse complement sequence of the insert sequence thereby producing the insert sequence. In some embodiments, the polypeptide with reverse transcriptase activity promotes integration of the insert sequence into the genomic DNA of the target cell. In some embodiments, the polypeptide with reverse transcriptase activity promotes integration of the insert sequence into the genomic DNA of the target cell via target primed reverse transcription (TPRT). In some embodiments, the polypeptide with reverse transcriptase activity promotes integration of the insert sequence into the genomic DNA of the target cell at the first target sequence of the genomic DNA of the target cell. In some embodiments, the polypeptide with reverse transcriptase activity promotes integration of the insert sequence into the genomic DNA of the target cell at the second target sequence of the genomic DNA of the target cell. In some embodiments, the polypeptide with reverse transcriptase activity promotes integration of the insert sequence into the genomic DNA of the target cell at the first target sequence of the genomic DNA of the target cell and integration of the insert sequence into the genomic DNA of the target cell at the second target sequence of the genomic DNA of the target cell. In some embodiments, the sequence encoding a polypeptide with reverse transcriptase activity comprises a mobile genetic element sequence. In some embodiments, the sequence encoding a polypeptide with reverse transcriptase activity comprises a human mobile genetic element sequence. In some embodiments, the mobile human mobile genetic element comprises LINE-1. In some embodiments, the polypeptide with reverse transcriptase activity comprises human ORF2p polypeptide or a functional fragment thereof. In some embodiments, the human ORF2p polypeptide or functional fragment thereof lacks endonuclease (EN) activity. In some embodiments, the human ORF2p polypeptide or functional fragment thereof comprises a mutant endonuclease domain that lacks endonuclease activity. In some embodiments, the human ORF2p polypeptide or functional fragment thereof comprises a D205A mutation relative to wild type human ORF2p. In some embodiments, the human ORF2p polypeptide or functional fragment thereof comprises an amino acid sequence with at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence identity to **SEQ ID NO: 160**. In some embodiments, the human ORF2p polypeptide or functional fragment thereof is encoded by a nucleic acid sequence with at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence identity to **SEQ ID NO: 129**. In some embodiments, the human ORF2p polypeptide or functional fragment thereof is a functional fragment of a human ORF2p polypeptide that lacks an endonuclease domain. In some embodiments, the RNA molecule further comprises a sequence encoding a human ORF1p polypeptide, wherein the sequence encoding the human ORF1p polypeptide is upstream of the 5' homology arm or downstream of the 3' homology arm.

[0258] In some embodiments the reverse transcriptase polypeptide is an avian R2 retroelement. In some embodiments the avian R2 retroelement is a wild type TaGu R2. In some embodiments, the wild type TaGu R2 comprises an amino acid sequence with at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence identity to a sequence:

MGTEKVMVTVDPKNPPCPCGTRVNSVLN~~IE~~HLKVSHGKRGVCFRCAKCGKENSNYHSV
VCHFPKCRGPETEKAPAGEWICEVCNRDFTTKIGLGQHKRLAHPAVRNQERIVASQPKETS
NRGAHKRCWTKEEEE~~ELL~~IRLEAQFEGNKNINKLIAEHITTKTAKQISDKRRLLSRKPAEEPRE
EPGTCHHTRRAAASLRTEPEMSHHAQAEDRDNGPGRRLPGRAAAGGRTMDEIRRHDPDKG
NGQQRPTKQKSEEQLQAYYKKTLEERLSAGALNTFPRAFKQVMEGRDIKLVINQTAQDCFG
CLESISQIRTATRDKKD~~TV~~TREKHPKKPFQKWMKDRAIKKGN~~YL~~RFQRLFYLD~~RG~~KLAKIIL
DDIECLSCDIPLSEIYSVFKTRWETTGSFKSLGDFKTYGKADNTAFRELITAKEIEKNVQEMS
KGSAPGPDGITLGDVVKMDPEFSRTMEIFNLWLTTGKIPDMV~~RG~~CRTVLIPKSSKPDRLKDI
NNWRPITIGSILLR~~LF~~SRIVTARLSKACPLNPRQ~~RG~~FIRAAGCSENKLLQTI~~WS~~AKREHRPL
GVV~~F~~VDIAKAFDTVSHQHIIHALQREVDPHIVGLVSNMYENISTYITTKRNTHTDKIQIRVG
VKQGDPMSP~~LL~~FN~~L~~AMD~~PL~~LCKLEESGKGYHRGQSSITAMAFADDLVLLSDSWENMNTNI
SILETFCNL~~T~~GLKTQGQKCHGFYIKPTKDSYTINDCAAWTINGTPLNMIDPGESEKYLGLQFD
PWIGIARSG~~L~~STKLDFWLQRIDQAPLKPLQKTDILKTYTIPRLIYADHSEVKTALLETLDQKI
RTAVKEWLHLPCTCDAILYSSTRDGG~~L~~GITKL~~AG~~LIPSVQARRLHRIAQSSDDTMKCFMEK
EKMEQLHKKLWIQAGGDRENIPSIWEAPPSSEPPNNVSTNSEWEAPTQKDKFPKPCNWRKN
EFKKWTKLASQGRGIVNFERDKISNHWIQYYRRIPHRKLLTALQLRANVYPTREFLAGRQ
DQYIKACRHCADIESCAHIIGNCPVTQDARIKRHNYICELLLEEAKKKDWVVFKEPHIRDS
NKELYKPD~~L~~IFVKDARALVVDVTVRYEAAKSSLEEAAAEKVRKYKHLETEVRHLTNAKDV
TFVGFPLGARGKWHQDNFKLLTELGLSKSRQVKMAET~~F~~STVALFSSVDIVHMFASRARKS
MVM* (SEQ ID NO: 1).

In some embodiments, the avian R2 retroelement is an endonuclease dead (ENdead) TaGu R2. In some embodiments, the ENdead TaGu R2 comprises an amino acid sequence with at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence identity to a sequence (mutations bolded):

MGTEKVMVTVDPKNPPCPCGTRVNSVLN~~IE~~HLKVSHGKRGVCFRCAKCGKENSNY
HSVVCHFPKCRGPETEKAPAGEWICEVCNRDFTTKIGLGQHKRLAHPAVRNQERIVASQ
PKETSNRGAHKRCWTKEEEE~~ELL~~IRLEAQFEGNKNINKLIAEHITTKTAKQISDKRRLLSR
KPAEEPREPGTCHHTRRAAASLRTEPEMSHHAQAEDRDNGPGRRLPGRAAAGGRTM
DEIRRHDPDKGNGQQRPTKQKSEEQLQAYYKKTLEERLSAGALNTFPRAFKQVMEGRDI
KLVINQTAQDCFGCLESISQIRTATRDKKD~~TV~~TREKHPKKPFQKWMKDRAIKKGN~~YL~~RF
QRLFYLD~~RG~~KLAKIILDDIECLSCDIPLSEIYSVFKTRWETTGSFKSLGDFKTYGKADNTA
FRELITAKEIEKNVQEMSKGSAPGPDGITLGDVVKMDPEFSRTMEIFNLWLTTGKIPDMV
RGCRTVLIPKSSKPDRLKDINNWRPITIGSILLR~~LF~~SRIVTARLSKACPLNPRQ~~RG~~FIRAAG

CSENLKLLQTIWSAKREHRPLGVVFVDIAKAFDTVSHQHIIHALQQREVDPHIVGLVSN
 MYENISTYITTKRNTHTDKIQRVGVKQGDPMSPLLFNAMDPLLCKLEESGKGYHRGQ
 SSITAMAFAAALVLLSDSWENMNTNISILETFCNLTGLKTQGQKCHGFYIKPTKDSYTIN
 DCAAWTINGTPLNMIDPGESEKYLGLQFDPWIGIARSGLSTKLDFWLQRIDQAPLKPLQ
 KTDILKTYTIPRLIYIADHSEVKTALLETLDQKIRTAVKEWLHLPPCTCDAILYSTRDGGGL
 GITKLAGLIPSVQARRLHRIAQSSDDTMKCFMEKEKMEQLHKKLWIQAGGDRENIPSIW
 EAPPSSEPPNNVSTNSEWEAPTQKDKFPKPCNWRKNEFKKWTKLASQGRGIVNFERDKI
 SNHWIQYYRRIPIHRKLLTALQLRANVYPTREFLARGRQDQYIKACRHCADIESCAHIIG
 NCPVTQDARIKRHNYICELLLEEAKKKDWVVFKEPHIRDSNKELYKPDILFVKDARALV
 VDVTVRYEAAKSSLEEAAAEKVRKYKHLETEVRHLTNAKDVTVFGLGARGKWHQD
 NFKLLTELGLSKSRQVKMAETFSTVALFSSVDIVHMFASRARKSMVM* (SEQ ID NO: 2).

In some embodiments, the avian R2 retroelement is a wild type ZoAI R2. In some embodiments, the wild type ZoAI R2 comprises an amino acid sequence with at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence identity to a sequence

MNIVKVTVPDKNPPCPCGVRNLNSVLALIEHLKGSHGRRRVCFRCAKCGRENFNHHST
 VCHYAKCKGPQIERPPVGEWICEVCGRDFTTKIGLGQHKRHHMHAMVRNQERIDASQPK
 ETSNRGAHKRCWTKEEEELMKLEVQFENHKNINKLIAEQLTTKTAKQISDKRRMLLK
 KGRGTTGNLETEPGMSHQSAKVKDNLGGDHLPGGPVVDKGTIGKPGQHLDTDNH
 QITAGKKKGGGLQARYRRRIMKRLAAGTINIFPKVFKELINDQEARPLINQTTEDCFGLL
 DSACQIRTALREKKGKSQEERPRKQYQKWMKKRAIKRGDYLRQRLFHLDRGKLARIILD
 NTELSCDISPSEIYSVFKARWETPGHFNGLDGFEIKGKANNKAFRDFITAKEIEKNVREM
 SKGSAPGPDGIALGDIKKMDPGYSRTAELFNWLTAGDIPDMVRGCRTVLIPKSTTPERL
 KDINNWRPITIGSILLRLFSRIITARMTKACPLNPRQRFISAPGCSENLKLLQSIIRTAKNE
 HKPLGVIFVDIAKAFDTVSHQHIIHVLQQRVDPHIVGLVNNMYKDISTYVTTKKNTH
 DTKIQRVGVKQGDPLSPLLFNAMDPLLCKLEESGKGFHRGQSSITAMAFADDLVLLSDS
 WENMKENIKILETFCNLTGLKTQGQKCHGFYIKPTKDSYTINNCPAWTINGTPLNMINPG
 ESEKYLGLQIDPWTGVAKYDLSTKLKIWLESIDRAPLKPLQKLDILKTYTIPRLTYLADHS
 EMKAGALEALDQQIRTAVKDWLHLPSCCTDAILYVSTRDGGGLGVTKLAGLIPSVQARRL
 HRIAQSPDETMKDFLEKAQMEKMYEKLWVQAGGKKKGMPsiWEALPMTVPPTNTGNL
 SEWEAPNPKSKYPKPCDWRRKELKKWTKLESQGRGVKNFRNDTISNDWIQYYRRIPIHR
 KLLTAIQLRANVYPTREFLARGRGDNYVKFCRHCEADLETGHIIGFCPVTKDARIKRH
 NRICDRLCEEAAKREWVVFKEPHLRDATTELFKPDVIFVKEDRALVVDVTVRYESA
 KTLEAAAMEKVDKYKHLEAEVKELTNAKDVFVMGFPLGARGKFKGNFNLLETGLPKT
 RQLSVAKTLSTYALMSSVDIVHMFASRSRKPNV* (SEQ ID NO: 3). In some embodiments, the avian R2 retroelement is an endonuclease dead (ENdead) ZoAI R2. In some embodiments, the

ENdead ZoAl R2 comprises an amino acid sequence with at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence identity to a sequence (mutations bolded):

MNIVKVTVPDKNPPCPCCGVRLNSVLALIEHLKGSHGRRRVCFRCAKCGRENFNHHSTVCH
YAKCKGPQIERPPVGEWICEVCGRDFTTKIGLGQHKRHHAMVRNQERIDASQPKETSNRG
AHKRCWTKEEEEELMKLEVQFENHKNINKLIAEQLTTKTAKQISDKRRMLLKKGRGTTGNL
ETEPGMSHQSQAKVKDNGLGGDHLPGGPVVDKGTIGKPGQHLDTDNSHQITAGKKKGGGL
QARYRRRIMKRLAAGTINIFPKVFKELINDQEARPLINQTTEDCFGLLDSACQIRTALREK GK
SQEERPRKQYQKWMKKRAIKRGDYLRQFLHLDRGKLARIILDNTESLSCDISPSEIYSVFK
ARWETPGHFNGLGDFEIKGKANNKAFRDFITAKEIEKNVREMSKGSAPGPDGIALGDIKKM
DPGYSRTAELFNLWLTAGDIPDMVRGCRTVLIPKSTTPERLKDINNWRPITIGSILLRLFSRIITA
RMTKACPLNPRQRGFISAPGCSENKLLQSIIRTAKEHKLPLGVIFVDIAKAFTVSHQHIIHV
LQQRVDPHIVGLVNNMYKDISTYVTTKKNTHTDKIQRVGVKQGDPLSPLLNLAMDPLLC
KLEESGKGFHRGQSSITAMAF~~AA~~LVLSDSWENMKENIKILETFCNLTGLKTQGQKCHGFY
IKPTKDSYTINNCPAWTINGTPLNMINPGESEKYLGLQIDPWTGVAKYDLSTKLKIWLESIDR
APLKPLQKLDILKTYTIPRLTYLADHSEMKAGALEALDQQIRTAVKDWLHLPSTCDAILYVS
TRDGGGLGVTKLAGLIPSVQARRLHRIAQSPDETMKDFLEKAQMEKMYEKLWVQAGGKKK
GMPSIWEALPMTVPPTNTGNLSEWEAPNPKSKYPKPCDWRRKELKKWTKLESQGRGVKNF
RNDTISNDWIQYYRRIPHRKLLTAIQLRANVYPTREFLARGRGDNYVKFCRHCEADLET CGH
IIGFCPVTKDARIKRHNRICDRLCEEAAKREWWVFKEPHLRDATTELFKPDVIFVKEDRALVV
DVTVRYESA~~KT~~TTLEAAAMEKV~~DK~~YKHLEAEVKELTNAKDVVFMGFPLGARGKFYKGNFN
LLETGLPKTRQLSVAKTLSTYALMSSVDIVHMFASRSRKNV* (SEQ ID NO: 4).

[0259] In some embodiments, the nucleic acid sequence encoding the avian R2 comprises a 3' UTR. In some embodiments, the 3' UTR comprises a ZoAl 3' UTR. In some embodiments, the ZoAl 3'UTR comprises a sequence with at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence identity to a sequence of

TAGGTAGTCACATTGCACTTTCTGTAACCTGCACTGGGTGTGGGATGTGGGCCTGGGGT
GTGGGTATGGGGTATATATGTGGGATATTCTGGTGGGAATGTCCATTCACTGTATGCC
ATCTTTTAAATAAAAAGACGGTAGCTAGGTTTCGCGAAGCAGCCACAAGCCAATAGCCA
GTTAGGTAGCTCATAGTGGGTAGGTGACAGGAACCTTTGACTCAGAACGCGTCCATTAA
CATCTAGAACGGACCAAACTTCGGACATGCACCGATTAACCGGATTTGTCCAAGGTGGA
CGGGCCACCTTTACTTAACCCGGAAGGGAACATATATAGTTATATGTGTTCGTAATA
(SEQ ID NO: 5).

[0260] In some embodiments, the 3' UTR comprises a TaGu 3' UTR. In some embodiments, the TaGu 3'UTR comprises a sequence with at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% identity to a sequence of

TAATTCAGGTTATTTAGATGCTTAGTTTTTGTACCTTTCTTGTGTTTGTGTTAGGATTTTGAT

AGTGTTAGTATTTTTATATTTTTGTACGATTGCATAATGTTCTTTTTTATACAGTTCTGTT
TTAATAAAATAGACGATAGCTAGAGACGTTAGGGCAGCCACAAGCCAGTTAGGTAGCG
GATAGTAGGTAGGAACAGACTTTTACTATTTTCATAACGCGTCAATTACCACCTGATTTG
GACCAATTCACGGGATTTGTCCAAGGTGGACGGGCCACCTTTACTTAACCCGGAAAAGG
AACATATATAATTTATGTGTGTTTCGATAAA (SEQ ID NO: 6).

[0261] In some embodiments, the polypeptide is a therapeutic polypeptide. In some embodiments, the polypeptide is a human polypeptide. In some embodiments, the polypeptide is a ligand. In some embodiments, the polypeptide is an antibody. In some embodiments, the polypeptide is a receptor. In some embodiments, the polypeptide is an enzyme. In some embodiments, the polypeptide is a transport protein. In some embodiments, the polypeptide is a structural protein. In some embodiments, the polypeptide is a hormone. In some embodiments, the polypeptide is a contractile protein. In some embodiments, the polypeptide is a storage protein. In some embodiments, the polypeptide is a transcription factor. In some embodiments, the polypeptide is a chimeric antigen receptor (CAR). In some embodiments, the polypeptide is a T cell receptor (TCR).

[0262] In some embodiments, the target cell is a mammalian cell. In some embodiments, the target cell is a human cell. In some cases, the human cell is a liver cell. In some embodiments, the target cell is an ex vivo cell. In some embodiments, the target cell is an in vivo cell. In some embodiments, the target cell is an immune cell. In some embodiments, the immune cell is a T cell. In some embodiments, the immune cell is a B cell. In some embodiments, the immune cell is a myeloid cell. In some embodiments, the immune cell is a monocyte. In some embodiments, the immune cell is a macrophage. In some embodiments, the immune cell is a dendritic cell.

[0263] In some embodiments, the composition further comprises one or more delivery vehicles for delivery of the RNA molecule, the endonuclease or the polynucleic acid sequence encoding the endonuclease, and the one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules into the target cell. In some embodiments, the one or more delivery vehicles is a nanoparticle delivery vehicle. In some embodiments, the one or more delivery vehicles is a plasmid vector. In some embodiments, the one or more delivery vehicles is a viral vector. In some embodiments, the viral vector is an adenoviral vector. In some embodiments, the viral vector is an adeno-associated viral vector. In some embodiments, the viral vector is a lentiviral vector. In some embodiments, the viral vector is a retroviral vector. In some embodiments, the nanoparticle delivery vehicle is a lipid nanoparticle. In some embodiments, the nanoparticle delivery vehicle is a polymeric nanoparticle.

[0264] Also provided herein is a pharmaceutical composition comprising any of the foregoing embodiments and a pharmaceutically acceptable excipient. The term “pharmaceutically acceptable” refers to approved or approvable by a regulatory agency of the Federal or a state government or listed in the U.S. Pharmacopeia (U.S.P.) or other generally recognized pharmacopeia for use in animals,

including humans. A “pharmaceutically acceptable excipient” refers to an excipient that can be administered to a subject, together with an active moiety, and which does not destroy the pharmacological activity thereof and is nontoxic when administered in doses sufficient to deliver a therapeutic amount of the moiety.

[0265] Also provided herein is a gene editing system for use in incorporating an exogenous sequence encoding a therapeutic polypeptide in a genome of a target cell, the system comprising a composition comprising an endonuclease or a polynucleic acid encoding the same and one or more polynucleotides, wherein the one or more polynucleotides comprise a first guide RNA sequence and a second guide RNA sequence or DNA sequences encoding the same, and an mRNA comprising a sequence encoding a transposon machinery comprising a sequence that is a reverse complement of the sequence encoding the exogenous therapeutic polypeptide, a 5' homology arm and a 3' homology arm flanking the sequence encoding the exogenous therapeutic polypeptide, and wherein each homology arm is complementary to a sequence comprising a target site in the genome, and a sequence encoding a recombinant human LINE1 ORF1p or functional fragment thereof, and a sequence encoding a recombinant human ORF2p reverse transcriptase (RT), wherein the endonuclease activity of ORF2p is inactivated. In some embodiments, the nucleic acid sequence flanked by the two homology arms does not comprise or overlap with the sequences encoding the endonuclease or the LINE1 ORF1p or the ORF2p, or a fragment thereof.

[0266] In some embodiments, the endonuclease is a nickase. In some embodiments, the nickase is a Cas nickase. In some embodiments the nickase is a Cas9 nickase. In some embodiments the nickase is a Cas12a nickase. In some embodiments the nickase is a Cas12b nickase. In some embodiments the nickase is a Cas13 nickase. In some embodiments the nickase is a CasX nickase. In some embodiments the nickase is a CasY nickase. In some embodiments, the nickase is a Cas9 nickase comprising an H840A mutation relative to the wild type Cas9. In some embodiments, the Cas9 nickase is encoded by a nucleic acid having at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% identity to a sequence of SEQ ID NO: 124.

[0267] In some embodiments, the EN is inactivated by introduction of one or more mutations. In some embodiments, the ORF2p EN comprises a single mutation. In some embodiments, the ORF2p EN comprises a double mutation. In some embodiments, the ORF2p EN comprises a triple mutation. In some embodiments, the ORF2p EN comprises a S228P mutation. In some embodiments, the ORF2p EN comprises a Y1180A mutation. In some embodiments, the ORF2p EN comprises a S228P and Y1180A mutation. In some embodiments, the activity of the ORF2p EN is inactivated by truncation mutation.

[0268] Also provided herein is a method of incorporating an exogenous sequence encoding a therapeutic polypeptide at a specific site in a genome of a target mammalian cell, the method comprising contacting the target mammalian cell with a composition comprising an endonuclease or

a polynucleic acid encoding the same and one or more polynucleic acids, wherein the one or more polynucleic acids comprise a first guide RNA sequence and a second guide RNA sequence, or DNA sequences encoding the same and a human mobile genetic element comprising a sequence encoding a polypeptide that promotes integration of the exogenous sequence encoding the therapeutic polypeptide into the genome of the target mammalian cell via target primed reverse transcription (TPRT), and expressing the exogenous sequence encoding the therapeutic polypeptide from a genomically incorporated sequence of the target mammalian cell, wherein only the exogenous sequence encoding the therapeutic polypeptide is the genomically incorporated sequence that is incorporated at the specific site in the genome of the target mammalian cell.

[0269] In some embodiments, the 5' homology arm and 3' homology arm flank the mRNA sequence that is a reverse complement of an exogenous sequence encoding the therapeutic polypeptide. In some embodiments, wherein the nucleic acid sequence flanked by the two homology arms does not comprise or overlap with the sequences encoding the endonuclease or the human mobile genetic element, or a fragment thereof. In some embodiments, the human mobile genetic element comprises one or more sequences encoding a human LINE1 ORF polypeptide. In some embodiments, the one or more sequences encoding a human LINE1 ORF polypeptide comprise a sequence encoding an ORF2p polypeptide or a functional fragment thereof. In some embodiments, the one or more sequences encoding a human LINE1 ORF polypeptide comprise a sequence encoding an ORF1p polypeptide or a functional fragment thereof. In some embodiments, the ORF2p polypeptide comprises a reverse transcriptase (RT). In some embodiments, the ORF2p polypeptide comprises an endonuclease (EN). In some embodiments, the ORF2p polypeptide comprises an RT and an EN. In some embodiments, the ORF2p polypeptide comprises an RT and an EN wherein the EN is functionally inactivated. In some embodiments, the EN is inactivated by introduction of one or more mutations. In some embodiments, the ORF2p EN comprises a single mutation. In some embodiments, the ORF2p EN comprises a double mutation. In some embodiments, the ORF2p EN comprises a triple mutation. In some embodiments, the ORF2p EN comprises a S228P mutation. In some embodiments, the ORF2p EN comprises a Y1180A mutation. In some embodiments, the ORF2p EN comprises a S228P and Y1180A mutation. In some embodiments, the activity of the ORF2p EN is inactivated by truncation mutation.

[0270] In some embodiments, the endonuclease is a nickase. In some embodiments, the nickase is a Cas nickase. In some embodiments the nickase is a Cas9 nickase. In some embodiments the nickase is a Cas12a nickase. In some embodiments the nickase is a Cas12b nickase. In some embodiments the nickase is a Cas13 nickase. In some embodiments the nickase is a CasX nickase. In some embodiments the nickase is a CasY nickase. In some embodiments, the nickase is a Cas9 nickase comprising an H840A mutation relative to the wild type Cas9.

[0271] In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 50 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 100 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 150 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 200 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 300 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 400 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 500 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 600 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 700 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 800 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 900 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the endonuclease cuts the genomic DNA at a site that is within 1000 nucleotides from the specific site in the genome of a target mammalian cell. In some embodiments, the method prevents double stranded breaks on the genomic DNA.

[0272] In some embodiments, the first guide RNA sequence and the second guide RNA sequence are on a single polynucleotide. In some embodiments, the first guide RNA sequence and the second guide RNA sequence are on a single polynucleotide and separated by an auto-cleavable sequence. In some embodiments the auto-cleavable sequence is T2A. In some embodiments, the auto-cleavable sequence is P2A. In some embodiments, the first and second guide RNA sequences are encoded on two separate RNA molecules. In some embodiments, the first the first guide RNA comprises a sequence that has Watson-Crick pairing with a plurality of contiguous nucleotides on a first strand of the genomic DNA. In some embodiments, the second guide RNA comprises a sequence that has Watson-Crick pairing with a plurality of contiguous nucleotides on the second and opposite strand of the genomic DNA. In some embodiments, the first guide RNA comprises a sequence that has Watson-Crick pairing with a plurality of contiguous nucleotides on a first strand of the genomic DNA, and the second guide RNA comprises a sequence that has Watson-Crick pairing with a plurality of contiguous nucleotides on the second and opposite strand of the genomic DNA. In some embodiments, the first guide RNA and the second guide RNA do not comprise sequences that have Watson-Crick pairing

with a plurality of contiguous nucleotides of the genomic DNA that are themselves complementary to each other.

[0273] In some embodiments, provided herein is a composition or a system comprising: (a) an RNA molecule comprising: (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; (ii) a 5' homology arm and a 3' homology arm; and (iii) a sequence encoding a human ORF2p polypeptide, wherein the sequence encoding a human ORF2p polypeptide is upstream of the 5' homology arm or downstream of the 3' homology arm; (b) an endonuclease or a polynucleic acid sequence encoding the endonuclease; and (c) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA sequence and a second guide RNA sequence, or a polynucleic acid encoding the one or more guide RNA molecules, wherein the first guide RNA sequence comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA sequence comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell. In some embodiments, the first guide RNA hybridizes with nucleotide sequences on a first genomic DNA strand and the second guide RNA hybridizes with nucleotide sequences on the complementary genomic DNA strand, at non-overlapping regions.

[0274] The composition may be referred to as a retrotransposition complex, or editing complex.

[0275] In some embodiments, the RNA molecule comprising the reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to the sequence that is a reverse complement of a promoter sequence; the 5' homology arm and the 3' homology arm; and the sequence encoding a human ORF2p polypeptide may comprise the CREATE mRNA as described herein. In some embodiments, the CREATE mRNA also comprises a sequence encoding ORF1p.

[0276] In some embodiments, contacting the target mammalian cell with a composition comprises delivering the composition into the target mammalian cell. In some embodiments, the one or more delivery vehicle is a nanoparticle delivery vehicle. In some embodiments, the one or more delivery vehicle is a plasmid vector. In some embodiments, the one or more delivery vehicle is a viral vector. In some embodiments, a polynucleic acid of the composition is delivered to the target mammalian cell using a viral vector. In some embodiments, the viral vector is an adenovirus. In some embodiments, the viral vector is an adeno-associated virus. In some embodiments, the viral vector is a lentivirus. In some embodiments, the viral vector is a retrovirus. In some embodiments, the composition is delivered

to the target mammalian cell using a lipid nanoparticle. In some embodiments, the composition is delivered to the target mammalian cell using a polymeric nanoparticle.

[0277] In some embodiments, the therapeutic polypeptide is a ligand. In some embodiments, the therapeutic polypeptide is an antibody. In some embodiments, the therapeutic polypeptide is an enzyme. In some embodiments, the therapeutic polypeptide is a transport protein. In some embodiments, the therapeutic polypeptide is a structural protein. In some embodiments, the therapeutic polypeptide is a hormone. In some embodiments, the therapeutic polypeptide is a contractile protein. In some embodiments, the therapeutic polypeptide is a storage protein. In some embodiments, the therapeutic polypeptide is a transcription factor. In some embodiments, the therapeutic polypeptide is a chimeric antigen receptor (CAR). In some embodiments, the therapeutic polypeptide is a T cell receptor (TCR).

[0278] In some embodiments, the target mammalian cell is an immune cell. In some embodiments, the target mammalian cell is a T cell. In some embodiments, the target mammalian cell is a B cell. In some embodiments, the target mammalian cell is a myeloid cell. In some embodiments, the target mammalian cell is a monocyte. In some embodiments, the target mammalian cell is a macrophage. In some embodiments, the target mammalian cell is a dendritic cell.

[0279] Immunotherapy using phagocytic cells involves making and using engineered myeloid cells, such as macrophages or other phagocytic cells that attack and kill diseased cells, such as cancer cells, or infected cells. Engineered myeloid cells, such as macrophages and other phagocytic cells are prepared by incorporating in them via recombinant nucleic acid technology, a synthetic, recombinant nucleic acid encoding an engineered protein, such as a chimeric antigen receptor, that comprises a targeted antigen binding extracellular domain that is designed to bind to specific antigens on the surface of a target, such as a target cell, such as a cancer cell. Binding of the engineered chimeric receptor to an antigen on a target, such as cancer antigen (or likewise, a disease target), initiates phagocytosis of the target. This may trigger a two-fold action: one, phagocytic engulfment and lysis of the target destroys the target and eliminates it as a first line of immune defense; two, antigens from the target are digested in the phagolysosome of the myeloid cell, are presented on the surface of the myeloid cell, which then leads to activation of T cells and further activation of the immune response and development of immunological memory. Chimeric receptors are engineered for enhanced phagocytosis and immune activation of the myeloid cell in which it is incorporated and expressed. Chimeric antigen receptors of the disclosure are variously termed herein as a chimeric fusion protein, CFP, phagocytic receptor (PR) fusion protein (PFP), or chimeric antigen receptor for phagocytosis (CAR-P), while each term is directed to the concept of a recombinant chimeric and/or fusion receptor protein. In some embodiments, genes encoding non-receptor proteins are also co-expressed in the myeloid cells, typically for an augmentation of the chimeric antigen receptor function. In summary, contemplated herein are various engineered receptor and non-receptor recombinant proteins that are

designed to augment phagocytosis and or immune response of a myeloid cell against a disease target, and methods and compositions for creating and incorporating recombinant nucleic acids that encode the engineered receptors or non-receptor recombinant protein, such that the methods and compositions are suitable for creating an engineered myeloid cell for immunotherapy.

[0280] In one aspect, the present disclosure provides compositions and methods for stable gene transfer into a cell, where the cell can be any somatic cell. In some embodiments, the compositions and methods are designed for cell-specific or tissue-specific delivery. In some cases, the methods described herein relate to supplying a functional protein or a fragment thereof to compensate for an absent or defective (mutated) protein *in vivo*, e.g., for a protein replacement therapy.

[0281] Incorporation of a recombinant nucleic acid in a cell can be accomplished by one or more gene transfer techniques that are available in the state of the art. However, incorporation of exogenous genetic (e.g., nucleic acid) elements into the genome for therapeutic purposes still faces several challenges. Achieving stable and specific integration in a safe and dependable manner, and efficient and prolonged expression are a few among them. Most of the successful gene transfer systems aimed at genomic integration of the cargo nucleic acid sequence rely on viral delivery mechanisms, which have some inherent safety and efficacy issues. Furthermore, certain systems result in non-specific integration of the payload into the genome. Efficient and safe delivery and specific integration of long nucleic acid sequences (e.g., sequences larger than 1kb) cannot be achieved by current gene editing systems.

[0282] Little attention has so far been devoted to making and using engineered myeloid cells for stable long-term gene transfer and expression of the transgene. For example, gene transfer to differentiated mammalian cells *ex vivo* for cell therapy can be accomplished via viral gene transfer mechanisms. However, there are several strategic disadvantages associated with the use of viral gene-transfer vectors, including an undesired potential for transgene silencing over time, the preferential integration into transcriptionally active sites of the genome with associated undesired activation of other genes (e.g., oncogenes) and genotoxicity. In addition to the safety issues increased expense and cumbersome effort of manufacturing, storing and handling integrating viruses often stand in the way of large-scale use of viral vector mediated of gene-modified cells in therapeutic applications. These persistent concerns associated with viral vectors regarding safety, as well as cost and scale of vector production necessitates alternative methods for effective therapy.

[0283] Integration of a transgene into the genome of a cell to be used for an immunotherapy can be advantageous in the sense that it is stable, and a lower number of cells is required for delivery during the therapy. On the other hand, integrating a transgene in a non-dividing cell can be challenging in both affecting the health and function of the cell as well as the ultimate lifespan of the cell *in vivo*, and therefore affects its overall utility as the therapeutic. In some embodiments, the methods described herein for generating a myeloid cell for immunotherapy can be a cumulative product of a number of

steps and compositions involving but not limited to, for example, selecting a myeloid cell for modifying; methods and compositions for incorporating a recombinant nucleic acid in a myeloid cell; methods and compositions for enhancing expression of the recombinant nucleic acid; methods and compositions for selecting and modifying vectors; methods of preparing a recombinant nucleic acid suitable for *in vivo* administration for uptake and incorporation of the recombinant nucleic acid by a myeloid cell *in vivo* and therefore generating a myeloid cell for therapy. In some aspects, one or more embodiments of the various inventions described herein are transferrable among each other, and one of skill in the art is expected to use them in alternatives, combinations or interchangeably without the necessity of undue experimentation. All such variations of the disclosed elements are contemplated and fully encompassed herein.

[0284] In one aspect, transposons, or transposable elements (Tes) are considered herein, for means of incorporating a heterologous, synthetic or recombinant nucleic acid encoding a transgene of interest in a myeloid cell. Transposon, or transposable elements are genetic elements that have the capability to transpose fragments of genetic material into the genome by use of an enzyme known as transposase. Mammalian genomes contain a high number of transposable element (TE)-derived sequences, and up to 70% of our genome represents TE-derived sequences (de Koning et al. 2011; Richardson et al. 2015). These elements could be exploited to introduce genetic material into the genome of a cell. The TE elements are capable of mobilization, often termed as “jumping” genetic material within the genome. Tes generally exist in eukaryotic genomes in a reversibly inactive, epigenetically silenced form. In the present disclosure methods and compositions for efficient and stable integration of transgenes into macrophages and other phagocytic cells. The method is based on use of a transposase and transposable elements mRNA-encoded transposase. In some embodiments, Long Interspersed Element-1 (L1) RNAs are used for stable integration and/or retrotransposition of the transgene into a cell (e.g., a macrophage or phagocytic cell).

[0285] Contemplated herein are methods for retrotransposon mediated stable and specific integration of an exogenous nucleic acid sequence into the genome of a cell. Methods described herein can be used for robust and versatile incorporation of an exogenous nucleic acid sequence into a cell, such that the exogenous nucleic acid is only incorporated at a safe locus within the genome and is expressed without being silenced by the cell's inherent defense mechanism. The method described herein can be used to incorporate an exogenous nucleic acid that is about 1 kb, about 1.5 kb, about 2 kb, about 2.5 kb, about 3 kb, about 3.5 kb, about 4 kb, about 4.5 kb, about 5 kb, about 5.5 kb, about 6 kb, about 6.5 kb, about 7kb, about 7.5 kb, about 8 kb, about 8.5 kb, about 9 kb, about 9.5 kb, about 10 kb, or more in size. In some embodiments, the exogenous nucleic acid is not incorporated within a ribosomal locus. In some embodiments, the exogenous nucleic acid is not incorporated within a ROSA26 locus, or another safe harbor locus. In some embodiments, the exogenous nucleic acid is incorporated within a ribosomal locus. In some embodiments, the exogenous nucleic acid is incorporated within a ROSA26

locus, or another safe harbor locus. In some embodiments, the methods and compositions described herein can incorporate an exogenous nucleic acid sequence at a specific locus anywhere within the genome of the cell. Furthermore, contemplated herein is a retrotransposition system that is developed to incorporate an exogenous nucleic acid sequence into a specific predetermined site within the genome of a cell, without creating an adverse effect. The disclosed methods and compositions incorporate several mechanisms of engineering the retrotransposons for highly specific incorporation of the exogenous nucleic acid into a cell with high fidelity. Retrotransposons chosen for this purpose may be a human retrotransposon.

[0286] In one aspect, provided herein are methods and compositions for delivery inside a cell, for example a myeloid cell, and stable incorporation of one or more nucleic acids, comprising nucleic acid sequences encoding one or more proteins, wherein the stable incorporation may be via non-viral mechanisms. In some embodiments, the delivery of a nucleic acid composition into a myeloid cell is via a non-viral mechanism. In some embodiments, the delivery of the nucleic acids may further bypass plasmid mediated delivery. A "plasmid," as used herein, refers to a non-viral expression vector, e.g., a nucleic acid molecule that encodes for genes and/or regulatory elements necessary for the expression of genes. A "viral vector," as used herein, refers to a viral-derived nucleic acid that is capable of transporting another nucleic acid into a cell. A viral vector is capable of directing expression of a protein or proteins encoded by one or more genes carried by the vector when it is present in the appropriate environment. Examples for viral vectors include, but are not limited to retroviral, adenoviral, lentiviral and adeno-associated viral vectors.

[0287] In some embodiments, provided herein is a method of delivering a composition inside a cell, such as in a myeloid cell, the composition comprising one or more nucleic acid sequences encoding one or more proteins, wherein the nucleic acid sequence is an RNA. In some embodiments, the RNA is mRNA. In some embodiments, the mRNA may comprise at least one modified nucleotide. The term "nucleotide," as used herein, refers to a base-sugar-phosphate combination. A nucleotide may comprise a synthetic nucleotide. A nucleotide may comprise a synthetic nucleotide analog. Nucleotides may be monomeric units of a nucleic acid sequence (e.g., deoxyribonucleic acid (DNA) and ribonucleic acid (RNA)). The term nucleotide may include ribonucleoside triphosphates adenosine triphosphate (ATP), uridine triphosphate (UTP), cytosine triphosphate (CTP), guanosine triphosphate (GTP) and deoxyribonucleoside triphosphates such as dATP, dCTP, dITP, dUTP, dGTP, or derivatives thereof. Such derivatives may include, for example, [aS]dATP, 7-deaza-dGTP and 7-deaza-dATP, and nucleotide derivatives that confer nuclease resistance on the nucleic acid molecule containing them. The term nucleotide as used herein may refer to dideoxyribonucleoside triphosphates (ddNTPs) and their derivatives. Illustrative examples of dideoxyribonucleoside triphosphates may include, but are not limited to, ddATP, ddCTP, ddGTP, ddITP, and ddTTP. A nucleotide may be unlabeled or detectably labeled by well-known techniques. Labeling may also be carried out with quantum dots. Detectable labels may

include, for example, radioactive isotopes, fluorescent labels, chemiluminescent labels, bioluminescent labels and enzyme labels. Fluorescent labels of nucleotides may include but are not limited fluorescein, 5-carboxyfluorescein (FAM), 2'7'-dimethoxy-4'5'-dichloro-6-carboxyfluorescein (JOE), rhodamine, 6-carboxyrhodamine (R6G), N,N,NcN'-tetramethyl-6-carboxyrhodamine (TAMRA), 6-carboxy-X-rhodamine (ROX), 4-(4'dimethylaminophenylazo) benzoic acid (DABCYL), Cascade Blue, Oregon Green, Texas Red, Cyanine and 5-(2'-aminoethyl)aminonaphthalene-1-sulfonic acid (EDANS). Specific examples of fluorescently labeled nucleotides may include [R6G]dUTP, [TAMRA]dUTP, [R110]dCTP, [R6G]dCTP, [TAMRA]dCTP, [JOE]ddATP, [R6G]ddATP, [FAM]ddCTP, [R110]ddCTP, [TAN1RA]ddGTP, [ROX]ddTTP, [dR6G]ddATP, [dR110]ddCTP, [dTAMRA]ddGTP, and [dROX]ddTTP available from Perkin Elmer, Foster City, Calif. FluoroLink DeoxyNucleotides, FluoroLink Cy3-dCTP, FluoroLink Cy5-dCTP, FluoroLink Fluor X-dCTP, FluoroLink Cy3-dUTP, and FluoroLink Cy5-dUTP available from Amersham, Arlington Heights, Ill.; Fluorescein-15-dATP, Fluorescein-12-dUTP, Tetramethyl-rhodamine-6-dUTP, TR770-9-dATP, Fluorescein-12-ddUTP, Fluorescein-12-UTP, and Fluorescein-15-2'-dATP available from Boehringer Mannheim, Indianapolis, Ind.; and Chromosome Labeled Nucleotides, BODIPY-FL-1 4-UTP, BODIPY-FL-4-UTP, BODIPY-TMR-14-UTP, BODIPY-TMR-14-dUTP, BODIPY-TR-14-UTP, BODIPY-TR-14-dUTP, Cascade Blue-7-UTP, Cascade Blue-7-dUTP, fluorescein-12-UTP, fluorescein-12-dUTP, Oregon Green 488-5-dUTP, Rhodamine Green-5-UTP, Rhodamine Green-5-dUTP, tetramethylrhodamine-6-UTP, tetramethylrhodamine-6-dUTP, Texas Red-5-UTP, Texas Red-5-dUTP, and Texas Red-12-dUTP available from Molecular Probes, Eugene, Oreg. Nucleotides may also be labeled or marked by chemical modification. A chemically-modified single nucleotide can be biotin-dNTP. Some non-limiting examples of biotinylated dNTPs can include, biotin-dATP (e.g., bio-N6-ddATP, biotin-14-dATP), biotin-dCTP (e.g., biotin-11-cICTP, biotin-14-dCTP), and biotin-dUTP (e.g. biotin-11-dUTP, biotin-1.6-dUTP, biotin-20-dUTP).

[0288] The terms "polynucleotide," "oligonucleotide," and "nucleic acid" are used interchangeably to refer to a polymeric form of nucleotides of any length, either deoxyribonucleotides or ribonucleotides, or analogs thereof, either in single-, double-, or multi-stranded form. A polynucleotide may be exogenous or endogenous to a cell. A polynucleotide may exist in a cell-free environment. A polynucleotide may be a gene or fragment thereof. A polynucleotide may be DNA. A polynucleotide may be RNA. A polynucleotide may be a mix of RNA and DNA. A polynucleotide may have any three-dimensional structure, and may perform any function, known or unknown. A polynucleotide may comprise one or more analogs (e.g., altered backbone, sugar, or nucleobase). If present, modifications to the nucleotide structure may be imparted before or after assembly of the polymer. Some non-limiting examples of modified nucleotides or analogs include: pseudouridine, 5-bromouracil, 5-methylcytosine, peptide nucleic acid, xeno nucleic acid, morpholinos, locked nucleic acids, glycol nucleic acids, threose nucleic acids, dideoxynucleotides, cordycepin, 7-deaza-GTP,

fluorophores (e.g., rhodamine or fluorescein linked to the sugar), thiol containing nucleotides, biotin linked nucleotides, fluorescent base analogs, CpG islands, methyl-7-guanosine, methylated nucleotides, inosine, thiouridine, pseudouridine, dihydrouridine, queuosine, and wyosine. Non-limiting examples of polynucleotides include coding or non-coding regions of a gene or gene fragment, loci (locus) defined from linkage analysis, exons, introns, messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), short interfering RNA (siRNA), short-hairpin RNA (shRNA), micro-RNA (miRNA), ribozymes, eDNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, cell-free polynucleotides including cell-free DNA (cfDNA) and cell-free RNA (cfRNA), nucleic acid probes, and primers. The sequence of nucleotides may be interrupted by non-nucleotide components.

[0289] In some embodiments, the nucleic acid composition may comprise one or more mRNA. The mRNA may comprise at least one mRNA encoding a transmembrane receptor implicated in an immune response function (e.g., a phagocytic receptor or synthetic chimeric antigen receptor) into human macrophage or dendritic cell or a suitable myeloid cell or a myeloid precursor cell. In some embodiments, the nucleic acid composition comprises one or more mRNA, and one or more lipids for delivery of the nucleic acid into a cell of hematopoietic origin, such as a myeloid cell or a myeloid cell precursor cell. In some embodiments, the one or more lipids may form a liposomal complex.

[0290] Accordingly, it is an object of the present invention to provide novel retrotransposon-based compositions useful in providing gene therapy to an animal. It is an object of the present invention to provide novel retrotransposon-based compositions for use in the preparation of a medicament useful in providing gene therapy to an animal or human. It is an object of the present invention to provide novel transposon-based compositions that produce a specific integration of a payload (such as desirable nucleic acids) to a genome at and only at a specific target locus. It is another object of the present invention to provide novel transposon-based compositions that encode for the production of desired proteins or peptides in cells. Yet another object of the present invention to provide novel transposon-based compositions that encode for the production of desired nucleic acids in cells. It is a further object of the present invention to provide methods for cell and tissue specific incorporation of transposon-based DNA or RNA constructs comprising targeting a selected gene to a specific cell or tissue of an animal. It is yet another object of the present invention to provide methods for cell and tissue specific expression of transposon-based DNA or RNA constructs comprising designing a DNA or RNA construct with cell specific promoters that enhance stable incorporation of the selected gene by the transposase and expressing the selected gene in the cell. It is an object of the present invention to provide gene therapy for generations through germ line administration of a transposon-based composition. Another object of the present invention is to provide gene therapy in animals through non germ line administration of a transposon-based compositions. Another object of the present invention is to provide gene therapy in animals through administration of a transposon-based

compositions, wherein the animals produce desired proteins, peptides or nucleic acids. Yet another object of the present invention is to provide gene therapy in animals through administration of a transposon-based compositions, wherein the animals produce desired proteins or peptides that are recognized by receptors on target cells. Still another object of the present invention is to provide gene therapy in animals through administration of a transposon-based composition, wherein the animals produce desired fusion proteins or fusion peptides, a portion of which are recognized by receptors on target cells, in order to deliver the other protein or peptide component of the fusion protein or fusion peptide to the cell to induce a biological response. Yet another object of the present invention is to provide a method for gene therapy of animals through administration of transposon-based compositions comprising tissue specific promoters and a gene of interest to facilitate tissue specific incorporation and expression of a gene of interest to produce a desired protein, peptide or nucleic acid. Another object of the present invention is to provide a method for gene therapy of animals through administration of transposon-based compositions comprising cell specific promoters and a gene of interest to facilitate cell specific incorporation and expression of a gene of interest to produce a desired protein, peptide or nucleic acid. Still another object of the present invention is to provide a method for gene therapy of animals through administration of transposon-based compositions comprising cell specific promoters and a gene of interest to facilitate cell specific incorporation and expression of a gene of interest to produce a desired protein, peptide or nucleic acid, wherein the desired protein, peptide or nucleic acid has a desired biological effect in the animal.

[0291] As used herein, the composition described herein may be used for delivery inside a cell. A cell may originate from any organism having one or more cells. Some non-limiting examples include: a prokaryotic cell, eukaryotic cell, a bacterial cell, an archaeal cell, a cell of a single-cell eukaryotic organism, a protozoa cell, a cell from a plant (e.g., cells from plant crops, fruits, vegetables, grains, soy bean, corn, maize, wheat, seeds, tomatoes, rice, cassava, sugarcane, pumpkin, hay, potatoes, cotton, cannabis, tobacco, flowering plants, conifers, gymnosperms, ferns, clubmosses, hornworts, liverworts, mosses), an algal cell, (e.g., *Botryococcus braunii*, *Chlamydomonas reinhardtii*, *Nannochloropsis gaditana*, *Chlorella pyrenoidosa*, *Sargassum patens*, *C. agardh*, and the like), seaweeds (e.g. kelp), a fungal cell (e.g., a yeast cell, a cell from a mushroom), an animal cell, a cell from an invertebrate animal (e.g. fruit fly, cnidarian, echinoderm, nematode, etc.), a cell from a vertebrate animal (e.g., fish, amphibian, reptile, bird, mammal), a cell from a mammal (e.g., a pig, a cow, a goat, a sheep, a rodent, a rat, a mouse, a non-human primate, a human, etc.), and etcetera. Sometimes a cell may not be originating from a natural organism (e.g., a cell may be a synthetically made, sometimes termed an artificial cell). In some embodiments, the cell referred to herein is a mammalian cell. In some embodiments, the cell is a human cell. The methods and compositions described herein relates to incorporating a genetic material in a cell, more specifically a human cell, wherein the human cell can be any human cell. As used herein, a human cell may be of any origin, for

example, a somatic cell, a neuron, a fibroblast, a muscle cell, an epithelial cell, a cardiac cell, or a hematopoietic cell. The methods and compositions described herein can also be applicable to and useful for incorporating exogenous nucleic acid in hard-to-transfect human cell. The methods are simple and universally applicable once a suitable exogenous nucleic acid construct has been designed and developed. The methods and compositions described herein are applicable to incorporate an exogenous nucleic acid in a cell *ex vivo*. In some embodiments, the compositions may be applicable for systemic administration in an organism, where the nucleic acid material in the composition may be taken up by a cell *in vivo*, whereupon it is incorporated in cell *in vivo*.

[0292] In some embodiments, the methods and compositions described herein may be directed to incorporating an exogenous nucleic acid in a human hematopoietic cell, for example, a human cell of hematopoietic origin, such as a human myeloid cell or a myeloid cell precursor. However, the methods and compositions described herein can be used or made suitable for use in any biological cell with minimum modifications. Therefore, a cell may refer to any cell that is a basic structural, functional and/or biological unit of a living organism.

[0293] In one aspect, provided herein are methods and compositions for utilizing transposable elements for stable incorporation of one or more nucleic acids into the genome of a cell, where the cell is a member of a hematopoietic cells, for example a myeloid cell. In some embodiments, the one or more nucleic acids comprise at least one nucleic acid sequence encoding a transmembrane receptor protein having a role in immune response. In some embodiments, the methods and compositions are directed to using a retrotransposable element for incorporating one or more nucleic acid sequences into a myeloid cell. The nucleic acid composition may comprise one or more nucleic sequences, such as a gene, where the gene is a transgene. The term "gene," as used herein, refers to a nucleic acid (e.g., DNA such as genomic DNA and cDNA) and its corresponding nucleotide sequence that is involved in encoding an RNA transcript. The term as used herein with reference to genomic DNA includes intervening, non-coding regions as well as regulatory regions and may include 5' and 3' ends. In some uses, the term encompasses the transcribed sequences, including 5' and 3' untranslated regions (5'-UTR and 3'-UTR), exons and introns. In some genes, the transcribed region will contain "open reading frames" that encode polypeptides. In some uses of the term, a "gene" comprises only the coding sequences (e.g., an "open reading frame" or "coding region") necessary for encoding a polypeptide. In some cases, genes do not encode a polypeptide, for example, ribosomal RNA genes (rRNA) and transfer RNA (tRNA) genes. In some cases, the term "gene" includes not only the transcribed sequences, but in addition, also includes non-transcribed regions including upstream and downstream regulatory regions, enhancers and promoters. A gene may refer to an "endogenous gene" or a native gene in its natural location in the genome of an organism. A gene may refer to an "exogenous gene" or a non-native gene. A non-native gene may refer to a gene not normally found in the host organism, but which is introduced into the host organism by gene transfer. A non-native gene may also refer to

a gene not in its natural location in the genome of an organism. A non-native gene may also refer to a naturally occurring nucleic acid or polypeptide sequence that comprises mutations, insertions and/or deletions (e.g., non-native sequence). In some uses, an exogenous polynucleic acid encoding a polypeptide may comprise an exogenous gene or part thereof. An exogenous polynucleic acid sequence may represent a sequence that is not originally present in the immediate surrounding with respect to the position in the genome where it is inserted, e.g., via retrotransposition (alternatively termed in some cases as editing). In some uses, the exogenous polynucleic acid encoding the therapeutic polypeptide, or the transgene may also be referred to as the "payload". For example, for simplicity of experimental review of the system, a sequence encoding green fluorescence protein, GFP, has been used in the examples as surrogate payload for demonstration of feasibility and effectivity of the designing of the system.

[0294] The term "transgene" refers to any nucleic acid molecule that is introduced into a cell, that may be intermittently termed herein as a recipient cell. The resultant cell after receiving a transgene may be referred to a transgenic cell. A transgene may include a gene that is partly or entirely heterologous (i.e., foreign) to the transgenic organism or cell, or may represent a gene homologous to an endogenous gene of the organism or cell. In some cases, transgenes include any polynucleotide, such as a gene that encodes a polypeptide or protein, a polynucleotide that is transcribed into an inhibitory polynucleotide, or a polynucleotide that is not transcribed (e.g., lacks an expression control element, such as a promoter that drives transcription). Transcripts and encoded polypeptides may be collectively referred to as "gene product." If the polynucleotide is derived from genomic DNA, expression may include splicing of the mRNA in a eukaryotic cell. "Up-regulated," with reference to expression, refers to an increased expression level of a polynucleotide (e.g., RNA such as mRNA) and/or polypeptide sequence relative to its expression level in a wild-type state while "down-regulated" refers to a decreased expression level of a polynucleotide (e.g., RNA such as mRNA) and/or polypeptide sequence relative to its expression in a wild-type state. Expression of a transfected gene may occur transiently or stably in a cell. During "transient expression" the transfected gene is not transferred to the daughter cell during cell division. Since its expression is restricted to the transfected cell, expression of the gene is lost over time. In contrast, stable expression of a transfected gene may occur when the gene is co-transfected with another gene that confers a selection advantage to the transfected cell. Such a selection advantage may be a resistance towards a certain toxin that is presented to the cell. Where a transfected gene is required to be expressed, the application envisages the use of codon-optimized sequences. An example of a codon optimized sequence may be a sequence optimized for expression in a eukaryote, e.g., humans (i.e., being optimized for expression in humans), or for another eukaryote, animal or mammal. Codon optimization for a host species other than human, or for codon optimization for specific organs is known. In some embodiments, the coding sequence encoding a protein may be codon optimized for expression in particular cells, such as eukaryotic cells. The eukaryotic cells may be those

of or derived from a particular organism, such as a plant or a mammal, including but not limited to human, or non-human eukaryote or animal or mammal as herein discussed, e.g., mouse, rat, rabbit, dog, livestock, or non-human mammal or primate. Codon optimization refers to a process of modifying a nucleic acid sequence for enhanced expression in the host cells of interest by replacing at least one codon (e.g., about or more than about 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, or more codons) of the native sequence with codons that are more frequently or most frequently used in the genes of that host cell while maintaining the native amino acid sequence. Various species exhibit particular bias for certain codons of a particular amino acid. Codon bias (differences in codon usage between organisms) often correlates with the efficiency of translation of messenger RNA (mRNA), which is in turn believed to be dependent on, among other things, the properties of the codons being translated and the availability of particular transfer RNA (tRNA) molecules. The predominance of selected tRNAs in a cell may generally reflect the codons used most frequently in peptide synthesis. Accordingly, genes may be tailored for optimal gene expression in a given organism based on codon optimization. Codon usage tables are readily available, for example, at the "Codon Usage Database" available at www.kazusa.or.jp/codon/ and these tables may be adapted in a number of ways. Computer algorithms for codon optimizing a particular sequence for expression in a particular host cell are also available, such as Gene Forge (Aptagen; Jacobus, PA), are also available.

[0295] A "multicistronic transcript" as used herein refers to an mRNA molecule that contains more than one protein coding region, or cistron. A mRNA comprising two coding regions is denoted a "bicistronic transcript." The "5'-proximal" coding region or cistron is the coding region whose translation initiation codon (usually AUG) is closest to the 5' end of a multicistronic mRNA molecule. A "5'-distal" coding region or cistron is one whose translation initiation codon (usually AUG) is not the closest initiation codon to the 5' end of the mRNA.

[0296] The terms "transfection" or "transfected" refer to introduction of a nucleic acid into a cell by non-viral or viral-based methods. The nucleic acid molecules may be gene sequences encoding complete proteins or functional portions thereof. See, e.g., Sambrook et al., 1989, *Molecular Cloning: A Laboratory Manual*, 18.1-18.88.

[0297] The term "promoter," as used herein, refers to a polynucleotide sequence capable of driving transcription of a coding sequence in a cell. Thus, promoters used in the polynucleotide constructs of the disclosure include cis-acting transcriptional control elements and regulatory sequences that are involved in regulating or modulating the timing and/or rate of transcription of a gene. For example, a promoter may be a cis-acting transcriptional control element, including an enhancer, a promoter, a transcription terminator, an origin of replication, a chromosomal integration sequence, 5' and 3' untranslated regions, or an intronic sequence, which are involved in transcriptional regulation. These cis-acting sequences typically interact with proteins or other biomolecules to carry out (turn on/off, regulate, modulate, etc.) gene transcription. A "constitutive promoter" is one that is capable of initiating transcription in nearly all tissue types, whereas

a "tissue-specific promoter" initiates transcription only in one or a few particular tissue types. An "inducible promoter" is one that initiates transcription only under particular environmental conditions, developmental conditions, or drug or chemical conditions. Exemplary inducible promoter may be a doxycycline or a tetracycline inducible promoter. Tetracycline regulated promoters may be both tetracycline inducible or tetracycline repressible, called the tet-on and tet-off systems. The *tet* regulated systems rely on two components, *i.e.*, a tetracycline-controlled regulator (also referred to as transactivator) (tTA or rtTA) and a tTA/rtTA-dependent promoter that controls expression of a downstream cDNA, in a tetracycline-dependent manner. tTA is a fusion protein containing the repressor of the Tn10 tetracycline-resistance operon of *Escherichia coli* and a carboxyl-terminal portion of protein 16 of herpes simplex virus (VP16). The tTA-dependent promoter consists of a minimal RNA polymerase II promoter fused to *tet* operator (*tetO*) sequences (an array of seven cognate operator sequences). This fusion converts the *tet* repressor into a strong transcriptional activator in eukaryotic cells. In the absence of tetracycline or its derivatives (such as doxycycline), tTA binds to the *tetO* sequences, allowing transcriptional activation of the tTA-dependent promoter. However, in the presence of doxycycline, tTA cannot interact with its target and transcription does not occur. The *tet* system that uses tTA is termed *tet*-OFF, because tetracycline or doxycycline allows transcriptional down-regulation. In contrast, in the tet-ON system, a mutant form of tTA, termed rtTA, has been isolated using random mutagenesis. In contrast to tTA, rtTA is not functional in the absence of doxycycline but requires the presence of the ligand for transactivation. The term "exon" refers to a nucleic acid sequence found in genomic DNA that is bioinformatically predicted and/or experimentally confirmed to contribute contiguous sequence to a mature mRNA transcript. The term "intron" refers to a sequence present in genomic DNA that is bioinformatically predicted and/or experimentally confirmed to not encode part of or all of an expressed protein, and which, in endogenous conditions, is transcribed into RNA (e.g. pre-mRNA) molecules, but which is spliced out of the endogenous RNA (e.g. the pre-mRNA) before the RNA is translated into a protein.

[0298] The term "splice acceptor site" refers to a sequence present in genomic DNA that is bioinformatically predicted and/or experimentally confirmed to be the acceptor site during splicing of pre-mRNA, which may include identified and unidentified natural and artificially derived or derivable splice acceptor sites.

[0299] An "internal ribosome entry site" or "IRES" refers to a nucleotide sequence that allows for 5' -end/cap-independent initiation of translation and thereby raises the possibility to express 2 proteins from a single messenger RNA (mRNA) molecule. IRESs are commonly located in the 5' UTR of positive-stranded RNA viruses with uncapped genomes. Another means to express 2 proteins from a single mRNA molecule is by insertion of a 2A peptide(-like) sequence in between their coding sequence. 2A peptide(-like) sequences mediate self-processing of primary translation products by a process variously referred to as "ribosome skipping", "stop-go" translation and "stop carry-on"

translation. 2A peptide(-like) sequences are present in various groups of positive- and double-stranded RNA viruses including Picornaviridae, Flaviviridae, Tetraviridae, Dicistroviridae, Reoviridae and Totiviridae.

[0300] The term "2A peptide" refers to a class of 18–22 amino-acid (AA)-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 (thossea asigna virus 2A) were also identified. The mechanism of 2A-mediated "self-cleavage" is believed to be ribosome skipping the formation of a glycyl-prolyl peptide bond at the C-terminus of the 2A sequence. 2A peptide(-like) sequences mediate self-processing of primary translation products by a process variously referred to as "ribosome skipping", "stop-go" translation and "stop carry-on" translation. 2A peptide(-like) sequences are present in various groups of positive- and double-stranded RNA viruses including Picornaviridae, Flaviviridae, Tetraviridae, Dicistroviridae, Reoviridae and Totiviridae.

[0301] As used herein, the term "operably linked" refers to a functional relationship between two or more segments, such as nucleic acid segments or polypeptide segments. Typically, it refers to the functional relationship of a transcriptional regulatory sequence to a transcribed sequence.

[0302] As used herein, the term "flank" refers to the relationship between three sequence elements, wherein a flanked sequence element is upstream from a first flanking sequence element and downstream from a second flanking sequence. The distance between the first or second flanking sequence elements and the flanked sequence can be at most 100 nucleotides, at most 75 nucleotides, at most 50 nucleotides, at most 25 nucleotides, at most 10 nucleotides, at most 5 nucleotides, at most 1 nucleotide.

[0303] The term "termination sequence" refers to a nucleic acid sequence which is recognized by the polymerase of a host cell and results in the termination of transcription. The termination sequence is a sequence of DNA that, at the 3' end of a natural or synthetic gene, provides for termination of mRNA transcription or both mRNA transcription and ribosomal translation of an upstream open reading frame. Prokaryotic termination sequences commonly comprise a GC-rich region that has a two-fold symmetry followed by an AT-rich sequence. A commonly used termination sequence is the T7 termination sequence. A variety of termination sequences are known in the art and may be employed in the nucleic acid constructs of the present invention, including the TINT3, TL13, TL2, TR1, TR2, and T6S termination signals derived from the bacteriophage lambda, and termination signals derived from bacterial genes, such as the trp gene of *E. coli*.

[0304] The terms "polyadenylation sequence" (also referred to as a "poly A site" or "poly A sequence") refers to a DNA sequence which directs both the termination and polyadenylation of the nascent RNA transcript. Efficient polyadenylation of the recombinant transcript is desirable, as

transcripts lacking a poly A tail are typically unstable and rapidly degraded. The poly A signal utilized in an expression vector may be "heterologous" or "endogenous". An endogenous poly A signal is one that is found naturally at the 3' end of the coding region of a given gene in the genome. A heterologous poly A signal is one which is isolated from one gene and placed 3' of another gene, e.g., coding sequence for a protein. A commonly used heterologous poly A signal is the SV40 poly A signal. The SV40 poly A signal is contained on a 237 bp BamHI/BclI restriction fragment and directs both termination and polyadenylation; numerous vectors contain the SV40 poly A signal. Another commonly used heterologous poly A signal is derived from the bovine growth hormone (BGH) gene; the BGH poly A signal is also available on a number of commercially available vectors. The poly A signal from the Herpes simplex virus thymidine kinase (HSV tk) gene is also used as a poly A signal on a number of commercial expression vectors. The polyadenylation signal facilitates the transportation of the RNA from within the cell nucleus into the cytosol as well as increases cellular half-life of such an RNA. The polyadenylation signal is present at the 3'-end of an mRNA.

[0305] The terms "complement," "complements," "complementary," and "complementarity," as used herein, refer to a sequence that is complementary to and hybridizable to the given sequence. In some cases, a sequence hybridized with a given nucleic acid is referred to as the "complement" of the given molecule if its sequence of bases over a given region is capable of complementarily binding those of its binding partner, such that, for example, A-T, A-U, G-C, and G-U base pairs are formed. In general, a first sequence that is hybridizable to a second sequence is specifically or selectively hybridizable to the second sequence, such that hybridization to the second sequence or set of second sequences is preferred (e.g., thermodynamically more stable under a given set of conditions, such as stringent conditions commonly used in the art) to hybridization with non-target sequences during a hybridization reaction. Typically, hybridizable sequences share a degree of sequence complementarity over all or a portion of their respective lengths, such as between 25%-100% complementarity, including at least 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, and 100% sequence complementarity. Sequence identity, such as for the purpose of assessing percent complementarity, may be measured by any suitable alignment algorithm, including but not limited to the Needleman-Wunsch algorithm (see e.g. the EMBOSS Needle aligner available at https://www.ebi.ac.uk/Tools/psa/emboss_needle/), the BLAST algorithm (see e.g. the BLAST alignment tool available at blast.ncbi.nlm.nih.gov/Blast.cgi, optionally with default settings), or the Smith-Waterman algorithm (see e.g. the EMBOSS Water aligner available at www.ebi.ac.uk/Tools/psa/emboss_water/nucleotide.html, optionally with default settings). In some cases, the sequence identity is measured by the EMBOSS Needle aligner with default settings. Optimal alignment can be assessed using any suitable parameters of a chosen algorithm, including default parameters.

[0306] Complementarity may be perfect or substantial/sufficient. Perfect complementarity between two nucleic acids may mean that the two nucleic acids may form a duplex in which every base in the duplex is bonded to a complementary base by Watson-Crick pairing. Substantial or sufficient complementary may mean that, a sequence in one strand is not completely and/or perfectly complementary to a sequence in an opposing strand, but that sufficient bonding occurs between bases on the two strands to form a stable hybrid complex in set of hybridization conditions (e.g., salt concentration and temperature). Such conditions may be predicted by using the sequences and standard mathematical calculations to predict the melting temperature (T_m) of hybridized strands, or by empirical determination of T_m by using routine methods.

[0307] "Transposons" as used herein are segments within the chromosome that can translocate within the genome, also known as "jumping gene". There are two different classes of transposons: class I, or retrotransposons, that mobilize via an RNA intermediate and a "copy-and-paste" mechanism, and class II, or DNA transposons, that mobilize via excision integration, or a "cut-and-paste" mechanism (Ivics Nat Methods 2009). Bacterial, lower eukaryotic (e.g. yeast) and invertebrate transposons appear to be largely species specific, and cannot be used for efficient transposition of DNA in vertebrate cells. "Sleeping Beauty" (Ivics Cell 1997), was the first active transposon that was artificially reconstructed by sequence shuffling of inactive TEs from fish. This made it possible to successfully achieve DNA integration by transposition into vertebrate cells, including human cells. Sleeping Beauty is a class II DNA transposon belonging to the Tc1/mariner family of transposons (Ni Genomics Proteomics 2008). In the meantime, additional functional transposons have been identified or reconstructed from different species, including *Drosophila*, frog and even human genomes, that all have been shown to allow DNA transposition into vertebrate and also human host cell genomes. Each of these transposons have advantages and disadvantages that are related to transposition efficiency, stability of expression, genetic payload capacity etc. Exemplary class II transposases that have been created include Sleeping Beauty, PiggyBac, Frog Prince, Himarl, Passport, Minos, hAT, Toll, To12, AciDs, PIF, Harbinger, Harbinger3- DR, and Hsmarl.

[0308] "Heterologous" as used herein, includes molecules such as DNA and RNA which may not naturally be found in the cell into which it is inserted. For example, when mouse or bacterial DNA is inserted into the genome of a human cell, such DNA is referred to herein as heterologous DNA. In contrast, the term "homologous" as used herein, denotes molecules such as DNA and RNA that are found naturally in the cell into which it is inserted. For example, the insertion of mouse DNA into the genome of a mouse cell constitutes insertion of homologous DNA into that cell. In the latter case, it is not necessary that the homologous DNA be inserted into a site in the cell genome in which it is naturally found; rather, homologous DNA may be inserted at sites other than where it is naturally found, thereby creating a genetic alteration (a mutation) in the inserted site.

[0309] A "transposase" is an enzyme that is capable of forming a functional complex with a transposon end-containing composition (e.g., transposons, transposon ends), and catalyze insertion or

transposition of the transposon end-containing composition into double stranded DNA which is incubated with an *in vitro* transposon reaction. The term "transposon end" means a double-stranded DNA that contains the nucleotide sequences (the "transposon end sequences") necessary to form the complex with the transposase or integrase enzyme that is functional in an *in vitro* transposition reaction.

[0310] A transposon end forms a complex or a synaptic complex or a transposon complex or a transposon composition with a transposase or integrase that recognizes and binds to the transposon end, and which complex is capable of inserting or transposing the transposon end into target DNA with which it is incubated in an *in vitro* transposition reaction. A transposon end exhibits two complementary sequences consisting of a transferred transposon end sequence or transferred strand and a non-transferred transposon end sequence, or non-transferred strand. For example, one transposon end that forms a complex with a hyperactive Tn5 transposase that is active in an *in vitro* transposition reaction comprises a transferred strand that exhibits a transferred transposon end sequence as follows: 5' AGATGTGTATAAGAGACAG 3' (SEQ ID NO: 7), and a non-transferred strand that exhibits a "non-transferred transposon end sequence" as follows: 5' CTGTCTCTTATACACATCT 3' (SEQ ID NO: 8). The 3'-end of a transferred strand is joined or transferred to target DNA in an *in vitro* transposition reaction. The non-transferred strand, which exhibits a transposon end sequence that is complementary to the transferred transposon end sequence, is not joined or transferred to the target DNA in an *in vitro* transposition reaction.

[0311] In some embodiments, the transferred strand and non-transferred strand are covalently joined. For example, in some embodiments, the transferred and non-transferred strand sequences are provided on a single oligonucleotide, e.g., in a hairpin configuration. As such, although the free end of the non-transferred strand is not joined to the target DNA directly by the transposition reaction, the non-transferred strand becomes attached to the DNA fragment indirectly, because the non-transferred strand is linked to the transferred strand by the loop of the hairpin structure. As used herein an "cleavage domain" refers to a nucleic acid sequence that is susceptible to cleavage by an agent, e.g., an enzyme.

[0312] In some cases, an RNA molecule or a DNA molecule or, for example, an mRNA molecule refers to an independent molecule that may comprise a sequence, for example, a sequence encoding a polypeptide. A sequence, for example, a polynucleic acid (e.g., DNA, RNA) sequence, on the other hand refers to a stretch of polynucleotides within a polynucleic acid molecule, wherein the stretch comprises of 2 or more nucleotides. A sequence may also refer to a polypeptide sequence as is understood by the adjacent context by one of skill in the art – wherein the polypeptide sequence comprises a series of amino acids on a polypeptide molecule.

[0313] A "restriction site domain" usually means a tag domain that exhibits a sequence for the purpose of facilitating cleavage using a restriction endonuclease. For example, in some embodiments,

the restriction site domain is used to generate di-tagged linear ssDNA fragments. In some embodiments, the restriction site domain is used to generate a compatible double-stranded 5'-end in the tag domain so that this end can be ligated to another DNA molecule using a template-dependent DNA ligase. In some embodiments, the restriction site domain in the tag exhibits the sequence of a restriction site that is present only rarely, if at all, in the target DNA (e.g., a restriction site for a rare-cutting restriction endonuclease such as NotI or AscI).

[0314] As used herein, the term "recombinant nucleic acid molecule" may refer to a recombinant DNA molecule or a recombinant RNA molecule. A recombinant nucleic acid molecule is any nucleic acid molecule containing joined nucleic acid molecules from different original sources and not naturally attached together. Recombinant RNA molecules may include RNA molecules transcribed from recombinant DNA molecules. A recombinant nucleic acid may be synthesized in the laboratory. A recombinant nucleic acid can be prepared by using recombinant DNA technology by using enzymatic modification of DNA, such as enzymatic restriction digestion, ligation, and DNA cloning. A recombinant DNA may be transcribed *in vitro*, to generate a messenger RNA (mRNA), the recombinant mRNA may be isolated, purified and used to transfect a cell. A recombinant nucleic acid may encode a protein or a polypeptide. A recombinant nucleic acid, under suitable conditions, can be incorporated into a living cell, and can be expressed inside the living cell. As used herein, "expression" of a nucleic acid may usually refer to transcription and/or translation of the nucleic acid. The product of a nucleic acid expression is usually a protein but can also be an mRNA. Detection of an mRNA encoded by a recombinant nucleic acid in a cell that has incorporated the recombinant nucleic acid, is considered positive proof that the nucleic acid is "expressed" in the cell. The process of inserting or incorporating a nucleic acid into a cell can be via transformation, transfection or transduction. In some cases, transformation relates the process of uptake of foreign nucleic acid by a bacterial cell. This process is adapted for propagation of plasmid DNA, protein production, and other applications. Transformation may relate to introduction of recombinant plasmid DNA into competent bacterial cells that take up extracellular DNA from the environment. Some bacterial species are naturally competent under certain environmental conditions, but competence is artificially induced in a laboratory setting. Transfection may refer to the forced introduction of small molecules such as DNA, RNA, or antibodies into eukaryotic cells. 'Transduction' is mostly used to describe the introduction of recombinant viral vector particles into target cells, while 'infection' refers to natural infections of humans or animals with wild-type viruses.

[0315] A "stem-loop" sequence may often refer to a nucleic acid sequence (e.g., RNA sequence) with sufficient self-complementarity to hybridize and form a stem and the regions of non-complementarity that bulges into a loop. The stem may comprise mismatches or bulges.

[0316] The term "vector" may refer to a nucleic acid molecule capable of transporting or mediating expression of a heterologous nucleic acid. A "vector sequence" as used herein, refers to a sequence of

nucleic acid comprising at least one origin of replication and at least one selectable marker gene. Vectors capable of directing the expression of genes and/or nucleic acid sequence to which they are operatively linked are referred to herein as "expression vectors".

[0317] A plasmid may be a species of the genus encompassed by the term "vector." In general, expression vectors of utility are often in the form of "plasmids" which refer to circular double stranded DNA molecules which, in their vector form are not bound to the chromosome, and typically comprise entities for stable or transient expression of the encoded DNA. Other expression vectors that can be used in the methods as disclosed herein include, but are not limited to plasmids, episomes, bacterial artificial chromosomes, yeast artificial chromosomes, bacteriophages or viral vectors, and such vectors can integrate into the host's genome or replicate autonomously in the cell. A vector can be a DNA or RNA vector. Other forms of expression vectors known by those skilled in the art which serve the equivalent functions can also be used, for example, self-replicating extrachromosomal vectors or vectors capable of integrating into a host genome. Exemplary vectors are those capable of autonomous replication and/or expression of nucleic acids to which they are linked. A safe harbor locus is a region within the genome where additional exogenous or heterologous nucleic acid sequence can be inserted, and the host genome is able to accommodate the inserted genetic material. Exemplary safe harbor sites include but are not limited to: AAVS1 site, GGTA1 site, CMAH site, B4GALNT2 site, B2M site, ROSA26 site, COLA1 site, and TIGRE site. For example, the heterologous nucleic acid described in this disclosure may be integrated at one or more sites in the genome of the cell, wherein the one or more locations is selected from the group consisting of: AAVS1 site, GGTA1 site, CMAH site, B4GALNT2 site, B2M site, ROSA26 site, COLA1 site, and TIGRE site. In some embodiments, the nucleic acid cargo comprising the transgene may be delivered to a R2D locus.

[0318] In some embodiments, the nucleic acid cargo comprising the transgene may be delivered to the genome in an intergenic or intragenic region. In some embodiments the nucleic acid cargo comprising the transgene is integrated into the genome 5' or 3' within 0.1 kb, 0.25 kb, 0.5 kb, 0.75, kb, 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 7.5 kb, 10 kb, 15 kb, 20 kb, 25 kb, 50, 75 kb, or 100 kb of an endogenous active gene. In some embodiments the nucleic acid cargo comprising the transgene is integrated into the genome 5' or 3' within 0.1 kb, 0.25 kb, 0.5 kb, 0.75, kb, 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 7.5 kb, 10 kb, 15 kb, 20 kb, 25 kb, 50 kb, 75 kb, or 100 kb of an endogenous promoter or enhancer. In some embodiments the nucleic acid cargo comprising the transgene is 50-50,000 base pairs, e.g., between 50-40,000 bp, between 500-30,000 bp between 500-20,000 bp, between 100-15,000 bp, between 500-10,000 bp, between 50-10,000 bp, between 50-5,000 bp. In some embodiments the nucleic acid cargo comprising the transgene is less than 1,000, 1,300, 1500, 2,000, 3,000, 4,000, 5,000, or 7,500 nucleotides in length.

L1 and Non-L1 Retrotransposon Systems

[0319] Retrotransposons can contain transposable elements that are active participants in reorganizing their resident genomes. Broadly, retrotransposons can refer to DNA sequences that are transcribed into RNA and translated into protein and have the ability to reverse-transcribe themselves back into DNA. Approximately 45% of the human genome is comprised of sequences that result from transposition events. Retrotransposition occasionally generates target site deletions or adds non-retrotransposon DNA to the genome by processes termed 5'- and 3'-transduction. Recombination between non-homologous retrotransposons causes deletions, duplications or rearrangements of gene sequence. Ongoing retrotransposition can generate novel splice sites, polyadenylation signals and promoters, and so builds new transcription modules.

[0320] Generally, retrotransposons may be grouped into two classes, the retrovirus-like LTR retrotransposons, and the non-LTR elements such as human L1 elements, *Neurospora* TAD elements (Kinsey, 1990, Genetics 126:317-326), I factors from *Drosophila* (Bucheton et al., 1984, Cell 38:153-163), and R2Bm from *Bombyx mori* (Luan et al., 1993, Cell 72: 595-605). These two types of retrotransposons are structurally different and also retrotranspose using radically different mechanisms. Exemplary, non-limiting examples of LINE-encoded polypeptides are found in GenBank Accession Nos. AAC51261, AAC51262, AAC51263, AAC51264, AAC51265, AAC51266, AAC51267, AAC51268, AAC51269, AAC51270, AAC51271, AAC51272, AAC51273, AAC51274, AAC51275, AAC51276, AAC51277, AAC51278 and AAC51279.

[0321] The decision to focus on LINE-1 to develop into a system as described in the disclosure for a number of reasons at least some of which are exemplified below: (a) LINE-1 (or L1-) elements are autonomous as they encode all of the machinery alone to complete this reverse transcription and integration process; (b) L1 elements are abundant in the human genome, such that these elements may be considered as a naturalized element of the genome; (c) L1 retrotransposon retrotransposes its own mRNA with high degree of specificity, compared to other mRNAs floating around in the cells.

[0322] The L1 expresses a 6-kb bicistronic RNA that encodes the 40 kDa Open Reading Frame-1 RNA-binding protein (ORF1p) of essential but uncertain function, and a 150 kDa ORF2 protein with endonuclease and reverse transcriptase (RT) activities. L1 retrotransposition is a complex process involving transcription of the L1, transport of its RNA to the cytoplasm, translation of the bicistronic RNA, formation of a ribonucleoprotein (RNP) particle, its re-import to the nucleus and target-primed reverse transcription at the integration site. A few transcription factors that interact with L1s have been identified. Transcribed L1 RNA forms an RNP *in cis* with the proteins that are translated from the transcript. L1 integrates into genomic DNA by target-site primer reverse transcription (TPRT) by ORF2p cleavage at the 5'-TTTT-3' where a poly A sequence of L1 RNA anneals and primes reverse transcriptase (RT) activity to make L1 cDNA.

[0323] Other mobile elements of the genome can "hijack" the L1 ORF for retrotransposition. For example, *Alu* elements are such mobile DNA elements that belong to the class of short interspersed elements (SINEs) that are non-autonomous retrotransposons and acquire trans-factors to integrate. *Alu* elements and SINE-1 elements can associate with the L1 ribonucleoproteins in trans to be also retrotransposed by ORF1p and ORF2p. Somewhat similar to the L1 RNA, the *Alu* element ends with a long A-run, often referred to as the A-tail, and it also has a smaller A-rich region (indicated by AA) separating the two halves of a diverged dimer structure. *Alu* elements are likely to have the internal components of an RNA polymerase III promoter (such as, commonly designated as an A box and a B box promoters), but they do not encode a terminator for RNA polymerase III. They may utilize a stretch of T nucleotides at various distances downstream of the *Alu* element to terminate a transcription. A typical *Alu* transcript encompasses the entire *Alu*, including the A-tail, and has a 3' region that is unique for each locus. The *Alu* RNA folds into separate structures for each monomer unit. The RNA has been shown to bind the 7SL RNA SRP9 and 14 heterodimer, as well as poly A-binding protein (PABP). The poly A tail of *Alu* primes with T rich (TTTT) region of the genome and attracts ORF2p to bind to the primed region and cleaves at the T rich region via its endonuclease activity. The T-rich region primes reverse transcription by ORF2p on the 3' A-tail region of the *Alu* element. This creates a cDNA copy of the body of the *Alu* element. A nick occurs by an unknown mechanism on the second strand and second-strand synthesis is primed. The new *Alu* element is then flanked by short direct repeats that are duplicates of the DNA sequence between the first and second nicks. *Alu* elements are extremely prevalent within RNA molecules, owing to their preference for gene-rich regions. A full-length *Alu* (~ 300 bp) is derived from the signal recognition particle RNA 7SL and consists of two similar monomers with an A-rich linker in-between, A- and B-boxes present in the 5' monomer, and a poly-A tail lacking the preceding polyadenylation signal resulting in an elongated tail (up to 100 bp in length). *Alu* can be transcribed by RNA polymerase III using the internal promoters within the A- and B-boxes; however, *Alu* contain no ORFs and therefore do not encode for protein products.

[0324] Other non-L1 transposons include SVAs and HERV-Ks. A full-length SVA (SINE-VNTR-*Alu*) element (~ 2–3 kb) is a composite unit that contains a CCCTCT repeat, two *Alu*-like sequences, a VNTR, a SINE-R region with *env* (envelope) gene, the 3' LTR of HERV-K10, and a polyadenylation signal followed by a poly-A tail. It is most likely that SVAs are transcribed by RNA polymerase II, although it is unknown whether SVA elements carry an internal promoter.

[0325] A full-length HERV-K element (~ 9–10 kb) is comprised of ancient remnants of endogenous retroviral sequences and includes two flanking LTR regions surrounding three retroviral ORFs: (1) *gag* encoding the structural proteins of a retroviral capsid; (2) *pol-pro* encoding the enzymes: protease, RT, and integrase; and (3) *env* encoding proteins allowing for horizontal transfer. The LTR

of HERV-K contains an internal, bidirectional promoter that appears to be under the transcriptional control of RNA polymerase II.

[0326] L1 retrotransposition and RNA binding can take place at or near poly-A tail. The 3'-UTR plays a role in the recognition of stringent-type LINE RNA of ORF1 protein (ORF1p). Stringent-type LINEs can contain a stem-loop structure located at the end of the 3'UTR. Branched molecules consisting of junctions between transposon 3'-end cDNA and the target DNA, as well as specific positioning of L1 RNA within ORF2 protein (ORF2p), were detected during initial stages of L1 retrotransposition *in vitro*. Secondary or tertiary RNA structure shared by L1 and Alu are likely to be responsible for recognition by and binding of ORF2, possibly along with a poly-A tail. In some embodiments, the stem-loop structure located downstream of the poly-A sequence correlates with cleavage intensity.

[0327] Mechanisms for restricting or resolving L1 integration have also evolved for the sake of maintaining genetic integrity and stability of the genome. Non-homologous end-joining repair proteins, such as XRCC1, Ku70 and DNA-PK, have been implicated in resolution of the L1 integrate at the time of insertion. In addition, the cell has evolved a number of proteins that stand against unrestricted retrotransposition, including the APOBEC3 family of cytosine deaminases, adenosine deaminase ADAR1, chromatin-remodeling factors and members of the piRNA pathway for post-transcription gene silencing that functions in the male germ line.

Cas endonucleases

[0328] The compositions disclosed herein can comprise at least one endonuclease. In some cases, the endonuclease is a Cas endonuclease such as Cas9 endonuclease. Cas9 endonuclease is active when it forms a complex with two naturally occurring RNA species, the tracrRNA and the crRNA. The first 20 nucleotides of the crRNA sequence define the specificity of the nuclease, which occurs by complementary base pairing with the target sequence within genomic DNA. Once activated, the nuclease generates a double-strand break (DSB) at the target site. Cas9 uses two distinct active sites, RuvC and HNH, generating site-specific nicks on opposite DNA strands (Gasiunas et al. 2012; Jinek et al. 2012). By simply specifying the targeting sequence of the crRNA, one can direct the CRISPR-Cas9 system to the appropriate genomic target site. An additional requirement for Cas9-mediated genome cleavage is the presence of a short and conserved protospacer adjacent motif (PAM) flanking the genomic target site. Functionality in mammalian cells was rapidly demonstrated and the native bacterial system was further simplified into a two-component system, with the crRNA and tracrRNA fused together to form a single-guide RNA (sgRNA). A Cas9 protein may be at least 80% identical (e.g., at least 90% identical, at least 95% identical or at least 98% identical or at least 99% identical) to a wild type Cas9 protein, e.g., to the *Streptococcus pyogenes* Cas9 protein.

[0329] CRISPR-Cas9 systems promote genome editing by inducing a DSB at a target genomic locus, which is quickly acted upon by the cell's DNA repair machinery. The generated ends of DNA can be

religated by non-homologous end joining (NHEJ), a process known to be quite precise but which can also introduce indel mutations at the DSB site (Lieber 2010), especially when the nucleases are active in the cell for a prolonged period. Alternatively, regions of homology flanking the DSB can lead to a process known as microhomology mediated end joining (MMEJ), also introducing indel mutations at the target site. These indel mutations provide a means of disrupting protein coding or other functional DNA sequences. An additional pathway, homology directed repairs (HDR), can be adopted by the cell to repair the lesion accurately if homologous DNA sequences are present. The HDR repair pathway enables intentional replacement of endogenous genomic sequence with information on a homologous donor molecule, presented to the cell. These templates can be delivered as single stranded oligodeoxynucleotide (ssODN) harboring desired nucleotide changes or as conventional double-stranded DNA targeting constructs, in both cases with regions of homologous sequence flanking the sequence change.

[0330] Cas12a is an additional class 2 CRISPR effector, originally described as Cpf1 and now reclassified as Cas12a (Shmakov et al. 2017). The Cas12a protein is structurally distinct from Cas9 effectors and possesses interesting new traits. Firstly, Cas12a binds to its target sequence by its association with a single RNA species, without the requirement of an additional tracrRNA. Secondly, the Cas12a-crRNA complex recognizes a T-rich PAM sequence, lying 5' of its target sequence, in contrast to the G-rich PAM sequence of the Cas9 systems that lies 3' of its target. The T-rich PAM is of note as it might facilitate genome engineering in organisms with particularly AT-rich genomes. The PAM motif was initially reported as being TTTN; however, recent more in depth investigations in mammalian cells have revealed that target sequences with a TTTT PAM motif are inefficiently cleaved and this study redefined the Cas12a PAM as being TTTV. Thirdly, Cas12a cleaves DNA via a staggered DSB, leaving a 4 or 5-nt 5' overhang (Fonfara et al. 2016; Zetsche et al. 2015a), an attribute which might facilitate the introduction of specific sequences into the genome. Moreover, the staggered cleavage site of Cas12a from *Francisella novicida* U112 occurs after the 18th base on the non-targeted (+) strand and after the 23rd base on the targeted (–) strand, which is quite distant from the PAM sequence. This distal cleavage, far away from both the seed region and the PAM, might preserve the target sequence for subsequent rounds of cleavage. This might be useful for encouraging repair via HDR for the targeted integration of exogenous DNA, since indel mutations caused by the dominant NHEJ repair pathway would be less likely to destroy the target site. Cas12a, orthologues of which have been identified in several bacterial and archeal genomes, contains only a RuvC-like endonuclease domain and lacks the HNH domain present in Cas9 proteins. A putative novel nuclease domain, Nuc domain, has been ascertained from the crystal structure, which provides the 2nd nuclease activity responsible for cleaving the target strand. Furthermore, a ribonuclease activity has been ascribed to this enzyme and has a role in processing of the precursor CRISPR RNA.

[0331] Cas9 has also been fused to other programmable DNA binding domains, such as Zinc Finger binding domains and TALE domains, designed against sequences lying downstream of the Cas9 target site. By tethering the Cas9 in this sequence-specific manner, the requirement for the 3' NGG PAM was partially overcome and cleavage of alternative PAM sequences was achieved. The same study combined this tethering idea with Cas9 mutations that compromised the key PAM recognition residues (Arg1333 and Arg1335). These attenuated Cas9-programmable DNA-binding domain (Cas9-pDBD) nucleases showed improved precision as judged by deep sequencing of previously characterized off-target sites for 3 genomic target sites. An unbiased assessment of off-target mutagenesis also revealed no new sites are generated by the Cas9- pDBD fusions.

Nickase

[0332] In some cases, the Cas endonuclease is a Cas nickase. Usually a nickase is an enzyme that cleaves one nucleic acid strand. As used herein, the term "Cas nickase" may refer to a Cas nuclease that mediates cleavage of only a single strand of a defined nucleotide sequence. The Cas nickase can cut the displaced strand of the nucleic acid sequence to introduce a nick. Suitable Cas nickases include, for example, Cas9 nickase and Cpf1 nickase. Cas nickases can be found in nature or developed by modification of natural CRISPR nucleases. Cas nickases found in nature can be isolated and purified using a variety of protein isolation and purification techniques known to those skilled in the art. The Cas nickase can be produced recombinantly, or produced *in vitro* (i.e., synthesized and purified) using a variety of protein production techniques known to those skilled in the art. The Cas nickase can be provided as a protein or a polynucleotide encoding such (e.g., an mRNA transcript). The Cas nickase can alternatively be provided as a nucleotide vector capable of expressing the Cas nickase. The Cas nickase can alternatively be provided from a cell expressing the transgene encoding the Cas nickase. A second protein can be fused to a Cas9 nickase as a fusion protein. In some cases, the second protein is a ligase.

[0333] A Cas9 nickase can bind DNA based on gRNA specificity, though nickases will only cut one of the DNA strands (i.e., generating a "nick"). The majority of CRISPR plasmids currently being used are derived from *S. pyogenes* and the RuvC domain can be inactivated by an amino acid substitution at position D10 (e.g., D10A) and the HNH domain can be inactivated by an amino acid substitution at position H840 (e.g., H840A), or at positions corresponding to those amino acids in other proteins. As is known, the D10 and H840 variants of Cas9 cleave a Cas9-induced bubble at specific sites on opposite strands of the DNA within the bubble. Depending on which mutant is used, the guide RNA-hybridized strand or the non-hybridized strand may be cleaved. Thus, one CAS9 nickase (e.g., the D10 or H840 variant) can be used to create a 3' overhang and the other nickase can be used to create a 5' overhang at the same locus by cleaving the opposite strand of DNA.

[0334] In one aspect, the instant disclosure discloses a fusion of a Cas endonuclease or a family member of these proteins, with a retrotransposon machinery, in particular with a retrotransposon ORF2

reverse transcriptase (RT), providing the ORF2 RT a distinct programmable advantage of targeting and cutting the genome at a specific predetermined sequence and incorporating a foreign genetic material (e.g., the exogenous human therapeutic polypeptide) in the predetermined sequence of the genome. In other aspect, the instant disclosure discloses a composition comprising a separate Cas endonuclease or a family member of these proteins and a separate retrotransposon machinery, in particular with a retrotransposon ORF2 reverse transcriptase (RT). In some cases, the composition comprising a first polynucleotide encoding the Cas endonuclease or a family member of these proteins, and a second polynucleotide comprising a sequence encoding the retrotransposon machinery. In some cases, the retrotransposon machinery comprises the ORF2. In some cases, the ORF2 comprises RT. In some cases, the ORF2 comprises EN. In some cases, the ORF2 comprises both the RT and the EN. In some cases, the retrotransposon comprises ORF1 and ORF2.

[0335] It is the objective of the endeavor to insert large polynucleotide sequences (e.g., therapeutic genes, the exogenous human therapeutic polypeptide) into a targeted sequence of the genome in a programmable manner. While it may have been possible to base edit at specific genomic locations using variations of CRISPR/Cas system fused with other enzymes, insertion of large sequence has proven to be a huge challenge to overcome, especially with a delivery of polynucleotide in a single shot, minimum intervention procedure to a cell. It was also an objective of the endeavor to use a single mRNA, or a single mRNA with two guide RNA sequences for delivery of the gene editing system. One of the last but not the least hurdles remained that retrotransposition is toxic to the cells, specifically, primary human cells. For example, data not shown, flow cytometry analysis indicated that highly active retrotransposition resulted in cell toxicity. There was substantial loss of cell viability observed at day 10 post transfection of LINE1-GFP constructs, especially at the higher ORF1:ORF2 ratios in split constructs described elsewhere in the disclosure. Applicants had observed loss of GFP+ cells (GFP is a surrogate for an exogenous human therapeutic polypeptide for genomic insertion at testing stages), between day 7 and 10 in hepatoma cell line Huh7.

[0336] Presence of several LINE1 suppression factors also play a key role in low integration efficiency of the constructs. The instant disclosure provides methods and compositions that is a cumulative of several improvements over a basic organization of a retrotransposition machinery.

[0337] Hence, in one aspect, provided herein is a single mRNA molecule comprising a genome editing system for incorporating an exogenous sequence encoding a therapeutic polypeptide in a genome of a target cell, comprising: (i) a Cas endonuclease or a functional fragment thereof; (ii) a human LINE1 retrotransposon reverse transcriptase or a fragment thereof, and (iii)(a) the exogenous sequence encoding the therapeutic polypeptide flanked by one or more homology arms comprising non-ribosomal genomic sequence; or (iii)(b) a sequence comprising reverse complement of (iii)(a).

[0338] In some embodiments, the Cas endonuclease of the single mRNA molecule is a Cas 9 endonuclease. In some embodiments, the Cas endonuclease is a Cas 9 nickase. In some

embodiments, the exogenous sequence encoding the therapeutic polypeptide is greater than 1kb in length. In some embodiments, the exogenous sequence encoding the therapeutic polypeptide is greater than 1.2kb, 1.5kb, 1.7kb, 1.8 kb, 1.9kb, or 2 kb in length. In some embodiments, the exogenous sequence encoding the therapeutic polypeptide is greater than 2.1kb, 2.5kb, 2.7kb, 2.8 kb, 2.9kb, or 3 kb in length. In some embodiments, the exogenous sequence encoding the therapeutic polypeptide is greater than 3.1kb, 4kb, or 5 kb in length.

[0339] In some embodiments, the human LINE1 retrotransposon reverse transcriptase or a fragment thereof comprises a sequence encoding LINE1 ORF2p reverse transcriptase or a fragment thereof.

[0340] In some embodiments, the single mRNA molecule further comprising a sequence encoding LINE1 ORF2p endonuclease or a fragment thereof. In some embodiments, the Cas endonuclease the sequence encoding LINE1 ORF2p endonuclease or a fragment thereof, further comprising a mutation in the sequence encoding the LINE1 ORF2p endonuclease, wherein, when translated, the endonuclease activity of the endonuclease is reduced or absent.

[0341] Provided herein is a composition comprising the single mRNA molecule described above and a nucleic acid delivery vehicle. In some embodiments, the nucleic acid delivery vehicle comprises one or more lipids. In some embodiments, the composition further comprises one or more guide RNAs. In some embodiments, the composition further comprises one or more siRNAs.

[0342] In some embodiments the single mRNA is 6, 7, 8, 9, 10, 11, 12 or more kilobases in length.

Guide RNAs

[0343] The present disclosure provides a composition comprising one or more guide RNAs, e.g., at least two guide RNAs (gRNA) for guiding the Cas enzyme (e.g., Cas9 nickase) to generate two cuts (e.g., two nicks) on the target genomic sequence. In some embodiments, the two cuts are optionally on two opposite strands, and are located at staggered distances so as to avoid double stranded breaks.

[0344] As used herein, “guide RNA” or “gRNA” refers to any one of many natural or modified nucleic acid sequences comprising a 5' region termed the “spacer” sequence, which hybridizes to a target strand of double-stranded nucleic acid (such as, for example, a genomic DNA sequence) and a scaffold nucleic acid sequence termed as “tracr” sequence. The spacer sequence can range from about 17 nucleotides in length to about 30 nucleotides in length, including about 17 nucleotides to about 20 nucleotides, including about 18 nucleotides to about 25 nucleotides, and including about 21 nucleotides to about 24 nucleotides. The guide RNAs can be a single guide RNA (“sgRNA”), where the spacer sequence and scaffold are in a single molecule. In some cases, the spacer sequence and scaffold may be in two molecules linked by hybridization (or the spacer sequence and part of the scaffold in one molecule while the remaining scaffold in a second molecule). The guide RNA scaffold region (also referred to herein as “a scaffold nucleic acid sequence”) is recognized by and provides a

binding site for a CRISPR protein (e.g., a Cas protein, including a Cas nickase (e.g., a Cas9 nickase)). The scaffold nucleic acid sequence of the CRISPR guide RNA forms a complex with the Cas nickase to make a single-strand cut (i.e., a nick). CRISPR guide RNA(s) can be produced *in vitro* (i.e., synthesized and purified) using a variety of RNA production techniques known to those skilled in the art, such as *in vitro* transcription or chemical synthesis. CRISPR guide RNA(s) can also be expressed from DNA, e.g., after transfection of cells with a plasmid that expresses the guide RNA(s). As used herein, “nick” refers to a single-stranded cut (or break) in a double-stranded nucleic acid sequence. Some Cas nickases nick the target strand, which is recognized and hybridized by the CRISPR guide RNA, while other Cas nickases nick the strand that opposes the target strand. The strand opposite from the target strand is known as the displaced strand and contains a protospacer adjacent motif (PAM) site. The composition for genomic integration disclosed herein can comprises two guide RNAs, wherein the first gRNA comprises a first spacer sequence that is capable of hybridizing to the first target sequence of the genomic DNA, and the second gRNA comprises a second spacer sequence that is capable of hybridizing to a second target sequence of the genomic DNA. In some cases, the first gRNA guides the Cas9 nickase to the first target sequence and effects a first nick, and the second gRNA guides the Cas9 nickase to the second target sequence and effects a second nick. In some cases, the distance between the first nick and the second nick (i.e., the length of the nucleic acids between the first nick and the second nick) is at least 10 bases, at least 20 bases, at least 30 bases, at least 40 bases, at least 50 bases, at least 60 bases, at least 70 bases, at least 80 bases, at least 90 bases, at least 100 bases, at least 110 bases, at least 120 bases, at least 130 bases, at least 140 bases, at least 150 bases, at least 160 bases, at least 170 bases, at least 180 bases, at least 190 bases, at least 200 bases, at least 210 bases, at least 220 bases, at least 230 bases, at least 240 bases, at least 250 bases, at least 260 bases, at least 270 bases, at least 280 bases, at least 290 bases, or at least 300 bases. In some cases, the distance between the first nick and the second nick is at most 50 bases, at most 75 bases, at most 100 bases, at most 125 bases, at most 150 bases, at most 175 bases, at most 200 bases, at most 225 bases, at most 250 bases, at most 275 bases, at most 300 bases, at most 325 bases, at most 350 bases, at most 375 bases, at most 400 bases, at most 450 bases, at most 500 bases, at most 550 bases, at most 600 bases, at most 650 bases, at most 700 bases, at most 750 bases, at most 800 bases, at most 850 bases, at most 900 bases, at most 950 bases, or at most 1000 bases. In some cases, the distance between the first nick and the second nick is about 10-50, 10-100, 10-150, 10-200, 10-300, 10-400, 10-500, 10-600, 10-700, 10-800, 10-900, 10-1000, 20-50, 30-50, 40-50, 50-100, 50-150, 50-200, 50-300, 50-400, 50-500, 50-600, 50-700, 50-800, 50-900, 50-1000, 60-100, 70-100, 80-100, 90-100, 100-150, 100-200, 100-300, 100-400, 100-500, 100-600, 100-700, 100-800, 100-900, 100-1000, 110-150, 120-150, 130-150, 140-150, 150-200, 150-300, 150-400, 150-500, 150-600, 150-700, 150-800, 150-900, 150-1000, 150-160, 160-200, 170-200, 180-200, 190-200, 200-300, 300-400, 400-500, 500-600, 600-700, 700-800, 800-900, or 900-1000 bases. In some cases, the distance between the first nick and the second

nick is about 10, about 20, about 30, about 40, about 50, about 60, about 70, about 80, about 90, about 100, about 110, about 120, about 130, about 140, about 150, about 160, about 170, about 180, about 190, about 200, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600, about 650, about 700, about 750, about 800, about 850, about 900, about 950, or about 1000 bases.

[0345] In some cases, the two guide RNAs are two separate molecules. In some cases, the two guide RNAs are encoded by a single polynucleotide. In some cases, the two guide RNAs are encoded by two separate polynucleotides.

[0346] Design of guide RNAs for a particular target sequence in a genome is well known in the art. In some cases, the first target sequence and the second target sequence comprises sequences from the same region of the genome of the target cell. In some cases, the first target sequence and the second target sequence comprise sequences from the same locus of the genome of the target mammalian cell. In some cases, the locus is a genomic safe harbor locus as disclosed herein. In some cases, the locus is non-ribosomal DNA. The locus can be any of the locus disclosed herein, for example, the human ortholog of the mouse Rosa26 locus, adeno-associated virus site 1 (AAVS1), or CCR5 gene. In some cases, the first target sequence and the second target sequence comprise sequences from a single gene within the genome of the target mammalian cell. In some cases, the single gene is a non-essential gene. In some cases, the target sequence comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity of the homology arm sequence disclosed herein. In some cases, the target sequence comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity of the reverse complement of the homology arm sequence disclosed herein. For instance, the first target sequence can comprise a sequence that is identical to the 5' homology arm of the polynucleotide disclosed herein, while the second target sequence can comprise a sequence that is identical to the 3' homology arm of the polynucleotide disclosed herein. For instance, the first target sequence can comprise a sequence that is identical to the reverse complement of the 5' homology arm of the polynucleotide disclosed herein, while the second target sequence can comprise a sequence that is identical to the 3' homology arm of the polynucleotide disclosed herein. For instance, the first target sequence can comprise a sequence that is identical to the 5' homology arm of the polynucleotide disclosed herein, while the second target sequence can comprise a sequence that is identical to the reverse complement of the 3' homology arm of the polynucleotide disclosed herein.

[0347] In some embodiments, a first guide RNA and a second guide RNA are designed for improving or enhancing the efficiency of the retrotransposition by LINE 1 elements. Retrotransposition by LINE1 elements as designed and described in the studies shown herein are useful in editing large polynucleic acid "insert sequences" into the genome, that may comprise a therapeutic gene or fragment thereof, encoding a therapeutic polypeptide. The first guide RNA helps direct the Cas 9 nickase (may be referred to as nCas9 or Cas9n) to a position on the genomic strand for exercising the nick and thereby

to effect the annealing of the first homology arm sequence to a sequence on the genomic DNA. The second guide RNA helps direct the nCas9 to a position on the opposite genomic strand to create a nick and thereby to effect the annealing of the second homology arm sequence to a sequence on the genomic DNA.

Polynucleic acids in the composition

[0348] In some embodiments, the composition for genomic integration disclosed herein comprises polynucleic acids encoding the components of the composition. For example, the composition may comprise a polynucleic acid encoding the first gRNA. The composition may comprise a polynucleic acid encoding the second gRNA. In some embodiments, the composition comprises a polynucleic acid encoding the Cas endonuclease such as a Cas9 nickase or a functional fragment thereof. In some embodiments, the composition comprises a polynucleic acid encoding the ORF1p or a functional fragment thereof. In some embodiments, the composition comprises a polynucleic acid encoding the ORF2p or a functional fragment thereof. In some embodiments, the composition comprises a polynucleotide encoding the ORF1p and ORF2p, or a functional fragment thereof. In some embodiments, the composition comprises a polynucleic acid comprising a first sequence encoding the ORF1p or a functional fragment thereof, a second sequence encoding the ORF2p or a functional fragment thereof, a third sequence that is reverse complement of a sequence encoding a protein of interest (e.g., a therapeutic peptide), wherein the third sequence is flanked by at least one homology arm. In some cases, the 3' end of the third sequence is flanked by the homology arm. In some embodiments, the 5' end of the third sequence is flanked by the homology arm. In some cases, both the 3' end and the 5' end of the third sequence are flanked by the homology arms.

[0349] In some embodiments, the polynucleic acid sequence encoding the ORF1p or a functional fragment thereof comprises a wildtype coding sequence for ORF1p. In some embodiments, the polynucleic acid sequence encoding the ORF1p or a functional fragment thereof comprises a nucleic acid sequence that is at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or 99.5% of a wildtype coding sequence for ORF1p. In some embodiments, the polynucleic acid sequence encoding the ORF2p or a functional fragment thereof comprises a wildtype coding sequence for ORF2p. In some embodiments, the polynucleic acid sequence encoding the ORF2p or a functional fragment thereof comprises a nucleic acid sequence that is at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or 99.5% of a wildtype coding sequence for ORF2p.

[0350] In some embodiments, the polynucleic acid encoding the Cas9 nickase or a functional fragment thereof comprises a wildtype coding sequence for a wildtype Cas9 endonuclease. In some embodiments, the polynucleic acid encoding the Cas9 nickase or a functional fragment thereof comprises a nucleic acid sequence that is at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or 99.5% of a wildtype coding sequence for wildtype Cas9 endonuclease.

[0351] In some embodiments, the polynucleic acids of the composition are mRNA. In some embodiments, the polynucleic acids comprise mRNA and DNA. For example, in one embodiment, the composition of the CREATE system comprises a CREATE mRNA (as shown in the illustrations in FIG. 1, FIG. 3 and the others, which encodes the retrotransposon machinery, including but not limited to sequences encoding ORF1, ORF2 and reverse complement payload sequences); and further comprises one or more plasmids encoding one or more guide RNA and/or a Cas endonuclease.

Retrotransposon

[0352] In some embodiments, provided herein is a polynucleotide construct comprising an mRNA wherein the mRNA comprises a sequence encoding a human retrotransposon, wherein, (i) the sequence of a human retrotransposon comprises a sequence encoding ORF1p, (ii) the mRNA does not comprise a sequence encoding ORF1p, or (iii) the mRNA comprises a replacement of the sequence encoding ORF1p with a 5' UTR sequence from the complement gene. In some embodiments, the mRNA comprises a first mRNA molecule encoding ORF1p, and a second mRNA molecule encoding an endonuclease and/or a reverse transcriptase. In some embodiments, the mRNA is an mRNA molecule comprising a first sequence encoding ORF1p, and a second sequence encoding an endonuclease and/or a reverse transcriptase. In some embodiments, the first sequence encoding ORF1p and the second sequence encoding an endonuclease and/or a reverse transcriptase are separated by a linker sequence.

[0353] In some embodiments, the linker sequence comprises an internal ribosome entry sequence (IRES). In some embodiments, the IRES is an IRES from CVB3 or EV71. In some embodiments, the linker sequence encodes a self-cleaving peptide sequence. In some embodiments, the linker sequence encodes a T2A, a E2A or a P2A sequence.

[0354] In some embodiments, the sequence of a human retrotransposon comprises a sequence that encodes ORF1p fused to an additional protein sequence and/or a sequence that encodes ORF2p fused to an additional protein sequence. In some embodiments, the ORF1p and/or the ORF2p is fused to a nuclear retention sequence. In some embodiments, the nuclear retention sequence is an Alu sequence. In some embodiments, the ORF1p and/or the ORF2p is fused to an MS2 coat protein. In some embodiments, the 5' UTR sequence or the 3' UTR sequence comprises at least one, two, three or more MS2 hairpin sequences. In some embodiments, the 5' UTR sequence or the 3' UTR sequence comprises a sequence that promotes or enhances interaction of a poly A tail of the mRNA with the endonuclease and/or a reverse transcriptase. In some embodiments, the 5' UTR sequence or the 3' UTR sequence comprises a sequence that promotes or enhances interaction of a poly-A-binding proteins (e.g., PABP) with the endonuclease and/or a reverse transcriptase. In some embodiments, the 5' UTR sequence or the 3' UTR sequence comprises a sequence that increases specificity of the endonuclease and/or a reverse transcriptase to the mRNA relative to another mRNA expressed by the

cell. In some embodiments, the 5' UTR sequence or the 3' UTR sequence comprises an Alu element sequence.

[0355] In some embodiments, the first sequence encoding ORF1p and the second sequence encoding an endonuclease and/or a reverse transcriptase have the same promoter. In some embodiments, the insert sequence has a promoter that is different from the promoter of the first sequence encoding ORF1p. In some embodiments, the insert sequence has a promoter that is different from the promoter of the second sequence encoding an endonuclease and/or a reverse transcriptase. In some embodiments, the first sequence encoding ORF1p and/or the second sequence encoding an endonuclease and/or a reverse transcriptase have a promoter or transcription initiation site selected from the group consisting of an inducible promoter, a CMV promoter or transcription initiation site, a T7 promoter or transcription initiation site, an EF1a promoter or transcription initiation site and combinations thereof. In some embodiments, the insert sequence has a promoter or transcription initiation site selected from the group consisting of an inducible promoter, a CMV promoter or transcription initiation site, a T7 promoter or transcription initiation site, an EF1a promoter or transcription initiation site and combinations thereof.

[0356] In some embodiments, the first sequence encoding ORF1p and the second sequence encoding an endonuclease and/or a reverse transcriptase are codon optimized for expression in a human cell.

[0357] In some embodiments, the mRNA comprises a WPRE element. In some embodiments, the mRNA comprises a selection marker. In some embodiments, the mRNA comprises a sequence encoding an affinity tag. In some embodiments, the affinity tag is linked to the sequence encoding an endonuclease and/or a reverse transcriptase.

[0358] In some embodiments, the 3' UTR comprises a poly A sequence or wherein a poly A sequence is added to the mRNA in vitro. In some embodiments, the poly A sequence is downstream of a sequence encoding an endonuclease and/or a reverse transcriptase. In some embodiments, the insert sequence is upstream of the poly A sequence.

[0359] In some embodiments, the 3' UTR sequence comprises the insert sequence. In some embodiments, the insert sequence comprises a sequence that is a reverse complement of the sequence encoding the exogenous polypeptide. In some embodiments, the insert sequence comprises a polyadenylation site. In some embodiments, the insert sequence comprises an SV40 polyadenylation site. In some embodiments, the insert sequence comprises a polyadenylation site upstream of the sequence that is a reverse complement of the sequence encoding the exogenous polypeptide. In some embodiments, the insert sequence is integrated into the genome at a locus that is not a ribosomal locus. In some embodiments, the insert sequence is integrated into the genome at a locus that is not a rDNA locus. In some embodiments, the insert sequence integrates into a gene or regulatory region of a gene, thereby disrupting the gene or downregulating expression of the gene. In some embodiments, the insert sequence integrates into a gene or regulatory region of a gene, thereby upregulating expression of the

gene. In some embodiments, the insert sequence integrates into the genome and replaces a gene. In some embodiments, the insert sequence is stably integrated into the genome. In some embodiments, the insert sequence is retrotransposed into the genome. In some embodiments, the insert sequence is integrated into the genome by cleavage of a DNA strand of a target site by an endonuclease encoded by the mRNA. In some embodiments, the insert sequence is integrated into the genome via target-primed reverse transcription (TPRT). In some embodiments, the insert sequence is integrated into the genome via reverse splicing of the mRNA into a DNA target site of the genome.

[0360] In some embodiments, the retrotransposon is a LINE-1 (L1) retrotransposon. In some embodiments, the retrotransposon is human LINE-1. Human LINE-1 sequences are abundant in the human genome. There are approximately 13,224 total human L1s, of which 480 are active, which make up about 3.6%. Therefore, human L1 proteins are well tolerated and non-immunogenic in humans. Moreover, a tight regulation of random transposition in human ensures that random transposase activity will not be triggered by introduction of the L1 system as described herein. In addition, the retrotransposable constructs designed herein may comprise targeted and specific incorporation of the insert sequence. In some embodiments, the retrotransposable genetic element may comprise designs intended to overcome the silencing machinery actively prevalent in human cells, while being careful that random integration resulting in genomic instability is not initiated.

[0361] Accordingly, the retrotransposable constructs may comprise a sequence encoding a human LINE-1 ORF1 protein; and a human LINE-1 ORF2 protein. In some embodiments, the construct comprises a nucleic acid sequence encoding an ORF1p protein with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to

MGKKQNRKTGNSKTQSASPPKERSSSPATEQSWMENDFDELREEGFRRSNYSELREDIQT
KGKEVENFEKNLEECITRITNTEKCLKELMELKTKARELREECRSLRSRCDQLEERVSAMED
EMNEMKREGKFREKRIKRNEQSLQEIWVYKRPNLRLIGVPESDVENGTKLENTLQDIIQEN
FPNLARQANVQIQEIQRTPQRYSSRRATPRHIIVRFTKVEMKEKMLRAAREKGRVTLKGKPI
RLTVDLAETLQARREWGPIFNILKEKNFQPRISYPAKLSFISEGEIKYFIDKQMLRDFVTTRP
ALKELLKEALNMERNRYQPLQNHAKM (SEQ ID NO: 9). In some embodiments, the

construct comprises a nucleic acid sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to

ATGGGCAAGAAGCAAAATCGCAAGACGGGAATTCCAAGACACAATCCGCTAGCCCAC
CACCTAAAGAGCGTTCTAGCTCCCCTGCTACTGAGCAGTCCTGGATGGAAAACGACTTC
GATGA ACTCCGGAAGAGGGATTTAGGCGATCCA ACTATT CAGAACTCCGCGAAGATA

TCCAGACAAAGGGGAAGGAAGTCGAGAATTTTCGAGAAGAACCTCGAGGAGTGCATCAC
 CCGTATCACAAACACTGAGAAATGTCTCAAAGAACTCATGGAACTTAAGACAAAAGCC
 AGGGAGCTTCGAGAGGAGTGTCTGGAGTCTGAGATCCAGGTGTGACCAGCTCGAGGAGC
 GCGTGAGCGCGATGGAAGACGAGATGAACGAGATGAAAAGAGAGGGCAAATTCAGGG
 AGAAGCGCATTAAAGAGGAACGAACAGAGTCTGCAGGAGATTTGGGATTACGTCAAGAG
 GCCTAACCTGCGGTTGATCGGCGTCCCCGAGAGCGACGTAGAAAACGGGACTAACTG
 GAGAATACACTTCAAGACATCATTCAAGAAAATTTTCCAAACCTGGCTCGGCAAGCTAA
 TGTGCAAATCCAAGAGATCCAACGCACACCCCAGCGGTATAGCTCTCGGCGTGCCACCC
 CTAGGCATATTATCGTGCGCTTTACTAAGGTGGAGATGAAAGAGAAGATGCTGCGAGCC
 GCTCGGGAAAAGGGAAGGGTGACTTTGAAGGGCAAACCTATTCGGCTGACGGTTGACC
 TTAGCGCCGAGACACTCCAGGCACGCCGGAATGGGGCCCCATCTTTAATATCCTGAAG
 GAGAAGAACTTCCAGCCACGAATCTCTTACCCTGCAAAGTTGAGTTTTATCTCCGAGGG
 TGAGATTAAGTATTTTCATCGATAAACAGATGCTGCGAGACTTCGTGACAACTCGCCCAG
 CTCTCAAGGAAGTCTCAAAGAGGCTCTTAATATGGAGCGCAATAATAGATATCAACCC
 TTGCAGAACCACGCAAAGATGTGA (SEQ ID NO: 10).

[0362] In some embodiments, the construct comprises a nucleic acid sequence encoding an ORF2p protein with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to

MTGSNSHITILTLNINGLNSAIKRHRLASWIKSQDPSVCCIQETHLTCRDTHRLKIKGWRKIY
 QANGKQKKAGVAILVSDKTDFKPTKIKRDKEGHYIMVKGSIQQUEELTILNIYAPNTGAPRFI
 KQVLSDLQRDLDSHTLMGDFNTPLSTLDRSTRQKVNKDTQELNSALHQADLIDIYRTLHPK
 STEYTFFSAPHHTYSKIDHIVGSKALLSKCKRTEITNYLSDHSAIKLELRIKNLTQSRSTTWK
 LNNLLLNNDYWVHNEMKAEIKMFFETNENKDDTTYQNLWDAFKAVCRGKFIALNAYKRKQE
 RSKIDTLTSQLKELEKQEQTHSKASRRQEITKIRAELEIETQKTLQKINESRSWFFERINKIDR
 PLARLIKKKREKNQIDTIKNDKGDITDPTEIQTIREYYKHLANKLENLEEMDTFLDITYTL
 PRLNQEEVESLNRPTGSEIVAIINSLPTKKSPPDGFTAIFYQRYMEELVPFLLKLFQSIEKEG
 ILPNSFYEASIIIPKPRDITTKENFRPISLMNIDAKILNKILANRIQQHIKKLIHHDQVGFIPG
 MQGWFNIRKSINVIQHINRAKDKNHMIISIDAEKAFDKIQQPFMLKTLNKLIGDGTIFYKIIIRAI
 YDKPTANIILNGQKLEAFPLKTGTRQGCPSPLLFNIVLEVLAIRAIRQEKEIKIGIQLGKEEVKL
 SLFADDMIVYLENPIVSAQNLLKLISNFSKVSQYKINVQKSQAFLYTNNRQTESQIMGELPFV
 IASKRIKYLGIQLTRDVKDLFKENYKPLLKEIKEDTNKWKNIKPCSWVGRINIVKMAILPKVIY
 RFNAIPIKLPMTFFTELEKTTLKFIWNQKRARIKSIKQKAGGITLPDFKLYYKATVTKT
 AWYWYQNRDIDQWNRTEPSEIMPHIYNLIFDKPEKNKQWGKDSL FNKWCWENWLAICR
 KLKLDPFLTPYTKINSRWIKDLNVKPKTIKLEENLGITIDIGVGKDFMSKTPKAMATKDKI

DKWDLIKLKSFACTAKETTIRVNRQPTTWEEKIFATYSSDKGLISRIYNELKQIYKKKTNNPIKK
WAKDMNRHFSKEDIYAAKKHMKKCSSSLAIREMQIKTTMRYHLTPVRMAIIKKSGNNRCW
RGCGEIGTLLHCWWDCKLVQPLWKSVMRFLRDLELEIPFDPAPILLGIYPNEYKSCCYKDTCT
TRMFIAALFTIAKTWNQPKCPTMIDWIKKMWHIYTMEYYAAIKNDEFISFVGTWMKLETIIL
SKLSQEQKTKHRIFSLIGGN (SEQ ID NO: 11). In some embodiments, the construct comprises a
nucleic acid sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at
least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least
92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%,
or 100% sequence identity to

ATGACCGGCTCTAACTCACATATCACCATCCTTACACTTAACATTAACGGCCTCAACTC
AGCTATCAAGCGCCATCGGCTGGCCAGCTGGATCAAATCACAGGATCCAAGCGTTTGT
GCATCCAAGAGACCCACCTGACCTGTAGAGATACTCACCGCCTCAAGATCAAGGGATG
GCGAAAGATTTATCAGGCGAACGGTAAGCAGAAGAAAGCCGGAGTCGCAATTCTGGTC
TCAGACAAGACGGATTTCAAGCCCACCAAAATTAAGCGTGATAAGGAAGGTCCTATA
TTATGGTGAAAGGCAGCATACAGCAGGAAGAACTTACCATATTGAACATCTACGCGCC
AAACACCGGCGCACCTCGCTTTATCAAACAGGTCCTGTCCGATCTGCAGCGAGATCTGG
ATTCTCATACGTTGATTATGGGTGATTTCAATACACCATTGAGCACCTGGATCGCAGC
ACCAGGCAAAAGGTAAATAAAGACACGCAAGAGCTCAATAGCGCACTGCATCAGGCAG
ATCTCATTGATATTTATCGCACTCTTCATCCTAAGAGTACCGAGTACACATTCTTCAGCG
CCCCACATCATATACTCAAAGATCGATCATATCGTCGGCTCAAAGGCTCTGCTGTCA
AAGTGCAAGCGCACAGAGATAATTACAAATTACCTGTCAGATCATAGCGCGATCAAGC
TCGAGCTGAGAATCAAGAACCTGACCCAGAGCCGGAGTACCACTTGGAAGCTTAATAA
CCTGCTGCTCAACGATTATTGGGTCCACAATGAGATGAAGGCAGAGATTAATAATGTTCT
TCGAAACAAATGAGAATAAGGATACTACCTATCAAAACCTTTGGGATGCCTTTAAGGCC
GTCTGCAGAGGCAAGTTCATCGCCCTCAACGCCTATAAAAGAAAACAAGAGAGATCTA
AGATCGATACTCTCACCTCTCAGCTGAAGGAGTTGGAGAAACAGGAACAGACCCACTC
CAAGGCGTCAAGACGGCAGGAGATCACAAAGATTTCGCGCCGAGTTGAAAGAGATCGAA
ACCCAAAAGACTCTTCAGAAAATTAACGAGTCTCGTAGTTGGTTCTTCGAGCGGATTAA
TAAGATAGACAGACCTCTGGCACGACTGATTAAGAAGAAGCGCGAAAAGAACCAGATT
GATACCATCAAGAACGACAAGGGCGACATCACTACTGACCCGACCGAGATCCAGACCA
CTATTCGGGAGTATTATAAGCATTTGTATGCTAACAAGCTTGAGAACCTGGAAGAGATG
GACACTTTTCTGGATACCTATACTCTGCCACGGCTTAATCAAGAGGAAGTCGAGTCCCT
CAACCGCCCAATTACAGGAAGCGAGATTGTGGCCATAATTAACCTCCCTGCCGACAAAG
AAATCTCCTGGTCCGGACGGGTTTACAGCTGAGTTTTATCAACGGTATATGGAAGAGCT
TGTACCGTTTCTGCTCAAGCTCTTTCAGTCTATAGAAAAGGAAGGCATCTTGCCCAATTC
CTTCTACGAAGCTTCTATAATACTTATTCCCAAACCAGGACGCGATACCACAAAGAAGG

AAAACCTCCGGCCCATTAGTCTCATGAATATCGACGCTAAAATATTGAACAAGATTCTC
GCCAACAGAATCCAACAACATATTAAGAAATTGATACATCACGACCAGGTGGGGTTTAT
ACCTGGCATGCAGGGCTGGTTTAACATCCGGAAGAGTATTAACGTCATTCAACACATTA
ATAGAGCTAAGGATAAGAATCATATGATCATCTCTATAGACGCGGAAAAGGCATTCTGA
TAAGATTCTAGCAGCCATTTATGCTCAAGACTCTGAACAACTCGGCATCGACGGAACAT
ATTTTAAGATTATTCGCGCAATTTACGATAAGCCGACTGCTAACATTATCCTTAACGGCC
AAAAGCTCGAGGCCTTTCCGCTCAAGACTGGAACCCGCCAAGGCTGTCCCCTCTCCCCG
CTTTTGTTTAATATTGTACTCGAGGTGCTGGCTAGGGCTATTCGTCAAGAGAAAGAGAT
TAAAGGGATACAGCTCGGGAAGGAAGAGGTCAAGCTTTCCTTGTTTCGCCGATGATATGA
TTGTGTACCTGGAGAATCCTATTGTGTCTGCTCAGAACCTTCTTAACTTATTTCTAACT
TTAGCAAGGTCAGCGGCTATAAGATTAACTGTCAGAAATCTCAGGCCTTTCTGTACACA
AATAATCGACAGACCGAATCCCAGATAATGGGTGAGCTTCCGTTTGTTCATAGCCAGCAA
AAGGATAAAGTATCTCGGAATCCAGCTGACACGAGACGTTAAAGATTTGTTTAAGGAA
AATTACAAGCCTCTCCTGAAAGAGATTAAAGGAAGATACTAATAAGTGGAAGAATATCC
CCTGTTTCATGGGTTGGCAGAATCAACATAGTGAAGATGGCAATACTTCCTAAAGTGATA
TATCGCTTTAACGCCATCCCAATTAAGCTGCCTATGACCTTCTTTACGGAGCTCGAGAAA
ACAACCCTTAAATTTATATGGAATCAAAAGAGAGCAAGAATAGCGAAGTCCATCTTGA
GCCAGAAGAATAAGGCCGGTGGGATTACTTTGCCTGATTTTAAGTTGTATTATAAAGCC
ACAGTAACTAAGACAGCCTGGTATTGGTATCAGAATAGAGACATCGACCAGTGGAATC
GGACCGAACCATCAGAGATAATGCCCCACATCTATAATTACCTTATATTGATAAGCCA
GAAAAGAATAAACAGTGGGGCAAAGACAGCCTCTTCAACAAGTGGTGTTGGGAGAATT
GGCTGGCCATATGCCGGAACTCAAGCTCGACCCCTTTCTTACACCCTACACTAAAATC
AACAGTAGGTGGATCAAGGACTTGAATGTCAAGCCAAAGACTATAAAGACACTGGAAG
AGAATCTTGGGATCACAATACAAGATATAGGCGTCGGCAAAGATTTTATGTCAAAGAC
GCCAAGGCCATGGCCACTAAGGATAAGATTGATAAGTGGGACCTTATTAAGCTCAAA
AGCTTCTGTACTGCCAAGGAGACCACGATCAGAGTTAATAGGCAGCCCACTACATGGG
AAAAGATTTTCGCCACTTATTCATCAGATAAGGGGGTTGATAAGCAGAATATATAACGAG
CTGAAGCAGATCTACAAGAAGAAAACGAATAATCCCATCAAGAAGTGGGCAAAAGATA
TGAACAGGCATTTTAGCAAAGAGGATATCTACGCCGCGAAGAAGCATATGAAGAAGTG
TAGTTCAAGCTTGGCCATTCGTGAGATGCAGATTAAGACGACCATGCGATACCACCTTA
CCCCAGTGAGGATGGCAATTATCAAGAAATCTGGCAATAATAGATGTTGGCGGGGCTGT
GGCGAGATTGGCACCTGCTCCATTGCTGGTGGGATTGCAAGCTGGTGCAGCCGCTTTG
GAAATCAGTCTGGCGCTTTCTGAGGGACCTCGAGCTTGAGATTCCCTTCGATCCCGCAA
TTCCCTTGCTCGGAATCTATCCTAACGAATACAAGAGCTGTTGTTACAAGGATACGTGT
ACCCGGATGTTTCATCGCGGCCTTGTTTACGATAGCTAAGACGTGGAATCAGCCTAAGTG
CCCCACAATGATCGATTGGATCAAGAAAATGTGGCATATTTATACCATGGAGTATTACG

CAGCAATTAAGAATGACGAATTTATTTCTTCGTTGGGACCTGGATGAAGCTGGAGACT
ATTATTCTGAGCAAGCTGTCTCAGGAGCAAAAGACAAAGCATAGAATCTTCTCTCTCAT
TGGTGGTAACTAA (SEQ ID NO: 12).

[0363] In some embodiments, the construct comprises a nucleic acid sequence encoding an ORF2p protein with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to

MVIGTYISITLNVNGLNAPTKRHRLAEWIKQDPYICCLQETHFRPRDTYRLKVRGWKKIF
HANGNQKKAGVAILISKIDFKIKNVTRDKEGHYIMIQGSIQEEDITIINIYAPNIGAPQYIRQL
LTAIKEEIDSNTIIVGDFNTSLTPMDRSSKMKINKETEALNDTIDQIDLIDIYRTFHPKTADYTF
FSSAHGTFSRIDHILGHKSSLSKFKKIEIISIFSDHNAMRLEMNHREKNVKKTNWRLNNTL
LNNQEITEEIKQEIKKYLETNDNENTTTQNLWDAKAVLRGKFIAIQAYLKKQEKSQVNNL
TLHLKKLEKEEQTKPKVSRKEIIRAEINEIETKKTIKINKTKSWFFEKINKIDKPLARLIK
KKRERTQINKIRNEKGEVTTDTAEIQNILRDYYKQLYANKMDNLEEMDKFLERYNLPRLNQ
EETENINRPITSNEIETVIKNLPTNKSPGPDGFTGEFYQTFREELTPILLKLFQKIAEEGTLPSNF
YEATITLIPKPKDPTTKKENYRPISLMNIDAKILNKILANRIQQHIKRHHHDQVGFIPGMQGGF
NIRKSINVIHHINKLKKKNHMIISIDAFAFDKIQHFMKTLQKVGIEGTYLNIKAIYDKPTA
NIILNGEKLKAFPLRSGTRQGCPLSPLLFNIVLEVLATAIREEKEIKGIQIGKEEVKLSLFADD
MILYIENPKTATRKLLELINEYGVAGYKINAQKSLAFLYTNDKSEREIMETLPFTIATKRIK
YLGINLPKETKDLYAENYKTLMKEIKDDTNRWRDIPCSWIGRINIVKMSILPKAIYRFNAIPIK
LPMAFFTELEQIILKFVWRHKRPRIKAVLRQKNGAGGIRLPDFRLYYKATVIKTIWYWHK
NRNIDQWNKIESPEINPRTYGQLIYDKGGKDIQWRKDSL FNKWCWENWTATCKRMKLEYS
LTPYTKINSKWIRDLNIRLDTIKLLEENIGRTLFDINH SKIFFDPPPRVMEIKTKINKWDLMKL
QSFCTAKETINKTKRQPSEWEKIFANESTDKGLISKIYKQLIQLNIKETNTPIQKWAEDLNRHF
SKEDIQTATKHMKRCSTSLIREMQIKTTMRYHLTPVRMGIIRKSTNNKCWRGCGEKGTLHLH
CWWECKLIQPLWRTIWRFLKLLKIELPYDPAIPLLLGIYPEKTVIQKDTCTRMFIAALFTIARS
WKQPKCPSTDEWIKMWYIYTMEYSSAIKRNEIGSFLETWMDLETVIQSEVSQKEKNKYRI
LTHICGTWKNGTDEPVCRTIEIETQM (SEQ ID NO: 13).

In some embodiments, the construct comprises a nucleic acid sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to

ATGGTCATAGGAACATACATATCGATAATTACCTTAAACGTGAATGGATTAAATGCCCC
AACCAAAAGACATAGACTGGCTGAATGGATACAAAAACAAGACCCATATATATGCTGT
CTACAAGAGACCCACTTCAGACCTAGGGACACATACAGACTGAAAGTGAGGGGATGGA

AAAAGATATTCCATGCAAATGGAAATCAAAAGAAAGCTGGAGTAGCTATACTCATATC
AGATAAAATAGACTTTTAAATAAAAGAATGTTACAAGAGACAAGGAAGGACACTACATA
ATGATCCAGGGATCAATCCAAGAAGAAGATATAACAATTATAAATATATATGCACCCA
ACATAGGAGCACCTCAATACATAAGGCAACTGCTAACAGCTATAAAAGAGGAAATCGA
CAGTAACACAATAATAGTGGGGGACTTTAACACCTCACTTACACCAATGGACAGATCAT
CCAAAATGAAAATAAATAAGGAAACAGAAGCTTTAAATGACACAATAGACCAGATAGA
TTTAATTGATATATATAGGACATTCCATCCAAAAACAGCAGATTACACGTTCTTCTCAA
GTGCGCACGGAACATTCTCCAGGATAGATCACATCTTGGGTCACAAATCAAGCCTCAGT
AAATTTAAGAAAATTGAAATCATATCAAGCATCTTTTCTGACCACAACGCTATGAGATT
AGAAATGAATCACAGGGAAAAAACGTAAAAAGACAAACACATGGAGGCTAAACAA
TACGTTACTAAATAACCAAGAGATCACTGAAGAAATCAAACAGGAAATAAAAAATAC
CTAGAGACAAATGACAATGAAAACACGACGACCCAAAACCTATGGGATGCAGCAAAA
GCGGTTCTAAGAGGGAAGTTTATAGCTATACAAGCCTACCTAAAGAAACAAGAAAAAT
CTCAAGTAAACAATCTAACCTTACACCTAAAGAACTAGAGAAAGAAGAACAAACAAA
ACCCAAAGTTAGCAGAAGGAAAGAAATCATAAAGATCAGAGCAGAAATAAATGAAAT
AGAAACAAAGAAAACAATAGCAAAGATCAATAAAACTAAAAGTTGGTTCTTTGAGAAG
ATAAACAAAATTGATAAGCCATTAGCCAGACTCATCAAGAAAAAGAGGGAGAGGACTC
AAATCAATAAAATCAGAAATGAAAAAGGAGAAGTTACAACAGACACCGCAGAAATAC
AAAACATCCTAAGAGACTACTACAAGCAACTTTATGCCAATAAAATGGACAACCTGGA
AGAAATGGACAAATTCTTAGAAAGGTATAACCTTCCAAGACTGAACCAGGAAGAAACA
GAAAATATCAACAGACCAATCACAAGTAATGAAATTGAACTGTGATTAAAAATCTTCC
AACAAACAAAAGTCCAGGACCAGATGGCTTCACAGGTGAATTCTATCAAACATTTAGA
GAAGAGCTAACCCCATCCTTCTCAAACCTCTTCCAAAAAATTGCAGAAGAAGGAACAC
TCCCAAACCTCATTCTATGAGGCCACCATCACCTGATACCAAACACAGACAAAGACACT
ACAAAAAAGAAAATTACAGACCAATATCACTGATGAATATAGATGCAAAAATCCTCA
ACAAAATACTAGCAAACAGAATCCAACAACACATTAAGGATCATAACCCACGATCA
AGTGGGATTTATCCCAGGGATGCAAGGATTCTTCAATATACGCAAATCAATCAATGTGA
TACACCATATTAACAAATTGAAGAAGAAAAACCATATGATCATCTCAATAGATGCAGA
AAAAGCTTTTGACAAAATTCAACACCCATTTATGATAAAAACCTCTCCAGAAAGTGGGCA
TAGAGGGAACCTACCTCAACATAATAAGGCCATATATGACAAACCCACAGCAAACAT
CATTCTCAATGGTGAAAAACTGAAAGCATTCCTCTAAGATCAGGAACGAGACAAGGA
TGTCCACTCTCACCCTATTATTCAACATAGTTCTGGAAGTCCTAGCCACGGCAATCAG
AGAAGAAAAAGAAATAAAAGGAATACAAATTGGAAAAGAAGTAAGTAAACTGTCACT
GTTTGCGGATGACATGATACTATACATAGAGAATCCTAAAACCTGCCACCAGAAAACCTGC
TAGAGCTAATTAATGAATATGGTAAAGTTGCAGGTTACAAAATTAATGCACAGAAATCT
CTTGCAATCCTATACACTAATGATGAAAAATCTGAAAGAGAAATTATGGAAACACTCCC

ATTTACCATTGCAACAAAAAGAATAAAATACCTAGGAATAAACCTACCTAAGGAGACA
AAAGACCTGTATGCAGAAAACATAAGACACTGATGAAAGAAATTAAGATGATACCA
ACAGATGGAGAGATATACCATGTTCTTGATTGGAAGAATCAACATTGTGAAAATGAGT
ATACTACCCAAAGCAATCTACAGATTCAATGCAATCCCTATCAAATTACCAATGGCATT
TTTTACGGAGCTAGAACAAATCATCTTAAAAATTTGTATGGAGACACAAAAGACCCCGAA
TAGCCAAAGCAGTCTTGAGGCAAAAAAATGGAGCTGGAGGAATCAGACTCCCTGACTT
CAGACTATACTACAAAGCTACAGTAATCAAGACAATATGGTACTGGCACAAAAACAGA
AACATAGATCAATGGAACAAGATAGAAAGCCCAGAGATTAAACCCACGCACCTATGGTC
AACTAATCTATGACAAAGGAGGCAAGATATACAATGGAGAAAAGACAGTCTCTTCAA
TAAGTGGTGCTGGGAAAACCTGGACAGCCACATGTAAAAGAATGAAATTAGAATACTCC
CTAACACCATACACAAAAATAAACTCAAAATGGATTAGAGACCTAAATATAAGACTGG
ACACTATAAACTCTTAGAGGAAAACATAGGAAGAACTCTTTGACATAAATCACAG
CAAGATCTTTTTCGATCCACCTCCTAGAGTAATGGAAATAAAAAACAAAAATAAACAAGT
GGGACCTAATGAACTTCAAAGCTTTTGCACAGCAAAGGAAACCATAAACAAGACGAA
AAGACAACCCTCAGAATGGGAGAAAATATTTGCAAATGAATCAACGGACAAAGGATTA
ATCTCCAAAATATATAAACAGCTCATTAGCTCAATATCAAAGAAACAAACACCCCAAT
CCAAAAATGGGCAGAAGACCTAAATAGACATTTCTCCAAAGAAGACATACAGACGGCC
ACGAAGCACATGAAAAGATGCTCAACATCACTAATTATTAGAGAAATGCAAATCAAAA
CTACAATGAGGTATCACCTCACTCCTGTTAGAATGGGCATCATCAGAAAATCTACAAAC
AACAAATGCTGGAGAGGGTGTGGAGAAAAGGGAACCCTCTTGCACTGTTGGTGGGAAT
GTAAATTGATACAGCCACTATGGAGAACAATATGGAGGTTCTTAAAAAACTAAAAAT
AGAATTACCATATGACCCAGCAATCCCACTACTGGGCATATACCCAGAGAAAACCGTA
ATTCAAAAAGACACATGCACCCGAATGTTCAATTGCAGCACTATTTACAATAGCCAGGTC
ATGGAAGCAACCTAAATGCCCATCGACAGACGAATGGATAAAGAAGATGTGGTACATA
TATACAATGGAATATTACTCAGCCATAAAAAAGGAACGAAATTGGGTCATTTTTAGAGAC
GTGGATGGATCTAGAGACTGTCATACAGAGTGAAGTAAGTCAGAAAGAGAAAAACAAA
TATCGTATATTAACGCATATATGTGGAACCTGGAAAAATGGTACAGATGAACCGGTCTG
CAGGACAGAAATTGAGACACAAATGTAA (SEQ ID NO: 14).

[0364] In some embodiments, the construct comprises a nucleic acid sequence encoding an ORF2 lacking an endonuclease domain. In some embodiments, the ORF2 lacking an endonuclease domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 151**. In some embodiments, the nucleic acid sequence encoding an ORF2 lacking an endonuclease domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at

least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 143**.

[0365] The constructs can comprise a nucleic acid sequence encoding an additional exogenous DNA binding domain. The additional exogenous DNA binding domain of the CREATE systems can increase ORF2p DNA binding activities. In some embodiments, the ORF2p DNA binding activity of the CREATE system is increased by about 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% as compared to a CREATE system without the additional exogenous DNA binding domain. In some embodiments, the ORF2p DNA binding activity of the CREATE system is increased by at least 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% as compared to a CREATE system without the additional exogenous DNA binding domain. In some embodiments, the ORF2p DNA binding activity of the CREATE system is increased by at most 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 100% as compared to a CREATE system without the additional exogenous DNA binding domain. In some embodiments, the ORF2p DNA binding activity of the CREATE system is increased in a range between any of the lower limits and any of the upper limits disclosed earlier (for example, between 5% and 100%, or between 10% and 50%) as compared to a CREATE system without the additional exogenous DNA binding domain. The ORF2p DNA binding activity can be measured by electrophoretic mobility shift assays (EMSAs), chromatin immunoprecipitation (ChIP), or other method for detecting DNA binding activities known in the art.

[0366] The additional exogenous DNA binding domain of the CREATE systems can increase TPRT processivity. In some embodiments, the TPRT processivity of the CREATE system is increased by about 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% as compared to a CREATE system without the additional exogenous DNA binding domain. In some embodiments, the TPRT processivity of the CREATE system is increased by at least 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% as compared to a CREATE system without the additional exogenous DNA binding domain. In some embodiments, the TPRT processivity of the CREATE system is increased by at most 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 100% as compared to a CREATE system without the additional exogenous DNA binding domain. In some embodiments, the TPRT processivity of the CREATE system is increased in a range between any of the lower limits and any of the upper limits disclosed earlier (for example, between 5% and 100%, or between 10% and 50%) as compared to a CREATE system without the additional exogenous DNA binding domain. The TPRT processivity can be determined by processivity assay known in the art, such as assays to measure premature termination of nucleotide or probability of chain elongation.

[0367] The additional exogenous DNA binding domain of the CREATE systems can increase direct recruitment of CREATE RNP complexes to single strand breaks induced by the Cas9 nickase of the

CREATE complex. In some embodiments, the recruitment of CREATE RNP complexes to single strand breaks is increased by about 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% as compared to a CREATE system without the additional exogenous DNA binding domain. In some embodiments, the recruitment of CREATE RNP complexes to single strand breaks is increased by at least 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% as compared to a CREATE system without the additional exogenous DNA binding domain. In some embodiments, the recruitment of CREATE RNP complexes to single strand breaks is increased by at most 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 100% as compared to a CREATE system without the additional exogenous DNA binding domain. In some embodiments, the recruitment of CREATE RNP complexes to single strand breaks is increased in a range between any of the lower limits and any of the upper limits disclosed earlier (for example, between 5% and 100%, or between 10% and 50%) as compared to a CREATE system without the additional exogenous DNA binding domain. The recruitment of CREATE RNP complexes to single strand breaks can be measured by chromatin immunoprecipitation (ChIP), or other method for detecting DNA binding activities known in the art.

[0368] In some embodiments, the construct comprises a nucleic acid sequence encoding a peptide comprising (i) an ORF2 lacking a functional endonuclease domain and (ii) an additional exogenous DNA binding domain, wherein the ORF2 lacking a functional endonuclease domain is operably linked to the additional exogenous DNA binding domain. In some embodiments, the additional exogenous DNA binding domain is derived from a Sso7d protein. In some embodiments, the additional exogenous DNA binding domain is a Sso7 DNA binding domain. In some embodiments, the ORF2 lacking a functional endonuclease domain is an ORF2 lacking an endonuclease domain. In some embodiments, the ORF2 lacking an endonuclease domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 151**. In some embodiments, the Sso7d DNA binding domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 152**. In some embodiments, the additional DNA binding domain is fused to the ORF2 lacking a functional endonuclease domain. In some embodiments, the additional exogenous DNA binding domain is operably linked to the ORF2 lacking a functional endonuclease domain via a linker. In some embodiments, the linker is a flexible linker. In some embodiments, the flexible linker is a 33XTEN linker. In some embodiments, the 33XTEN linker comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at

least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 153**. In some embodiments the nucleic acid sequence encoding the ORF2 lacking a functional endonuclease domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 143**. In some embodiments, the nucleic acid sequence encoding the Sso7d DNA binding domain is codon optimized. In some embodiments, the nucleic acid sequence encoding the Sso7d DNA binding domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 141**. In some embodiments, the nucleic acid sequence encoding the 33XTEN linker comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 142**.

[0369] In some embodiments, the construct comprises a nucleic acid sequence encoding a peptide comprising (i) an ORF2 lacking a functional endonuclease domain and (ii) an additional exogenous DNA binding domain, wherein the ORF2 lacking a functional endonuclease domain is operably linked to the additional exogenous DNA binding domain. In some embodiments, the additional exogenous DNA binding domain is derived from a High Mobility Group Protein D (HMGD) protein. In some embodiments, the additional exogenous DNA binding domain is an HMGD domain. In some embodiments, the ORF2 lacking a functional endonuclease domain is an ORF2 lacking an endonuclease domain. In some embodiments, the ORF2 lacking a functional endonuclease domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 151**. In some embodiments, the HMGD domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 154**. In some embodiments, the HMGD domain is fused to the ORF2 lacking a functional endonuclease domain. In some embodiments, the HMGD domain is operably linked to the ORF2 lacking a functional endonuclease domain via a linker. In some embodiments, the linker is a flexible linker. In some embodiments, the flexible linker is a 33XTEN linker. In some embodiments,

the 33XTEN linker comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 153**. In some embodiments the nucleic acid sequence encoding the ORF2 lacking a functional endonuclease domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 143**. In some embodiments, the nucleic acid sequence encoding the HMGD domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 145**. In some embodiments, the nucleic acid sequence encoding the 33XTEN linker comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 142**.

[0370] In some embodiments, the construct comprises a nucleic acid sequence encoding a peptide comprising (i) an ORF2 lacking a functional endonuclease domain and (ii) an additional exogenous DNA binding domain, wherein the ORF2 lacking a functional endonuclease domain is operably linked to the additional exogenous DNA binding domain. In some embodiments, the additional exogenous DNA binding domain is derived from a human Poly(ADP-ribose) polymerase 1 (hPARP1) protein. In some embodiments, the additional exogenous DNA binding domain comprises the ZnF1-ZnF2 domains of PARP-1. In some embodiments, the ORF2 lacking a functional endonuclease domain is an ORF2 lacking an endonuclease domain. In some embodiments, the ORF2 lacking a functional endonuclease domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 151**. In some embodiments, the hPARP1-derived DNA binding domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 155**. In some embodiments, the hPARP1-derived DNA binding domain is fused to the ORF2 lacking a functional endonuclease domain. In some embodiments, the hPARP1-derived DNA binding domain is operably linked to the ORF2 lacking a functional endonuclease domain via a linker. In some embodiments, the linker is a flexible linker. In

some embodiments, the flexible linker is a 33XTEN linker. In some embodiments, the 33XTEN linker comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 153**. In some embodiments the nucleic acid sequence encoding the ORF2 lacking a functional endonuclease domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 143**. In some embodiments, the nucleic acid sequence encoding the hPARP1-derived DNA binding domain comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 145**. In some embodiments, the nucleic acid sequence encoding the 33XTEN linker comprises a sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to **SEQ ID NO: 142**.

[0371] In some embodiments, the construct comprises a nucleic acid sequence encoding a nuclear localization sequence with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to PAAKRVKLD (**SEQ ID NO: 15**). In some embodiments, the nuclear localization sequence is fused to the ORF2p sequence. In some embodiments, the construct comprises a nucleic acid sequence encoding a flag tag having the sequence DYKDDDDK (**SEQ ID NO: 16**). In some embodiments, the flag tag is fused to the ORF2p sequence. In some embodiments, the flag tag is fused to the nuclear localization sequence.

[0372] In some embodiments, the construct comprises a nucleic acid sequence encoding an MS2 coat protein with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to

ASNFTQFVLVDNNGGTGDVTVPSPNFANGIAEWISSNSRSQAYKVTCSVRQSSAQNRKYTIK
VEVPKGAWRSYLNMEITIPFATNSDCELVKAMQGLLDGNIPIPSAIAANSIGIYAMASNFT
QFVLVDNNGGTGDVTVPSPNFANGIAEWISSNSRSQAYKVTCSVRQSSAQNRKYTIKVEVPK

GAWRSYLNMEITIPFATNSDCELVKAMQGLLKDGNPIPSAIAANSIY (SEQ ID NO: 17).

In some embodiments, the MS2 coat protein sequence is fused to the ORF2p sequence.

[0373] In some embodiments, the transgene may comprise a flanking sequence which comprises an Alu ORF2p recognition sequence.

[0374] In some embodiments, additional elements may be introduced into the mRNA. In some embodiments, the additional elements may be an IRES element or a T2A element. In some embodiments, the mRNA transcript comprises one, two, three or more stop codons at the 3'-end.

[0375] In some embodiments, the one, two, three or more stop codons are designed to be in tandem. In some embodiments, the one, two, three or more stop codons are designed to be in all three reading frames. In some embodiments, the one, two, three or more stop codons may be designed to be both in multiple reading frames and in tandem.

[0376] In some embodiments, one or more target specific nucleotides may be added at the priming end of the L1 or the Alu RNA priming region.

[0377] In some embodiments, the 5' UTR sequence or the 3' UTR sequence in addition to be able to bind the ORF protein may also be capable of binding to one or more endogenous proteins that regulate gene retrotransposition and/or stable integration. In some embodiments, the flanking sequence is capable of binding to a PABP protein.

[0378] In some embodiments, the 5' region flanking the transgene may comprise a strong promoter. In some embodiments, the promoter is a CMV promoter. In some embodiments, the promoter is EF1. In some embodiments, the transgene (e.g., also variously designated as an exogenous polynucleic acid, a heterologous polynucleic acid, a polynucleic acid sequence comprising a therapeutic gene or polynucleic acid encoding a therapeutic polypeptide and grammatical equivalents thereof) may be placed in the retrotransposition construct as a reverse complement sequence of the sequence encoding the therapeutic polypeptide. In some embodiments, a sequence that is the reverse complement sequence of a promoter is incorporated in the construct, downstream of the reverse complement sequence of the sequence encoding the therapeutic polypeptide, such that when the sequences are further reverse complemented, the sequence encoding the polypeptide is operably linked to the promoter, and its expression is driven by the promoter.

[0379] In some embodiments, an additional nucleic acid encoding L1 ORF2p is introduced into the cell. In some embodiments, the sequence encoding L1 ORF1 is omitted, and only L1-ORF2 is included. In some embodiments, the nucleic acid encoding the transgene with the flanking elements is mRNA. In some embodiments, the endogenous L1-ORF1p function may be suppressed or inhibited.

[0380] In some embodiments, the nucleic acid encoding the transgene with the retrotransposition flanking elements comprise one or more nucleic acid modifications. In some embodiments, the nucleic acid encoding the transgene with the retrotransposition flanking elements comprises one or more nucleic acid modifications in the transgene. In some embodiments, the modifications comprise codon

optimization of the transgene sequence. In some embodiments, the codon optimization is for more efficient recognition by the human translational machinery, leading to more efficient expression in a human cell. In some embodiments, the one or more nucleic acid modification is performed in the 5'-flanking sequence or the 3'-flanking sequence including one or more stem-loop regions. the nucleic acid encoding the transgene with the retrotransposition flanking elements comprise one, two, three, four, five, six, seven eight, nine, ten or more nucleic acid modifications.

[0381] In some embodiments, the retrotransposed transgene is stably expressed for the life of the cell. In some embodiments, the cell is a myeloid cell. In some embodiments, the myeloid cell is a monocyte precursor cell. In some embodiments, the myeloid cell is an immature monocyte. In some embodiments, the monocyte is an undifferentiated monocyte. In some embodiments, the myeloid cell is a CD14⁺ cell. In some embodiments, the myeloid cell does not express CD16 marker. In some embodiments, the myeloid cell is capable of remaining functionally active for a desired period of greater than 3 days, greater than 4 days, greater than 5 days, greater than 6 days, greater than 7 days, greater than 8 days, greater than 9 days, greater than 10 days, greater than 11 days, greater than 12 days, greater than 13 days, greater than 14 days or more under suitable conditions. A suitable condition may denote an *in vitro* condition, or an *in vivo* condition or a combination of both.

[0382] In some embodiments, the retrotransposed transgene may be stably expressed in the cell for about 2 days, about 3 days, about 4 days, about 5 days, about 6 days, about 7 days, about 8 days, about 9 days or about 10 days. In some embodiments, the retrotransposed transgene is stably expressed in the cell for more than 10 days. In some embodiments, the retrotransposed transgene is stably expressed in the cell for more than 2 weeks. In some embodiments, the retrotransposed transgene is stably expressed in the cell for about 1 month.

[0383] In some embodiments, the retrotransposed transgene may be modified for stable expression. In some embodiments, the retrotransposed transgene may be modified for resistant to *in vivo* silencing.

[0384] In some embodiments, the expression of the retrotransposed transgene may be controlled by a strong promoter. In some embodiments, the expression of the retrotransposed transgene may be controlled by a moderately strong promoter. In some embodiments, the expression of the retrotransposed transgene may be controlled by a strong promoter that can be regulated in an *in vivo* environment. In some embodiments, the promoter is a CMV promoter. In some embodiments, the promoter is a L1-Ta promoter.

[0385] In some embodiments, the ORF1p may be overexpressed. In some embodiments, the ORF2 may be overexpressed. In some embodiments, the ORF1p or ORF2p or both are overexpressed. In some embodiments, upon overexpression of an ORF1, ORF1p is at least 1.1 fold, 1.5 fold, 2 fold, 3 fold, 4 fold, 5 fold, 6 fold, 7 fold, 8 fold, 9 fold, 10 fold, 12 fold, 14 fold, 16 fold, 18 fold, 20 fold, 30 fold, 40 fold, 50 fold, 60 fold, 70 fold, 80 fold, 90 fold, or at least 100 fold higher than a cell not overexpressing and ORF1.

[0386] In some embodiments, upon overexpression of an ORF2 sequence, ORF2p is at least 1.1 fold, 1.5 fold, 2 fold, 3 fold, 4 fold, 5 fold, 6 fold, 7 fold, 8 fold, 9 fold, 10 fold, 12 fold, 14 fold, 16 fold, 18 fold, 20 fold, 30 fold, 40 fold, 50 fold, 60 fold, 70 fold, 80 fold, 90 fold, or at least 100 fold higher than a cell not overexpressing and ORF2p.

[0387] In some embodiments, the exogenous sequence comprises a sequence encoding an exogenous polypeptide. In some embodiments, the sequence encoding an exogenous polypeptide is not in frame with a sequence encoding an endonuclease and/or a reverse transcriptase. In some embodiments, the sequence encoding an exogenous polypeptide is not in frame with a sequence encoding an endonuclease and/or a reverse transcriptase. In some embodiments, the exogenous sequence does not comprise introns. In some embodiments, the exogenous sequence comprises a sequence encoding an exogenous polypeptide selected from the group consisting of an enzyme, a receptor, a transport protein, a structural protein, a hormone, an antibody, a contractile protein and a storage protein. In some embodiments, the exogenous sequence comprises a sequence encoding an exogenous polypeptide selected from the group consisting of a chimeric antigen receptor (CAR), a ligand, an antibody, a receptor, and an enzyme. In some embodiments, the exogenous sequence comprises a regulatory sequence. In some embodiments, the regulatory sequence comprises a cis-acting regulatory sequence. In some embodiments, the regulatory sequence comprises a cis-acting regulatory sequence selected from the group consisting of an enhancer, a silencer, a promoter or a response element. In some embodiments, the regulatory sequence comprises a trans-acting regulatory sequence. In some embodiments, the regulatory sequence comprises a trans-acting regulatory sequence that encodes a transcription factor.

[0388] In some embodiments, the exogenous sequence encodes a polypeptide with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to one or more sequences presented in **Table 1**.

[0389] In some embodiments, the insert sequence integrates into the genome of a cell when introduced into the cell. In some embodiments, the insert sequence integrates into a gene associated a condition or disease, thereby disrupting the gene or downregulating expression of the gene. In some embodiments, the insert sequence integrates into a gene, thereby upregulating expression of the gene.

[0390] In some embodiments, the constructs described herein comprise one or more sequences with at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to one or more sequences presented in **Table 11**. In some embodiments, the constructs described herein comprise one or more sequences encoding one or more peptides with at least 80%, at least 81%, at

least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% sequence identity to one or more sequences presented in **Table 12**.

Homology Arm

[0391] The sequence that is a reverse complement of an exogenous sequence encoding a therapeutic polypeptide can be flanked by at least one homology arm at either 3' end or 5' end of the sequence. In some cases, the sequence that is a reverse complement of an exogenous sequence encoding a therapeutic polypeptide is flanked by two homology arms at both ends.

[0392] In some cases, the homology arm has a length of at least 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 bases. In some cases, the homology arm has a length of no more than 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, or 250 bases. In some cases, the homology arm has a length of about 10-300, 10-250, 10-200, 10-150, 10-100, 20-250, 20-200, 20-150, 20-100, 30-250, 30-200, 30-150, 30-100, 40-250, 40-200, 40-150, 40-100, 10-90, 10-80, 10-70, 10-60, 10-50, 20-90, 20-80, 20-70, 20-60, 20-50, 30-90, 30-80, 30-70, 30-60, or 30-50 bases. In some cases, the homology arm has a length of about 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, or 250 bases.

[0393] The first homology arm sequence and the second homology arm sequence may each be 20 nucleotides or less. The first homology arm sequence and the second homology arm sequence may each be 19 nucleotides or less. The first homology arm sequence and the second homology arm sequence may each be 18 nucleotides or less. The first homology arm sequence and the second homology arm sequence may each be about 17 nucleotides or less.

[0394] In some embodiments, the first and second homology arm sequence may each be about 20 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 25 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 30 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 35 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 40 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 45 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 50 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 55 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 60 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 65 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 70 nucleotides. In some embodiments, the first and second homology arm sequence may each be about 75 nucleotides.

[0395] The homology arm can have substantial sequence identity to the target DNA sequence. A target DNA sequence may be a sequence of 15, 16, 17, 18, 19, 20 or more nucleotides on a strand of genomic DNA of a target mammalian cell at or adjacent to, or adjoining a specific sequence. In some embodiments, the target DNA sequence may be a sequence in the safe harbor regions of the genome, e.g., the Rosa 36 locus, AAVS1 locus, or ribosomal DNA locus etc. In some cases, the homology arm is substantially identical to at least about 15, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, or 110 nucleotides of the target DNA sequence. In some cases, the homology arm is substantially identical to no more than about 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, or 250 nucleotides of the target DNA sequence. In some cases, the homology arm is substantially identical to between 35 to 120 nucleotides of the target DNA sequence. In some cases, the 5' homology arm is substantially identical to the target DNA sequence on one side of the cut created by the endonuclease disclosed herein and the 3' homology arm is substantially identical to the target DNA sequence on the other side of the cut created by the endonuclease disclosed herein.

[0396] As used herein, "substantially identical to" or "substantial identity" when referring to polynucleotide sequences of the homology arm means polynucleotide sequence identity of at least 40%, 50%, 60%, or 70%. Suitable polynucleotide identity can be any value between 40% and 100%, between 50% and 100%, between 60% and 100%, or between 70% and 100%. Preferably, polynucleotide identity of the homology arm is 100%.

[0397] The length of the homology arm (or the total length of the two homology arms) can be shorter than the sequence that is the reverse complement of an exogenous sequence. In some cases, the ratio of the length of the sequence that is the reverse complement of an exogenous sequence to the total length of the homology arms is about 1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1, 15:1, 20:1, 30:1, 40:1, 50:1, 60:1, 70:1, 80:1, 90:1, 100:1, 150:1, 200:1 or within a ratio range bounded by any two of these values (e.g., a ratio within a range of 2:1 to 200:1).

[0398] In some cases, the homology arm comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity of the target sequence disclosed herein. For instance, the 5' homology arm can comprise a sequence that is identical to the first target sequence disclosed herein, while the 3' homology arm can comprise a sequence that is identical to the second target sequence disclosed herein.

[0399] In some cases, the homology arm comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity of the target sequence disclosed herein. In some cases, the homology arm comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity of the reverse complement of the target sequence disclosed herein. For instance, the 5' homology arm can comprise a sequence that is identical to the reverse complement of the first target sequence disclosed herein, while the 3' homology arm can comprise a sequence that is identical to the second target sequence disclosed herein.

[0400] In some embodiments, the 5' homology arm comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity to a sequence of **SEQ ID NO: 49**. In some embodiments, the 5' homology arm comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity to a sequence of **SEQ ID NO: 51**. In some embodiments, the 5' homology arm comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity to a sequence of **SEQ ID NO: 53**.

[0401] In some embodiments, the 3' homology arm comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity to a sequence of **SEQ ID NO: 48**. In some embodiments, the 3' homology arm comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity to a sequence of **SEQ ID NO: 50**. In some embodiments, the 3' homology arm comprises a sequence with at least 80%, 85%, 90%, 95%, or 100% sequence identity to a sequence of **SEQ ID NO: 52**.

Modifications and enhancements of the retrotransposition-based CREATE gene editing system

[0402] In one aspect, it is an objective of the current work to further improve retrotransposition based genome editing machinery described herein to increase efficiency of genome integration of the exogenous payload sequence. In one embodiment, efficiency of genome integration may be considered as genomic integration of the exogenous payload sequence in a population of cells that are exposed to an event of gene editing using the CREATE gene editing system described herein. In some embodiments, the efficiency of genome integration of the exogenous payload sequence is measurable by the percentage of cells in a cell population that comprise in their genome the sequence encoding the exogenous payload sequence when the cell population is exposed to an event of gene editing using the CREATE gene editing system described herein. In some embodiments, the improvement in efficiency does not include an increase in copy numbers per cell; e.g., incorporation of a plurality of the exogenous payload sequence within the genome of a single cell of the cell population. In some embodiments, an increase in efficiency is considered to have been achieved when about 0.3%, 0.4%, 0.5%, 0.6%, 0.7%, 0.8%, 0.9% or 1% more cells in the population of cells express the exogenous payload sequence each from a single integration in the genome of the cell. In some embodiments, an increase in efficiency is considered to have been achieved when about 1.1%, 1.2%, 1.3%, 1.4%, 1.5%, 1.6%, 1.7%, 1.8%, 1.9% or about 2% more cells in the population of cells express the exogenous payload sequence from the specific site of integration in the genome of a cell. In some embodiments, the integration is at the specific site targeted in the cell, and no substantial off-target effects occur.

[0403] In some embodiments, provided herein is a composition comprising one or more RNA molecules with modifications in the CREATE system, wherein the CREATE system may comprise one, two or more RNA molecules, with, for example, one, two or more different retrotransposon elements, one or more different combinations of LINE1 retrotransposon elements (e.g., the LINE1 ORF1p, LINE1 ORF2p); one or more different combinations of exonucleases and reverse transcriptases, one or more mutated forms thereof, one or more functionally deactivated forms thereof, and one or more sequences enhanced for nuclear import (e.g., with

addition of NLS sequences at the N- or C-termini). In some embodiments, provided herein is a composition comprising one or more RNA molecules with modifications in the CREATE system, wherein the CREATE system may comprise one, two or more RNA molecules, which comprise a 5' methylated guanosyl cap structure. In some embodiments, translational start site of an RNA (e.g., mRNA) encoding an ORF2p polypeptide is designed to be within 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 nucleobases from the 5' cap structure. In some embodiments, modifications in the CREATE system may further comprise incorporation of one or more stop codons at the end of polypeptide coding sequences, for example, a sequence encoding ORF1p, or a sequence encoding ORF2p, etc. In some embodiments, modifications in the CREATE system may further comprise incorporation of one or more ribosomal entry sites for efficient translation of the encoded polypeptide. In some embodiments, modifications in the CREATE system may further comprise incorporation of one or more sequences encoding an NLS, at the N terminus, the C terminus or both at N- and C-terminus of polypeptide encoded therein, e.g., an ORF1p polypeptide, an ORF2p polypeptide or an endonuclease such as a Cas endonuclease. In some embodiments, modifications in the CREATE system may further comprise incorporation of one or more cleavable sequences, for example autocleavable P2A or T2A sequences, for example between sequences encoding two polypeptides, such as ORF1p and ORF2p, or Cas endonuclease from the same RNA molecule. In some embodiments, modifications in the CREATE system may further comprise incorporation of interORF regions, flexible linkers, dimerization domains or linkers with cleavable sequences between two adjoining sequences encoding two different polypeptides, such as ORF1p and ORF2p, or Cas endonuclease from the same RNA molecule.

[0404] Accordingly, provided herein are compositions comprising RNA molecules, for example, a composition, comprising (a) an RNA molecule comprising: (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; (ii) a 5' homology arm and a 3' homology arm; and (iii) a sequence encoding a modified human LINE1 sequence or fragment thereof, wherein the sequence encoding the modified human LINE1 sequence or fragment thereof is upstream of the 5' homology arm or downstream of the 3' homology arm; (b) an RNA molecule encoding an endonuclease, wherein the endonuclease is a nickase; (c) an RNA molecule comprising a sequence encoding a human ORF2p polypeptide comprising an endonuclease wherein the endonuclease is functionally deficient; and (d) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one

or more guide RNA molecules comprise a first guide RNA sequence and a second guide RNA sequence, wherein the first guide RNA sequence comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA sequence comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell. In some embodiments, the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF2p polypeptide, wherein the human ORF2p polypeptide comprises an endonuclease and a reverse transcriptase, wherein the endonuclease comprises a mutation. In some embodiments, the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF1p polypeptide upstream of the sequence encoding the human ORF2p polypeptide, and is separated by a LINE1 interORF sequence. In some embodiments, the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF1p polypeptide upstream of the sequence encoding the human ORF2p polypeptide and lacks a LINE1 interORF sequence. In some embodiments, the sequence encoding the human ORF1p polypeptide and the sequence encoding the human ORF2p polypeptide are separated by a linker sequence. In some embodiments, the sequence encoding the human ORF1p polypeptide and the sequence encoding the human ORF2p polypeptide are separated by a GSG linker sequence. In some embodiments, the sequence encoding the human ORF1p polypeptide and the sequence encoding the human ORF2p polypeptide are separated by a T2A cleavage sequence. In some embodiments, the sequence encoding the human ORF1p polypeptide and the sequence encoding the human ORF2p polypeptide are separated by a linker and a T2A cleavage sequence. In some embodiments, the sequence encoding the human ORF1p polypeptide and the sequence encoding the human ORF2p polypeptide are separated by a GSG linker and a T2A cleavage sequence. In some embodiments, the sequence encoding the modified human LINE1 sequence as well as the sequence encoding the human ORF2p polypeptide comprise a mutation in ORF2p endonuclease domains, wherein the mutation in the ORF2p endonuclease domains render the endonuclease functionally deficient. In some embodiments, none of the ORF2p endonucleases have endonuclease activity. In some embodiments, the RNA molecule comprising a sequence encoding a human ORF2p polypeptide comprises a 5'- methylated guanosyl cap structure (m7G cap). In some embodiments, the sequence encoding the human ORF2p polypeptide in the modified human LINE1 sequence comprises an N-terminal NLS and a C-terminal NLS. In some embodiments, the nickase is a functional nickase, and comprises an N-terminal NLS, a C-terminal NLS or both N-terminal and C-terminal NLS. In some embodiments, the sequence encoding the modified human LINE1 sequence or fragment thereof comprises one or more STOP codons at the end of the sequence encoding the human ORF2p polypeptide; the sequence encoding the human ORF1p

polypeptide or both. In some embodiments, any of the RNA molecules comprise one or more ribosomal entry sites. In some embodiments, any of the RNA molecules comprise one or more self-cleavage sites. In some embodiments, any of the RNA molecules comprise one or more oligomerization domains. In some embodiments, each of the RNA molecules comprise a poly A sequence at the 3' end.

Methods of Genomic Integration

[0405] Further disclosed herein are methods for genomic integration, comprising introducing any of the compositions disclosed herein to a target cell.

[0406] The methods for genomic integration disclosed herein can integrate the exogenous sequence to the genomic DNA with high specificity. For example, in some cases, the off-target integration rate is less than 1%, less than 2%, less than 3%, less than 4%, less than 5%, less than 6%, less than 7%, less than 8%, less than 9%, less than 10%, less than 15%, less than 20%, or less than 25%. In some cases, the on-target integration rate is at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, or at least 95%. The on-target/off target integration is measured by PCT or nano-sequencing.

[0407] The method for genomic integration disclosed herein can integrate the exogenous sequence to the genomic DNA with high efficiency. For example, in some cases, the genomic integration is at least 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, or 70%. In some cases, the genomic integration is about 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, or 70%, or any range between the values referenced above (for example, the genomic integration is about 5% - 55%). The integration efficiency is measured by flow cytometry.

[0408] In some cases, the method disclosed herein comprises two steps. In some case, the two-step method comprises: step (1) contacting the target mammalian cells with a first composition comprising: an RNA molecule comprising (i) an mRNA sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide, (ii) a 5' homology arm and a 3' homology arm, and (iii) a human mobile genetic element comprising a sequence encoding a polypeptide that promotes integration of the exogenous sequence encoding the therapeutic polypeptide into the genome of the target mammalian cell via target primed reverse transcription (TPRT); and step (2) contacting the target mammalian cells with a second composition comprising: a first guide RNA sequence a polynucleic acid encoding the same and a second guide RNA sequence a polynucleic acid encoding

the same. In some cases, the first composition further comprises an endonuclease or a polynucleic acid encoding the same. In some cases, the endonuclease is a Cas9 nickase.

[0409] In some cases, step (1) and step (2) occur at different time points. For example, step (1) can occur before or after step (2). In some cases, step (1) occurs before step (2). In some cases, step (2) occurs about 0.5 hour, about 1 hour, about 1.5 hours, about 2 hours, about 2.5 hours, about 3 hours, about 3.5 hours, about 4 hours, about 4.5 hours, about 5 hours, about 5.5 hours, about 6 hours, about 6.5 hours, about 7 hours, about 7.5 hours, or about 8 hours after step (1). In some cases, step (2) occurs about 4 hours after step (1). In some cases, step (2) occurs about 5 hours after step (1). In some cases, step (2) occurs about 6 hours after step (1).

[0410] In some cases, the method disclosed herein comprises three steps. In some case, the three-step method comprises: step (1) contacting the target mammalian cells with a first composition comprising: an RNA molecule comprising (i) an mRNA sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide, (ii) a 5' homology arm and a 3' homology arm, and (iii) a human mobile genetic element comprising a sequence encoding a polypeptide that promotes integration of the exogenous sequence encoding the therapeutic polypeptide into the genome of the target mammalian cell via target primed reverse transcription (TPRT); step (2) contacting the target mammalian cells with a second composition comprising: a first guide RNA sequence a polynucleic acid encoding the same and a second guide RNA sequence a polynucleic acid encoding the same; and step (3) contacting the target mammalian cells with an endonuclease or a polynucleic acid encoding the same. In some cases, the endonuclease is a Cas9 nickase.

[0411] All three steps can occur at the same time or at different time points. For example, step (1) and step (3) can occur at about the same time, but before step (2). In some cases, step (1) occurs first, followed by step (3), and further followed by step (2). In some cases, step (3) occurs first, followed by step (1), and further followed by step (2). The time interval between these steps can be about 0.5 hour, about 1 hours, about 2 hours, about 3 hours, about 4 hours, about 5 hours, about 6 hours, about 7 hours, or about 8 hours. In some cases, the time interval between step (1) and step (3) is about 0.1 hour, about 0.2 hour, about 0.3 hour, about 0.4 hour, about 0.5 hour, about 0.6 hour, about 0.7 hour, about 0.8 hour, about 0.9 hour, about 1 hour, about 1.1 hours, about 1.2 hours, about 1.3 hours, about 1.4 hours, about 1.5 hours, about 1.6 hours, about 1.7 hours, about 1.8 hours, about 1.9 hours, about 2.0 hours, about 2.1 hours, about 2.2 hours, about 2.3 hours, about 2.4 hours, or about 2.5 hours. In some cases, the time interval between step (1) and step (2) is about 0.5 hour, about 1 hour, about 1.5 hours, about 2 hours, about 2.5 hours, about 3 hours, about 3.5 hours, about 4 hours, about 4.5 hours, about 5 hours, about 5.5 hours, about 6 hours, about 6.5 hours, about 7 hours, about 7.5 hours, or about 8 hours.

[0412] In some cases, the method disclosed herein comprises four steps. In some case, the four-step method comprises: step (1) contacting the target mammalian cells with a first composition comprising:

an RNA molecule comprising (i) an mRNA sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide, (ii) a 5' homology arm and a 3' homology arm, and (iii) a human mobile genetic element comprising a sequence encoding a polypeptide that promotes integration of the exogenous sequence encoding the therapeutic polypeptide into the genome of the target mammalian cell via target primed reverse transcription (TPRT); step (2) contacting the target mammalian cells with a second composition comprising: a first guide RNA sequence a polynucleic acid encoding the same and a second guide RNA sequence a polynucleic acid encoding the same; step (3) contacting the target mammalian cells with a third composition comprising: a human mobile genetic element comprising a sequence encoding a polypeptide that promotes integration of the exogenous sequence encoding the therapeutic polypeptide into the genome of the target mammalian cell via target primed reverse transcription (TPRT); and step (4) contacting the target mammalian cells with an endonuclease or a polynucleic acid encoding the same. In some cases, the endonuclease is a Cas9 nickase. In some cases, steps (1)-(4) occur simultaneously.

[0413] In some embodiments, the human mobile genetic element sequence of the third composition and the human mobile genetic element sequence of the first composition are the same. In some embodiments, the human mobile genetic element sequence of the third composition and the human mobile genetic element sequence of the second composition are different. In some embodiments, the human mobile genetic element sequence of the first composition encodes any one of the ORF2 polypeptides described herein. In some embodiments, the human mobile genetic element sequence of the third composition encodes any one of the ORF2 polypeptides described herein. In some embodiments, the ORF2 polypeptide encoded by the human mobile genetic element sequence of the first composition lacks an endonuclease domain. In some embodiments, the ORF2 polypeptide encoded by the human mobile genetic element sequence of the third composition lacks an endonuclease domain. In some embodiments, the ORF2 polypeptide lacking an endonuclease domain comprises an additional exogenous DNA binding domain. In some embodiments, the additional DNA binding domain is derived from a Sso7d protein. In some embodiments, the additional DNA binding domain is an Sso7d DNA binding domain. In some embodiments, the additional DNA binding domain is derived from a HMGD protein. In some embodiments, the additional DNA binding domain is derived from a hPARP1 protein. In some embodiments, the ORF2 polypeptide is operably linked to the additional DNA binding domain via a linker. In some embodiments, the linker is a 33XTEN linker. All four steps can occur at the same time or at different time points.

[0414] In some cases, the ratio of the amount of guide RNAs (total amount of both guide RNAs) and the amount of RNA molecule used in the method disclosed herein is about 1:1, about 1:2, about 1:3, about 1:4, about 1:5, about 1:6, about 1:7, about 1:8, about 1:9, about 1:10, about 1:11, about 1:12, about 1:13, about 1:14, about 1:15, about 1:16, about 1:17, about 1:18, about 1:19, about 1:20, about 1:21, about 1:22, about 1:23, about 1:24, about 1:25, about 1:26, about 1:27, about 1:28, about 1:29,

or about 1:30. In some cases, the ratio is at least 1:3. In some cases, the ratio is at least 1:4. In some cases, the ratio is at least 1:5. In some cases, the ratio is at most 1:30. In some cases, the ratio is at most 1:20. In some cases, the ratio is at most 1:10. In some cases, the ratio is about 1:3. In some cases, the ratio is about 1:4. In some cases, the ratio is about 1:5. In some cases, the ratio is about 1:6. In some cases, the ratio is about 1:7. In some cases, the ratio is about 1:8. In some cases, the ratio is about 1:9. In some cases, the ratio is about 1:10.

Methods of Treatment

[0415] Also provided herein is a method of treating a disease in a subject, the method comprising administering any of the compositions or pharmaceutical compositions described herein. In some embodiments, the subject expresses the therapeutic polypeptide from a specific sequence of the genome, thereby treating the disease. In some embodiments, the disease is cancer. In some embodiments, the cancer is selected from the group consisting of T cell lymphoma, cutaneous lymphoma, a B cell cancer, multiple myeloma, Waldenstrom's macroglobulinemia, benign monoclonal gammopathy, immunocytic amyloidosis, melanoma, breast cancer, lung cancer, bronchus cancer, colorectal cancer, prostate cancer, metastatic hormone refractory prostate cancer, pancreatic cancer, stomach cancer, ovarian cancer, urinary bladder cancer, brain or central nervous system cancer, peripheral nervous system cancer, esophageal cancer, cervical cancer, uterine or endometrial cancer, cancer of the oral cavity or pharynx, liver cancer, kidney cancer, testicular cancer, biliary tract cancer, small bowel or appendix cancer, salivary gland cancer, thyroid gland cancer, adrenal gland cancer, osteosarcoma, chondrosarcoma, carcinomas, fibrosarcoma, myxosarcoma, liposarcoma, chondrosarcoma, osteogenic sarcoma, chordoma, angiosarcoma, endotheliosarcoma, lymphangiosarcoma, lymphangioendotheliosarcoma, synovioma, mesothelioma, Ewing's tumor, leiomyosarcoma, rhabdomyosarcoma, colon carcinoma, colorectal cancer, pancreatic cancer, breast cancer, ovarian cancer, squamous cell carcinoma, basal cell carcinoma, adenocarcinoma, sweat gland carcinoma, sebaceous gland carcinoma, papillary carcinoma, papillary adenocarcinomas, cystadenocarcinoma, medullary carcinoma, bronchogenic carcinoma, renal cell carcinoma, hepatoma, bile duct carcinoma, liver cancer, choriocarcinoma, seminoma, embryonal carcinoma, Wilms' tumor, cervical cancer, bone cancer, brain tumor, testicular cancer, lung carcinoma, small cell lung carcinoma, bladder carcinoma, epithelial carcinoma, glioma, astrocytoma, medulloblastoma, craniopharyngioma, ependymoma, pinealoma, hemangioblastoma, acoustic neuroma, oligodendroglioma, meningioma, melanoma, neuroblastoma, retinoblastoma; leukemias, e.g., acute lymphocytic leukemia and acute myelocytic leukemia, myeloblastic leukemia, promyelocytic leukemia, myelomonocytic leukemia, monocytic leukemia, erythroleukemia, chronic leukemia, chronic myelocytic, granulocytic leukemia, chronic lymphocytic leukemia, polycythemia vera, Hodgkin's lymphoma and non-Hodgkin's lymphoma, mantle cell lymphoma, multiple myeloma, Waldenstrom's macroglobulinemia, bladder cancer, cervical cancer, colon cancer, gynecologic cancers, renal cancer, laryngeal cancer, lung cancer,

oral cancer, head and neck cancer, ovarian cancer, pancreatic cancer, prostate cancer, brain cancer, lymphoma, leukemia, and skin cancer. In some embodiments, the cancer is a T cell cancer.

[0416] In some embodiments, the therapeutic polypeptide is a chimeric antigen receptor (CAR). In some embodiments, the therapeutic polypeptide is a T cell receptor (TCR). In some embodiments, the therapeutic polypeptide is a chimeric fusion protein (CFP) that is expressed in myeloid cells. In some embodiments, myeloid cells expressing the CFP exhibit enhanced phagocytosis of disease cells.

[0417] In some embodiments, the CFP comprises (i) an extracellular antigen binding domain, (ii) a transmembrane domain and (iii) an intracellular signaling domain. In some embodiments, the extracellular antigen binding domain binds to an antigen on a disease cell. In some embodiments, the disease cell is a cancer cell. In some embodiments, the extracellular antigen binding domain binds to a cancer antigen. In some embodiments, the cancer antigen is selected from a group consisting of Thymidine Kinase (TK1), Hypoxanthine-Guanine Phosphoribosyltransferase (HPRT), Receptor Tyrosine Kinase-Like Orphan Receptor 1 (ROR1), Mucin-1, Mucin-16 (MUC16), MUC1, Epidermal Growth Factor Receptor vIII (EGFRvIII), Mesothelin, Human Epidermal Growth Factor Receptor 2 (HER2), Mesothelin, EBNA-1, LEMD1, Phosphatidyl Serine, Carcinoembryonic Antigen (CEA), B-Cell Maturation Antigen (BCMA), Glypican 3 (GPC3), Follicular Stimulating Hormone receptor, Fibroblast Activation Protein (FAP), Erythropoietin-Producing Hepatocellular Carcinoma A2 (EphA2), EphB2, a Natural Killer Group 2D (NKG2D) ligand, Disialoganglioside 2 (GD2), CD2, CD3, CD4, CD5, CD7, CD8, CD19, CD20, CD22, CD24, CD30, CD33, CD38, CD44v6, CD45, CD56CD79b, CD97, CD117, CD123, CD133, CD138, CD171, CD179a, CD213A2, CD248, CD276, PSMA, CS-1, CLECL1, GD3, PSMA, FLT3, TAG72, EPCAM, IL-1, an integrin receptor, PRSS21, VEGFR2, PDGFR- β , SSEA-4, EGFR, NCAM, prostase, PAP, ELF2M, GM3, TEM7R, CLDN6, TSHR, GPRC5D, ALK, IGLL1 and combinations thereof. In some embodiments, the cancer antigen is an ovarian cancer antigen. In some embodiments, the cancer antigen is a T lymphoma antigen. In some embodiments, the transmembrane domain comprises a sequence from the transmembrane domain of FcR- α . In some embodiments, the transmembrane domain comprises a sequence from the transmembrane domain of FcR β . In some embodiments, the transmembrane domain comprises a sequence from the transmembrane domain of FcR ϵ . In some embodiments, the transmembrane domain comprises a sequence from the transmembrane domain of FcR γ . In some embodiments, the intracellular signaling domain comprises a PI3 Kinase recruitment domain. In some embodiments, the therapeutic polypeptide activates immune response of the subject to the cancer, thereby treating the disease.

[0418] In some embodiments, the method further comprises administering one or more additional therapeutic or helper components. In some embodiments, the method comprises administering an inhibitor of an inhibitor of LINE1 retrotransposition. In some embodiments, the helper element is an SAMHD1 inhibitor. In some embodiments, the helper element is an inhibitory nucleic acid. In some

embodiments, the inhibitory nucleic acid is an siRNA. In some embodiments, the inhibitory nucleic acid is a miRNA. In some embodiments, the mRNA comprises unmodified nucleotides.

[0419] In one aspect, the instant methods and systems may be useful in treating genetic diseases. It may be understood that about 30% of such diseases are caused by transition point mutation, about 20% result from transversion point mutations, approximately 26% from deletion mutations, about 8.2% from gene duplication, and others from copy number gain or loss, insertion mutations, insertion deletion (indel) mutations and others. Moreover, most of the human pathogenic alleles cannot be treated by current genetic approaches. Base editing and prime editing are capable of as much as a few bases, less than about 40 bases in the genome. Attempts are also undertaken by many others to generate site-specific incisions to prepare a landing pad for incorporating a gene segment that may be longer than the existing capabilities of base and prime editing technologies. However, among other drawbacks, these attempts suffer from introducing multiple gene manipulation steps which evokes concerns related genetic stability and safety issues.

[0420] In some embodiments, the methods, compositions, and systems described herein are used in treating monogenic disorders. Monogenic disorders of single-gene diseases, a mutation in a single gene is responsible for the disease. In several cases, such as cystic fibrosis and Duchenne Muscular dystrophy, there can be several modifiers that contribute to the disease, however, there still exists a single gene disfunction that would lead to the disease in the subject. Candidate diseases that may fall under this category and may be treated using the methodologies may include, but are not limited to phenylketonuria (PKU), in which the gene responsible is phenylalanine hydrolase (PAH); Hemophilia (the gene responsible is coagulation factor 8); sickle cell anemia (the gene responsible is beta hemoglobin). Muscular dystrophy (Dystrophin gene) or Huntington's disease (Huntingtin gene). In some embodiments, the methods, compositions, and systems described herein may be used for developing enzyme replacement therapy (ERT), for example for lysosomal storage diseases (e.g., Gaucher's disease, Fabry's disease, Hunter's Disease or Pompe's disease). Major concerns with current technology for ERT is durable expression of a therapeutic enzyme at a therapeutically effective amount at a constant rate. Currently there are no effective technologies for genome editing large segments of DNA ranging from hundred to several thousand bases, such that the edited gene segment would be capable of producing a steady level of the encoded enzyme. The retrotransposon-based system is non-immunogenic, and has the capability of delivering and expressing large payload, as well as enable constant prolonged expression, and therefore meets the criteria for therapeutic development in this field in need. Similarly, rare diseases can often be devastating to patients and their families, for which remedies may be devised using the retrotransposon technology disclosed herein. For example, the involvement of C9orf72 in Angelman's syndrome, facioscapulohumeral muscular dystrophy (FHMD), Spinal Muscular Atrophy (SMA), and Amyotrophic Lateral Sclerosis (ALS) and familial Frontotemporal dementia (FTD) are all diseases that may have life-long effects, such as mental

retardation (Angelman syndrome), cognitive deficits (e.g., FTD), and/or muscle weakness (FHMD, SMA, and ALS).

[0421] In some embodiments, the genetically inherited disease is Meier-Gorlin syndrome. In some embodiments, the genetically inherited disease is Seckel syndrome 4. In some embodiments, the genetically inherited disease is Joubert syndrome 5. In some embodiments, the genetically inherited disease is Leber congenital amaurosis. In some embodiments, the genetically inherited disease is Charcot-Marie-Tooth disease, type 2. In some embodiments, the genetically inherited disease is leukoencephalopathy. In some embodiments, the genetically inherited disease is Usher syndrome, type 2C. In some embodiments, the genetically inherited disease is spinocerebellar ataxia 28. In some embodiments, the genetically inherited disease is glycogen storage disease type III. In some embodiments, the genetically inherited disease is primary hyperoxaluria, type I. In some embodiments, the genetically inherited disease is long QT syndrome 2. In some embodiments, the genetically inherited disease is Sjogren-Larsson syndrome. In some embodiments, the genetically inherited disease is hereditary fructosuria. In some embodiments, the genetically inherited disease is neuroblastoma. In some embodiments, the genetically inherited disease is amyotrophic lateral sclerosis type 9. In some embodiments, the genetically inherited disease is Kalimam syndrome 1. In some embodiments, the genetically inherited disease is limb-girdle muscular dystrophy, type 2L. In some embodiments, the genetically inherited disease is familial adenomatous polyposis 1. In some embodiments, the genetically inherited disease is familial type 3 hyperlipoproteinemia. In some embodiments, the genetically inherited disease is Alzheimer's disease, type 1. In some embodiments, the genetically inherited disease is metachromatic leukodystrophy. In some embodiments, the genetically inherited disease is cancer. In some embodiments, the genetically inherited disease is Uveitis. In some embodiments, the genetically inherited disease is SCA1. In some embodiments, the genetically inherited disease is SCA2. In some embodiments, the genetically inherited disease is FUS-Amyotrophic Lateral Sclerosis (ALS). In some embodiments, the genetically inherited disease is MAPT-Frontotemporal Dementia (FTD). In some embodiments, the genetically inherited disease is Myotonic Dystrophy Type 1 (DM1). In some embodiments, the genetically inherited disease is Diabetic Retinopathy (DR/DME). In some embodiments, the genetically inherited disease is Oculopharyngeal Muscular Dystrophy (OPMD). In some embodiments, the genetically inherited disease is SCAB. In some embodiments, the genetically inherited disease is C9ORE72-Amyotrophic Lateral Sclerosis (ALS). In some embodiments, the genetically inherited disease is SOD1-Amyotrophic Lateral Sclerosis (ALS). In some embodiments, the genetically inherited disease is SCA6. In some embodiments, the genetically inherited disease is SCA3 (Machado-Joseph Disease). In some embodiments, the genetically inherited disease is Multiple system Atrophy (MSA). In some embodiments, the genetically inherited disease is Treatment-resistant Hypertension. In some embodiments, the genetically inherited disease is Myotonic Dystrophy Type 2 (DM2). In some

embodiments, the genetically inherited disease is Fragile X-associated Tremor Ataxia Syndrome (FXTAS). In some embodiments, the genetically inherited disease is West Syndrome with ARX Mutation. In some embodiments, the genetically inherited disease is Age-related Macular Degeneration (AMD)/Geographic Atrophy (GA). In some embodiments, the genetically inherited disease is C9ORF72-Frontotemporal Dementia (FTD). In some embodiments, the genetically inherited disease is Facioscapulohumeral Muscular Dystrophy (FSHD). In some embodiments, the genetically inherited disease is Fragile X Syndrome (FXS). In some embodiments, the genetically inherited disease is Huntington's Disease. In some embodiments, the genetically inherited disease is Glaucoma. In some embodiments, the genetically inherited disease is Acromegaly. In some embodiments, the genetically inherited disease is Achromatopsia (total color blindness). In some embodiments, the genetically inherited disease is Ullrich congenital muscular dystrophy. In some embodiments, the genetically inherited disease is Hereditary myopathy with lactic acidosis. In some embodiments, the genetically inherited disease is X-linked spondyloepiphyseal dysplasia tarda. In some embodiments, the genetically inherited disease is Neuropathic pain (Target: CPEB). In some embodiments, the genetically inherited disease is Persistent Inflammation and injury pain (Target: PABP). In some embodiments, the genetically inherited disease is Neuropathic pain (Target: miR-30c-5p). In some embodiments, the genetically inherited disease is Neuropathic pain (Target: miR-195). In some embodiments, the genetically inherited disease is Friedreich's Ataxia. In some embodiments, the genetically inherited disease is Uncontrolled gout. In some embodiments, the genetically inherited disease is Inflammatory pain (Target: Nav1.7 and Nav1.8). In some embodiments, the genetically inherited disease is Choroideremia. In some embodiments, the genetically inherited disease is Focal epilepsy. In some embodiments, the genetically inherited disease is Alpha-1 Antitrypsin deficiency (AATD). In some embodiments, the genetically inherited disease is Androgen Insensitivity Syndrome. In some embodiments, the genetically inherited disease is Opioid-induced hyperalgesia (Target: Raf-1). In some embodiments, the genetically inherited disease is Neurofibromatosis type 1. In some embodiments, the genetically inherited disease is Stargardt's Disease. In some embodiments, the genetically inherited disease is Dravet Syndrome. In some embodiments, the genetically inherited disease is Retinitis Pigmentosa. In some embodiments, the genetically inherited disease is Hemophilia A (factor VIII). In some embodiments, the genetically inherited disease is Hemophilia B (factor IX). In some embodiments, the genetically inherited disease is Parkinson's Disease.

[0422] **Table 1.** Exemplary therapeutic polypeptide sequences. Underlined sequences designate the CDRs (successively CDR1, 2 and 3) in Kabat system.

CFP /domain	Sequence	SEQ ID NO:
Anti-CD5 _{scF} v-Fcγ-PI3K CFP	MWLQSLLLLGTVACSISEIQLVQSGGGLVKPGGSVRISCAASGYTFT <u>NYGMN</u> WVRQAPGKGLEWMGWINTHTGEPTYADSFKGRFTFSLDD SKNTAYLQINSLRAEDTAVYFCTRRGYDWYFDVWGQGTTVTVSSG GGSGGGGGSGGGGSDIQMTQSPSSLSASVGDRVITITCRASQDINSYL <u>SWFQQKPGKAPKTLIYRANRLES</u> GVPSRFSGSGSGTDYTLTISSLQY EDFGIYYCQQYDESPWTFGGGKLEIKSGGGGSGALSNSIMYFSHFV PVFLPAKPTTTPAPRPPTPAPTIASQPLSLRPEACRPAAGGAVHTRGL DIYIWAPLAGTCGVLLLSLVITLYCRLKIQVRKAAITSYEKSDGVYTG LSTRNQETYETLKHEKPPQGSYSYEDMRGILYAAPQLRSIRGQPGPN HEEDADSYENM	18
Anti-CD5 _{scF} v	EIQLVQSGGGLVKPGGSVRISCAASGYTFTNYGMN <u>WVRQAPGKGL</u> EWMGWINTHTGEPTYADSFKGRFTFSLDDSKNTAYLQINSLRAEDT AVYFCTRRGYDWYFDVWGQGTTVTVSSGGGGSGGGGSGGGGSDI QMTQSPSSLSASVGDRVITITCRASQDINSYLSWFQQKPGKAPKTLIY <u>RANRLES</u> GVPSRFSGSGSGTDYTLTISSLQYEDFGIYYCQQYDESPW TFGGGKLEIK	19
Anti-CD5 VH	EIQLVQSGGGLVKPGGSVRISCAASGYTFTNYGMN <u>WVRQAPGKGL</u> EWMGWINTHTGEPTYADSFKGRFTFSLDDSKNTAYLQINSLRAEDT AVYFCTRRGYDWYFDVWGQGTTVTVSS	20
Anti-CD5 VL	DIQMTQSPSSLSASVGDRVITITCRASQDINSYLSWFQQKPGKAPKTLI <u>YRANRLES</u> GVPSRFSGSGSGTDYTLTISSLQYEDFGIYYCQQYDESP WTFGGGKLEIK	21
Anti-CD5 _{scF} v- FcRγ-CD40 CFP	MWLQSLLLLGTVACSISEIQLVQSGGGLVKPGGSVRISCAASGYTFTN <u>YGMN</u> WVRQAPGKGLEWMGWINTHTGEPTYADSFKGRFTFSLDDSK NTAYLQINSLRAEDTAVYFCTRRGYDWYFDVWGQGTTVTVSSGGGG SGGGGSGGGGSDIQMTQSPSSLSASVGDRVITITCRASQDINSYLSWFQ QKPGKAPKTLIYRANRLES <u>GVPSRFSGSGSGTDYTLTISSLQYEDFGIY</u> <u>YCQQYDESPWTFGGGKLEIKSGGGGSGALSNSIMYFSHFVPVFLPA</u> PTTTPAPRPPTPAPTIASQPLSLRPEACRPAAGGAVHTRGLDIYIWAP AGTCGVLLLSLVITLYCRLKIQVRKAAITSYEKSDGVYTG LSTRNQETYETLKHEKPPQKKVAKKPTNKAPHPKQEPQEINFPDDLPGSNTAAPV QETLHGCQPVTQEDGKESRISVQERQ	22
Anti-CD5 _{scF} v- FcRγ-MDA5	MWLQSLLLLGTVACSISEIQLVQSGGGLVKPGGSVRISCAASGYTFTN <u>YGMN</u> WVRQAPGKGLEWMGWINTHTGEPTYADSFKGRFTFSLDDSK NTAYLQINSLRAEDTAVYFCTRRGYDWYFDVWGQGTTVTVSSGGGG SGGGGSGGGGSDIQMTQSPSSLSASVGDRVITITCRASQDINSYLSWFQ QKPGKAPKTLIYRANRLES <u>GVPSRFSGSGSGTDYTLTISSLQYEDFGIY</u> <u>YCQQYDESPWTFGGGKLEIKSGGGGSGALSNSIMYFSHFVPVFLPA</u> KPTTTPAPRPPTPAPTIASQPLSLRPEACRPAAGGAVHTRGLDIYIWAP LAGTCGVLLLSLVITLYCRLKIQVRKAAITSYEKSDGVYTG LSTRNQETYETLKHEKPPQGSMSNGYSTDENFRYLISCFRARKMYIQVEPV LDYLTFLPAEVKEQIQRTVATSGNMQAVELLSTLEKGVWHLGWTR EFVEALRRTGSPLAARYMNPETDLPSPSFENAHDEYLQLLNLLQPTL VDKLLVRDVLDKCMEELLTIEDRNRIAAENNGNESGVRELLKRIV QKENWFS AFLNVL RQTGNNELVQELTGSDCSESNAEIEN	23
Anti-HER2 CFP	MWLQSLLLLGTVACSIDIQMTQSPSSLSASVGDRVITITCRASQDVNT <u>AVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSRSGTDFTLTISLQP</u> EDFATYYCQQHYTTPPTFGQGKVEIKRTGSTSGSGKPGSGEGSEVQL	24

CFP /domain	Sequence	SEQ ID NO:
	VESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDVWGQGTLLTVSSSSGGGGSGAALSNSIMYFSHFVPVFLPAKPTTTPAPRPPTPAPTIASQPLSLRPEACRPAAGGAVHTRGLDIYIWAFLAGTCGVLLLSLVITLYCKKVAKKPTNKAPHPKQEPQEIFPDDLPGSNTAAPVQETLHGCQPVTQEDGKESRISVQERQGTRLKIQVRKAAITSYEKSDGVYTGLSTRNQETIETLKHEKPPQGSYEDMRGILYAAPQLRSIRGQPGPNHEEDADSYENM	
Anti-HER2 ScFv-VH	EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDVWGQGTLLTVSS	25
Anti-HER2 ScFv-VL	DIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGRSGTDFTLTISLQPEDFATYYCQQHYTTPPTFGQGTKVEIK	26
Anti-HER2 ScFv	DIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGRSGTDFTLTISLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTGSTSGSGKPGSGEGSEVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDVWGQGTLLTVSS	27
Anti-HER2 scFv	DIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGRSGTDFTLTISLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTGSTSGSGKPGSGEGSEVQLVESGGGGSGGGGSGGGSLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDVWGQGTLLTV	28
Anti-GPC3 ScFv	QVQLVQSGAEVKKPGASVKVSCKASGYTFTDYEMHWVRQAPGQGLEWMGALDPKTGDTAYSQKFKGKATLTADKSTSTAYMELSSLTSEDTA VYYCTRFYSYTYWGQGTLLTVSSGGGGSGGGGSGGGGSDVMTQSPLSLPVTGPGEPAISCRSSQSLVHSNRNTYLHWYLQKPGQSPQLLIYK VSNRFSGVPDRFSGSGTDFTLKISRVEAEDVGVYYCSQNTHPPTFGQGTKLEIK	29
Anti-GPC3 binder (1)	MWLQSLLLLGTVAACISQVQLVQSGAEVKKPGASVKVSCKASGYTFDYEMHWVRQAPGQGLEWMGALDPKTGDTAYSQKFKGKATLTADKSTSTAYMELSSLTSEDTA VYYCTRFYSYTYWGQGTLLTVSSGGGGSGGGGSGGGGSDVMTQSPLSLPVTGPGEPAISCRSSQSLVHSNRNTYLHWYLQKPGQSPQLLIYK VSNRFSGVPDRFSGSGTDFTLKISRVEAEDVGVYYCSQNTHPPTFGQGTKLEIKGSGGSDSIHQDYTTQNLIRMAVAGLVLVALLAILVENWHSHTALNKEASADV AEPWSQQMCQPGLTFA RTPSVCK	30
Anti-GPC3 Binder (2)	MWLQSLLLLGTVAACISQVQLVQSGAEVKKPGASVKVSCKASGYTFDYEMHWVRQAPGQGLEWMGALDPKTGDTAYSQKFKGRVTLTADKSTSTAYMELSSLTSEDTA VYYCTRFYSYTYWGQGTLLTVSSGGGGSGGGGSGGGGSDVMTQSPLSLPVTGPGEPAISCRSSQSLVHSNRNTYLHWYLQKPGQSPQLLIYK VSNRFSGVPDRFSGSGTDFTLKISRVEAEDVGVYYCSQNTHPPTFGQGTKLEIKGSGGSDSIHQDYTTQNLIRMAV	31

CFP /domain	Sequence	SEQ ID NO:
	AGLVLVALLAILVENWHSHTALNKEASADVAEPSWSQQMCQPGLTFA RTPSVCK	
Anti-TROP2 scFv-VH	QVQLQQSGSELKKPGASVKVSCKASGYTFTNYGMN WVKQAPGQGL KWMGWINTYTGEPTYTDDFKGRFAFSLDTSVSTAYLQISSLKADDTA VYFCARGGFGSSYWYFDVWGQGS�TVSS	32
Anti-TROP2 scFv-VL	DIQLTQSPSSLSASVGDRVSI TC KASQDV SI AVAWYQQKPGKAPKLLI Y S ASYRY T GVPDRFSGSGSGTDFTLTISSLQPEDFAVYYC Q QHYIT PLT FGAGTKVEIKR	33
Anti-TROP2 CFP (1)	MWLQSLLLLGTVAC SISQVQLQQSGSELKKPGASVKVSCKASGYTFT NYGMN WVKQAPGQGLKWMGWINTYTGEPTYTDDFKGRFAFSLDTS VSTAYLQISSLKADDTAVYFCARGGFGSSYWYFDVWGQGS�TVSS GGGGSGGGSGGGGSDIQLTQSPSSLSASVGDRVSI TC KASQDV SI AV A WYQQKPGKAPKLLI Y SASYRY T GVPDRFSGSGSGTDFTLTISSLQPE DFAVYYC Q QHYIT PLT FGAGTKVEIKRSGGSDSIHQDYTTQNLIRM AVAGLVLVALLAILVENWHSHTALNKEASADVAEPSWSQQMCQPGL TFARTPSVCK	34
Anti-TROP2 CFP (2)	MWLQSLLLLGTVAC SISQVQLQQSGSELKKPGASVKVSCKASGYTFT NYGMN WVKQAPGQGLKWMGWINTYTGEPTYTDDFKGRFAFSLDTS VSTAYLQISSLKADDTAVYFCARGGFGSSYWYFDVWGQGS�TVSS GGGGSGGGSGGGGSDIQLTQSPSSLSASVGDRVSI TC KASQDV SI AV A WYQQKPGKAPKLLI Y SASYRY T GVPDRFSGSGSGTDFTLTISSLQPE DFAVYYC Q QHYIT PLT FGAGTKVEIKRSGGGGAAAGSDSIHQDYTTQ NLIRMAVAGLVLVALLAILVENWHSHTALNKEASADVAEPSWSQQM CQPGLTFARTPSVCK	35
Anti-GPC3 scFv VH	QVQLVQSGAEVKKPGASVKVSCKASGYTFTDYEMHWVRQAPGQGL EWMGALDPKTGDTAYSQKF K GRVTLTADKSTSTAYMELSSLTSED T AVYYCTRFYSYTYWGQGLTVSS	165
Anti-GPC3 scFv VL	DVVM T QSP L SLPVTPGEPASISCRSSQSLVHSNRNTYLHWYLQKPGQS PQLLIYK V SNRFS G VPDRFSGSGSGTDFTLKISRVEAEDVGVYYC SQ N THVPPTFGQGTKLEIK	166
Anti-GPC3 scFv	QVQLVQSGAEVKKPGASVKVSCKASGYTFTDYEMHWVRQAPGQGL EWMGALDPKTGDTAYSQKF K GRVTLTADKSTSTAYMELSSLTSED T AVYYCTRFYSYTYWGQGLTVSSGGGGSGGGSGGGGSDVVM T Q SPLSLPVTPGEPASISCRSSQSLVHSNRNTYLHWYLQKPGQSPQLLIYK V SNRFS G VPDRFSGSGSGTDFTLKISRVEAEDVGVYYC SQ N THVPPTF GQGTKLEIK	167

EXAMPLES

[0423] It should be appreciated that the invention should not be construed to be limited to the examples which are now described; rather, the invention should be construed to include any and all applications provided herein and all equivalent variations within the skill of the ordinary artisan.

Example 1. Exemplary Constructs for Genomic Integration

[0424] mRNA constructs are generated to direct CRISPR-guided Retroelement Activation and Editing (CREATE) comprising, from 5' to 3', sequences encoding ORF1, endonuclease deactivated ORF2, a 5' homology arm, a sequence that is the reverse complement for a GFP reporter payload under the control of the CMV promoter, and a 3' homology arm (**FIG. 1**). Other components include a Cas9 nickase bearing an H840A mutation and two guide RNAs specific for a safe harbor locus such as AAVS1. When the construct is delivered to a cell and expressed, the two guide RNAs direct single strand nicking at two different sites within the AAVS1 region. Then the 3' homology arm directs first strand synthesis by the ORF2p target primed reverse transcription. The 5' homology arm similarly directs second strand synthesis. As the 5' and 3' ends meet, synthesis is terminated and results in integration of only the GFP reporter payload gene and its promoter (**FIG. 2**).

Example 2. Expression of Genomically Integrated Payload Sequence After Construct Mediated Transposition

[0425] An mRNA construct was generated comprising, from 5' to 3', sequences encoding ORF1, endonuclease deactivated ORF2, a 5' homology arm matching safe harbor site AAVS1, a sequence that is the reverse complement for a GFP reporter payload under the control of the CMV promoter, and a 3' homology arm matching safe harbor site AAVS1. The construct was delivered in a lipid nanoparticle to HEK293 cells engineered to endogenously express a Cas9 nickase bearing an H840A mutation. To determine the efficacy of genomic integration and expression of the GFP reporter payload, flow cytometry was performed to detect GFP expression as well as PCR and sequencing of the GFP positive cells to confirm AAVS1 integration and the absence of integration of the ORF1 and ORF2 sequences (**FIG. 3**).

[0426] Flow cytometric analysis was performed and demonstrated successful integration and expression of GFP, resulting in about 2% integration efficiency (**FIG. 4A**). Following sorting for GFP positive cells, this further enriched the population to more than 70% GFP expressing cells (**FIG. 4B**). Genetic analysis utilizing PCR comparing HEK293 cells receiving the construct with control cells was performed. As shown in **FIG. 5**, cells receiving the construct and undergoing CREATE demonstrated effective integration of the GFP payload, indicated by the 1500bp band. Control cells demonstrated no bands at 1500bp, indicating no integration. This was further confirmed by nanopore sequencing of GFP+ cells, the results of which are shown in **FIGs 6A and 6B**. Sequencing demonstrated that, when integrating into the genome of the target cell, the CREATE process only results in integration of homology arms and the GFP payload gene. No sequences were detected that are derived from the ORF1 or ORF2 components.

Example 3. Validating requirement of Cas9 nickase activity, matched dual sgRNA, and the CREATE mRNA construct for genomic integration

[0427] In this example, the requirements for three components for an effective payload nucleic acid delivery design to genomic DNA was validated for stable integration and expression of an exogenous sequence encoding a polypeptide. The three components are (1) Cas9 H840A nickase activity, (2) matched dual sgRNAs that target and nick AAVS1 site on each DNA strand, and (3) the CREATE mRNA carrying the payload flanked by two PBS sites (i.e., homology arms) matching to Cas9-guided nicking sites. A sequence encoding GFP (payload) was targeted to the AAVS1 safe harbor locus as illustrated graphically in **FIG. 7A- FIG. 7C**. A sequence encoding ORF1 and a sequence encoding endonuclease inactivated ORF2 (mutated) were incorporated on the same CREATE mRNA molecule, along with 3'- homology arm sequence PBS1, reverse complement of 5' homology sequence PBS2 (RC-PBS2), reverse complements of GFP encoding sequence and promoter sequence (EF1) (**FIG. 7A, FIG. 8A**). GFP expression was used as readouts for successful retrotransposition (editing). Experimental sets include supplementing non-targeting sgRNA controls; and controls lacking sgRNA guide to PBS1 or sgRNA guide to PBS2 were evaluated. On 24-well plates 8×10^4 cells were plated and cultured in DMEM + 2% FBS, and transfected with the CREATE mRNA design shown in **FIG. 8A**. GFP positive cells were observed using flow cytometric analysis – on Day 3 and Day 8. Percentages of GFP-positive cells indicate the CREATE editing and payload insertion efficiency.

[0428] Experimental results are shown in **FIG. 8B**. In the absence of target specific guide RNA sequences or one of the guide RNAs, no GFP expression was observed. Experiment was repeated in HEK293T cells and in HEK293T cell line expressing Cas9 nickase (nCas293T). From this study, it was observed that only the combination of nCas9, matched dual sgRNA, and CREATE mRNA produces stable GFP expression. Also observed, optimizing transfection conditions by (1) using 24-well format with 8×10^4 cells per well, and (2) reducing FBS in culture medium from 10% to 2%, enhance CREATE editing efficiency.

Example 4. Evaluating split design CREATE mRNA construct to experimental cells for improved genomic integration

[0429] In this example, the LINE1 mobile genetic elements were delivered as two separate mRNA molecules, one comprising a sequence encoding ORF1 (RET50) and the other mRNA molecule comprising the CREATE mRNA except the sequence encoding ORF1. This design is designated as the split ORF1/ORF2 (or Design 2, or CREATE 2, **FIG. 9A**). Following transfection and analysis by Cytex flow cytometry, it was seen that the split CREATE system increased the initial GFP expression by about 7.2 fold (**FIGs. 9B and 9C**). The section highlighted with bold rectangle within **FIG. 9B** was further subjected to GFP sorting to validate GFP expression and stability by culturing and detecting

GFP positive cells by flow cytometry at day 8. However, the lack of effective enrichment after GFP+ cell sorting (**FIG. 9D**) may suggest that the initial high GFP expression was not stable.

Example 5. Evaluating timing and order of delivery and proportions of CREATE mRNA for improving retrotransposition efficiency

[0430] To further improve the efficiency of CREATE editing, CREATE mRNA performance was tested by varying the conditions of transfection methods. For most of the studies in the following section, unless otherwise mentioned, CREATE mRNA design 1 was used. Initially the following three protocols were tested:

[0431] Protocol #1: the CREATE mRNA and sgRNAs were delivered at the same time (**One Shot**) to the nCas9-expressing cells, followed by Flow analysis on Day 3.

[0432] Protocol #2: the CREATE mRNA was delivered to the cells first (**mRNA first**). Four (4) hours after the delivery, the medium was refreshed and dual sgRNA transfection was performed. Flow analysis was conducted on Day 3.

[0433] Protocol #3: the sgRNAs were delivered to the cells first (**sgRNA first**). Four (4) hours after the delivery, the medium was refreshed and mRNA transfection was performed. Flow analysis was conducted on Day 3.

[0434] Upon testing, it was observed that Protocol 2 yielded the best results, shown in **FIG. 10B**. The integration was stable. Further investigation on varying the time interval between delivery of the mRNA and the delivery of sgRNA showed that the efficiency peaks between 4 and 6 hours, but extending it beyond 8 hours drastically reduces the efficiency (**FIGs. 11A-11B**). The improved two step delivery protocol (Protocol #2 at 4 hours interval) was adopted for the later studies. Keeping the time interval as 4h under Protocol #2, the amounts of CREATE mRNA and guide RNAs were then varied for finding the best outcome as shown **FIG. 12A-12B**. It was observed that the editing efficiency plateaus when CREATE mRNA was above 3 μg per 8×10^4 cells. mRNA above 10 μg started to cause significant cytotoxicity. Optimal amounts of sgRNAs (dual sgRNA combined) are 0.484 μg per 8×10^4 cells.

Example 6. Delivery of Cas9 nickase as mRNA

[0435] Delivery of Cas9 nickase as mRNA was tested using the two-step protocol. The mRNA designs used are depicted graphically in **FIG. 13A**. The CREATE mRNA was co-transfected with an mRNA encoding Cas9 H840A nickase in the cells. The dual sgRNAs were transfected in the cells 4 hours later. GFP + cells detected at day 3 is shown in **FIG. 13B**. This data support that CREATE editing can be delivered with only RNA-based components (mRNA and sgRNA).

Example 7. Deletion of endonuclease domain of ORF2 reduces retrotransposon efficiency

[0436] Using the improvements in the delivery protocol for genomic integration described above, it was tested whether deletion of the endonuclease domain would affect the retrotransposition. The ORF2 endonuclease is a mutated inactive form, but whether it was structurally necessary for the

functioning of the ORFs would be addressed by deleting the endonuclease EN domain (depicted as CREATE 3 in **FIG. 14A, row II**). Results shown in **FIG. 14B**, third bar (CREATE 3), compared to second bar (CREATE 1), thereby using the successful CREATE mRNA design 1 (CREATE1) as control. Surprisingly, deletion of the EN domain drastically reduced (nearly abolished) the function of the ORF2, even though the EN domain of ORF2p does not contribute to the endonuclease function in the successful CREATE designs. This indicates that the EN domain is structurally necessary for the functioning of LINE1.

Example 8. Determining an optimal length of homology arm

[0437] In this example, length of homology arm domains were varied to see if there was an effect on the retrotransposition efficiency. Instead of 17 bases PBS1 and reverse complement PBS2 sequences of 3' homology domain and 5' homology domain respectively in the CREATE 1 designs, longer homology arm regions of 50 bases each were generated (CREATE 4 mRNA, **FIG. 14A, row III**). Such alteration also led to a reduction in the efficiency of retrotransposition, denoted by a drop in percentage (%) of GFP positive cells (**FIG. 14B**, comparing bar#4 with bar #2). This indicates that increasing the size of the homology arm sequences is not conducive to an improvement of the design. In each case, all other parameters were kept constant.

Example 9. Necessity and orientation of 5' – and 3' - homology arms

[0438] In this example, it was tested whether each of the homology arms were required, or one or both of them were dispensable for the optimal design. First, the RC- PBS2 sequence was deleted from a CREATE 1 design while keeping the PBS1 sequence intact (CREATE 5, **FIG. 14A, row IV**). In another test, the RC-PBS2 structure was left intact, while PBS1 sequence was deleted from the CREATE 1 mRNA (CREATE 6, **FIG. 14A, row V**). Either of these alterations led to a reduction in percentage (%) GFP positive cells (**FIG. 14B**, bars 6-8 compared to bar 2). However, inversion of the PBS orientation led to successful expression of GFP in the reverse orientation (**FIG. 14B**, bar 9 compared to bar 2). Therefore, these studies indicated that both PBS sites and both sgRNAs are required for successful editing, clearly demonstrating the exquisite specificity of this approach. PBS orientation can be reversed, enabling insertion of payload at both orientation.

[0439] This study depicted in **FIGs. 14A-14B** led to key understanding for generating an effective CREATE system, summarized as follows. An endonuclease domain is needed for the genome DNA binding and scanning to identify nicking site and deletion of this domain reduces the functionality of the system. Increase in length of the homology arms, at least in case of the PBS homology arms beyond 20 bp reduces the efficiency of TPRT. Both PBS sites and both sgRNAs are required for successful editing. Also understood is that the orientation of the homology arms (PBS1 and 2) can be reversed, enabling insertion of payload at both orientations.

Example 10. Demonstration of feasibility of CREATE editing in liver cell line.

[0440] In this example, Huh7 cell lines were used as a model for hepatocytes. CREATE editing was performed in Huh7 cell line with GFP as the payload, and the cell line expresses Cas9 nickase. Following the protocols described in the previous examples, using two gRNAs targeting to the AAVS1 site, successful incorporation and expression of GFP was observed in these cells, as shown in the graph showing flow cytometry data (**FIG. 15**, bottom panel). Compared to control, CREATE system resulted in GFP expression in 1.31% cells at day 4 after transfection. This indicates, that in addition to the epithelial cells (HEK293 cell line) demonstrated above, the system can also edit liver cells such as a liver cell line.

Example 11. RNA-mediated gene editing and delivery: CRISPR-Enabled Autonomous Transposable Element (CREATE)

[0441] To realize the potential of genome editing for broad therapeutic applications, tools that can selectively deliver multi-kilobase gene payloads are necessary. CRISPR-Enabled Autonomous Transposable Element (CREATE), an RNA-based genome editing system that merges CRISPR/Cas9 with the capabilities of the human L1 retrotransposon to insert gene-sized payloads without DNA donors or double strand breaks is disclosed in this document. This is achieved by endowing the L1 mRNA with features that enables L1-mediated reverse transcription and integration of the RNA-encoded payload gene specifically at two Cas9-induced nick sites. As demonstrated herein, CREATE can be delivered with only RNA components and achieves integration of >1 kb payload guided by Cas9 in mammalian cells.

[0442] **A. Design of the CREATE editing system:** The CREATE system encodes, on a single mRNA, the codon optimized L1 components ORF1p and ORF2p, separated by the native inter-ORF sequence which mediates bicistronic expression (**FIG. 16A**, **FIG. 16B** and **FIG. 16C**, **FIG. 16D**). To facilitate nuclear import, the ORF2p is fused with an N-terminal SV40 nuclear localization signal (NLS) and a C-terminal nucleoplasmin NLS. A key design element implemented to prevent nonspecific DNA cleavage and integration was the inactivation of the catalytic residue within the EN domain of ORF2p (mutation D205A). The payload cassette consisting of a promoter and the gene of interest is placed at the 3'UTR of the CREATE mRNA (**FIG. 16B**). The CREATE mRNA when delivered into the cells will be translated into ORF1p and ORF2pENdead, which then co-assemble with the mRNA to form CREATE RNP and enter cell nucleus (**FIG. 16D**, step 1). Natural L1 retrotransposition relies on nicking of the non-targeting DNA strand by the EN domain of ORF2p. To harness this mechanism, a Cas9H840A non-target strand nickase was employed to introduce a single strand nick guided by sgRNA1 (**FIG. 16D**, step 2). The liberated DNA 3'-flap then hybridizes with the PBS1 sequence on CREATE mRNA, serving as the primer for the RT domain of ORF2p for first strand cDNA synthesis (**FIG. 16D**, step 3). The newly synthesized cDNA strand includes both the payload and the PBS2 sequences in the sense orientation (**FIG.**

16D, step 4). Subsequently, PBS2 hybridizes with the 3' flap released from the second nicked site generated by sgRNA2, invoking the L1 template jump mechanism to complete the second strand synthesis and integration (**FIG. 16D**, step 5-6). The outcome of the editing cycle is the successful integration of the payload cassette between the PBS sites, discarding the L1 components, resulting in safe non-replicative gene insertion. To ensure that the payload cannot be expressed by direct translation, the payload cassette is encoded in the anti-sense orientation (**FIG. 16A and FIG. 16B**). Critically, the payload cassette is flanked by sequences designed to hybridize with single-guide RNA (sgRNA) target sites in the genome to prime reverse transcription. These sequences are referred to as Primer binding site 1 (PBS1) and reverse complement primer binding site 2 (RC-PBS2). Natural L1 retrotransposition starts with nicking of the non-targeting strand by the EN domain of ORF2p. To harness this mechanism, a Cas9^{H840A} non-target strand nickase was employed to introduce a single strand nick guided by sgRNA1. The liberated DNA 3'-flap then hybridizes with the PBS1 sequence on CREATE mRNA, serving as the primer for the RT domain of ORF2p for first strand cDNA synthesis (**FIG. 16C and FIG. 16D**). The newly synthesized cDNA strand includes both the payload and the PBS2 sequences in the sense orientation at its 3'-end. Subsequently, PBS2 hybridizes with the single strand DNA released from the second nicked site generated by sgRNA2, invoking the L1 template jump mechanism to complete the second strand synthesis and integration (**FIG. 16C and FIG. 16D**). The outcome of the editing cycle is the successful integration of the payload cassette between the PBS sites, discarding the L1 components, resulting in safe non-replicative gene insertion (**FIG. 16C and FIG. 16D**).

[0443] B. CREATE-mediated large payload delivery in mammalian cells: To demonstrate that CREATE can be used to deliver functional genes, a reporter gene payload was designed consisting of an EF1 α core promoter-driven GFP with an SV40 poly(A) signal (totaling 1.1 kb). The AAVS1 site, a well-characterized genomic safe harbor site, was chosen to show programmable targeted insertion. Two sgRNAs were chosen that have been validated previously, targeting sites within AAVS1 that are 90 bp apart. Successful editing integration will result in replacement of the 90 bp DNA segment with the 1.1 kb payload cassette (**FIG. 16A and FIG. 16B**). PBS1 and RC-PBS2 were designed to anneal with the DNA flaps released from two sgRNA cut sites, and positioned them on either side of the payload. Both Cas9 and ORF2p are large multidomain proteins, each exceeding 100 kD. It was possible that direct Cas9-ORF2p fusions may compromise activity and the ability to enter cell nucleus. Since ORF2p has the inherent ability to scan genomic DNA for entry sites, it was hypothesized that it could independently identify sites nicked by Cas9, eliminating the need for a direct fusion. To test this concept, a 293T cell line engineered to stably express a SpCas9 non-targeting strand nickase (293T-nCas9^{H840A}) was developed. Flow cytometry was used to quantify the percentage of cells expressing GFP as a measurement of successful payload integration. Transfection of cells with CREATE mRNA and dual AAVS1-targeting sgRNAs resulted in GFP expression on Day 3 and Day 8 (**FIG. 8B, FIG.**

16E). No GFP expression was observed in nCas9^{H840A}-293T cells transfected with non-targeting control sgRNAs or in wildtype 293T cells. This confirms sgRNA-guided integration and excludes the possibility of Cas9-independent, nonspecific insertion. Furthermore, it is demonstrated that payload expression requires co-transfection of both sgRNAs, as delivery of either sgRNA alone did result in integration (**FIG. 8B, FIG. 16E**). This showed that dual nicking events on opposite DNA strands are essential for complete TPRT and successful integration. This inherently enhances specificity, as a single independent sgRNA off-target binding is insufficient to cause payload integration.

[0444] To further appreciate the temporal requirements for each of the CREATE components, sequential transfection of RNA for the individual CREATE elements was performed. The specific experiments included (1) Simultaneous transfection of CREATE mRNA and sgRNAs (One-shot Protocol); (2) Transfecting CREATE mRNA four hours prior to sgRNAs (Two-step, mRNA-first Protocol); (3) Transfecting sgRNAs four hours before CREATE mRNA (Two-step, sgRNA-first Protocol). In Protocols #2 and #3, the culture media was replaced immediately prior to the second transfection. Observations support superior performance of the two-step protocols when compared to the one-shot approach (**FIG. 10B, FIG. 16F**). Notably, the mRNA-first strategy yielded the highest CREATE editing efficiency, reaching ~1.2% GFP expression on day 3 and remained stable on day 8. The observed improvement is likely due to the necessary time for the ORF1p and ORF2p mRNA to be translated, co-assemble with CREATE mRNA and subsequently enter the cell nucleus. Delivery of sgRNAs before the CREATE RNP complex forms might result in the single strand nicks being repaired by the cell before the CREATE editing process can occur. Further optimization of the time interval between mRNA and sgRNA transfections, along with adjustments in their concentrations, culminated in around 1.5% of the cells expressing GFP payload (**FIGs. 11B, and 12B**).

[0445] **C. Characterization of CREATE editing outcome by amplicon sequencing:** Fluorescence-activated cell sorting was employed to enrich the population to approximately 70% GFP positive and then passaged for an additional two weeks (**FIG. 17A**). During this time GFP expression remained stable (>70%), indicative of successful and stable genomic integration of the payload. Specific integration into the AAVS1 site was determined using PCR on the post-sort GFP-positive cells. A band corresponding to expected size associated with insertion at the AAVS1 site was observed along with unedited allele (**FIG. 17B**). The lower diagram shows the feature of the payload. Prime editing and related approaches have been observed to incur indel formation at the sgRNA target sites. To further confirm specific integration at AAVS1 site and assess the editing accuracy at the payload-genome junctions, amplicon sequencing on CREATE-edited GFP-positive cells were performed from two independent experiments. At the PBS2-AAVS1 junction, 95% of sequencing reads showed expected editing outcome without indels in both experiments (**FIG. 17C**). To validate the site-specificity of the insertion and search for potential off-target editing, next generation sequencing was performed with target enrichment by using hybridization probes to capture EF1a-GFP payload

sequences from genomic DNA extracted from edited cells. Insertion site analysis across the whole genome showed highly specific editing at the intended AAVS1 locus at chromosome 19 (**FIG. 17E**). Detailed sequencing results are shown in **Table 6**.

[0446] Interestingly, PBS1-AAVS1 junction showed increased indels at the sgRNA cut site, and in one experiment insertion of poly-A tract was also observed (**FIGs. 17C and 17D**). This suggested a novel mechanism that governs initiation of the ORF2p TPRP process at PBS1 site that influenced indel formation.

[0447] To characterized the editing accuracy at the junction sites between payload sequences and genomic sequences, amplicon sequencing was performed by designing PCR primers to cover the expected junctions (**FIG. 17C**). Primer information is summarized in **Table 4**. For AAVS1 edited cells, varying levels of indel formation was observed at the PBS1 junction, while the PBS 2 junction was >95% without indels (**FIG. 17C**). Changing the length of PBS from 17 to 30bp reduced indel formation at PBS1 (**FIG. 17F**). This analysis was also performed on junction regions for cells with targeted insertion at HEK3, PRNP, and IDS sites. This analysis demonstrated that for HEK3 and IDS, the insertions were more than 95% accurate; while the pRNP showed indel formation only at PBS1 junction but not PBS2 (**FIG. 17G**). Taken together, these results demonstrated that the formation of indels occurs only at the junction site and are likely dependent on the sgRNA sequence for the corresponding target locus.

[0448] **D. L1 components are essential for CREATE editing:** To validate the CREATE editing mechanism as shown in FIG. 16A, mutational analysis was performed to abolish key steps of the editing cycle. Whether CREATE editing requires non-target strand (Cas9^{H840A}) rather than target strand nickase (Cas9^{D10A}) activity. When CREATE mRNA and dual AAVS1-targeting sgRNAs were transfected into HEK293T cells stably expressing either Cas9^{D10A} or Cas9^{H840A} nickases, GFP expression and integration exclusively in the presence of Cas9^{H840A} nickase were observed (**FIG. 18A**), confirming natural L1 replication mechanism that initiates with non-target strand cleavage. In addition, the activity of Cas9^{H840A} also destroys the PAM sequence after successful editing, mitigating concerns about continuous nicking of the edited genome. Next the reverse transcriptase (RT) domain of L1 ORF2p was inactivated by introducing the D205A mutation (CREATE^{RTdead}). When CREATE mRNA and dual AAVS1-targeting sgRNAs were transfected into HEK293T cells stably expressing either Cas9^{D10A} or Cas9^{H840A} nickases, we observed GFP expression and integration exclusively in the presence of Cas9^{H840A} nickase (**FIG. 18A**), confirming the natural L1 replication mechanism that initiates with non-target strand cleavage. In addition, the activity of Cas9^{H840A} also destroys the PAM sequence after successful editing, mitigating concerns about continuous nicking of the edited genome. Next we introduced D702A mutation to inactivate the reverse transcriptase (RT) domain of L1 ORF2p (CREATE^{RTdead}). This modification entirely abolished the integration and expression of the GFP payload. This modification entirely abolished

the integration and expression of the GFP payload (**FIG. 18A**). This findings also discounted the possibility of contaminating plasmid DNA-mediated integration or episomal expression, which would occur independent of the RT activity of L1. Similarly, the removal of the entire EN domain (CREATE mRNA^{ΔEN}) resulted in a marked decrease in GFP insertion efficiency, indicating that while the EN domain must be rendered inactive to avert non-specific editing, its complete removal is detrimental to the editing process. Complete removal of the EN domain may sabotage the required structural interactions between the DNA strand and the L1 retrotransposon disabling the DNA binding/scanning necessary for full function of ORF2p.

[0449] E. CREATE paired sgRNAs and matched PBS sequences impart high fidelity and precision, PBS length and sgRNA sites influence editing efficiency: To determine if PBS length impacts insertion efficiency, CREATE mRNA was designed with different PBS1 and RC-PBS2 hybridization sequence lengths (17-bp, 30-bp, and 50-bp) and measured their capacity to insert GFP payload at the AAVS1 sites targeted with the same dual sgRNAs. It was further observed that 17-30bp appear to be ideal for efficient editing, and longer length led to a reduction of GFP integration (**FIG. 18C**). To demonstrate the importance of the dual PBS sites flanking the payload we designed CREATE mRNA with either PBS1 or a RC-PBS2 sequence. In the presence of AAVS1 targeting sgRNAs, the removal of either PBS1 or RC-PBS2 resulted in no GFP expression on day 3 or day 8 (**FIG. 18C**). Additionally, swapping the position of PBS1 and PBS2 while simultaneously encoding them in a reverse-complementary sequence (“RC-PBS1” and “PBS2”) restored the GFP expression, suggesting correct insertion of the payload cassette in the opposite orientation. The requirement for dual PBS flanking sites and dual sgRNA imparts a level of specificity, which was confirmed further when we reviewed the top 5 predicted off-target sites for sgRNA1 and sgRNA2, by PCR. We observed no PCR product for any of the predicted off target sites (**FIGs. 11A, 11B, and FIGs. 12A, 12B**). Together these data not only validated the proposed mechanism of CREATE but also showed the editing process has a number of inherent checkpoints that prevent off-target activity. These include 1) correct base-pairing of PBS1 with sgRNA1 cut sites; 2) correct base-pairing of reverse transcribed PBS2 with sgRNA2 cut site and 3) adjacent sgRNA1 and sgRNA2 nicking.

[0450] CREATE can be used for exogenous gene delivery and search-and-replacement, as the segment between sgRNA1 and sgRNA2 will be deleted and replaced by the payload (**FIG. 16A**). To explore this capability, the influence of the distance between the two sgRNA cut sites on editing efficiency was investigated. By fixing sgRNA1 and selecting different sgRNA2 with cut sites located further away, Applicants designed several CREATE mRNA/sgRNAs to replace 90bp, 481bp and 976bp DNA segment with the GFP expression cassette. Our result demonstrated that CREATE can achieve successful replacement in all three cases, although with lower efficiencies when the target

sites are further apart (**FIG. 18D**). Notably this greatly exceeds the capability of prime editing-based approach, which can only replace < 100bp sequences.

[0451] F. Paired sgRNAs and matched PBS sequences imparts high fidelity and precision to CREATE To demonstrate the importance of the dual PBS sites flanking the payload CREATE mRNA were designed with either PBS1 or a RC-PBS2 sequence. In the presence of AAVS1 targeting sgRNAs, the removal of either PBS1 or RC-PBS2 resulted in no GFP expression on day 3 or day 8 (**FIG. 18B**). Additionally, swapping the position of PBS1 and PBS2 while simultaneously encoding them in a reverse-complementary sequence (“RC-PBS1” and “PBS2”) restored the editing and lead to inversion of the payload cassette. The requirement for dual PBS flanking sites and dual sgRNA imparts a level of specificity, which was confirmed further when the top 5 predicted off-target sites for sgRNA1 and sgRNA2 by PCR were examined. No PCR product indicative of off-target insertion for any of the predicted off target sites were observed (**FIGs. 20A and 20B**). **FIG. 20C** shows additional results analyzing predicted sgRNA2 off-target sites. Together these data not only validate the proposed mechanism of CREATE but also show the editing process has a number of inherent checkpoints that prevent off-target activity. These include 1) correct base-pairing of PBS1 with sgRNA1 cut sites; 2) correct base-pairing of reverse transcribed PBS2 with sgRNA2 cut site and 3) adjacent sgRNA1 and sgRNA2 nicking.

[0452] G. Targeted insertion of large DNA fragments into different genomic loci: To determine the ability of harnessing the CREATE system to edit and deliver genes specifically into sites in the genome beyond AAVS1, CREATE is engineered to deliver GFP encoding sequence into other genomic loci including *HEK3*, *PRNP*, and *IDS*. The *HEK3* locus, situated on chromosome 1 is frequently chosen to benchmark gene editing tools. The *PRNP* locus on chromosome 20 encodes the prion protein implicated in multiple neurodegenerative diseases and *IDS* gene on chromosome X is associated with Hunter’s syndrome. For each targeted site, a pair of sgRNAs targeting the site, on opposite strands of the genome 70-90bp apart that were validated before were selected. CREATE mRNA, which carried the corresponding 30-bp PBS1 and RC-PBS2 flanking the GFP expression cassette, was deployed for each target site. Flow cytometric analysis on day 3 demonstrated all three sites exhibited comparable editing efficiency, as determined by GFP expression (**FIG. 19A**). FACS was used to enrich PRNP locus edited cells to >60% GFP positivity (**FIG. 19B**). To confirm specific integration, the expected junction region were amplified and TA-cloning was performed of the PCR product. Sanger sequencing of the individual colonies from the TA-cloning revealed correct insertion at the expected junction at PRNP site, validated the programmable, targeted editing capability of CREATE (**FIG. 19C**).

[0453] H. CREATE as a fully RNA-based large-scale genome editing system: Finally, it was demonstrated that CREATE editing can be achieved by delivering all components as RNA. Electroporation was used to co-delivery all mRNA and sgRNA to the liver cell line Huh7, resulting in

~1% GFP gene re-expression on day 3 post transfection. Similar result can also be achieved in 293T (FIG. 19D). Together these findings confirm the ability to use RNA to deliver all necessary CREATE components resulting in stable integration of cargos not possible with other available approaches.

[0454] CREATE as a fully RNA-based large-scale genome editing system in human liver cells Finally, an important attribute of CREATE gene editing is the ability to deliver all components as RNA into mammalian cells (FIG. 19E). This was achieved in HEK293T and the immortalized liver cell line Huh7, whereby RNA transfection resulted in up to 1% GFP gene re-expression (FIG. 19E). Together these findings confirmed the ability to use RNA to deliver all necessary CREATE components resulting in stable integration of cargos not possible with other available approaches.

[0455] CREATE represents a significant step forward in gene editing. The system not only significantly increases the cargo size that can be inserted into the genome via a fully RNA-based approach, but its in-built regulatory mechanisms reduce the potential for off-target genotoxicity. By using the highly processive RT activity from L1 ORF2p, CREATE is able to complete cDNA synthesis larger than what can be achieved with the M-MLV RT domain used in PE or base editing. The ability to insert an 1.1kb GFP cassette is demonstrated herein in multiple mammalian cell lines at ~2% integration efficiency, while the largest payload delivered via PE related approach is ~1kb with only 0.002% integration.

[0456] Current approaches for targeted genome engineering in vivo relies on adeno-associated virus (AAV) or other vectors for delivery of the editing machinery as well as a DNA payload. Viral-based delivery approaches are costly to manufacture and can trigger severe immunogenic toxicity. Being a fully RNA-based system, all the components of CREATE can be packaged into LNP and delivered in vivo. Future optimization of the RNA components in CREATE system and delivery method will lead to improvement of delivery and editing efficiency in target tissues. By integrating CREATE with RNA delivery technologies, the development of a versatile and cost-effective method for large-scale gene delivery and genome editing is envisioned.

Table 2. Sequences of sgRNAs used in this study.

Nicking sgRNA	Spacer sequence (5'-3')	SEQ ID NO:
AAVS1 nicking PBS1	GAUGGAGCCAGAGAGGAUCC	36
AAVS1 nicking PBS2	GCAGCUCAGGUUCUGGAGA	37
AAVS1_nicking_PBS1_481-bp replacement	GCUCUCCAGCCCCUGUCA	38
AAVS1_nicking_PBS1_976-bp replacement	CCUUCCCUGCCGCCUCCUUC	39
HEK3 nicking PBS1	GUCAACCAGUAUCCCGGUGC	40
HEK3 nicking PBS2	GGCCAGACUGAGCACGUGA	41
PRNP nicking PBS1	GCAGUGGUGGGGGGCCUUGG	42
PRNP nicking PBS2	GCAUGUUUUCACGAUAGUAA	43

IDS nicking PBS1	GCAUUUUCGAUCCGUGACU	44
IDS nicking PBS2	ACUGAGGGAUGUCUGAAGGC	45
Negative Control, non-targeting	AAAUGUGAGAUCAGAGUAAU	46

Table 3A: Sequences of payload used in this study.

EF1 α -GFP payload in sense direction (1083 bp)	SEQ ID NO:
EF1α core promoter-GFP-SV40 polyA signal GGGCAGAGCGCACATCGCCACAGTCCCCGAGAAGTTGGGGGGAGGGGTCTGGCA ATTGATCCGGTGCCTAGAGAAGGTGGCGCGGGGTAACTGGGAAAAGTGATGTCGT GTACTGGCTCCGCCTTTTTCCCGAGGGTGGGGGAGAACCGTATATAAGTGACAGTA GTCGCCGTGAACGTTCTTTTTCGCAACGGGTTTGCCGCCAGAACACAGGGTTTAGT GAACCGTCAGATCCCGCCACCATGGTGAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTG CCCATCTTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTACGCGTGTCGGCGAGG GCGAGGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTG CCCGTGCCCTGGCCACCCCTCGTGACCACCTGACCTACGGCGTGCACTGCTTACGCCGCTA CCCCGACCACATGAAGCAGCAGCACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGG AGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTTCGAG GGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACA TCCTGGGGCACAAGCTGGAGTACAACACTACAACAGCCACAACGTCTATATCATGGCCGACAAGC AGAAGAACGGCATCAAGGTGAAGTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAG CTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACA ACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATG GTCTGCTGGAGTTCTGTACCGCCGCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTA <u>AAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTAC</u> <u>AAATAAAGCATTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATC</u> <u>TTA</u>	47

Table 3B: Sequences of PBS sites

CREATE constructs	PBS1 sequence (5'-3')	SEQ ID NO:	RC-PBS2 sequence (5'-3')	SEQ ID NO:
AAVS1_90-bp replacement_17-bp PBS	TCCTCTCTGGCTC	48	GCAGCTCAGGTTCTGGG	49
AAVS1_90-bp replacement_30-bp PBS	TCCTCTCTGGCTCCATC GTAAGCAAACCTT	50	CAGCCGCGTCAGAGCAGCTC AGGTTCTGGG	51
AAVS1_90-bp replacement_50-bp PBS	TCCTCTCTGGCTCCA TCGTAAGCAAACCTT AGAGGTTCTGGCAA GGAGAG	52	GATCAGTGAAACGCACCAGA CAGCCGCGTCAGAGCAGCTC AGGTTCTGGG	53
AAVS1_90-bp replacement_17-bp PBS reverse	CCCAGAACCTGAGCTG C	54	GAGCCAGAGAGGA	55
AAVS1_481-bp replacement_17-bp PBS	CAGGGGGCTGGAAGA GC	56	GCAGCTCAGGTTCTGGG	57
AAVS1_976-bp replacement_17-bp PBS	TCCTCTCTGGCTC	48	GCAGCTCAGGTTCTGGG	57
HEK3_90-bp replacement_30-bp PBS	CGTGCTCAGTCTGGGC CCCAAGGATTGACC	58	CCCAGCCAAACTTGTCAACCA GTATCCCGG	59
PRNP_72-bp replacement_30-bp PBS	AGGCCCCCCCACCACTG CCCCAGCTGCTGCA	60	TTGGGGTAACGGTGCATGTTT TCACGATAG	61
IDS_70-bp replacement_30-bp PBS	CACGGAATCGAAAATG CTTCAGAAGGTCT	62	TTGTCAGAATTCCACTGAGGG ATGTCTGAA	63

Table 4: Sequences of primers used for genomic DNA junction amplification

Target	Forward primer (5'-3')	SE Q ID NO:	Reverse primer (5'-3')	SE Q ID NO:
AAVS1 -PBS1	CGGAGCCAGTACACGACATC	64	TCCCAGGGCCGGTTAATGTG	65
AAVS1 -PBS2	GGGCTGGCTACTGGCCTTAT	66	ATCACTCTCGGCATGGACGA	67
PRNP- PBS1	CGGAGCCAGTACACGACATC	68	CTGGAGGCAACCGCTACC	69
PRNP- PBS2	GTGGTTGTGGTGACCGTGT	70	ATCACTCTCGGCATGGACGA	71
HEK3- PBS1	CGGAGCCAGTACACGACATC	72	TGATGTGGGCTGCCTAGAAA	73
HEK3- PBS2	GCCCTGAGATCTTTTCCTCTGT	74	CAACGAGAAGCGCGATCAC A	75
IDS- PBS1	AAAGAACGTTACGGCGACT	76	CTGTTTCAGGCAGGCAATCC	77
IDS- PBS2	GTTGGCAAAACTCAAGGCATC A	78	ACATGGTCCTGCTGGAGTTC G	79

Table 5: Sequences of primers used for PCR to detect off target insertion

Name	Target sequence	SEQ ID NO:	Location	GFP insertion forward primer (5'-3')	SEQ ID NO:	Genomic reverse primer (5'-3')	SEQ ID NO:	Genomic forward primer (5'-3')	SEQ ID NO:	Genomic amplicon size (bp) [#]	Predicted off-target amplicon (bp) ^{##}
AAVS1_nickin_g_PBS1	GATGGAGCCA GAGAGGATCC	80	chr19: 5511 5573								
PBS1_off_target_1	GATGAAGTCA GAGAGGATCC	81	chr20: 2451 191	GCTTGCCGTA GGTGGCAT	92	CACTCTGGC GTAAAGGA GCA	102	GAAGTCACC CCCGAATCA GC	112	649	427
PBS1_off_target_2	CACGGAGCCA GGGAGGATCC	82	chr9: 1275 9518 6	GCTTGCCGTA GGTGGCAT	93	CAGGATCTT ACCCTGCCA TGA	103	AGACTCGGA GCTCAAAC GC	113	439	453
PBS1_off_target_3	GCTGGAGGCA AGAGAGGATC C	83	chr11: 7554 6340	GCTTGCCGTA GGTGGCAT	94	CATCACAAT GCCCCAGGA CT	104	GTCCCTTCA CAGACCATG CG	114	757	513
PBS1_off_target_4	GATGAATGCT GAGAGGATCC	84	chr1: 9356 7284	GCTTGCCGTA GGTGGCAT	95	GCCGCTTGT GACGTTAAT GG	105	ACCTCCTCA TCTTGGCAC ATC	115	779	529
PBS1_off_target_5	GTGGGAGGCA GAGAGAATCC	85	chr1: 1157 9676	GCTTGCCGTA GGTGGCAT	96	GGCATTCCC TGTGAGAGT GT	106	GCCCTCTGT CAATAACCG CT	116	741	667
AAVS1_nickin_g_PBS2	GCAGCTCAGG TTCTGGGAGA	86	chr19: 551 1546 9								
PBS2_off_target_1	GAGGCTCCGG TTCTGGGAGA	87	chr1: +198 4949 71	CGGGATCACT CTCGGCATGG	97	GCAGAGCTA CCGGCTCCA TG	107	CCTTTCACT GTGAACCTC ATTGTTAGC	117	559	460
PBS2_off_target_2	TCAGTTCAGG TTCTGGAAGA	88	chr10: +648 3525	GCATGGACGA GCTGTACAAG TAAAA	98	GCCCTCGGG GTCCCTATT TA	108	CCTTGATGT GGCATTATT AACAGTGC	118	662	556
PBS2_off_target_3	GCTGGCTCTG GTTCTGGGAG A	89	chr11: +132 4500 11	GCATGGACGA GCTGTACAAG TAAAA	99	GCACCAACA GGTTCAGAG GT	109	GTGCTTTGG AGGACGAC CAG	119	429	521
PBS2_off_target_4	GCCCTCGGG TCCTGGGAGA	90	chr19: +445 1690 4	GCATGGACGA GCTGTACAAG TAAAA	100	CGATTCCCC ATATAGCTC ACTCC	110	GGTGCATCA CGTTTGGCT TT	120	953	809
PBS2_off_target_5	GCATTTCTGTT TCTGGGAGA	91	chr2: +307 6507 7	CGGGATCACT CTCGGCATGG	101	GGCACGTGA CCATCCCTT AAT	111	GGAACGCTT CCAATTCAC CC	121	567	577

Table 6: Target Hybridization Sequencing Results for AAVS1 edited samples

Chromosome	Location	Strand	Total Reads	Annotation
chr17	38905912	+	7	Endogenous Human EF1 α sequence
chr19	55115486	-	26904	On-target insertion at AAVS1 (chr19: 5515280 -- 5515880)
chr19	55115589	+	15692	
chr19	55115800	-	132	
chr19	55115148	-	8	
chr19	55115690	-	8	
chr19	55115285	-	5	
chr2	32916556	-	21	Mis-alignment to repetitive sequence
chr2	32916333	-	8	Mis-alignment to repetitive sequence
chr5	14653390	+	8	Endogenous Human EF1 α pseudo gene 13 sequence
chr6	73521212	+	134	Endogenous Human EF1 α sequence
chr6	73521001	-	117	

Identified potential insertion with >5 reads at a specific location are analyzed

Example 12. Modifications of the CREATE system

[0457] CREATE Avian single RNA and Split designs with multiple RNAs and with modifications: Upon success of RNA delivery of retrotransposition constructs and genomic integration using mammalian and non-mammalian TPRT systems, further improvement in retrotransposition efficiency was sought. In this example, CREATE retrotransposition designs for improved genome editing efficiency are contemplated, and accordingly several constructs are designed and tested. The Avian R2 retrotransposition construct design is further modified to obtain a few variant constructs. For example, as shown in **FIG. 21**, two different construct systems designs are shown. In the construct shown on top, the retrotransposon including the R2 endonuclease ("R2 (ENdead)"), and the payload sequences ("GFP") are on the same RNA molecule, while Cas endonuclease nickase ("Cas9^{H840A}") is on a separate RNA molecule. The sequence encoding the Cas endonuclease nickase further comprises a sequence of an NLS at the N-terminal side, and a sequence of an NLS at the C terminal side of the Cas9 nickase. In the design shown at the bottom, a split delivery system with 2 RNA molecules, wherein the system comprises one RNA molecule comprising a sequence encoding Cas9 nickase ("Cas9^{H840A}") and a sequence encoding the R2 endonuclease ("R2 (ENdead)"), wherein the endonuclease is mutated and deficient in enzyme activity, and the endonuclease is separated from the sequence encoding the Cas9 nickase by a self-cleavable moiety – e.g., a P2A sequence. Additionally, NLS sequences are present flanking the sequence encoding the Cas9 nickase and the non-functional endonuclease domain. Modifications are introduced in the RNA structure, where the polyadenylation is controlled to introduce shorter poly A sequences to the RNA molecule comprising the payload sequence, such as 20 A residues. The UTRs and poly A sequences are introduced as indicated.

Additionally, Stop codons (“STOP”) were introduced at the end of the non-functional endonuclease as shown.

CREATE L1 retrotransposon single RNA and split designs with modifications

[0458] FIG. 22 shows constructs with human ORF2p modifications. CREATE 24 and CREATE 69 constructs comprise RNA encoding both ORF1p and ORF2p sequences and the payload sequence in the same RNA molecule. CREATE 27 construct lacks an ORF1p coding sequence as compared to CREATE 24 as shown in the figure. In CREATE 24 as well as in the CREATE 69 constructs, the presence of an interORF region spaced between the ORF1p and ORF2p sequences, wherein the ORF2p endonuclease is functionally deficient. The ORF2p sequences contain a sequence encoding an NLS sequence on each of the 5’ and the 3’ of the ORF2p coding sequence. The GFP coding sequence is followed by a SV40 poly A sequence, and is flanked by homology arm sequences (“RC-PBS2” and “PBS1”) for AAVS1 targeting. All mRNAs are generated by in vitro transcription and a poly A sequence is added by controlled polyadenylation reaction before or after purification.

[0459] In the CREATE 24, CREATE 27 and CREATE 69 constructs, the Cas endonuclease is delivered as a second RNA molecule. The Cas 9 endonuclease is a nickase having a H840A mutation. In each case, ORF2p comprises an ORF2p endonuclease that is functionally deficient, e.g., deficient by mutation. Each RNA are polyadenylated after RNA generation via in vitro transcription. CREATE 27 does not comprise ORF1 sequence (Δ ORF1); ORF2 is translated via canonical cap-dependent mechanism.

[0460] Additional modifications include a short EF1alpha (EF1a) promoter operably linked to and driving expression of the exogenous sequence (in this case the surrogate GFP coding sequence) in CREATE 24 and CREATE 27 constructs. Conversely CREATE 69 has full length (FL) EF1a promoter operably linked with the sequence encoding GFP. Two sgRNAs each binding to opposite strands and about 70 nucleotides apart are designed to drive a Cas nickase molecule to each strand to the position on the genomic DNA to perform a single strand nick. The ORF2p reverse transcriptase (RT) enzyme then finds and binds to the nicked end and reverse transcribes its own payload up to or into the homology arm sequences.

CREATE 24 and CREATE 27 codelivery

[0461] In this example, the above discussed improved sequences were tested for improvement of retrotransposition efficiency. The following experiments were carried out using primary human cells. Human primary T cells were isolated from donor leukopak by negative immunomagnetic selection using EasySep™ Human T Cell Isolation Kit (STEMCELL Technologies). T cells were activated using ImmunoCult™ Human CD3/CD28 T Cell Activator (STEMCELL Technologies) for two days in TexMACS media with 5% heat-inactivated human AB serum, 10 ng/mL IL-2, 5 ng/mL IL-7 and 5 ng/mL IL-15 (PeproTech) prior to electroporation. Electroporation (EP) of activated human T cells were performed using the MaxCyte Atx System with Extended T cell-1 program. GFP expression was

determined by flow cytometry analysis on Day 3 and Day 7 post electroporation. Genomically expressed GFP persisting over Day 7 is one conclusive proof of stable genomic expression and low genotoxicity.

[0462] The assay included testing the CREATE 24, CREATE 27 and CREATE 69 sequences individually, when codelivered with Cas9 nickase; and additionally, a combination of CREATE 24 and CREATE 27 constructs with Cas9 nickase. Both guide RNAs (sgRNAs) were administered in a ratio of 1:1 as shown in **Table 7**.

Table 7. The details of the electroporated amounts in the above experiment.

Sample	Total amount in EP (pmole)			
	CREATE 24 mRNA	CREATE 27 mRNA	nCas9 ^{H840A} mRNA	AAVS1 sgRNAs
1 Mock	-	-	-	-
2 CREATE 24 alone	6.4	-	-	-
3 CREATE 27 alone	-	8.5	-	-
4 CREATE 24+CAS9 (H480) nickase +sgRNAs	6.4	-	4.4	150 (1:1)
5 CREATE 27+CAS9 (H480) nickase +sgRNAs	-	8.5	4.4	150 (1:1)
6 CREATE 24:CREATE 27 at 1.5:1 ratio	4.3	2.8	4.4	150 (1:1)

[0463] As shown in the data in **FIG. 23A**, at day 3 post transcription (“Day 3”), some improvement was shown upon codelivery of CREATE 24 and CREATE 27 constructs. Day 7 did not show further enrichment of GFP cells. Various different ratios of CREATE 24: CREATE 27 were experimented, with slightly higher expression levels at CREATE 24:9 being at 1.5:1, and it was tolerated, although no enrichment was observed on Day 7 (**FIG. 23B**).

[0464] The range of ratios of CREATE 24:9 was further expanded. In one exemplary assay, a range of CREATE 24: CREATE 27 was titrated from 0.37:1 to 1.5:1. The details of the electroporated amounts in the above experiment are shown in **Table 8** below:

Table 8: The details of the electroporated amounts in the above experiment.

	Total amount in EP (pmole)			
Sample	CREATE 24 mRNA	CREATE 27 mRNA	nCas9 ^{H840A} mRNA	AAVS1 sgRNAs
1 Mock	-	-	-	-
2 CREATE 24 alone	6.3	-	-	-
3 CREATE 24+CAS9 (H480) nickase +sgRNAs	6.3	-	4.4	150 (1:1)
4 CREATE 24:CREATE 27 at 1.5:1 ratio	6.3	4.3	4.4	150 (1:1)
5 CREATE 24:CREATE 27 at 0.75:1 ratio	4.2	5.7	4.4	150 (1:1)
6 CREATE 24:CREATE 27 at 0.37:1 ratio	3.1	8.6	4.4	150 (1:1)

[0465] Results from this experiment are shown in **FIG. 24A** (at day 3 post electroporation) and **FIG. 24B** at day 7 post electroporation. While codelivery of CREATE 24 and CREATE 27 showed some improvement in the expression of the exogenous sequence (GFP), which is understood to be the readout for retrotransposition efficiency for these experiments, significant differences were not observed with the variation of ranges. The trend appeared to direct that higher CREATE 24 amounts over lower was preferable. Further expansion of range was sought. In another exemplary experimental set up, CREATE 24: CREATE 27 were used as follows: CREATE 24: CREATE 27 at 1.5:1, 2:1 and 5:1. The experimental set up with electroporation amounts are shown in **Table 9**.

Table 9. The details of the electroporated amounts in the above experiment

Sample	Total amount in EP (μM)			
	CREATE 24 mRNA	CREATE 27 mRNA	nCas9 ^{H840A} mRNA	AAVS1 sgRNAs
1 Mock	-	-	-	-
2 CREATE 24 alone	7.5	-	-	-
3 CREATE 24+CAS9 (H480) nickase +sgRNAs	7.5	-	4.5	100 (1:1)
4 CREATE 24:CREATE 27 at 1.5:1 ratio	4.5	3	4.5	100 (1:1)
5 CREATE 24:CREATE 27 at 2:1 ratio	5	2.5	4.5	100 (1:1)
6 CREATE 24:CREATE 27 at 5:1 ratio	6.25	1.25	4.5	100 (1:1)

[0466] Results from this experiment are shown in **FIG. 25A** and **FIG. 25B**. At day 3 post electroporation, significant improvement in GFP expression was noted when CREATE 24:CREATE 27 was at 5:1 ratio. The increase was greater than 3 fold compared to the average of results from CREATE 24:CREATE 27 at 1.5:1 ratio over the three example titrations experiments shown. At day 7, a further enrichment to about 0.9 % of total cells showing GFP fluorescence was noted. This was encouraging improvement over the course of attempts in adjusting to the application of RNA mediated delivery and incorporation of longer sequences than a few bases using L1 retrotransposon machinery and in site-specific targeting of the genome.

[0467] Therefore, in summary, codelivery of CREATE 24:CREATE 27 effectively increases the retrotransposition (editing) efficiency as measure by expression of the exogenous sequence that is to be retrotransposed, in this case GFP expression. ORF2p is translated via 5'Cap-Kozak sequence. Without wishing to be bound by theory, it could be possible that the design of CREATE 27 offers advantages to ORF2p translation and increases the availability of ORF2p at the genomic site for editing.

Example 13. CREATE 24, 29 and 80 codelivery: Test co-transfection of ORF2 mRNA to enhance its expression and integration efficiency

[0468] In this example, several new gene editing system designs are set forth for testing and improving integration efficiency. In our experience, every time one parameter is changed in the retrotransposon mediated gene editing system, CREATE, it alters the paradigm greatly and many further adjustments and optimizations are required on the basic designs and application to get a functional outcome. **FIGs. 25C, 25D and 25E** represent various CREATE designs for testing, in which each design comprises delivery of the principal mRNA, comprising the ORF elements and the payload (sequence encoding the GFP) along with the adjoining homology arms (PBS1 and RC-PBS2),

an EF-S (EF1alpha-short) promoter or where indicated, full length EF1alpha promoter, as shown and in the orientation shown in the figures. Additionally in each case, an RNA sequence encoding Cas nickase is co-delivered (CREATE 29; as demonstrated in the exemplary codelivery system of **FIG. 26A**, and **FIG. 34** top diagram, for examples). For example, a model principal retrotransposon-payload CREATE mRNA structure is shown in **FIG. 26A** and **FIG. 26B** shows the structure in greater detail by enlarging a portion as shown. **FIG. 27** shows CREATE-Cas9 combination mRNA designs, with self-cleavage sequences P2A and T2A as shown in the figure, and may replace part of the entirety of the interORF sequence. In the model shown at the bottom, endogenously naturally expressed ORF1p may be utilized. **FIGs. 28** and **29** involve additional/designer linkers as shown adjacent to the Cas9 nickase that help in associating with the genomic DNA in a flexible and sensitive manner, such that steric hindrance or crowding of the mRNA-DNA hybrid intermediate where retrotransposition occurs is avoided. The two guide RNAs (sgRNA1 and sgRNA2) are included for delivery for Cas9 or Cas9 nickase function, whether or not they are mentioned in the description or included in the figures. **FIGs. 30-33** represent partial system diagrams to be tested.

[0469] **FIG. 34** demonstrates a retrotransposon mediated CREATE systems comprising the principal RNA encoding ORF1, ORF2 (endonuclease deficient), and the payload (CREATE 24, 6.4kb); an RNA encoding Cas9 nickase (CREATE 29) and an additional RNA encoding endonuclease deficient ORF2 RNA (CREATE 80) simultaneously. The objective of this study is to test if cotransfection of CREATE 80 to enhance ORF2p expression could enhance retrotransposon efficiency. The system targets the AAVS1 locus of the genomic DNA (homology arms PBS1 and RC-PBS2 are designed for AAVS1 targeting, along with sgRNAs as discussed before).

[0470] CREATE 80, CREATE 24 and CREATE 29 mRNA were made using RNA team optimized IVT and polyA tailing conditions. Engineered HiCap T7 polymerase and IVT incubation were added at 27°C to create capped and tailed mRNA for each construct. mRNA concentration, IVT buffer and rNTPs were optimized for in vitro transcription.

[0471] Donor derived pan-T cells were used. (Donor PD00G-001, Rotea-isolated lymphocyte fraction, isolated by Negative selection of CD3+ cells). Primary donor T cells were thawed and allowed to recover for 2 hours in TexMACS media (5% HI-HABS, 10 ng/mL IL-2, 5 ng/mL IL-7 and 5 ng/mL IL-15).

[0472] The cells were activated using CD3/CD28 antibody complex activator (STEMCELL Technologies) for 68 hours before electroporation (EP). 2 million T cells were electroporated via MaxCyte ET-2 program. An experimental set designated as Mock set was used as EP control. CREATE 24: CREATE 80 were used at Molar ratio of 5:1. A total of 7.5µM of CREATE construct mRNA (with cargo) was used for transfection.

Table 10. mRNA components for electroporation each condition or sets in the study.

Sample #	Condition	Total amount in EP (μM)			
		CREATE mRNA	CREATE 80 mRNA	nCas9 ^{H840A} mRNA	AAVS1 sgRNAs
1	Mock	-	-	-	-
2	CREATE 24 alone	7.5	-	-	-
3	CREATE 24+ CREATE 29	7.5	-	4.5	100 (1:1)
4	CREATE 24: CREATE 80 5:1 molar ratio	7.5	1.5	4.5	100 (1:1)

[0473] GFP expressing cells, indicating successful retrotransposition and payload delivery to genomic DNA was detected on Day 3 and Day 7 post-EP via Flow Cytometry. Table 10 provides details of the electroporation components described above in the study.

[0474] **FIGs. 35A-35B** and **36** indicate the results from this study. On observing GFP positive cells by flow cytometry on day 3 after electroporation, it was the combination of the ORF2p-encoding mRNA in addition of the mRNA encoding both ORF1 and ORF2 and the payload (CREATE 24 + CREATE 80) with the Cas nickase system led to slight increase in the integration efficiency over electroporating CREATE 24 and the nickase. At day 7 there was maintenance and increase of the GFP expressing cells in the CREATE 24 + CREATE 80 set, indicating that the incorporation was stable, and tolerated.

Example 14. Modification of interORF region

[0475] Without being bound by any particular theory, **FIG. 37** shows the translation process of ORF2p from an mRNA comprising a coding sequence for ORF1p ("ORF1") and a coding sequence for the modified ORF2p ("modORF2"), which are separated by an inter-ORF sequence. The translation of ORF2p from such mRNA can be highly inefficient, resulting in less formation of CREATE RNP complex. The interORF sequence between the ORF1 and ORF2 coding regions thus were modified to improve the translation efficiency for ORF2p. A construct was designed, termed CREATE-81, which substitutes the interORF sequence in construct CREATE-24 with a GSG linker sequence and a T2A sequence (**FIG. 38** and **FIG. 39**). The performance of the construct CREATE-81 was tested compared to CREATE-24, which comprises the native interORF sequence (**FIG. 39**). To form the CREATE RNP complex, both construct CREATE-81 and CREATE-24 were introduced to the cells with CREATE-29 which encodes Cas9 H840A nickase as well as two guide RNAs targeting

AAVS1 locus respectively. Construct nucleotide sequences are provided in **Table 11**. Construct amino acid sequences are provided in **Table 12**.

[0476] Cells were electroporated as described in the above example with each of the constructs alone (as a control) and in combination with CREATE-29, encoding the Cas9 nickase and providing the 2 guide RNAs. As shown in **FIGs. 40A-40C**, cells treated with CREATE-29 and CREATE-81 demonstrated the highest levels of GFP payload integration on both Day 3 and Day 8. This indicated that modification of the interORF region to introduce a GSG linker and T2A sequence can enhance the integration of the payload as compared to the mRNA comprising the interORF region.

Example 15. ORF2p Binding Domain Modifications

[0477] Next, the natural endonuclease domain of ORF2 was replaced with other DNA binding domains to measure the ability of the modified ORF2p to bind target sites nicked by the Cas9 nickase. One construct (CREATE-82), where the endonuclease domain of ORF2 is replaced with a Sso7d DNA binding domain which binds to dsDNA. An exemplary construct design is provided in **FIG. 31**. One construct (CREATE-83) was designed, whose endonuclease domain of ORF2 is replaced with a high mobility group protein D (HMGD), which is a chromosomal protein found in *Drosophila melanogaster* which has DNA binding and chromatin organization activity, with an architectural preference for distorted DNA regions. An exemplary design is provided in **FIG. 41**. One construct (CREATE-84) was designed to replace the endonuclease domain of ORF2 with zinc finger domains ZnF1 and ZnF2 from a Poly(ADP-ribose) polymerase (PARP) enzyme. PARP enzymes are involved in the DNA damage response and can detect strand breaks in DNA. PARP-1 binds to the ends of DNA strand break damage via the function of two zinc finger domains ZnF1 and ZnF2. This construct (shown in **FIG. 32**) was tested to measure whether the modified ORF2p with ZnF1 and ZnF2 can promote direct recruitment of the CREATE ribonucleoprotein to the single strand break induced by the Cas nickase. Construct nucleotide sequences are provided in **Table 11**. Construct amino acid sequences are provided in **Table 12**.

[0478] To test the efficacy of constructs with the above described ORF2 modifications, the constructs shown in **FIG. 42** were used. In these experiments, each of CREATE constructs 80-84 (encoding ORF2 only) was co-delivered with CREATE-24 and CREATE-29. Cells were treated as in **Example 13**, using a 5:1 molar ratio of CREATE-24 to the ORF2 only constructs (CREATE-80, CREATE-82, CREATE-83, or CREATE-84) (**FIG. 43B**). **FIG. 43A** shows the results of the control groups, including the composition comprising CREATE-24 and CREATE-29 with guide RNAs, with no additional ORF2-only construct (CREATE-80, CREATE-82, CREATE-83, or CREATE-84) co-delivered. As shown in **FIG. 43C**, codelivery with CREATE-82 (containing an Sso7d domain) performed the best, demonstrating the highest levels of payload integration and expression.

Example 16. Modification of PBS Sequence Length

[0479] In order to test the effect of varying PBS sequence length on the integration efficiency, the constructs shown in **FIG. 44** were designed. These constructs utilized PBS sequences ranging in length from 13 to 50 nucleotides. Cells were treated as in **Example 13**. As shown in **FIG. 45**, the highest performing construct was found to be CREATE-37, which included longer PBS sequences (50 nucleotides). Construct nucleotide sequences are provided in **Table 11**. Construct amino acid sequences are provided in **Table 12**.

[0480] The combined effect of the additional codelivery of ORF2-only constructs with 50-nucleotide PBS sequences were investigated using the constructs described in **FIG. 46**. Cells were treated as in **Example 13**. As shown in **FIG. 47**, the combination of CREATE-37 containing 50-nucleotide PBS sequences and CREATE-80 containing an additional copy of ORF2 demonstrated the highest integration efficiency of the GFP payload.

Table 11. CREATE Construct Nucleotide Sequences

Name (size)	Description	Sequence	SEQ ID NO:
CREATE-29 (4.5kb)	3X FLAG	GACTATAAGGACCACGACGGAGACTACAAGGATCAT GATATTGATTACAAAGACGATGACGATAAG	122
	SV40 NLS	CCAAAGAAGAAGCGGAAGGTC	123
	Cas9 H840A Nickase (H840A mutation in bold)	GACAAGAAGTACAGCATCGGCCTGGACATCGGCACC AACTCTGTGGGCTGGGCCGTGATCACCGACGAGTAC AAGGTGCCCAGCAAGAAATTC AAGGTGCTGGGCAAC ACCGACCGGCACAGCATCAAGAAGAACCTGATCGGA GCCCTGCTGTTGACAGCGGCGAAACAGCCGAGGCC ACCCGGCTGAAGAGAACCGCCAGAAGAAGATACACC AGACGGAAGAACCGGATCTGCTATCTGCAAGAGATC TTCAGCAACGAGATGGCCAAGGTGGACGACAGCTTC TTCCACAGACTGGAAGAGTCCTTCCTGGTGG AAGAG GATAAGAAGCACGAGCGGCACCCCATCTTCGGCAAC ATCGTGGACGAGGTGGCCTACCACGAGAAGTACCCC ACCATCTACCACCTGAGAAAGAACTGGTGGACAGC ACCGACAAGGCCGACCTGCGGCTGATCTATCTGGCC CTGGCCACATGATCAAGTTCCGGGGCCACTTCCTGA TCGAGGGCGACCTGAACCCCGACAACAGCGACGTGG ACAAGCTGTTCATCCAGCTGGTGCAGACCTACAACC AGCTGTTGAGGAAAACCCCATCAACGCCAGCGGCG	124

Name (size)	Description	Sequence	SEQ ID NO:
		TGGACGCCAAGGCCATCCTGTCTGCCAGACTGAGCA AGAGCAGACGGCTGGAAAATCTGATCGCCCAGCTGC CCGGCGAGAAGAAGAATGGCCTGTTTCGAAACCTGA TTGCCCTGAGCCTGGGCCTGACCCCCAACTTCAAGAG CAACTTCGACCTGGCCGAGGATGCCAAACTGCAGCT GAGCAAGGACACCTACGACGACGACCTGGACAACCT GCTGGCCCAGATCGGCGACCAGTACGCCGACCTGTTT CTGGCCGCCAAGAACCTGTCCGACGCCATCCTGCTGA GCGACATCCTGAGAGTGAACACCGAGATCACCAAGG CCCCCTGAGCGCCTCTATGATCAAGAGATACGACG AGCACCACCAGGACCTGACCCTGCTGAAAGCTCTCG TGCGGCAGCAGCTGCCTGAGAAGTACAAAGAGATTT TCTTCGACCAGAGCAAGAACGGCTACGCCGGCTACA TTGACGGCGGAGCCAGCCAGGAAGAGTTCTACAAGT TCATCAAGCCCATCCTGGAAAAGATGGACGGCACCG AGGAACTGCTCGTGAAGCTGAACAGAGAGGACCTGC TGCGGAAGCAGCGGACCTTCGACAACGGCAGCATCC CCCACCAGATCCACCTGGGAGAGCTGCACGCCATTCT GCGGCGGCAGGAAGATTTTTACCCATTCTGAAGGA CAACCGGGAAAAGATCGAGAAGATCCTGACCTTCCG CATCCCCTACTACGTGGGCCCTCTGGCCAGGGGAAA CAGCAGATTTCGCTGGATGACCAGAAAGAGCGAGGA AACCATCACCCCCTGGA ACTTCGAGGAAGTGGTGG CAAGGGCGCTTCCGCCCAGAGCTTCATCGAGCGGAT GACCAACTTCGATAAGAACCTGCCCAACGAGAAGGT GCTGCCCAAGCACAGCCTGCTGTACGAGTACTTCACC GTGTATAACGAGCTGACCAAAGTGAAATACGTGACC GAGGGAATGAGAAAGCCCGCCTTCCTGAGCGGCGAG CAGAAAAAGGCCATCGTGGACCTGCTGTTCAAGACC AACCGGAAAGTGACCGTGAAGCAGCTGAAAGAGGAC TACTTCAAGAAAATCGAGTGCTTCGACTCCGTGGAAA TCTCCGGCGTGGAAGATCGGTTCAACGCCTCCCTGGG CACATACCACGATCTGCTGAAAATTATCAAGGACAA	

Name (size)	Description	Sequence	SEQ ID NO:
		GGACTTCCTGGACAATGAGGAAAACGAGGACATTCT GGAAGATATCGTGCTGACCCTGACACTGTTTGAGGAC AGAGAGATGATCGAGGAACGGCTGAAAACCTATGCC CACCTGTTTCGACGACAAAGTGATGAAGCAGCTGAAG CGGCGGAGATACACCGGCTGGGGCAGGCTGAGCCGG AAGCTGATCAACGGCATCCGGGACAAGCAGTCCGGC AAGACAATCCTGGATTTCTGAAGTCCGACGGCTTCG CCAACAGAACTTCATGCAGCTGATCCACGACGACA GCCTGACCTTTAAAGAGGACATCCAGAAAGCCCAGG TGTCGGGCCAGGGCGATAGCCTGCACGAGCACATTG CCAATCTGGCCGGCAGCCCCGCCATTAAGAAGGGCA TCCTGCAGACAGTGAAGGTGGTGGACGAGCTCGTGA AAGTGATGGGCCGGCACAAGCCCGAGAACATCGTGA TCGAAATGGCCAGAGAGAACCAGACCACCCAGAAGG GACAGAAGAACAGCCGCGAGAGAATGAAGCGGATC GAAGAGGGCATCAAAGAGCTGGGCAGCCAGATCCTG AAAGAACACCCCGTGGAACACCCAGCTGCAGAAC GAGAAGCTGTACCTGTACTACCTGCAGAATGGGCGG GATATGTACGTGGACCAGGAACTGGACATCAACCGG CTGTCCGACTACGATGTGGACGCCATCGTGCCTCAG AGCTTTCTGAAGGACGACTCCATCGACAACAAGGTG CTGACCAGAAGCGACAAGAACCGGGGCAAGAGCGA CAACGTGCCCTCCGAAGAGGTCTGTAAGAAGATGAA GAACTACTGGCGGCAGCTGCTGAACGCCAAGCTGAT TACCCAGAGAAAAGTTCGACAATCTGACCAAGGCCGA GAGAGGCGGCCTGAGCGAACTGGATAAGGCCGGCTT CATCAAGAGACAGCTGGTGGAAACCCGGCAGATCAC AAAGCACGTGGCACAGATCCTGGACTCCCGGATGAA CACTAAGTACGACGAGAATGACAAGCTGATCCGGGA AGTGAAAGTGATCACCTGAAGTCCAAGCTGGTGTC CGATTTCCGGAAGGATTTCCAGTTTTACAAAGTGCGC GAGATCAACAACCTACCACCACGCCCACGACGCCTAC CTGAACGCCGTCGTGGGAACCGCCCTGATCAAAAAG	

Name (size)	Description	Sequence	SEQ ID NO:
		TACCCTAAGCTGGAAAGCGAGTTCGTGTACGGCGAC TACAAGGTGTACGACGTGCGGAAGATGATCGCCAAG AGCGAGCAGGAAATCGGCAAGGCTACCGCCAAGTAC TTCTTCTACAGCAACATCATGAACTTTTTCAAGACCG AGATTACCCTGGCCAACGGCGAGATCCGGAAGCGGC CTCTGATCGAGACAAACGGCGAAACCGGGGAGATCG TGTGGGATAAAGGGCCGGGATTTTGCCACCGTGCGGA AAGTGCTGAGCATGCCCCAAGTGAATATCGTGAAAA AGACCGAGGTGCAGACAGGCGGCTTCAGCAAAGAGT CTATCCTGCCAAGAGGAACAGCGATAAGCTGATCG CCAGAAAGAAGGACTGGGACCCTAAGAAGTACGGCG GCTTCGACAGCCCCACCGTGGCCTATTCTGTGCTGGT GGTGGCCAAAGTGGAAGGGCAAGTCCAAGAACT GAAGAGTGTGAAAGAGCTGCTGGGGATCACCATCAT GGAAAGAAGCAGCTTCGAGAAGAATCCCATCGACTT TCTGGAAGCCAAGGGCTACAAAGAAGTGAAAAAGGA CCTGATCATCAAGCTGCCTAAGTACTCCCTGTTTCGAG CTGGAAAACGGCCGGAAGAGAATGCTGGCCTCTGCC GGCGAACTGCAGAAGGGAAACGAACTGGCCCTGCCC TCCAAATATGTGAACTTCCTGTACCTGGCCAGCCACT ATGAGAAGCTGAAGGGCTCCCCGAGGATAATGAGC AGAAACAGCTGTTTGTGGAACAGCACAAGCACTACC TGGACGAGATCATCGAGCAGATCAGCGAGTTCTCCA AGAGAGTGATCCTGGCCGACGCTAATCTGGACAAAG TGCTGTCCGCCTACAACAAGCACCGGGATAAGCCCA TCAGAGAGCAGGCCGAGAATATCATCCACCTGTTTAC CCTGACCAATCTGGGAGCCCCTGCCGCCTTCAAGTAC TTTGACACCACCATCGACCGGAAGAGGTACACCAGC ACCAAAGAGGTGCTGGACGCCACCCTGATCCACCAG AGCATCACCGGCCTGTACGAGACACGGATCGACCTG TCTCAGCTGGGAGGCGAC	
	Nucleoplasm in NLS	AAAAGGCCGGCGGCCACGAAAAAGGCCGGCCAGGC AAAAAAGAAAAAG	125

Name (size)	Description	Sequence	SEQ ID NO:
	Full Length	ATGGACTATAAGGACCACGACGGAGACTACAAGGAT CATGATATTGATTACAAAGACGATGACGATAAGATG GCCCCAAAGAAGAAGCGGAAGGTCGGTATCCACGGA GTCCCAGCAGCCGACAAGAAGTACAGCATCGGCCTG GACATCGGCACCAACTCTGTGGGCTGGGCCGTGATC ACCGACGAGTACAAGGTGCCAGCAAGAAATTCAAG GTGCTGGGCAACACCGACCGGCACAGCATCAAGAAG AACCTGATCGGAGCCCTGCTGTTCGACAGCGGCGAA ACAGCCGAGGCCACCCGGCTGAAGAGAACCGCCAGA AGAAGATACACCAGACGGAAGAACCGGATCTGCTAT CTGCAAGAGATCTTCAGCAACGAGATGGCCAAGGTG GACGACAGCTTCTTCCACAGACTGGAAGAGTCCTTCC TGGTGGAAAGAGGATAAGAAGCACGAGCGGCACCCCA TCTTCGGCAACATCGTGGACGAGGTGGCCTACCACG AGAAGTACCCCAACCATCTACCACCTGAGAAAGAAAC TGGTGGACAGCACCGACAAGGCCGACCTGCGGCTGA TCTATCTGGCCCTGGCCCACATGATCAAGTTCCGGGG CCACTTCCTGATCGAGGGCGACCTGAACCCCGACAA CAGCGACGTGGACAAGCTGTTCATCCAGCTGGTGCA GACCTACAACCAGCTGTTCGAGGAAAACCCCATCAA CGCCAGCGGCGTGGACGCCAAGGCCATCCTGTCTGC CAGACTGAGCAAGAGCAGACGGCTGGAAAATCTGAT CGCCCAGCTGCCCCGGCGAGAAGAAGAATGGCCTGTT CGGAAACCTGATTGCCCTGAGCCTGGGCCTGACCCCC AACTTCAAGAGCAACTTCGACCTGGCCGAGGATGCC AAAGTGCAGCTGAGCAAGGACACCTACGACGACGAC CTGGACAACCTGCTGGCCCAGATCGGCGACCAGTAC GCCGACCTGTTTCTGGCCGCCAAGAACCTGTCCGACG CCATCCTGCTGAGCGACATCCTGAGAGTGAACACCG AGATCACCAAGGCCCCCTGAGCGCCTCTATGATCA AGAGATACGACGAGCACCACCAGGACCTGACCTGTC TGAAAGCTCTCGTGCGGCAGCAGCTGCCTGAGAAGT ACAAAGAGATTTTCTTCGACCAGAGCAAGAACGGCT	126

Name (size)	Description	Sequence	SEQ ID NO:
		ACGCCGGCTACATTGACGGCGGAGCCAGCCAGGAAG AGTTCTACAAGTTCATCAAGCCCATCCTGGAAAAGAT GGACGGCACCGAGGAACTGCTCGTGAAGCTGAACAG AGAGGACCTGCTGCGGAAGCAGCGGACCTTCGACAA CGGCAGCATCCCCACCAGATCCACCTGGGAGAGCT GCACGCCATTCTGCGGCGGCAGGAAGATTTTACCCA TTCCTGAAGGACAACCGGGAAAAGATCGAGAAGATC CTGACCTTCCGCATCCCCTACTACGTGGGCCCTCTGG CCAGGGGAAACAGCAGATTTCGCCTGGATGACCAGAA AGAGCGAGGAAACCATCACCCCTGGAACCTTCGAGG AAGTGGTGGACAAGGGCGCTTCCGCCAGAGCTTCA TCGAGCGGATGACCAACTTCGATAAGAACCTGCCCA ACGAGAAGGTGCTGCCCAAGCACAGCCTGCTGTACG AGTACTTCACCGTGTATAACGAGCTGACCAAAGTGA AATACGTGACCGAGGGAATGAGAAAGCCCGCCTTCC TGAGCGGCGAGCAGAAAAAGGCCATCGTGGACCTGC TGTTCAAGACCAACCGGAAAGTGACCGTGAAGCAGC TGAAAGAGGACTACTTCAAGAAAATCGAGTGCTTCG ACTCCGTGGAATCTCCGGCGTGGAAGATCGGTTCA ACGCCTCCCTGGGCACATACCACGATCTGCTGAAAAT TATCAAGGACAAGGACTTCCTGGACAATGAGGAAAA CGAGGACATTCTGGAAGATATCGTGCTGACCTGAC ACTGTTTGAGGACAGAGAGATGATCGAGGAACGGCT GAAAACCTATGCCACCTGTTCGACGACAAAGTGAT GAAGCAGCTGAAGCGGCGGAGATACACCGGCTGGGG CAGGCTGAGCCGGAAGCTGATCAACGGCATCCGGGA CAAGCAGTCCGGCAAGACAATCCTGGATTTCCTGAA GTCCGACGGCTTCGCCAACAGAACTTCATGCAGCT GATCCACGACGACAGCCTGACCTTTAAAGAGGACAT CCAGAAAGCCCAGGTGTCCGGCCAGGGCGATAGCCT GCACGAGCACATTGCCAATCTGGCCGGCAGCCCCGC CATTAGAAGGGCATCCTGCAGACAGTGAAGGTGGT GGACGAGCTCGTGAAAGTGATGGGCCGGCACAAGCC	

Name (size)	Description	Sequence	SEQ ID NO:
		CGAGAACATCGTGATCGAAATGGCCAGAGAGAACCA GACCACCCAGAAGGGACAGAAGAACAGCCGCGAGA GAATGAAGCGGATCGAAGAGGGCATCAAAGAGCTGG GCAGCCAGATCCTGAAAGAACACCCCGTGGAACA CCCAGCTGCAGAACGAGAAGCTGTACCTGTACTACC TGCAGAATGGGCGGGATATGTACGTGGACCAGGAAC TGGACATCAACCGGCTGTCCGACTACGATGTGGACG CCATCGTGCCTCAGAGCTTTCTGAAGGACGACTCCAT CGACAACAAGGTGCTGACCAGAAGCGACAAGAACCG GGGCAAGAGCGACAACGTGCCCTCCGAAGAGGTCGT GAAGAAGATGAAGAACTACTGGCGGCAGCTGCTGAA CGCCAAGCTGATTACCCAGAGAAAAGTTCGACAATCT GACCAAGGCCGAGAGAGGGCGGCCTGAGCGAACTGG ATAAGGCCGGCTTCATCAAGAGACAGCTGGTGGA CCCGGCAGATCACAAAGCACGTGGCACAGATCCTGG ACTCCCGGATGAACACTAAGTACGACGAGAATGACA AGCTGATCCGGGAAGTGAAAGTGATCACCTGAAGT CCAAGCTGGTGTCCGATTTCCGGAAGGATTTCCAGTT TTACAAAGTGCGCGAGATCAACAACCTACCACCACGC CCACGACGCCTACCTGAACGCCGTCGTGGGAACCGC CCTGATCAAAAAGTACCCTAAGCTGGAAAGCGAGTT CGTGTACGGCGACTACAAGGTGTACGACGTGCGGAA GATGATCGCCAAGAGCGAGCAGGAAATCGGCAAGGC TACCGCCAAGTACTTCTTCTACAGCAACATCATGAAC TTTTTCAAGACCGAGATTACCCTGGCCAACGGCGAGA TCCGGAAGCGGCCTCTGATCGAGACAAACGGCGAAA CCGGGGAGATCGTGTGGGATAAGGGCCGGGATTTTG CCACCGTGCGGAAAGTGCTGAGCATGCCCCAAGTGA ATATCGTGAAAAAGACCGAGGTGCAGACAGGCGGCT TCAGCAAAGAGTCTATCCTGCCCCAAGAGGAACAGCG ATAAGCTGATCGCCAGAAAGAAGGACTGGGACCCTA AGAAGTACGGCGGCTTCGACAGCCCCACCGTGGCCT ATTCTGTGCTGGTGGTGGCCAAAGTGGAAGGGCA	

Name (size)	Description	Sequence	SEQ ID NO:
		AGTCCAAGAACTGAAGAGTGTGAAAGAGCTGCTGG GGATCACCATCATGGAAAGAAGCAGCTTCGAGAAGA ATCCCATCGACTTTCTGGAAGCCAAGGGCTACAAAG AAGTGAAAAAGGACCTGATCATCAAGCTGCCTAAGT ACTCCCTGTTCGAGCTGGAAAACGGCCGGAAGAGAA TGCTGGCCTCTGCCGGCGAACTGCAGAAGGGAAACG AACTGGCCCTGCCCTCCAAATATGTGAACTTCCTGTA CCTGGCCAGCCACTATGAGAAGCTGAAGGGCTCCCC CGAGGATAATGAGCAGAAACAGCTGTTTGTGGAACA GCACAAGCACTACCTGGACGAGATCATCGAGCAGAT CAGCGAGTTCTCCAAGAGAGTGATCCTGGCCGACGC TAATCTGGACAAAGTGCTGTCCGCCTACAACAAGCA CCGGGATAAGCCCATCAGAGAGCAGGCCGAGAATAT CATCCACCTGTTTACCCTGACCAATCTGGGAGCCCCT GCCGCCTTCAAGTACTTTGACACCACCATCGACCGGA AGAGGTACACCAGCACCAAAGAGGTGCTGGACGCCA CCCTGATCCACCAGAGCATCACCGGCCTGTACGAGA CACGGATCGACCTGTCTCAGCTGGGAGGCGACAAAA GGCCGGCGGCCACGAAAAAGGCCGGCCAGGCCAAAA AAGAAAAAGTAA	
CREATE-24 (6.4kb)	ORF1	ATGGGCAAGAAGCAAAATCGCAAGACGGGGAATTCC AAGACACAATCCGCTAGCCCACCACCTAAAGAGCGT TCTAGCTCCCCTGCTACTGAGCAGTCCTGGATGGAAA ACGACTTCGATGAACTCCGGGAAGAGGGATTTAGGC GATCCAATATTGAGAACTCCGCGAAGATATCCAGA CAAAGGGGAAGGAAGTCGAGAATTTGAGAAGAACC TCGAGGAGTGCATCACCCGTATCACAACACTGAGA AATGTCTCAAAGAACTCATGGAACCTAAGACAAAAG CCAGGGAGCTTCGAGAGGAGTGTGCGGAGTCTGAGAT CCAGGTGTGACCAGCTCGAGGAGCGCGTGAGCGCGA TGGAAGACGAGATGAACGAGATGAAAAGAGAGGGC AAATTCAGGGAGAAGCGCATTAAGAGGAACGAACAG AGTCTGCAGGAGATTTGGGATTACGTCAAGAGGCCT	10

Name (size)	Description	Sequence	SEQ ID NO:
		AACCTGCGGTTGATCGGCGTCCCCGAGAGCGACGTA GAAAACGGGACTAAACTGGAGAATACACTTCAAGAC ATCATTCAAGAAAATTTTCCAAACCTGGCTCGGCAAG CTAATGTGCAAATCCAAGAGATCCAACGCACACCCC AGCGGTATAGCTCTCGGCGTGCCACCCCTAGGCATAT TATCGTGCGCTTTACTAAGGTGGAGATGAAAGAGAA GATGCTGCGAGCCGCTCGGGAAAAGGGAAGGGTGAC TTTGAAGGGCAAACCTATTTCGGCTGACGGTTGACCTT AGCGCCGAGACACTCCAGGCACGCCGGAATGGGGC CCCATCTTTAATATCCTGAAGGAGAAGAACTTCCAGC CACGAATCTCTTACCCTGCAAAGTTGAGTTTATCTC CGAGGGTGAGATTAAGTATTTTCATCGATAAACAGAT GCTGCGAGACTTCGTGACAACTCGCCCAGCTCTCAAG GAACTGCTCAAAGAGGCTCTTAATATGGAGCGCAAT AATAGATATCAACCCTTGCAAGAACCACGCAAAGATG TGA	
	InterORF Sequence	AGACCATCAAGACTAGGAAGAACTGCATCAACTAA TGAGCAAAATCACCAGCTAACATCATA	127
	SV40 NLS	ATGCCAAAGAAGAAGCGGAAGGTTCGGCGGCGGCAG C	128
	ORF2 (EN dead, D205A mutation in bold)	ACCGGCTCTAACTCACATATCACCATCCTTACACTTA ACATTAACGGCCTCAACTCAGCTATCAAGCGCCATCG GCTGGCCAGCTGGATCAAATCACAGGATCCAAGCGT TTGTTGCATCCAAGAGACCCACCTGACGTGCAGGGA CACCCATCGGCTGAAAATAAAAGGCTGGCGGAAGAT CTACCAAGCCAATGGAAAACAAAAGAAGGCTGGGGT GGCGATACTTGTAAGCGATAAAACAGACTTTAAACC AACTAAGATCAAACGGGACAAAGAGGGCCATTACAT CATGGTAAAGGGTAGTATTCAACAAGAGGAGCTGAC TATCCTGAATATTTATGCACCTAATACTGGAGCCCCC AGATTCATAAAGCAAGTGTTGAGTGACCTTCAACGC GACCTCGACTCCCACACTCTGATCATGGGAGACTTTA ACACCCCGCTGTCCACTCTCGACAGATCTACTAGACA	129

Name (size)	Description	Sequence	SEQ ID NO:
		GAAAGTCAACAAGGATACACAGGAACTGAACAGTGC TCTCCACCAAGCGGACCTTATCGACATCTACAGAACA CTCCACCCCAAAAGCACAGAATATACCTTCTTTTCAG CCCCTCACCACACCTATTCCAAAATTGCCACATTGT GGGGAGTAAAGCCCTTCTCTCCAAATGTAAACGGAC CGAAATTATCACTAACTATCTCTCCGACCACAGTGCA ATAAACTTGAATTGCGAATTAAGAATCTCACTCAA GTAGATCCACGACATGGAACTGAACAATCTCCTCTT GAATGACTACTGGGTGCATAACGAAATGAAGGCTGA AATAAAGATGTTCTTTGAGACCAACGAAACAAAGA CACCACGTACCAGAATCTCTGGGACGCTTTCAAAGC AGTGTGTCGAGGAAAATTTATTGCACTGAATGCTTAC AAGCGGAAGCAGGAAAGATCCAAAATAGACACCCTG ACTAGCCAACTTAAAGAACTGGAAAAGCAAGAGCAA ACTCATAGCAAAGCTAGCCGTCGCCAAGAAATTACG AAAATCAGAGCTGAAGTGAAGGAAATTGAGACACAG AAAACCCTGCAAAAGATAAATGAAAGCCGCAGCTGG TTCTTTGAACGCATCAACAAAATCGATAGGCCACTTG CTCGCCTTATCAAGAAGAAAAGGGAGAAGAATCAAA TCGACACTATAAAGAATGATAAAGGCGATATAACCA CCGATCCCACAGAAATTCAAACAACCATACGCGAAT ACTACAAACACCTCTACGCCAATAAACTCGAAAATC TCGAGGAAATGGATACATTCCTCGACACGTACACCCT TCCCAGGCTGAACCAGGAAGAAGTTGAATCACTGAA TCGGCCTATCACGGGGAGTGAAATAGTAGCTATCATC AATTCCTCCCTACCAAGAAGTCACCCGGACCTGATG GATTCACCGCCGAATTCTACCAGAGATACATGGAAG AACTGGTGCCCTTCTTGCTGAACTTTTCCAAAGTAT TGAGAAAGAGGGAATACTTCCAACTCATTTTATGA GGCATCCATCATTCTGATCCCGAAGCCCGGCAGGGA CACGACCAAGAAAGAGAATTTTCGACCAATCTCATT GATGAACATTGATGCAAAGATCCTCAATAAAATACT GGCAAATCGGATTCAGCAGCACATAAAGAAGCTGAT	

Name (size)	Description	Sequence	SEQ ID NO:
		CCACCATGATCAAGTAGGCTTCATCCCCGGTATGCAA GGTTGGTTCAATATACGAAAATCAATCAATGTTATCC AGCATATAAACCGGGCCAAAGACAAGAACCACATGA TTATTAGTATCGATGCTGAGAAAGCCTTTGACAAAAT ACAACAACCCTTCATGCTGAAAACATTGAATAAGCT GGGAATTGATGGCACCTACTTCAAAATCATCAGAGC CATATATGACAAACCAACAGCAAATATCATTCTGAAT GGTCAGAAATTGGAAGCATTCCCCTTGAAAACCGGC ACACGGCAGGGTTGCCCTCTGTCACCACTCCTCTTCA ACATCGTGTTGGAAGTTCTTGCCCGCGCAATCCGGCA GGAAAAGGAAATCAAGGGCATTCAACTGGGCAAAGA GGAAGTTAAATTGAGCCTGTTTGCAGACGACATGATC GTCTATTTGGAAAACCCCATAGTTAGTGCACAAAATC TGCTGAAGTTGATCAGTAATTTCTCCAAAGTGAGTGG GTACAAAATCAATGTGCAAAAGAGCCAAGCTTTCTT GTACACCAACAACAGGCAAAGTCTCAAATCAT GGGCGAACTCCCCTTCGTGATTGCATCCAAGCGGATC AAATACCTGGGGATTCAATTGACTCGTGATGTGAAG GACCTCTTCAAGGAGAACTACAAACCCCTGCTCAAG GAAATCAAAGAGGACACAAACAAATGGAAGAACATT CCATGCTCTTGGGTGGGAAGGATCAATATCGTCAAA ATGGCCATCCTGCCCAAGGTAATTTACAGGTTCAATG CTATACCCATCAAGCTCCCCATGACATTCTTCACAGA ACTTGAAAAGACGACGCTGAAGTTCATTTGGAACCA GAAACGTGCCAGGATTGCTAAATCTATTCTCTCCCAA AAGAACAAAGCTGGCGGAATCACACTCCCAGACTTC AACTTTACTACAAGGCGACCGTGACGAAAACGGCT TGGTACTGGTACCAAAACAGGGATATAGATCAATGG AACCGAACGGAGCCCAGCGAAATTATGCCTCATATA TACAACTATCTGATCTTTGACAAACCGGAGAAGAAC AAGCAATGGGGAAAGGATAGTCTGTTTAATAAATGG TGCTGGGAAAACCTGGCTCGCAATCTGTAGGAAGCTG AACTGGATCCATTCTTGACGCCTTATACAAAGATAA	

Name (size)	Description	Sequence	SEQ ID NO:
		ATTCCCGATGGATTAAAGATCTCAACGTGAAACCCA AAACAATTAAAACCCTCGAGGAAAACCTGGGTATTA CGATTTCAGGACATTGGGGTGGGAAAGGACTTCATGT CCAAAACCCCAAAAGCGATGGCAACCAAAGACAAA ATCGACAAATGGGATCTCATAAACTTAAGTCATTTT GCACAGCTAAAGAAACGACAATTAGGGTGAACCGAC AACCGACCACTTGGGAGAAAATCTTCGCAACATACA GTTCTGACAAAGGCCTGATTTCAGGATCTACAATGA ATTGAAACAAATTTACAAGAAGAAGACGAACAACCC TATAAGAAATGGGCCAAGGACATGAACAGACACTT CTCTAAGGAAGACATTTATGCAGCCAAGAAACACAT GAAGAAATGCAGCTCTTCACTGGCAATCAGGGAAAT GCAAATCAAAACAACAATGAGATATCATCTCACACC CGTCAGAATGGCCATCATTAAAGAAGAGCGGAAACAA CCGGTGCTGGCGTGGTTGCGGAGAAATCGGTACTCTC CTTCCTGTTGGTGGGACTGTAAACTCGTTCAACCAC TGTGGAAGTCTGTGTGGCGGTTCTCAGAGATCTGGA ACTCGAAATCCCATTGACCCAGCCATCCCTCTCCTG GGTATATACCCGAATGAGTATAAATCCTGCTGCTATA AAGACACCTGCACAAGGATGTTTATTGCAGCTCTCTT CACAATCGCGAAGACGTGGAACCAACCCAAATGTCC GACTATGATTGACTGGATTAAGAAGATGTGGCACAT ATACACTATGGAATACTATGCTGCGATCAAGAACGA TGAGTTCATATCATTTGTGGGCACATGGATGAAACTC GAAACCATCATACTCTCTAAATTGAGTCAAGAACAG AAAACTAAACACCGTATATTTTCCCTGATCGGTGGGA AT	
	Nucleoplasm in NLS	AAAAGGCCGGCGGCCACGAAAAAGGCCGGCCAGGC AAAAAAGAAAAAG	125
	3' UTR	CTACAAAGACGATGACGACAAGGACCATGGAGACGG TGAGAGACACAAAAAATTCCAACACACTATTGCAAT GAAAATAAATTTCCCTTTATTAGCCAGAAGTCAGATGC TCAAGGGGCTTCATGATGTCCCCATAATTTTTGGCAG	130

Name (size)	Description	Sequence	SEQ ID NO:
		AGGGAAAAAGATCTCAGTGGTATTTGTGAGCCAGGG CATTGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGT GCGGCCGCT	
	PBS2	GCAGCTCAGGTTCTGGG	49
	SV40 PolyA	TAAGATACATTGATGAGTTTGGACAAACCACAATA GAATGCAGTGAAAAAATGCTTTATTTGTGAAATTTG TGATGCTATTGCTTTATTTGTAACCATTATAAGCTGC AATAACAAGTT	131
	GFP	TTACTTGTACAGCTCGTCCATGCCGAGAGTGATCCCG GCGGCGGTACGAACTCCAGCAGGACCATGTGATCG CGCTTCTCGTTGGGGTCTTTGCTCAGGGCGGACTGGG TGCTCAGGTAGTGGTTGTCTGGGCAGCAGCACGGGGC CGTCGCCGATGGGGGTGTTCTGCTGGTAGTGGTCGGC GAGCTGCACGCTGCCGTCCTCGATGTTGTGGCGGATC TTGAAGTTCACCTTGATGCCGTTCTTCTGCTTGTCTGGC CATGATATAGACGTTGTGGCTGTTGTAGTTGTACTCC AGCTTGTGCCCCAGGATGTTGCCGTCCTCCTTGAAGT CGATGCCCTTCAGCTCGATGCGGTTACCAGGGTGTC GCCCTCGAACTTCACCTCGGCGCGGGTCTTGTAGTTG CCGTCGTCCTTGAAGAAGATGGTGCGCTCCTGGACGT AGCCTTCGGGCATGGCGGACTTGAAGAAGTCGTGCT GCTTCATGTGGTCTGGGGTAGCGGCTGAAGCACTGCA CGCCGTAGGTACAGGGTGGTCACGAGGGTGGGCCAGG GCACGGGCAGCTTGCCGGTGGTGCAGATGAACTTCA GGGTCAGCTTGCCGTAGGTGGCATCGCCCTCGCCCTC GCCGGACACGCTGAACTTGTGGCCGTTTACGTCGCCG TCCAGCTCGACCAGGATGGGCACCACCCCGGTGAAC AGCTCCTCGCCCTTGCTCACCAT	132
	EF-1a Promoter	CTGTGTTCTGGCGGCAAACCCGTTGCGAAAAAGAAC GTTACGCGCGACTACTGCACTTATATACGGTTCTCCC CCACCCTCGGGAAAAAGGCGGAGCCAGTACACGACA TCACTTTCCCAGTTTACCCCGCGCCACCTTCTCTAGG	133

Name (size)	Description	Sequence	SEQ ID NO:
		CACCGGATCAATTGCCGACCCCTCCCCCAACTTCTC GGGGACTGTGGGCGATGTGCGCTCTGCCC	
	PBS1	TCCTCTCTGGCTC	48
	Full Length	ATGGGCAAGAAGCAAAATCGCAAGACGGGGAATTCC AAGACACAATCCGCTAGCCCACCACCTAAAGAGCGT TCTAGCTCCCCTGCTACTGAGCAGTCCTGGATGGAAA ACGACTTCGATGAACTCCGGGAAGAGGGATTTAGGC GATCCAATAATTCAGAACTCCGCGAAGATATCCAGA CAAAGGGGAAGGAAGTCGAGAATTTGAGAAGAACC TCGAGGAGTGCATACCCGTATCACAAACACTGAGA AATGTCTCAAAGAACTCATGGAACCTAAGACAAAAG CCAGGGAGCTTCGAGAGGAGTGTCGGAGTCTGAGAT CCAGGTGTGACCAGCTCGAGGAGCGCGTGAGCGCGA TGGAAGACGAGATGAACGAGATGAAAAGAGAGGGC AAATTCAGGGAGAAGCGCATTAAAGAGGAACGAACAG AGTCTGCAGGAGATTTGGGATTACGTCAAGAGGCCT AACCTGCGGTTGATCGGCGTCCCCGAGAGCGACGTA GAAAACGGGACTAACTGGAGAATACACTTCAAGAC ATCATTCAAGAAAATTTTCCAAACCTGGCTCGGCAAG CTAATGTGCAAATCCAAGAGATCCAACGCACACCCC AGCGGTATAGCTCTCGGCGTGCCACCCCTAGGCATAT TATCGTGCGCTTTACTAAGGTGGAGATGAAAGAGAA GATGCTGCGAGCCGCTCGGGAAAAGGGAAGGGTGAC TTTGAAGGGCAAACCTATTTCGGCTGACGGTTGACCTT AGCGCCGAGACACTCCAGGCACGCCGGAATGGGGC CCCATCTTTAATATCCTGAAGGAGAAGAACTTCCAGC CACGAATCTCTTACCCTGCAAAGTTGAGTTTTATCTC CGAGGGTGAGATTAAGTATTTTCATCGATAAACAGAT GCTGCGAGACTTCGTGACAACTCGCCCAGCTCTCAAG GAACTGCTCAAAGAGGCTCTTAATATGGAGCGCAAT AATAGATATCAACCCTTGCAGAACCACGCAAAGATG TGAGACAGCCGTCAGACCATCAAGACTAGGAAGAAA CTGCATCAACTAATGAGCAAAATCACCAGCTAACAT	134

Name (size)	Description	Sequence	SEQ ID NO:
		CATAGTATACATGCCAAAGAAGAAGCGGAAGGTCGG CGGCGGCAGCACCGGCTCTAACTCACATATCACCATC CTTAACTTAACATTAACGGCCTCAACTCAGCTATCA AGCGCCATCGGCTGGCCAGCTGGATCAAATCACAGG ATCCAAGCGTTTGTGTCATCCAAGAGACCCACCTGAC GTGCAGGGACACCCATCGGCTGAAAATAAAAGGCTG GCGGAAGATCTACCAAGCCAATGGAAAACAAAAGAA GGCTGGGGTGGCGATACTTGTAAGCGATAAAACAGA CTTTAAACCAACTAAGATCAAACGGGACAAAGAGGG CCATTACATCATGGTAAAGGGTAGTATTCAACAAGA GGAGCTGACTATCCTGAATATTTATGCACCTAATACT GGAGCCCCCAGATTCATAAAGCAAGTGTTGAGTGAC CTTCAACGCGACCTCGACTCCCACACTCTGATCATGG GAGACTTTAACACCCCGCTGTCCACTCTCGACAGATC TACTAGACAGAAAGTCAACAAGGATACACAGGAAT GAACAGTGCTCTCCACCAAGCGGACCTTATCGACATC TACAGAACACTCCACCCCAAAAGCACAGAATATACC TTCTTTTCAGCCCCTCACCACACCTATTCCAAAATTG CCCACATTGTGGGGAGTAAAGCCCTTCTCTCCAAATG TAAACGGACCGAAATTATCACTAACTATCTCTCCGAC CACAGTGCAATAAACTTGAATTGCGAATTAAGAAT CTCCTCAAAGTAGATCCACGACATGGAACTGAAC AATCTCCTCTTGAATGACTACTGGGTGCATAACGAAA TGAAGGCTGAAATAAAGATGTTCTTTGAGACCAACG AAAACAAAGACACCACGTACCAGAATCTCTGGGACG CTTTCAAAGCAGTGTGTCGAGGAAAATTTATTGCACT GAATGCTTACAAGCGGAAGCAGGAAAGATCCAAAAT AGACACCCTGACTAGCCAACTTAAAGAACTGGAAAA GCAAGAGCAAACCTCATAGCAAAGCTAGCCGTCGCCA AGAAATTACGAAAATCAGAGCTGAACTGAAGGAAAT TGAGACACAGAAAACCCTGCAAAAGATAAATGAAAG CCGCAGCTGGTTCTTTGAACGCATCAACAAAATCGAT AGGCCACTTGCTCGCCTTATCAAGAAGAAAAGGGAG	

Name (size)	Description	Sequence	SEQ ID NO:
		AAGAATCAAATCGACACTATAAAGAATGATAAAGGC GATATAACCACCGATCCACAGAAATTCAAACAACC ATACGCGAATACTACAAACACCTCTACGCCAATAAA CTCGAAAATCTCGAGGAAATGGATACATTCCTCGAC ACGTACACCCTTCCCAGGCTGAACCAGGAAGAAGTT GAATCACTGAATCGGCCTATCACGGGGAGTGAAATA GTAGCTATCATCAATTCCTCCCTACCAAGAAGTCAC CCGGACCTGATGGATTACCCGCCGAATTCTACCAGA GATACATGGAAGAACTGGTGCCCTTCTTGCTGAAACT TTTCCAAAGTATTGAGAAAGAGGGAATACTTCCAAA CTCATTTTATGAGGCATCCATCATTCTGATCCCGAAG CCCGGCAGGGACACGACCAAGAAAGAGAATTTTCGA CCAATCTCATTGATGAACATTGATGCAAAGATCCTCA ATAAAATACTGGCAAATCGGATTCAGCAGCACATAA AGAAGCTGATCCACCATGATCAAGTAGGCTTCATCCC CGGTATGCAAGGTTGGTTCAATATACGAAAATCAATC AATGTTATCCAGCATATAAACC GGCCAAAGACAAG AACCACATGATTATTAGTATCGATGCTGAGAAAGCCT TTGACAAAATACAACAACCCTTCATGCTGAAAACATT GAATAAGCTGGGAATTGATGGCACCTACTTCAAAT CATCAGAGCCATATATGACAAACCAACAGCAAATAT CATTCTGAATGGTCAGAAATTGGAAGCATTCCCCTTG AAAACCGGCACACGGCAGGGTTGCCCTCTGTCACCA CTCCTCTTCAACATCGTGTGGAAGTTCTTGCCCGCG CAATCCGGCAGGAAAAGGAAATCAAGGGCATTCAAC TGGGCAAAGAGGAAGTTAAATTGAGCCTGTTTGCAG ACGACATGATCGTCTATTTGGAAAACCCCATAGTTAG TGCACAAAATCTGCTGAAGTTGATCAGTAATTTCTCC AAAGTGAGTGGGTACAAAATCAATGTGCAAAGAGC CAAGCTTTCTTGACACCAACAACAGGCAAACCTGAGT CTCAAATCATGGGCGAACTCCCCTTCGTGATTGCATC CAAGCGGATCAAATACCTGGGGATTCAATTGACTCGT GATGTGAAGGACCTCTTCAAGGAGAACTACAAACCC	

Name (size)	Description	Sequence	SEQ ID NO:
		CTGCTCAAGGAAATCAAAGAGGACACAAACAAATGG AAGAACATTCCATGCTCTTGGGTGGGAAGGATCAAT ATCGTCAAAATGGCCATCCTGCCCAAGGTAATTTACA GGTTCAATGCTATACCCATCAAGCTCCCCATGACATT CTTCACAGAACTTGAAAAGACGACGCTGAAGTTCATT TGGAACCAGAAACGTGCCAGGATTGCTAAATCTATTC TCTCCCAAAAGAACAAGCTGGCGGAATCACACTCC CAGACTTCAAACCTTACTACAAGGCGACCGTGACGA AAACGGCTTGGTACTGGTACCAAAACAGGGATATAG ATCAATGGAACCGAACGGAGCCCAGCGAAATTATGC CTCATATATACAACTATCTGATCTTTGACAAACCGGA GAAGAACAAGCAATGGGGAAAGGATAGTCTGTTTAA TAAATGGTGCTGGGAAAACCTGGCTCGCAATCTGTAG GAAGCTGAAACTGGATCCATTCTTGACGCCTTATACA AAGATAAATTCCCGATGGATTAAAGATCTCAACGTG AAACCCAAAACAATTAACCCCTCGAGGAAAACCTG GGTATTACGATTCAGGACATTGGGGTGGGAAAGGAC TTCATGTCCAAAACCCCAAAAGCGATGGCAACCAAA GACAAAATCGACAAATGGGATCTCATAAACTTAAG TCATTTTGCACAGCTAAAGAAACGACAATTAGGGTG AACCGACAACCGACCACTTGGGAGAAAATCTTCGCA ACATACAGTTCTGACAAAGGCCTGATTTCCAGGATCT ACAATGAATTGAAACAAATTTACAAGAAGAAGACGA ACAACCTATAAAGAAATGGGCCAAGGACATGAACA GACACTTCTCTAAGGAAGACATTTATGCAGCCAAGA AACACATGAAGAAATGCAGCTCTTCACTGGCAATCA GGGAAATGCAAATCAAAACAACAATGAGATATCATC TCACACCCGTCAGAATGGCCATCATTAAGAAGAGCG GAAACAACCGGTGCTGGCGTGGTTGCGGAGAAATCG GTA CTCTCCTTCACTGTTGGTGGGACTGTAACTCGT TCAACCACTGTGGAAGTCTGTGTGGCGGTTCTCAGA GATCTGGAACTCGAAATCCCATTTGACCCAGCCATCC CTCTCCTGGGTATATACCCGAATGAGTATAAATCCTG	

Name (size)	Description	Sequence	SEQ ID NO:
		CTGCTATAAAGACACCTGCACAAGGATGTTTATTGCA GCTCTCTTCACAATCGCGAAGACGTGGAACCAACCC AAATGTCCGACTATGATTGACTGGATTAAGAAGATGT GGCACATATACACTATGGAATACTATGCTGCGATCAA GAACGATGAGTTCATATCATTTGTGGGCACATGGATG AAACTCGAAACCATCATACTCTCTAAATTGAGTCAAG AACAGAAAACCTAAACACCGTATATTTTCCCTGATCGG TGGGAATGGAGGCAAAAGGCCGCGGCCACGAAAA AGGCCGGCCAGGCAAAAAAGAAAAAGTAGCTACAA AGACGATGACGACAAGGACCATGGAGACGGTGAGA GACACAAAAAATTCCAACACACTATTGCAATGAAAA TAAATTTCTTTATTAGCCAGAAGTCAGATGCTCAAG GGGCTTCATGATGTCCCCATAATTTTTTGGCAGAGGGA AAAAGATCTCAGTGGTATTTGTGAGCCAGGGCATTG GCCTTCTGATAGGCAGCCTGCACCTGAGGAGTGCGG CCGCTGCAGCTCAGGTTCTGGGTAAGATACATTGATG AGTTTGGACAAACCACAACCTAGAATGCAGTGAAAAA AATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTA TTGTAAACCATTATAAGCTGCAATAAACAAGTTTTAC TTGTACAGCTCGTCCATGCCGAGAGTGATCCCGGCCG CGGTCACGAACCTCCAGCAGGACCATGTGATCGCGCT TCTCGTTGGGGTCTTTGCTCAGGGCGGACTGGGTGCT CAGGTAGTGGTTGTCGGGCAGCAGCACGGGGCCGTC GCCGATGGGGGTGTTCTGCTGGTAGTGGTCGGCGAG CTGCACGCTGCCGTCCTCGATGTTGTGGCGGATCTTG AAGTTCACCTTGATGCCGTTCTTCTGCTTGTGCGCCAT GATATAGACGTTGTGGCTGTTGTAGTTGTACTCCAGC TTGTGCCCCAGGATGTTGCCGTCCTCCTTGAAGTCGA TGCCCTTCAGCTCGATGCGGTTACCAGGGTGTGCC CTCGAACTTCACCTCGGCGCGGGTCTTGTAGTTGCCG TCGTCCTTGAAGAAGATGGTGCGCTCCTGGACGTAGC CTTCGGGCATGGCGGACTTGAAGAAGTCGTGCTGCTT CATGTGGTCGGGGTAGCGGCTGAAGCACTGCACGCC	

Name (size)	Description	Sequence	SEQ ID NO:
		GTAGGTCAGGGTGGTCACGAGGGTGGGCCAGGGCAC GGGCAGCTTGCCGGTGGTGCAGATGAACTTCAGGGT CAGCTTGCCGTAGGTGGCATCGCCCTCGCCCTCGCCG GACACGCTGAACTTGTGGCCGTTTACGTCGCCGTCCA GCTCGACCAGGATGGGCACCAACCCCGGTGAACAGCT CCTCGCCCTTGCTCACCATGGTGGCGGGATCTGACGG TTCATAAACCCTGTGTTCTGGCGGCAAACCCGTTGC GAAAAAGAACGTTACGGCGACTACTGCACTTATAT ACGGTTCTCCCCCACCCTCGGGAAAAAGGCGGAGCC AGTACACGACATCACTTTCCAGTTTACCCCGCGCCA CCTTCTCTAGGCACCGGATCAATTGCCGACCCCTCCC CCCAACTTCTCGGGGACTGTGGGCGATGTGCGCTCTG CCCTCCTCTCTGGCTC	
CREATE-81 (6.3kb)	ORF1	ATGGGCAAGAAGCAAAATCGCAAGACGGGGAATTCC AAGACACAATCCGCTAGCCCACCACCTAAAGAGCGT TCTAGCTCCCCTGCTACTGAGCAGTCCTGGATGGAAA ACGACTTCGATGAACTCCGGGAAGAGGGATTTAGGC GATCCAATAATTCAGAACTCCGCGAAGATATCCAGA CAAAGGGGAAGGAAGTCGAGAATTTGAGAAGAACC TCGAGGAGTGCATCACCCGTATCACAAACACTGAGA AATGTCTCAAAGAACTCATGGAACTTAAGACAAAAG CCAGGGAGCTTCGAGAGGAGTGTGCGGAGTCTGAGAT CCAGGTGTGACCAGCTCGAGGAGCGCGTGAGCGCGA TGGAAGACGAGATGAACGAGATGAAAAGAGAGGGC AAATTCAGGGAGAAGCGCATTAAAGAGGAACGAACAG AGTCTGCAGGAGATTTGGGATTACGTCAAGAGGCCT AACCTGCGGTTGATCGGCGTCCCCGAGAGCGACGTA GAAAACGGGACTAACTGGAGAATACACTTCAAGAC ATCATTCAAGAAAATTTTCCAAACCTGGCTCGGCAAG CTAATGTGCAAATCCAAGAGATCCAACGCACACCCC AGCGGTATAGCTCTCGGCGTGCCACCCCTAGGCATAT TATCGTGCGCTTTACTAAGGTGGAGATGAAAGAGAA GATGCTGCGAGCCGCTCGGGAAAAAGGGAAGGGTGAC	10

Name (size)	Description	Sequence	SEQ ID NO:
		TTTGAAGGGCAAACCTATTCGGCTGACGGTTGACCTT AGCGCCGAGACACTCCAGGCACGCCGGAATGGGGC CCCATCTTTAATATCCTGAAGGAGAAGAACTTCCAGC CACGAATCTCTTACCCTGCAAAGTTGAGTTTTATCTC CGAGGGTGAGATTAAGTATTTTCATCGATAAACAGAT GCTGCGAGACTTCGTGACAACTCGCCCAGCTCTCAAG GAACTGCTCAAAGAGGCTCTTAATATGGAGCGCAAT AATAGATATCAACCCTTGCAAGAACCACGCAAAGATG	
	GSG Linker	GGTAGCGGC	135
	T2A Sequence	GAGGGCAGGGGAAGTCTTCTAACATGCGGGGACGTG GAGGAAAATCCCGGCCCA	136
	SV40 NLS	ATGCCAAAGAAGAAGCGGAAGGTGCGCGGCGGCAG C	128
	ORF2 (EN dead, D205A)	ACCGGCTCTAACTCACATATCACCATCCTTACACTTA ACATTAACGGCCTCAACTCAGCTATCAAGCGCCATCG GCTGGCCAGCTGGATCAAATCACAGGATCCAAGCGT TTGTTGCATCCAAGAGACCCACCTGACGTGCAGGGA CACCCATCGGCTGAAAATAAAAGGCTGGCGGAAGAT CTACCAAGCCAATGGAAAACAAAAGAAGGCTGGGGT GGCGATACTTGTAAGCGATAAAACAGACTTTAAACC AACTAAGATCAAACGGGACAAAGAGGGGCCATTACAT CATGGTAAAGGGTAGTATTCAACAAGAGGAGCTGAC TATCCTGAATATTTATGCACCTAATACTGGAGCCCC AGATTTCATAAAGCAAGTGTTGAGTGACCTTCAACGC GACCTCGACTCCCACACTCTGATCATGGGAGACTTTA ACACCCCGCTGTCCACTCTCGACAGATCTACTAGACA GAAAGTCAACAAGGATACACAGGAACTGAACAGTGC TCTCCACCAAGCGGACCTTATCGACATCTACAGAACA CTCCACCCCAAAAGCACAGAATATACCTTCTTTTCAG CCCCTCACCACACCTATTCCAAAATTGCCACATTGT GGGGAGTAAAGCCCTTCTCTCCAAATGTAAACGGAC CGAAATTATCACTAACTATCTCTCCGACCACAGTGCA ATAAACTTGAATTGCGAATTAAGAATCTCACTCAA	129

Name (size)	Description	Sequence	SEQ ID NO:
		GTAGATCCACGACATGGAACTGAACAATCTCCTCTT GAATGACTACTGGGTGCATAACGAAATGAAGGCTGA AATAAAGATGTTCTTTGAGACCAACGAAAACAAAGA CACCACGTACCAGAATCTCTGGGACGCTTTCAAAGC AGTGTGTCGAGGAAAATTTATTGCACTGAATGCTTAC AAGCGGAAGCAGGAAAGATCCAAAATAGACACCCTG ACTAGCCAACTTAAAGAACTGGAAAAGCAAGAGCAA ACTCATAGCAAAGCTAGCCGTCGCCAAGAAATTACG AAAATCAGAGCTGAACTGAAGGAAATTGAGACACAG AAAACCCTGCAAAAGATAAATGAAAGCCGCAGCTGG TTCTTTGAACGCATCAACAAAATCGATAGGCCACTTG CTCGCCTTATCAAGAAGAAAAGGGAGAAGAATCAAA TCGACACTATAAAGAATGATAAAGGCGATATAACCA CCGATCCCACAGAAATTCAAACAACCATACGCGAAT ACTACAAACACCTCTACGCCAATAAACTCGAAAATC TCGAGGAAATGGATACATTCCTCGACACGTACACCCT TCCCAGGCTGAACCAGGAAGAAGTTGAATCACTGAA TCGGCCTATCACGGGGAGTGAAATAGTAGCTATCATC AATTCACCTCCCTACCAAGAAGTCACCCGGACCTGATG GATTCACCGCCGAATTCTACCAGAGATACATGGAAG AACTGGTGCCCTTCTTGCTGAACTTTTCCAAAGTAT TGAGAAAGAGGGAATACTTCCAAACTCATTTTATGA GGCATCCATCATTCTGATCCCGAAGCCCGGCAGGGA CACGACCAAGAAAGAGAATTTTCGACCAATCTCATT GATGAACATTGATGCAAAGATCCTCAATAAAATACT GGCAAATCGGATTCAGCAGCACATAAAGAAGCTGAT CCACCATGATCAAGTAGGCTTCATCCCCGGTATGCAA GGTTGGTTCAATATACGAAAATCAATCAATGTTATCC AGCATATAAACCGGGCCAAAGACAAGAACCACATGA TTATTAGTATCGATGCTGAGAAAGCCTTTGACAAAAT ACAACAACCCTTCATGCTGAAAACATTGAATAAGCT GGGAATTGATGGCACCTACTTCAAAATCATCAGAGC CATATATGACAAACCAACAGCAAATATCATTCTGAAT	

Name (size)	Description	Sequence	SEQ ID NO:
		GGTCAGAAATTGGAAGCATTCCCCTTGAAAACCGGC ACACGGCAGGGTTGCCCTCTGTCACCACTCCTCTTCA ACATCGTGTTGGAAGTTCTTGCCCGCGCAATCCGGCA GGAAAAGGAAATCAAGGGCATTCAACTGGGCAAAGA GGAAGTTAAATTGAGCCTGTTTGCAGACGACATGATC GTCTATTTGAAAACCCCATAGTTAGTGCACAAAATC TGCTGAAGTTGATCAGTAATTTCTCCAAAGTGAGTGG GTACAAAATCAATGTGCAAAAGAGCCAAGCTTTCTT GTACACCAACAACAGGCAAAGTCTCAAATCAT GGGCGAACTCCCCCTTCGTGATTGCATCCAAGCGGATC AAATACCTGGGGATTCAATTGACTCGTGATGTGAAG GACCTCTTCAAGGAGAACTACAAACCCCTGCTCAAG GAAATCAAAGAGGACACAAACAAATGGAAGAACATT CCATGCTCTTGGGTGGGAAGGATCAATATCGTCAAA ATGGCCATCCTGCCCAAGGTAATTTACAGGTTCAATG CTATACCCATCAAGCTCCCCATGACATTCTTCACAGA ACTTGAAAAGACGACGCTGAAGTTCATTTGGAACCA GAAACGTGCCAGGATTGCTAAATCTATTCTCTCCCAA AAGAACAAAGCTGGCGGAATCACACTCCCAGACTTC AACTTTACTACAAGGCGACCGTGACGAAAACGGCT TGGTACTGGTACCAAAACAGGGATATAGATCAATGG AACCGAACGGAGCCCAGCGAAATTATGCCTCATATA TACAACTATCTGATCTTTGACAAACCGGAGAAGAAC AAGCAATGGGGAAAGGATAGTCTGTTTAATAAATGG TGCTGGGAAAACCTGGCTCGCAATCTGTAGGAAGCTG AACTGGATCCATTCTTGACGCCTTATACAAAGATAA ATTCCCGATGGATTAAAGATCTCAACGTGAAACCCA AAACAATTAAAACCCTCGAGGAAAACCTGGGTATTA CGATTCAGGACATTGGGGTGGGAAAGGACTTCATGT CCAAAACCCAAAAGCGATGGCAACCAAAGACAAA ATCGACAAATGGGATCTCATAAACTTAAGTCATTTT GCACAGCTAAAGAAACGACAATTAGGGTGAACCGAC AACCGACCACTTGGGAGAAAATCTTCGCAACATACA	

Name (size)	Description	Sequence	SEQ ID NO:
		GTTCTGACAAAGGCCTGATTTCCAGGATCTACAATGA ATTGAAACAAATTTACAAGAAGAAGACGAACAACCC TATAAAGAAATGGGCCAAGGACATGAACAGACACTT CTCTAAGGAAGACATTTATGCAGCCAAGAAACACAT GAAGAAATGCAGCTCTTCACTGGCAATCAGGGAAAT GCAAATCAAAACAACAATGAGATATCATCTCACACC CGTCAGAATGGCCATCATTAAAGAAGAGCGGAAACAA CCGGTGCTGGCGTGGTTGCGGAGAAATCGGTACTCTC CTTCACTGTTGGTGGGACTGTAAACTCGTTCAACCAC TGTGGAAGTCTGTGTGGCGGTTCCCTCAGAGATCTGGA ACTCGAAATCCCATTGACCCAGCCATCCCTCTCCTG GGTATATACCCGAATGAGTATAAATCCTGCTGCTATA AAGACACCTGCACAAGGATGTTTATTGCAGCTCTCTT CACAATCGCGAAGACGTGGAACCAACCCAAATGTCC GACTATGATTGACTGGATTAAGAAGATGTGGCACAT ATACACTATGGAATACTATGCTGCGATCAAGAACGA TGAGTTCATATCATTTGTGGGCACATGGATGAAACTC GAAACCATCATACTCTCTAAATTGAGTCAAGAACAG AAAACATAACACCGTATATTTCCCTGATCGGTGGGA AT	
	Nucleoplasm in NLS	AAAAGGCCGGCGGCCACGAAAAAGGCCGGCCAGGC AAAAAAGAAAAAG	125
	3' UTR	CTACAAAGACGATGACGACAAGGACCATGGAGACGG TGAGAGACACAAAAAATTCCAACACACTATTGCAAT GAAAATAAATTTCTTTATTAGCCAGAAGTCAGATGC TCAAGGGGCTTCATGATGTCCCCATAATTTTTGGCAG AGGGAAAAAGATCTCAGTGGTATTTGTGAGCCAGGG CATTGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGT GCGGCCGCT	130
	PBS2	GCAGCTCAGGTTCTGGG	49
	SV40 PolyA	TAAGATACATTGATGAGTTTGGACAAACCACAATA GAATGCAGTGAAAAAATGCTTTATTTGTGAAATTTG	131

Name (size)	Description	Sequence	SEQ ID NO:
		TGATGCTATTGCTTTATTTGTAACCATTATAAGCTGC AATAAACAAGTT	
	GFP	TTACTTGTACAGCTCGTCCATGCCGAGAGTGATCCCG GCGGCGGTACGAACTCCAGCAGGACCATGTGATCG CGCTTCTCGTTGGGGTCTTTGCTCAGGGCGGACTGGG TGCTCAGGTAGTGGTTGTCGGGCAGCAGCACGGGGC CGTCGCCGATGGGGGTGTTCTGCTGGTAGTGGTCGGC GAGCTGCACGCTGCCGTCTCGATGTTGTGGCGGATC TTGAAGTTCACCTTGATGCCGTTCTTCTGCTTGTGGC CATGATATAGACGTTGTGGCTGTTGTAGTTGTACTCC AGCTTGTGCCCCAGGATGTTGCCGTCTCTCTGAAGT CGATGCCCTTCAGCTCGATGCGGTTACCAGGGTGTC GCCCTCGAACTTCACCTCGGCGCGGGTCTTGTAGTTG CCGTCGTCCTTGAAGAAGATGGTGCGCTCCTGGACGT AGCCTTCGGGCATGGCGGACTTGAAGAAGTCGTGCT GCTTCATGTGGTCGGGGTAGCGGCTGAAGCACTGCA CGCCGTAGGTACAGGGTGGTCACGAGGGTGGGCCAGG GCACGGGCAGCTTGCCGGTGGTGCAGATGAACTTCA GGGTCAGCTTGCCGTAGGTGGCATCGCCCTCGCCCTC GCCGGACACGCTGAACTTGTGGCCGTTTACGTCGCCG TCCAGCTCGACCAGGATGGGCACCACCCCGGTGAAC AGCTCCTCGCCCTTGCTCACCAT	132
	EF-1a Promoter	CTGTGTTCTGGCGGCAAACCCGTTGCGAAAAAGAAC GTTACGGCGACTACTGCACTTATATACGGTTCTCCC CCACCCTCGGGAAGGCGGAGCCAGTACACGACA TCACTTTCCCAGTTTACCCCGCGCCACCTTCTCTAGG CACCGGATCAATTGCCGACCCCTCCCCCAACTTCTC GGGGACTGTGGGCGATGTGCGCTCTGCCC	133
	PBS1	TCCTCTCTGGCTC	48
	Full Length	ATGGGCAAGAAGCAAAATCGCAAGACGGGGAATTCC AAGACACAATCCGCTAGCCCACCACCTAAAGAGCGT TCTAGCTCCCCTGCTACTGAGCAGTCCTGGATGGAAA ACGACTTCGATGAACTCCGGGAAGAGGGATTTAGGC	137

Name (size)	Description	Sequence	SEQ ID NO:
		GATCCAAC TATTCAGAACTCCGCGAAGATATCCAGA CAAAGGGGAAGGAAGTCGAGAATTTTCGAGAAGAACC TCGAGGAGTGCATCACCCGTATCACAAACACTGAGA AATGTCTCAAAGAACTCATGGAAC TTAAGACAAAAG CCAGGGAGCTTCGAGAGGAGTGTCTGGAGTCTGAGAT CCAGGTGTGACCAGCTCGAGGAGCGCGTGAGCGCGA TGGAAGACGAGATGAACGAGATGAAAAGAGAGGGC AAATTCAGGGAGAAGCGCATTAAAGAGGAACGAACAG AGTCTGCAGGAGATTTGGGATTACGTCAAGAGGCCT AACCTGCGGTTGATCGGCGTCCCCGAGAGCGACGTA GAAAACGGGACTAAACTGGAGAATACACTTCAAGAC ATCATTCAAGAAAATTTTCCAAACCTGGCTCGGCAAG CTAATGTGCAAATCCAAGAGATCCAACGCACACCCC AGCGGTATAGCTCTCGGCGTGCCACCCCTAGGCATAT TATCGTGCGCTTTACTAAGGTGGAGATGAAAGAGAA GATGCTGCGAGCCGCTCGGGAAAAGGGAAGGGTGAC TTTGAAGGGCAAACCTATTTCGGCTGACGGTTGACCTT AGCGCCGAGACACTCCAGGCACGCCGGGAATGGGGC CCCATCTTTAATATCCTGAAGGAGAAGAACTTCCAGC CACGAATCTCTTACCCTGCAAAGTTGAGTTTATCTC CGAGGGTGAGATTAAGTATTTTCATCGATAAACAGAT GCTGCGAGACTTCGTGACAACTCGCCCAGCTCTCAAG GAACTGCTCAAAGAGGCTCTTAATATGGAGCGCAAT AATAGATATCAACCCTTGCAGAACCACGCAAAGATG GGTAGCGGCGAGGGCAGGGGAAGTCTTCTAACATGC GGGGACGTGGAGGAAAATCCCGGCCCAATGCCAAAG AAGAAGCGGAAGGTCGGCGGCGGCAGCACCGGCTCT AACTCACATATCACCATCCTTACACTTAACATTAACG GCCTCAACTCAGCTATCAAGCGCCATCGGCTGGCCA GCTGGATCAAATCACAGGATCCAAGCGTTTGTTCAT CCAAGAGACCCACCTGACGTGCAGGGACACCCATCG GCTGAAAATAAAAGGCTGGCGGAAGATCTACCAAGC CAATGGAAAACAAAAGAAGGCTGGGGTGGCGATACT	

Name (size)	Description	Sequence	SEQ ID NO:
		TGTAAGCGATAAAACAGACTTTAAACCAACTAAGAT CAAACGGGACAAAGAGGGCCATTACATCATGGTAAA GGGTAGTATTCAACAAGAGGAGCTGACTATCCTGAA TATTTATGCACCTAATACTGGAGCCCCCAGATTCATA AAGCAAGTGTTGAGTGACCTTCAACGCGACCTCGACT CCCACACTCTGATCATGGGAGACTTTAACACCCCGCT GTCCACTCTCGACAGATCTACTAGACAGAAAGTCAA CAAGGATACACAGGAAGTGAACAGTGCTCTCCACCA AGCGGACCTTATCGACATCTACAGAACACTCCACCCC AAAAGCACAGAATATACCTTCTTTTCAGCCCCTCACC ACACCTATTCCAAAATTGCCACATTGTGGGGAGTAA AGCCCTTCTCTCCAAATGTAAACGGACCGAAAATTATC ACTAACTATCTCTCCGACCACAGTGCAATAAAACTTG AATTGCGAATTAAGAATCTCACTCAAAGTAGATCCAC GACATGGAAACTGAACAATCTCCTCTTGAATGACTAC TGGGTGCATAACGAAATGAAGGCTGAAATAAAGATG TTCTTTGAGACCAACGAAAACAAAGACACCACGTAC CAGAATCTCTGGGACGCTTTCAAAGCAGTGTGTCGAG GAAAATTTATTGCACTGAATGCTTACAAGCGGAAGC AGGAAAGATCCAAAATAGACACCCTGACTAGCCAAC TTAAAGAACTGGAAAAGCAAGAGCAAATCATAGCA AAGCTAGCCGTCGCCAAGAAATTACGAAAATCAGAG CTGAACTGAAGGAAATTGAGACACAGAAAACCTGC AAAAGATAAATGAAAGCCGCAGCTGGTTCTTTGAAC GCATCAACAAAATCGATAGGCCACTTGCTCGCCTTAT CAAGAAGAAAAGGGAGAAGAATCAAATCGACACTAT AAAGAATGATAAAGGCGATATAACCACCGATCCCAC AGAAATTCAAACAACCATACGCGAATACTACAAACA CCTCTACGCCAATAAACTCGAAAATCTCGAGGAAAT GGATACATTCTCGACACGTACACCTTCCCAGGCTG AACCAGGAAGAAGTTGAATCACTGAATCGGCCTATC ACGGGGAGTGAAATAGTAGCTATCATCAATCACTCC CTACCAAGAAGTCACCCGGACCTGATGGATTACCG	

Name (size)	Description	Sequence	SEQ ID NO:
		CCGAATTCTACCAGAGATACATGGAAGAACTGGTGC CCTTCTTGCTGAAACTTTTCCAAAGTATTGAGAAAGA GGGAATACTTCCAAACTCATTTTATGAGGCATCCATC ATTCTGATCCCGAAGCCCGGCAGGGACACGACCAAG AAAGAGAATTTTCGACCAATCTCATTGATGAACATTG ATGCAAAGATCCTCAATAAAATACTGGCAAATCGGA TTCAGCAGCACATAAAGAAGCTGATCCACCATGATC AAGTAGGCTTCATCCCCGGTATGCAAGGTTGGTCAA TATACGAAAATCAATCAATGTTATCCAGCATATAAAC CGGGCCAAAGACAAGAACCACATGATTATTAGTATC GATGCTGAGAAAGCCTTTGACAAAATACAACAACCC TTCATGCTGAAAACATTGAATAAGCTGGGAATTGATG GCACCTACTTCAAAATCATCAGAGCCATATATGACAA ACCAACAGCAAATATCATTCTGAATGGTCAGAAATT GGAAGCATTCCCCTTGAAAACCGGCACACGGCAGGG TTGCCCTCTGTCACCACTCCTCTTCAACATCGTGTGG AAGTTCTTGCCCGCGCAATCCGGCAGGAAAAGGAAA TCAAGGGCATTCAACTGGGCAAAGAGGAAGTTAAAT TGAGCCTGTTTGCAGACGACATGATCGTCTATTTGGA AAACCCCATAGTTAGTGCACAAAATCTGCTGAAGTTG ATCAGTAATTTCTCCAAAGTGAGTGGGTACAAAATCA ATGTGCAAAAGAGCCAAGCTTTCTTGTACACCAACA ACAGGCAAACCTGAGTCTCAAATCATGGGCGAACTCC CCTTCGTGATTGCATCCAAGCGGATCAAATACCTGGG GATTCAATTGACTCGTGATGTGAAGGACCTCTTCAAG GAGAACTACAAACCCCTGCTCAAGGAAATCAAAGAG GACACAAACAAATGGAAGAACATTCCATGCTCTTGG GTGGGAAGGATCAATATCGTCAAAATGGCCATCCTG CCCAAGGTAATTTACAGGTTCAATGCTATACCCATCA AGCTCCCCATGACATTCTTACAGAACTTGAAAAGAC GACGCTGAAGTTCATTTGGAACCAGAAACGTGCCAG GATTGCTAAATCTATTCTCTCCCAAAGAACAAGCT GGCGGAATCACACTCCAGACTTCAAACCTTACTACA	

Name (size)	Description	Sequence	SEQ ID NO:
		AGGCGACCGTGACGAAAACGGCTTGGTACTGGTACC AAAACAGGGATATAGATCAATGGAACCGAACGGAGC CCAGCGAAATTATGCCTCATATATACAACCTATCTGAT CTTTGACAAACCGGAGAAGAACAAGCAATGGGGAAA GGATAGTCTGTTTAATAAATGGTGCTGGGAAAACCTG GCTCGCAATCTGTAGGAAGCTGAAACTGGATCCATTC TTGACGCCTTATACAAAGATAAATTCCCGATGGATTA AAGATCTCAACGTGAAACCCAAAACAATTAACCC TCGAGGAAAACCTGGGTATTACGATTCAGGACATTG GGGTGGGAAAGGACTTCATGTCCAAAACCCCAAAAG CGATGGCAACCAAAGACAAAATCGACAAATGGGATC TCATAAACTTAAGTCATTTTGCACAGCTAAAGAAAC GACAATTAGGGTGAACCGACAACCGACCACTTGGGA GAAAATCTTCGCAACATACAGTTCTGACAAAGGCCT GATTTCCAGGATCTACAATGAATTGAAACAAATTTAC AAGAAGAAGACGAACAACCCTATAAAGAAATGGGCC AAGGACATGAACAGACACTTCTCTAAGGAAGACATT TATGCAGCCAAGAAACACATGAAGAAATGCAGCTCT TCACTGGCAATCAGGGAAATGCAAATCAAAACAACA ATGAGATATCATCTCACACCCGTCAGAATGGCCATCA TTAAGAAGAGCGGAAACAACCGGTGCTGGCGTGGTT GCGGAGAAATCGGTACTCTCCTTCACTGTTGGTGGGA CTGTAACTCGTTCAACCACTGTGGAAGTCTGTGTGG CGGTTCTCAGAGATCTGGAACCTCGAAATCCCATTG ACCCAGCCATCCCTCTCCTGGGTATATACCCGAATGA GTATAAATCCTGCTGCTATAAAGACACCTGCACAAG GATGTTTATTGCAGCTCTCTTCACAATCGCGAAGACG TGGAACCAACCCAAATGTCCGACTATGATTGACTGG ATTAAGAAGATGTGGCACATATACACTATGGAATAC TATGCTGCGATCAAGAACGATGAGTTCATATCATTTG TGGGCACATGGATGAACTCGAAACCATCATACTCTC TAAATTGAGTCAAGAACAGAAAACCTAAACACCGTAT ATTTCCCTGATCGGTGGGAATGGAGGCAAAAGGCC	

Name (size)	Description	Sequence	SEQ ID NO:
		GGCGGCCACGAAAAAGGCCGGCCAGGCAAAAAAGA AAAAGTAGCTACAAAGACGATGACGACAAGGACCAT GGAGACGGTGAGAGACACAAAAAATTCCAACACACT ATTGCAATGAAAATAAATTTCCCTTTATTAGCCAGAAG TCAGATGCTCAAGGGGCTTCATGATGTCCCCATAATT TTTGGCAGAGGGAAAAAGATCTCAGTGGTATTTGTG AGCCAGGGCATTGGCCTTCTGATAGGCAGCCTGCAC CTGAGGAGTGCGGCCGCTGCAGCTCAGGTTCTGGGT AAGATACATTGATGAGTTTGGACAAACCACAACCTAG AATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGT GATGCTATTGCTTTATTTGTAACCATTATAAGCTGCA ATAAACAAGTTTTACTTGTACAGCTCGTCCATGCCGA GAGTGATCCCGGCGGCGGTACGAACTCCAGCAGGA CCATGTGATCGCGCTTCTCGTTGGGGTCTTTGCTCAG GGCGGACTGGGTGCTCAGGTAGTGGTTGTCGGGCAG CAGCACGGGGCCGTCGCCGATGGGGGTGTTCTGCTG GTAGTGGTTCGGCGAGCTGCACGCTGCCGTCTCGATG TTGTGGCGGATCTTGAAGTTCACCTTGATGCCGTTCT TCTGCTTGTCGGCCATGATATAGACGTTGTGGCTGTT GTAGTTGTACTCCAGCTTGTGCCCCAGGATGTTGCCG TCCTCCTTGAAGTCGATGCCCTTCAGCTCGATGCGGT TCACCAGGGTGTCGCCCTCGAACTTCACCTCGGCGCG GGTCTTGTAGTTGCCGTCGTCCTTGAAGAAGATGGTG CGCTCCTGGACGTAGCCTTCGGGCATGGCAGGACTTGA AGAAGTCGTGCTGCTTCATGTGGTTCGGGGTAGCGGCT GAAGCACTGCACGCCGTAGGTCAGGGTGGTCACGAG GGTGGGCCAGGGCACGGGCAGCTTGCCGGTGGTGCA GATGAACTTCAGGGTCAGCTTGCCGTAGGTGGCATCG CCCTCGCCCTCGCCGGACACGCTGAACTTGTGGCCGT TTACGTCGCCGTCCAGCTCGACCAGGATGGGCACCA CCCCGGTGAACAGCTCCTCGCCCTTGCTCACCATGGT GGCGGGATCTGACGGTTCATAAACCCCTGTGTTCTGG CGGCAAACCCGTTGCGAAAAAGAACGTTACGGCGA	

Name (size)	Description	Sequence	SEQ ID NO:
		CTACTGCACTTATATACGGTTCTCCCCACCCTCGGG AAAAAGGCGGAGCCAGTACACGACATCACTTTCCCA GTTTACCCCGCGCCACCTTCTCTAGGCACCGGATCAA TTGCCGACCCCTCCCCCAACTTCTCGGGGACTGTGG GCGATGTGCGCTCTGCCCTCCTCTCTGGCTC	
	SV40 NLS	ATGCCAAAGAAGAAGCGGAAGGTCGGCGGCGGCAG C	128
CREATE-80 (4.2kb)	ORF2 (EN dead, D205A)	ACCGGCTCTAACTCACATATCACCATCCTTACACTTA ACATTAAACGGCCTCAACTCAGCTATCAAGCGCCATCG GCTGGCCAGCTGGATCAAATCACAGGATCCAAGCGT TTGTTGCATCCAAGAGACCCACCTGACGTGCAGGGA CACCCATCGGCTGAAAATAAAAGGCTGGCGGAAGAT CTACCAAGCCAATGGAAAACAAAAGAAGGCTGGGGT GGCGATACTTGTAAGCGATAAAACAGACTTTAAACC AACTAAGATCAAACGGGACAAAGAGGGCCATTACAT CATGGTAAAGGGTAGTATTCAACAAGAGGAGCTGAC TATCCTGAATATTTATGCACCTAATACTGGAGCCCC AGATTCATAAAGCAAGTGTTGAGTGACCTTCAACGC GACCTCGACTCCCACACTCTGATCATGGGAGACTTTA ACACCCCGCTGTCCACTCTCGACAGATCTACTAGACA GAAAGTCAACAAGGATACACAGGAACTGAACAGTGC TCTCCACCAAGCGGACCTTATCGACATCTACAGAACA CTCCACCCCAAAAGCACAGAATATACCTTCTTTTCAG CCCCTCACCACACCTATTCCAAAATTGCCACATTGT GGGGAGTAAAGCCCTTCTCTCCAAATGTAAACGGAC CGAAATTATCACTAACTATCTCTCCGACCACAGTGCA ATAAACTTGAATTGCGAATTAAGAATCTCACTCAA GTAGATCCACGACATGGAAACTGAACAATCTCCTCTT GAATGACTACTGGGTGCATAACGAAATGAAGGCTGA AATAAAGATGTTCTTTGAGACCAACGAAAACAAAGA CACCACGTACCAGAATCTCTGGGACGCTTTCAAAGC AGTGTGTCGAGGAAAATTTATTGCACTGAATGCTTAC AAGCGGAAGCAGGAAAGATCCAAAATAGACACCCTG	129

Name (size)	Description	Sequence	SEQ ID NO:
		ACTAGCCAACTTAAAGAACTGGAAAAGCAAGAGCAA ACTCATAGCAAAGCTAGCCGTCGCCAAGAAATTACG AAAATCAGAGCTGAACTGAAGGAAATTGAGACACAG AAAACCCTGCAAAAGATAAATGAAAGCCGCAGCTGG TTCTTTGAACGCATCAACAAAATCGATAGGCCACTTG CTCGCCTTATCAAGAAGAAAAGGGAGAAGAATCAAA TCGACACTATAAAGAATGATAAAGGCGATATAACCA CCGATCCCACAGAAATTCAAACAACCATACGCGAAT ACTACAAACACCTCTACGCCAATAAACTCGAAAATC TCGAGGAAATGGATACATTCCCTCGACACGTACACCCT TCCCAGGCTGAACCAGGAAGAAGTTGAATCACTGAA TCGGCCTATCACGGGGAGTGAAATAGTAGCTATCATC AATTCACCTCCCTACCAAGAAGTCACCCGGACCTGATG GATTCACCGCCGAATTCTACCAGAGATACATGGAAG AACTGGTGCCCTTCTTGCTGAAACTTTTCCAAAGTAT TGAGAAAGAGGGAATACTTCCAAACTCATTTTATGA GGCATCCATCATTCTGATCCCGAAGCCCGGCAGGGA CACGACCAAGAAAGAGAATTTTCGACCAATCTCATT GATGAACATTGATGCAAAGATCCTCAATAAAATACT GGCAAATCGGATTCAGCAGCACATAAAGAAGCTGAT CCACCATGATCAAGTAGGCTTCATCCCCGGTATGCAA GGTTGGTTCAATATACGAAAATCAATCAATGTTATCC AGCATATAAACCGGGCCAAAGACAAGAACCACATGA TTATTAGTATCGATGCTGAGAAAGCCTTTGACAAAAT ACAACAACCCTTCATGCTGAAAACATTGAATAAGCT GGGAATTGATGGCACCTACTTCAAAATCATCAGAGC CATATATGACAAACCAACAGCAAATATCATTCTGAAT GGTCAGAAATTGGAAGCATTCCCCTTGAAAACCGGC ACACGGCAGGGTTGCCCTCTGTCACCACTCCTCTTCA ACATCGTGTTGGAAGTTCTTGCCCGCGCAATCCGGCA GGAAAAGGAAATCAAGGGCATTCAACTGGGCAAAGA GGAAGTTAAATTGAGCCTGTTTGCAGACGACATGATC GTCTATTTGAAAACCCCATAGTTAGTGCACAAAATC	

Name (size)	Description	Sequence	SEQ ID NO:
		TGCTGAAGTTGATCAGTAATTTCTCCAAAGTGAGTGG GTACAAAATCAATGTGCAAAAGAGCCAAGCTTTCTT GTACACCAACAACAGGCAAAGTCTCAAATCAT GGGCGAACTCCCCTTCGTGATTGCATCCAAGCGGATC AAATACCTGGGGATTCAATTGACTCGTGATGTGAAG GACCTCTTCAAGGAGAACTACAAACCCCTGCTCAAG GAAATCAAAGAGGACACAAACAAATGGAAGAACATT CCATGCTCTTGGGTGGGAAGGATCAATATCGTCAAA ATGGCCATCCTGCCCAAGGTAATTTACAGGTTCAATG CTATACCCATCAAGCTCCCCATGACATTCTTCACAGA ACTTGAAAAGACGACGCTGAAGTTCATTTGGAACCA GAAACGTGCCAGGATTGCTAAATCTATTCTCTCCCAA AAGAACAAAGCTGGCGGAATCACACTCCCAGACTTC AACTTTACTACAAGGCGACCGTGACGAAAACGGCT TGGTACTGGTACCAAAACAGGGATATAGATCAATGG AACCGAACGGAGCCCAGCGAAATTATGCCTCATATA TACAACTATCTGATCTTTGACAAACCGGAGAAGAAC AAGCAATGGGGAAAGGATAGTCTGTTTAATAAATGG TGCTGGGAAAACCTGGCTCGCAATCTGTAGGAAGCTG AACTGGATCCATTCTTGACGCCTTATACAAAGATAA ATTCCCGATGGATTAAAGATCTCAACGTGAAAACCCA AAACAATTAAAACCCCTCGAGGAAAACCTGGGTATTA CGATTACAGGACATTGGGGTGGGAAAGGACTTCATGT CCAAAACCCCAAAAGCGATGGCAACCAAAGACAAA ATCGACAAATGGGATCTCATAAACTTAAGTCATTTT GCACAGCTAAAGAAACGACAATTAGGGTGAACCGAC AACCGACCACTTGGGAGAAAATCTTCGCAACATACA GTTCTGACAAAGGCCTGATTTCAGGATCTACAATGA ATTGAAACAAATTTACAAGAAGAAGACGAACAACCC TATAAAGAAATGGGCCAAGGACATGAACAGACACTT CTCTAAGGAAGACATTTATGCAGCCAAGAAACACAT GAAGAAATGCAGCTCTTCACTGGCAATCAGGGAAAT GCAAATCAAAACAACAATGAGATATCATCTCACACC	

Name (size)	Description	Sequence	SEQ ID NO:
		CGTCAGAATGGCCATCATTAAAGAAGAGCGGAAACAA CCGGTGCTGGCGTGGTTGCGGAGAAATCGGTACTCTC CTTCACTGTTGGTGGGACTGTAAACTCGTTCAACCAC TGTGGAAGTCTGTGTGGCGGTTTCCTCAGAGATCTGGA ACTCGAAATCCCATTGACCCAGCCATCCCTCTCCTG GGTATATACCCGAATGAGTATAAATCCTGCTGCTATA AAGACACCTGCACAAGGATGTTTATTGCAGCTCTCTT CACAATCGCGAAGACGTGGAACCAACCCAAATGTCC GACTATGATTGACTGGATTAAGAAGATGTGGCACAT ATACACTATGGAATACTATGCTGCGATCAAGAACGA TGAGTTCATATCATTTGTGGGCACATGGATGAAACTC GAAACCATCATACTCTCTAAATTGAGTCAAGAACAG AAAATAAACACCGTATATTTTCCCTGATCGGTGGGA AT	
	Nucleoplasm in NLS	AAAAGGCCGCGGCCACGAAAAAGGCCGGCCAGGC AAAAAAGAAAAAG	125
	FLAG	GACTACAAAGACGATGACGACAAG	138
	3' UTR	CTACAAAGACGATGACGACAAGGACCATGGAGACGG TGAGAGACACAAAAAATTCCAACACACTATTGCAAT GAAAATAAATTTCTTTATTAGCCAGAAGTCAGATGC TCAAGGGGCTTCATGATGTCCCCATAATTTTGGCAG AGGGAAAAAGATCTCAGTGGTATTTGTGAGCCAGGG CATTGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGT GCGGCCGCTTAAGCTGCAATAAACAAGTT	139
	Full Length	ATGCCAAAGAAGAAGCGGAAGGTCGGCGGCGGCAG CACCGGCTCTAACTCACATATCACCATCCTTACACTT AACATTAACGGCCTCAACTCAGCTATCAAGCGCCATC GGCTGGCCAGCTGGATCAAATCACAGGATCCAAGCG TTTGTTGCATCCAAGAGACCCACCTGACGTGCAGGGA CACCCATCGGCTGAAAATAAAAGGCTGGCGGAAGAT CTACCAAGCCAATGGAAAACAAAAGAAGGCTGGGGT GGCGATACTTGTAAGCGATAAAACAGACTTTAAACC AACTAAGATCAAACGGGACAAAGAGGGCCATTACAT	140

Name (size)	Description	Sequence	SEQ ID NO:
		CATGGTAAAGGGTAGTATTCAACAAGAGGAGCTGAC TATCCTGAATATTTATGCACCTAATACTGGAGCCCC AGATTCATAAAGCAAGTGTTGAGTGACCTTCAACGC GACCTCGACTCCCACACTCTGATCATGGGAGACTTTA ACACCCCGCTGTCCACTCTCGACAGATCTACTAGACA GAAAGTCAACAAGGATACACAGGAACTGAACAGTGC TCTCCACCAAGCGGACCTTATCGACATCTACAGAACA CTCCACCCCAAAAGCACAGAATATACCTTCTTTTCAG CCCCTCACCACACCTATTCCAAAATTGCCACATTGT GGGGAGTAAAGCCCTTCTCTCCAAATGTAAACGGAC CGAAATTATCACTAACTATCTCTCCGACCACAGTGCA ATAAAACTTGAATTGCGAATTAAGAATCTCACTCAA GTAGATCCACGACATGGAACTGAACAATCTCCTCTT GAATGACTACTGGGTGCATAACGAAATGAAGGCTGA AATAAAGATGTTCTTTGAGACCAACGAAAACAAAGA CACCACGTACCAGAATCTCTGGGACGCTTTCAAAGC AGTGTGTCGAGGAAAATTTATTGCACTGAATGCTTAC AAGCGGAAGCAGGAAAGATCCAAAATAGACACCCTG ACTAGCCAACTTAAAGAACTGGAAAAGCAAGAGCAA ACTCATAGCAAAGCTAGCCGTCGCCAAGAAATTACG AAAATCAGAGCTGAACTGAAGGAAATTGAGACACAG AAAACCCTGCAAAAGATAAATGAAAGCCGCAGCTGG TTCTTTGAACGCATCAACAAAATCGATAGGCCACTTG CTCGCCTTATCAAGAAGAAAAGGGAGAAGAATCAAA TCGACACTATAAAGAATGATAAAGGCGATATAACCA CCGATCCCACAGAAATTCAAACAACCATACGCGAAT ACTACAAACACCTCTACGCCAATAAACTCGAAAATC TCGAGGAAATGGATACATTCCTCGACACGTACACCCT TCCCAGGCTGAACCAGGAAGAAGTTGAATCACTGAA TCGGCCTATCACGGGGAGTGAAATAGTAGCTATCATC AATTCCTCCCTACCAAGAAGTCACCCGGACCTGATG GATTCACCGCCGAATTCTACCAGAGATACATGGAAG AACTGGTGCCCTTCTTGCTGAAACTTTTCCAAAGTAT	

Name (size)	Description	Sequence	SEQ ID NO:
		TGAGAAAGAGGGAATACTTCCAAACTCATTTTATGA GGCATCCATCATTCTGATCCCGAAGCCCGGCAGGGA CACGACCAAGAAAGAGAATTTTCGACCAATCTCATT GATGAACATTGATGCAAAGATCCTCAATAAAATACT GGCAAATCGGATTCAGCAGCACATAAAGAAGCTGAT CCACCATGATCAAGTAGGCTTCATCCCCGGTATGCAA GGTTGGTTCAATATACGAAAATCAATCAATGTTATCC AGCATATAAACCGGGCCAAAGACAAGAACCACATGA TTATTAGTATCGATGCTGAGAAAGCCTTTGACAAAAT ACAACAACCCTTCATGCTGAAAACATTGAATAAGCT GGGAATTGATGGCACCTACTTCAAATCATCAGAGC CATATATGACAAACCAACAGCAAATATCATTCTGAAT GGTCAGAAATTGGAAGCATTCCCCTTGAAAACCGGC ACACGGCAGGGTTGCCCTCTGTCACCACTCCTCTTCA ACATCGTGTTGGAAGTTCTTGCCCGCGCAATCCGGCA GGAAAAGGAAATCAAGGGCATTCAACTGGGCAAAGA GGAAGTTAAATTGAGCCTGTTTGCAGACGACATGATC GTCTATTTGGAAAACCCCATAGTTAGTGCACAAAATC TGCTGAAGTTGATCAGTAATTTCTCCAAAGTGAGTGG GTACAAAATCAATGTGCAAAAGAGCCAAGCTTTCTT GTACACCAACAACAGGCAAACCTGAGTCTCAAATCAT GGGCGAACTCCCCTTCGTGATTGCATCCAAGCGGATC AAATACCTGGGGATTCAATTGACTCGTGATGTGAAG GACCTCTTCAAGGAGAACTACAAACCCCTGCTCAAG GAAATCAAAGAGGACACAAACAAATGGAAGAACATT CCATGCTCTTGGGTGGGAAGGATCAATATCGTCAAA ATGGCCATCCTGCCCAAGGTAATTTACAGGTTCAATG CTATACCCATCAAGCTCCCCATGACATTCTTCACAGA ACTTGAAAAGACGACGCTGAAGTTCATTTGGAACCA GAAACGTGCCAGGATTGCTAAATCTATTCTCTCCCAA AAGAACAAAGCTGGCGGAATCACACTCCAGACTTC AACTTTACTACAAGGCGACCGTGACGAAAACGGCT TGGTACTGGTACCAAAACAGGGATATAGATCAATGG	

Name (size)	Description	Sequence	SEQ ID NO:
		AACCGAACGGAGCCCAGCGAAATTATGCCTCATATA TACAACTATCTGATCTTTGACAAACCGGAGAAGAAC AAGCAATGGGGAAAGGATAGTCTGTTTAATAAATGG TGCTGGGAAAACCTGGCTCGCAATCTGTAGGAAGCTG AAACTGGATCCATTCTTGACGCCTTATACAAAGATAA ATTCCCGATGGATTAAGATCTCAACGTGAAACCCA AAACAATTAAAACCTCGAGGAAAACCTGGGTATTA CGATTCAGGACATTGGGGTGGGAAAGGACTTCATGT CCAAAACCCCAAAAGCGATGGCAACCAAAGACAAA ATCGACAAATGGGATCTCATAAACTTAAGTCATTTT GCACAGCTAAAGAAACGACAATTAGGGTGAACCGAC AACCGACCACTTGGGAGAAAATCTTCGCAACATACA GTTCTGACAAAGGCCTGATTTCCAGGATCTACAATGA ATTGAAACAAATTTACAAGAAGAAGACGAACAACCC TATAAAGAAATGGGCCAAGGACATGAACAGACACTT CTCTAAGGAAGACATTTATGCAGCCAAGAAACACAT GAAGAAATGCAGCTCTTCACTGGCAATCAGGGAAAT GCAAATCAAAACAACAATGAGATATCATCTCACACC CGTCAGAATGGCCATCATTAAAGAAGAGCGGAAACAA CCGGTGCTGGCGTGGTTGCGGAGAAATCCGTACTCTC CTTCACTGTTGGTGGGACTGTAACTCGTTCAACCAC TGTGGAAGTCTGTGTGGCGGTTTCCTCAGAGATCTGGA ACTCGAAATCCCATTTGACCCAGCCATCCCTCTCCTG GGTATATACCCGAATGAGTATAAATCCTGCTGCTATA AAGACACCTGCACAAGGATGTTTATTGCAGCTCTCTT CACAATCGCGAAGACGTGGAACCAACCCAAATGTCC GACTATGATTGACTGGATTAAGAAGATGTGGCACAT ATACACTATGGAATACTATGCTGCGATCAAGAACGA TGAGTTCATATCATTTGTGGGCACATGGATGAAACTC GAAACCATCATACTCTCTAAATTGAGTCAAGAACAG AAAATAAACACCGTATATTTTCCCTGATCGGTGGGA ATGGAGGCAAAAGGCCGGCGGCCACGAAAAAGGCC GGCCAGGCAAAAAAGAAAAAGGACTACAAAGACGA	

Name (size)	Description	Sequence	SEQ ID NO:
		TGACGACAAGTAACTACAAAGACGATGACGACAAGG ACCATGGAGACGGTGAGAGACACAAAAAATTCCAAC ACACTATTGCAATGAAAATAAATTTCTTTATTAGCC AGAAGTCAGATGCTCAAGGGGCTTCATGATGTCCCC ATAATTTTTGGCAGAGGGAAAAAGATCTCAGTGGTA TTTGTGAGCCAGGGCATTGGCCTTCTGATAGGCAGCC TGCACCTGAGGAGTGCGGCCGCTTAAGCTGCAATAA ACAAGTT	
CREATE-82 (4.5kb)	SV40 NLS	ATGCCAAAGAAGAAGCGGAAGGTCTGGCGGCGGCAG C	128
	Sso7d (human codon optimized)	ATGGCCACCGTGAAGTTCAAGTACAAGGGCGAGGAG AAGGAGGTGGACATCAGCAAGATCAAGAAGGTGTGG AGGGTGGGCAAGATGATCAGCTTCACCTACGACGAG GGCGGCGGCAAGACCGGCAGGGGCGCCGTGAGCGA GAAGGACGCCCCAAGGAGCTGCTGCAGATGCTGGA GAAGCAGAAGAAG	141
	33XTEN	TCTGGAGGATCTAGCGGAGGATCCTCTGGCAGCGAG ACACCAGGAACAAGCGAGTCAGCAACACCAGAGAG CAGTGGCGGCAGCAGCGGCGGCAGCAGC	142
	ORF2 (aa. 239-1275)	ATTAAGAATCTCACTCAAAGTAGATCCACGACATGG AAACTGAACAATCTCCTCTTGAATGACTACTGGGTGC ATAACGAAATGAAGGCTGAAATAAAGATGTTCTTTG AGACCAACGAAAACAAAGACACCACGTACCAGAATC TCTGGGACGCTTTCAAAGCAGTGTGTCGAGGAAAATT TATTGCACTGAATGCTTACAAGCGGAAGCAGGAAAG ATCCAAAATAGACACCCTGACTAGCCAACTTAAAGA ACTGGAAAAGCAAGAGCAAATCATAGCAAAGCTAG CCGTCGCCAAGAAATTACGAAAATCAGAGCTGAACT GAAGGAAATTGAGACACAGAAAACCTGCAAAAAGAT AAATGAAAGCCGCAGCTGGTTCTTTGAACGCATCAA CAAAATCGATAGGCCACTTGCTCGCCTTATCAAGAAG AAAAGGGAGAAGAATCAAATCGACACTATAAAGAAT GATAAAGGCGATATAACCACCGATCCACAGAAATT	143

Name (size)	Description	Sequence	SEQ ID NO:
		CAAACAACCATACGCGAATACTACAAACACCTCTAC GCCAATAAACTCGAAAATCTCGAGGAAATGGATACA TTCCTCGACACGTACACCCTTCCCAGGCTGAACCAGG AAGAAGTTGAATCACTGAATCGGCCTATCACGGGGA GTGAAATAGTAGCTATCATCAATTCACCTCCCTACCAA GAAGTCACCCGGACCTGATGGATTACCGCCGAATT CTACCAGAGATACATGGAAGAACTGGTGCCCTTCTTG CTGAAACTTTTCCAAAGTATTGAGAAAGAGGGAATA CTTCCAACTCATTTTATGAGGCATCCATCATTCTGA TCCCGAAGCCCGGCAGGGACACGACCAAGAAAGAG AATTTTCGACCAATCTCATTGATGAACATTGATGCAA AGATCCTCAATAAAAATACTGGCAAATCGGATTCAGC AGCACATAAAGAAGCTGATCCACCATGATCAAGTAG GCTTCATCCCCGGTATGCAAGGTTGGTTCAATATACG AAAATCAATCAATGTTATCCAGCATATAAACCGGGC CAAAGACAAGAACCACATGATTATTAGTATCGATGC TGAGAAAGCCTTTGACAAAATACAACAACCCTTCAT GCTGAAAACATTGAATAAGCTGGGAATTGATGGCAC CTACTTCAAAATCATCAGAGCCATATATGACAAACCA ACAGCAAATATCATTCTGAATGGTCAGAAATTGGAA GCATTCCCCTTGAAAACCGGCACACGGCAGGGTTGC CCTCTGTCACCACTCCTCTTCAACATCGTGTGGAAG TTCTTGCCCGCGCAATCCGGCAGGAAAAGGAAATCA AGGGCATTCAACTGGGCAAAGAGGAAGTTAAATTGA GCCTGTTTGCAGACGACATGATCGTCTATTTGGAAAA CCCCATAGTTAGTGCACAAAATCTGCTGAAGTTGATC AGTAATTTCTCCAAAGTGAGTGGGTACAAAATCAAT GTGCAAAGAGCCAAGCTTTCTTGTACACCAACAAC AGGCAAATGAGTCTCAAATCATGGGCGAACTCCCC TTCGTGATTGCATCCAAGCGGATCAAATACCTGGGGA TTCAATTGACTCGTGATGTGAAGGACCTCTTCAAGGA GAACTACAAACCCCTGCTCAAGGAAATCAAAGAGGA CACAAACAAATGGAAGAACATTCCATGCTCTTGGGT	

Name (size)	Description	Sequence	SEQ ID NO:
		GGGAAGGATCAATATCGTCAAAATGGCCATCCTGCC CAAGGTAATTTACAGGTTCAATGCTATACCCATCAAG CTCCCCATGACATTCTTCACAGAACTTGAAAAGACGA CGCTGAAGTTCATTTGGAACCAGAAACGTGCCAGGA TTGCTAAATCTATTCTCTCCCAAAAGAACAAGCTGG CGGAATCACACTCCCAGACTTCAAACCTTACTACAAG GCGACCGTGACGAAAACGGCTTGGTACTGGTACCAA AACAGGGATATAGATCAATGGAACCGAACGGAGCCC AGCGAAATTATGCCTCATATATACAACTATCTGATCT TTGACAAACCGGAGAAGAACAAGCAATGGGGAAAG GATAGTCTGTTTAATAAATGGTGCTGGGAAACTGGC TCGCAATCTGTAGGAAGCTGAAACTGGATCCATTCTT GACGCCTTATACAAAGATAAATTCCCGATGGATTAA AGATCTCAACGTGAAACCCAAAACAATTAAAACCTT CGAGGAAAACCTGGGTATTACGATTCAGGACATTGG GGTGGGAAAGGACTTCATGTCCAAAACCCCAAAAGC GATGGCAACCAAAGACAAAATCGACAAATGGGATCT CATAAACTTAAGTCATTTTGCACAGCTAAAGAAAC GACAATTAGGGTGAACCGACAACCGACCACTTGGGA GAAAATCTTCGCAACATACAGTTCTGACAAAGGCCT GATTTCCAGGATCTACAATGAATTGAAACAAATTTAC AAGAAGAAGACGAACAACCTATAAAGAAATGGGCC AAGGACATGAACAGACACTTCTCTAAGGAAGACATT TATGCAGCCAAGAAACACATGAAGAAATGCAGCTCT TCACTGGCAATCAGGGAAATGCAAATCAAAACAACA ATGAGATATCATCTCACACCCGTCAGAATGGCCATCA TTAAGAAGAGCGGAAACAACCGGTGCTGGCGTGGTT GCGGAGAAATCGGTACTCTCCTTCACTGTTGGTGGA CTGTAACTCGTTCAACCACTGTGGAAGTCTGTGTGG CGGTTCCCTCAGAGATCTGGAACCTCGAAATCCCATTG ACCCAGCCATCCCTCTCCTGGGTATATACCCGAATGA GTATAAATCCTGCTGCTATAAAGACACCTGCACAAG GATGTTTATTGCAGCTCTCTTCACAATCGCGAAGACG	

Name (size)	Description	Sequence	SEQ ID NO:
		TGGAACCAACCCAAATGTCCGACTATGATTGACTGG ATTAAGAAGATGTGGCACATATACACTATGGAATAC TATGCTGCGATCAAGAACGATGAGTTCATATCATTTG TGGGCACATGGATGAAACTCGAAACCATCATACTCTC TAAATTGAGTCAAGAACAGAAAATAAACACCGTAT ATTTCCCTGATCGGTGGGAAT	
	Nucleoplasm in NLS	AAAAGGCCGGCGGCCACGAAAAAGGCCGGCCAGGC AAAAAAGAAAAAG	125
	FLAG	GACTACAAAGACGATGACGACAAG	138
	3' UTR	CTACAAAGACGATGACGACAAGGACCATGGAGACGG TGAGAGACACAAAAAATTCCAACACACTATTGCAAT GAAAATAAATTTTCCTTTATTAGCCAGAAGTCAGATGC TCAAGGGGCTTCATGATGTCCCCATAATTTTTGGCAG AGGGAAAAAGATCTCAGTGGTATTTGTGAGCCAGGG CATTGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGT GCGGCCGCTTAAGCTGCAATAAACAAGTT	139
	Full Length	ATGCCAAAGAAGAAGCGGAAGGTCTGGCGGCGGCAG CATGGCCACCGTGAAGTTCAAGTACAAGGGCGAGGA GAAGGAGGTGGACATCAGCAAGATCAAGAAGGTGTG GAGGGTGGGCAAGATGATCAGCTTCACCTACGACGA GGGCGGCGGCAAGACCGGCAGGGGCGCCGTGAGCG AGAAGGACGCCCCCAAGGAGCTGCTGCAGATGCTGG AGAAGCAGAAGAAGTCTGGAGGATCTAGCGGAGGAT CCTCTGGCAGCGAGACACCAGGAACAAGCGAGTCAG CAACACCAGAGAGCAGTGGCGGCAGCAGCGGCGGC AGCAGCATTAAGAATCTCACTCAAAGTAGATCCACG ACATGGAACTGAACAATCTCCTCTTGAATGACTACT GGGTGCATAACGAAATGAAGGCTGAAATAAAGATGT TCTTTGAGACCAACGAAAACAAAGACACCACGTACC AGAATCTCTGGGACGCTTTCAAAGCAGTGTGTGAG GAAAATTTATTGCACTGAATGCTTACAAGCGGAAGC AGGAAAGATCCAAAATAGACACCCTGACTAGCCAAC TTAAAGAACTGGAAAAGCAAGAGCAAACCTCATAGCA	144

Name (size)	Description	Sequence	SEQ ID NO:
		AAGCTAGCCGTCGCCAAGAAATTACGAAAATCAGAG CTGAAGTGAAGGAAATTGAGACACAGAAAACCTGC AAAAGATAAATGAAAGCCGCAGCTGGTTCTTTGAAC GCATCAACAAAATCGATAGGCCACTTGCTCGCCTTAT CAAGAAGAAAAGGGAGAAGAATCAAATCGACACTAT AAAGAATGATAAAGGCGATATAACCACCGATCCCAC AGAAATTCAAACAACCATACGCGAATACTACAAACA CCTCTACGCCAATAAACTCGAAAATCTCGAGGAAAT GGATACATTCTCGACACGTACACCCTTCCCAGGCTG AACCAGGAAGAAGTTGAATCACTGAATCGGCCTATC ACGGGGAGTGAAATAGTAGCTATCATCAATCACTCC CTACCAAGAAGTCACCCGGACCTGATGGATTCACCG CCGAATTCTACCAGAGATACATGGAAGAACTGGTGC CCTTCTTGCTGAAACTTTTCCAAAGTATTGAGAAAGA GGGAATACTTCCAAACTCATTTTATGAGGCATCCATC ATTCTGATCCCGAAGCCCGGCAGGGACACGACCAAG AAAGAGAATTTTCGACCAATCTCATTGATGAACATTG ATGCAAAGATCCTCAATAAAATACTGGCAAATCGGA TTCAGCAGCACATAAAGAAGCTGATCCACCATGATC AAGTAGGCTTCATCCCCGGTATGCAAGGTTGGTTCAA TATACGAAAATCAATCAATGTTATCCAGCATATAAAC CGGGCCAAAGACAAGAACCACATGATTATTAGTATC GATGCTGAGAAAGCCTTTGACAAAATACAACAACCC TTCATGCTGAAAACATTGAATAAGCTGGGAATTGATG GCACCTACTTCAAAAATCATCAGAGCCATATATGACAA ACCAACAGCAAATATCATTCTGAATGGTCAGAAATT GGAAGCATTCCCCTTGAAAACCGGCACACGGCAGGG TTGCCCTCTGTCACCACTCCTCTTCAACATCGTGTGG AAGTTCTTGCCCGCGCAATCCGGCAGGAAAAGGAAA TCAAGGGCATTCAACTGGGCAAAGAGGAAGTTAAAT TGAGCCTGTTTGCAGACGACATGATCGTCTATTTGGA AAACCCCATAGTTAGTGCACAAAATCTGCTGAAGTTG ATCAGTAATTTCTCCAAAGTGAGTGGGTACAAAATCA	

Name (size)	Description	Sequence	SEQ ID NO:
		ATGTGCAAAAGAGCCAAGCTTTCTTGTACACCAACA ACAGGCAAACTGAGTCTCAAATCATGGGCGAACTCC CCTTCGTGATTGCATCCAAGCGGATCAAATACCTGGG GATTCAATTGACTCGTGATGTGAAGGACCTCTTCAAG GAGAACTACAAACCCCTGCTCAAGGAAATCAAAGAG GACACAAACAAATGGAAGAACATTCCATGCTCTTGG GTGGGAAGGATCAATATCGTCAAAATGGCCATCCTG CCCAAGGTAATTTACAGGTTCAATGCTATACCCATCA AGCTCCCCATGACATTCTTCACAGAACTTGAAAAGAC GACGCTGAAGTTCATTTGGAACCAGAAACGTGCCAG GATTGCTAAATCTATTCTCTCCCAAAGAACAAGCT GGCGGAATCACACTCCCAGACTTCAAACCTTTACTACA AGGCGACCGTGACGAAAACGGCTTGGTACTGGTACC AAAACAGGGATATAGATCAATGGAACCGAACGGAGC CCAGCGAAATTATGCCTCATATATACAACCTATCTGAT CTTTGACAAACCGGAGAAGAACAAGCAATGGGGAAA GGATAGTCTGTTTAATAAATGGTGCTGGGAAAACCTG GCTCGCAATCTGTAGGAAGCTGAAACTGGATCCATTC TTGACGCCTTATACAAAGATAAATTCCCAGTGGATTA AAGATCTCAACGTGAAACCCAAAACAATTAACCC TCGAGGAAAACCTGGGTATTACGATTCAGGACATTG GGGTGGGAAAGGACTTCATGTCCAAAACCCCAAAAG CGATGGCAACCAAAGACAAAATCGACAAATGGGATC TCATAAACTTAAGTCATTTTGCACAGCTAAAGAAAC GACAATTAGGGTGAACCGACAACCGACCACTTGGGA GAAAATCTTCGCAACATACAGTTCTGACAAAGGCCT GATTTCCAGGATCTACAATGAATTGAAACAAATTTAC AAGAAGAAGACGAACAACCTATAAAGAAATGGGCC AAGGACATGAACAGACACTTCTCTAAGGAAGACATT TATGCAGCCAAGAAACACATGAAGAAATGCAGCTCT TCACTGGCAATCAGGGAAATGCAAATCAAAACAACA ATGAGATATCATCTCACACCCGTCAGAATGGCCATCA TTAAGAAGAGCGGAAACAACCGGTGCTGGCGTGGTT	

Name (size)	Description	Sequence	SEQ ID NO:
		GCGGAGAAATCGGTACTCTCCTTCACTGTTGGTGGGA CTGTAAACTCGTTCAACCACTGTGGAAGTCTGTGTGG CGGTTCTCAGAGATCTGGAACCTCGAAATCCCATTG ACCCAGCCATCCCTCTCCTGGGTATATACCCGAATGA GTATAAATCCTGCTGCTATAAAGACACCTGCACAAG GATGTTTATTGCAGCTCTCTTCACAATCGCGAAGACG TGGAACCAACCCAAATGTCCGACTATGATTGACTGG ATTAAGAAGATGTGGCACATATACTATGGAATAC TATGCTGCGATCAAGAACGATGAGTTCATATCATTG TGGGCACATGGATGAACTCGAAACCATCATACTCTC TAAATTGAGTCAAGAACAGAAAATAAACACCGTAT ATTTTCCCTGATCGGTGGGAATGGAGGCCAAAAGGCC GGCGGCCACGAAAAGGCCGGCCAGGCCAAAAAAGA AAAAGGACTACAAAGACGATGACGACAAGTAACTAC AAAGACGATGACGACAAGGACCATGGAGACGGTGA GAGACACAAAAAATTCCAACACACTATTGCAATGAA AATAAATTTCCTTTATTAGCCAGAAGTCAGATGCTCA AGGGGCTTCATGATGTCCCCATAATTTTGGCAGAGG GAAAAAGATCTCAGTGGTATTTGTGAGCCAGGGCAT TGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGTGCG GCCGCTTAAGCTGCAATAAACAAGTT	
CREATE-83 (4.4kb)	SV40 NLS	ATGCCAAAGAAGAAGCGGAAGGTCTGGCGGCGGCAGC	128
	HMGD	ATGAGCGACAAGCCCAAGAGACCCCTGAGCGCCTAC ATGCTGTGGCTGAACAGCGCTAGAGAGAGCATCAAG AGAGAGAACCCCGGCATCAAGGTGACCGAGGTGGCG AAGCGCGGGGGCGAGCTGTGGAGAGCCATGAAGGAC AAGAGCGAGTGGGAGGCCAAGGCCGCCAAGGCCAA GGACGACTACGACAGAGCCGTGAAGGAGTTCGAGGC CAATGGCGGCAGCAGCGCCGCCAACGGCGGCGGCGC CAAGAAGAGAGCCAAGCCCGCCAAGAAAGTGGCCA AGAAAAGTAAGAAGGAGGAGAGCGACGAGGACGAC GACGACGAGAGCGAG	145

Name (size)	Description	Sequence	SEQ ID NO:
	33XTEN	TCTGGAGGATCTAGCGGAGGATCCTCTGGCAGCGAG ACACCAGGAACAAGCGAGTCAGCAACACCAGAGAG CAGTGGCGGCAGCAGCGGCGGCAGCAGC	142
	ORF2 (aa. 239-1275)	ATTAAGAATCTCACTCAAAGTAGATCCACGACATGG AAACTGAACAATCTCCTCTTGAATGACTACTGGGTGC ATAACGAAATGAAGGCTGAAATAAAGATGTTCTTTG AGACCAACGAAAACAAAGACACCACGTACCAGAATC TCTGGGACGCTTTCAAAGCAGTGTGTGCGAGGAAAATT TATTGCACTGAATGCTTACAAGCGGAAGCAGGAAAG ATCCAAAATAGACACCCTGACTAGCCAACCTTAAAGA ACTGGAAAAGCAAGAGCAAATCATAGCAAAGCTAG CCGTCGCCAAGAAATTACGAAAATCAGAGCTGAACT GAAGGAAATTGAGACACAGAAAACCCTGCAAAAGAT AAATGAAAGCCGCAGCTGGTTCTTTGAACGCATCAA CAAAATCGATAGGCCACTTGCTCGCCTTATCAAGAAG AAAAGGGAGAAGAATCAAATCGACACTATAAAGAAT GATAAAGGCGATATAACCACCGATCCACAGAAATT CAAACAACCATACGCGAATACTACAAACACCTCTAC GCCAATAAACTCGAAAATCTCGAGGAAATGGATACA TTCCTCGACACGTACACCCTTCCCAGGCTGAACCAGG AAGAAGTTGAATCACTGAATCGGCCTATCACGGGGA GTGAAATAGTAGCTATCATCAATTCCTCCCTACCAA GAAGTCACCCGGACCTGATGGATTCACCGCCGAATT CTACCAGAGATACATGGAAGAACTGGTGCCCTTCTTG CTGAAACTTTTCCAAAGTATTGAGAAAGAGGGAATA CTTCCAAACTCATTTTATGAGGCATCCATCATTCTGA TCCCGAAGCCCGGCAGGGACACGACCAAGAAAGAG AATTTTCGACCAATCTCATTGATGAACATTGATGCAA AGATCCTCAATAAAATACTGGCAAATCGGATTCAGC AGCACATAAAGAAGCTGATCCACCATGATCAAGTAG GCTTCATCCCCGGTATGCAAGGTTGGTTCAATATACG AAAATCAATCAATGTTATCCAGCATATAAACCGGGC CAAAGACAAGAACCACATGATTATTAGTATCGATGC	143

Name (size)	Description	Sequence	SEQ ID NO:
		TGAGAAAGCCTTTGACAAAATACAACAACCCTTCAT GCTGAAAACATTGAATAAGCTGGGAATTGATGGCAC CTACTTCAAAATCATCAGAGCCATATATGACAAACCA ACAGCAAATATCATTCTGAATGGTCAGAAATTGGAA GCATTCCCCTTGAAAACCGGCACACGGCAGGGTTGC CCTCTGTCACCACTCCTCTTCAACATCGTGTGGAAG TTCTTGCCCGCGCAATCCGGCAGGAAAAGGAAATCA AGGGCATTCAACTGGGCAAAGAGGAAGTTAAATTGA GCCTGTTTGCAGACGACATGATCGTCTATTTGGAAAA CCCCATAGTTAGTGCACAAAATCTGCTGAAGTTGATC AGTAATTTCTCCAAAGTGAGTGGGTACAAAATCAAT GTGCAAAAGAGCCAAGCTTTCTTGTACACCAACAAC AGGCAAATGAGTCTCAAATCATGGGCGAACTCCCC TTCGTGATTGCATCCAAGCGGATCAAATACCTGGGGA TTCAATTGACTCGTGATGTGAAGGACCTCTTCAAGGA GAACTACAAACCCCTGCTCAAGGAAATCAAAGAGGA CACAAACAAATGGAAGAACATTCCATGCTCTTGGGT GGGAAGGATCAATATCGTCAAAATGGCCATCCTGCC CAAGGTAATTTACAGGTTCAATGCTATACCCATCAAG CTCCCCATGACATTCTTCACAGAACTTGAAAAGACGA CGCTGAAGTTCATTTGGAACCAGAAACGTGCCAGGA TTGCTAAATCTATTCTCTCCCAAAAGAACAAGCTGG CGGAATCACACTCCCAGACTTCAAACTTTACTACAAG GCGACCGTGACGAAAACGGCTTGGTACTGGTACCAA AACAGGGATATAGATCAATGGAACCGAACGGAGCCC AGCGAAATTATGCCTCATATATACAACATCTGATCT TTGACAAACCGGAGAAGAACAAGCAATGGGGAAAG GATAGTCTGTTTAATAAATGGTGCTGGGAAAACCTGGC TCGCAATCTGTAGGAAGCTGAAACTGGATCCATTCTT GACGCCTTATACAAAGATAAATTCCCGATGGATTAA AGATCTCAACGTGAAACCCAAAACAATTAAAACCTT CGAGGAAAACCTGGGTATTACGATTCAGGACATTGG GGTGGGAAAGGACTTCATGTCCAAAACCCCAAAAGC	

Name (size)	Description	Sequence	SEQ ID NO:
		GATGGCAACCAAAGACAAAATCGACAAATGGGATCT CATAAACTTAAGTCATTTTGCACAGCTAAAGAAAC GACAATTAGGGTGAACCGACAACCGACCACTTGGGA GAAAATCTTCGCAACATACAGTTCTGACAAAGGCCT GATTTCCAGGATCTACAATGAATTGAAACAAATTTAC AAGAAGAAGACGAACAACCCCTATAAAGAAATGGGCC AAGGACATGAACAGACACTTCTCTAAGGAAGACATT TATGCAGCCAAGAAACACATGAAGAAATGCAGCTCT TCACTGGCAATCAGGGAAATGCAAATCAAAACAACA ATGAGATATCATCTCACACCCGTCAGAATGGCCATCA TTAAGAAGAGCGGAAACAACCGGTGCTGGCGTGGTT GCGGAGAAATCGGTACTCTCCTTCACTGTTGGTGGGA CTGTAACTCGTTCAACCACTGTGGAAGTCTGTGTGG CGGTTCCCTCAGAGATCTGGAACTCGAAATCCCATTG ACCCAGCCATCCCTCTCCTGGGTATATACCCGAATGA GTATAAATCCTGCTGCTATAAAGACACCTGCACAAG GATGTTTATTGCAGCTCTCTTCACAATCGCGAAGACG TGGAACCAACCCAAATGTCCGACTATGATTGACTGG ATTAAGAAGATGTGGCACATATACTATGGAATAC TATGCTGCGATCAAGAACGATGAGTTCATATCATTTG TGGGCACATGGATGAACTCGAAACCATCATACTCTC TAAATTGAGTCAAGAACAGAAAATAAACACCGTAT ATTTCCCTGATCGGTGGGAAT	
	Nucleoplasm in NLS	AAAAGGCCGCGGCCACGAAAAAGGCCGGCCAGGC AAAAAAGAAAAAG	125
	FLAG	GACTACAAAGACGATGACGACAAG	138
	3' UTR	CTACAAAGACGATGACGACAAGGACCATGGAGACGG TGAGAGACACAAAAAATTCCAACACACTATTGCAAT GAAAATAAATTTCTTTATTAGCCAGAAGTCAGATGC TCAAGGGGCTTCATGATGTCCCATAATTTTGGCAG AGGGAAAAAGATCTCAGTGGTATTTGTGAGCCAGGG CATTGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGT GCGGCCGCTTAAGCTGCAATAAACAAGTT	139

Name (size)	Description	Sequence	SEQ ID NO:
	Full Length	ATGCCAAAGAAGAAGCGGAAGGTCGGCGGCGGCAG CATGAGCGACAAGCCCAAGAGACCCCTGAGCGCCTA CATGCTGTGGCTGAACAGCGCTAGAGAGAGCATCAA GAGAGAGAACCCCGGCATCAAGGTGACCGAGGTGGC GAAGCGCGGGGGCGAGCTGTGGAGAGCCATGAAGG ACAAGAGCGAGTGGGAGGCCAAGGCCGCCAAGGCC AAGGACGACTACGACAGAGCCGTGAAGGAGTTCGAG GCCAATGGCGGCAGCAGCGCCGCCAACGGCGGCGGC GCCAAGAAGAGAGCCAAGCCCGCCAAGAAAGTGGC CAAGAAAAGTAAGAAGGAGGAGAGCGACGAGGACG ACGACGACGAGAGCGAGTCTGGAGGATCTAGCGGAG GATCCTCTGGCAGCGAGACACCAGGAACAAGCGAGT CAGCAACACCAGAGAGCAGTGGCGGCAGCAGCGGC GGCAGCAGCATTAAGAATCTCACTCAAAGTAGATCC ACGACATGGAACTGAACAATCTCCTCTTGAATGACT ACTGGGTGCATAACGAAATGAAGGCTGAAATAAAGA TGTTCTTTGAGACCAACGAAAACAAAGACACCACGT ACCAGAATCTCTGGGACGCTTTCAAAGCAGTGTGTCG AGGAAAATTTATTGCACTGAATGCTTACAAGCGGAA GCAGGAAAGATCCAAAATAGACACCCTGACTAGCCA ACTTAAAGAACTGGAAAAGCAAGAGCAAATCATAG CAAAGCTAGCCGTCGCCAAGAAATTACGAAAATCAG AGCTGAACTGAAGGAAATTGAGACACAGAAAACCCT GCAAAAGATAAATGAAAGCCGCAGCTGGTTCTTTGA ACGCATCAACAAAATCGATAGGCCACTTGCTCGCCTT ATCAAGAAGAAAAGGGAGAAGAATCAAATCGACACT ATAAAGAATGATAAAGGCGATATAACCACCGATCCC ACAGAAATTCAAACAACCATACGCGAATACTACAAA CACCTCTACGCCAATAAACTCGAAAATCTCGAGGAA ATGGATACATTCTCGACACGTACACCCTTCCAGGC TGAACCAGGAAGAAGTTGAATCACTGAATCGGCCTA TCACGGGGAGTGAAATAGTAGCTATCATCAATTCACT CCCTACCAAGAAGTCACCCGGACCTGATGGATTAC	164

Name (size)	Description	Sequence	SEQ ID NO:
		CGCCGAATTCTACCAGAGATACATGGAAGAACTGGT GCCCTTCTTGCTGAAACTTTTCCAAAGTATTGAGAAA GAGGGAATACTTCCAAACTCATTATGAGGCATCCA TCATTCTGATCCCGAAGCCCGGCAGGGACACGACCA AGAAAGAGAATTTTCGACCAATCTCATTGATGAACAT TGATGCAAAGATCCTCAATAAAATACTGGCAAATCG GATTCAGCAGCACATAAAGAAGCTGATCCACCATGA TCAAGTAGGCTTCATCCCCGGTATGCAAGGTTGGTTC AATATACGAAAATCAATCAATGTTATCCAGCATATAA ACCGGGCCAAAGACAAGAACCACATGATTATTAGTA TCGATGCTGAGAAAGCCTTTGACAAAATACAACAAC CCTTCATGCTGAAAACATTGAATAAGCTGGGAATTGA TGGCACCTACTTCAAAATCATCAGAGCCATATATGAC AAACCAACAGCAAATATCATTCTGAATGGTCAGAAA TTGGAAGCATTCCCCTTGAAAACCGGCACACGGCAG GGTTGCCCTCTGTCACCACTCCTCTTCAACATCGTGTT GGAAGTTCTTGCCCGCGCAATCCGGCAGGAAAAGGA AATCAAGGGCATTCAACTGGGCAAAGAGGAAGTTAA ATTGAGCCTGTTTGCAGACGACATGATCGTCTATTTG GAAAACCCCATAGTTAGTGCACAAAATCTGCTGAAG TTGATCAGTAATTTCTCCAAAGTGAGTGGGTACAAAA TCAATGTGCAAAAAGAGCCAAGCTTTCTTGACACCAA CAACAGGCAAACCTGAGTCTCAAATCATGGGCGAACT CCCCTTCGTGATTGCATCCAAGCGGATCAAATACCTG GGGATTCAATTGACTCGTGATGTGAAGGACCTCTTCA AGGAGAACTACAAACCCCTGCTCAAGGAAATCAAAG AGGACACAAACAAATGGAAGAACATTCCATGCTCTT GGGTGGGAAGGATCAATATCGTCAAAATGGCCATCC TGCCCAAGGTAATTTACAGGTTCAATGCTATACCCAT CAAGCTCCCCATGACATTCTTCACAGAACTTGAAAAG ACGACGCTGAAGTTCATTTGGAACCAGAAACGTGCC AGGATTGCTAAATCTATTCTCTCCCAAAAGAACAAAG CTGGCGGAATCACACTCCAGACTTCAAACCTTTACTA	

Name (size)	Description	Sequence	SEQ ID NO:
		CAAGGCGACCGTGACGAAAACGGCTTGGTACTGGTA CCAAAACAGGGATATAGATCAATGGAACCGAACGGA GCCCAGCGAAATTATGCCTCATATATACTATCTG ATCTTTGACAAACCGGAGAAGAACAAGCAATGGGGA AAGGATAGTCTGTTTAATAAATGGTGCTGGGAAAAC TGGCTCGCAATCTGTAGGAAGCTGAAACTGGATCCAT TCTTGACGCCTTATACAAAGATAAATTCCCGATGGAT TAAAGATCTCAACGTGAAACCCAAAACAATTA AAC CCTCGAGGAAAACCTGGGTATTACGATTCAGGACATT GGGGTGGGAAAGGACTTCATGTCCAAAACCCAAAA GCGATGGCAACCAAAGACAAAATCGACAAATGGGAT CTCATAAACTTAAGTCATTTTGCACAGCTAAAGAAA CGACAATTAGGGTGAACCGACAACCGACCACTTGGG AGAAAATCTTCGCAACATACAGTTCTGACAAAGGCC TGATTTCCAGGATCTACAATGAATTGAAACAAATTTA CAAGAAGAAGACGAACAACCCTATAAAGAAATGGGC CAAGGACATGAACAGACACTTCTCTAAGGAAGACAT TTATGCAGCCAAGAAACACATGAAGAAATGCAGCTC TTC ACTGGCAATCAGGGAAATGCAAATCAAAACAAC AATGAGATATCATCTCACACCCGTCAGAATGGCCATC ATTAAGAAGAGCGGAAACAACCGGTGCTGGCGTGGT TGCGGAGAAATCGGTACTCTCCTTCACTGTTGGTGGG ACTGTAACTCGTTCAACCACTGTGGAAGTCTGTGTG GCGGTTCTCAGAGATCTGGAATCGAAATCCCATT GACCCAGCCATCCCTCTCCTGGGTATATACCCGAATG AGTATAAATCCTGCTGCTATAAAGACACCTGCACAA GGATGTTTATTGCAGCTCTCTTCACAATCGCGAAGAC GTGGAACCAACCCAAATGTCCGACTATGATTGACTG GATTAAGAAGATGTGGCACATATACACTATGGAATA CTATGCTGCGATCAAGAACGATGAGTTCATATCATTT GTGGGCACATGGATGAACTCGAAACCATCATACTC TCTAAATTGAGTCAAGAACAGAAAATAAACACCGT ATATTTTCCCTGATCGGTGGGAATGGAGGCAAAAGG	

Name (size)	Description	Sequence	SEQ ID NO:
		CCGGCGGCCACGAAAAAGGCCGGCCAGGCAAAAAA GAAAAAGGACTACAAAGACGATGACGACAAGTAACT ACAAAGACGATGACGACAAGGACCATGGAGACGGTG AGAGACACAAAAAATTCCAACACACTATTGCAATGA AAATAAATTTCTTTATTAGCCAGAAGTCAGATGCTC AAGGGGCTTCATGATGTCCCCATAATTTTGGCAGAG GGAAAAAGATCTCAGTGGTATTTGTGAGCCAGGGCA TTGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGTGC GGCCGCTTAAGCTGCAATAAACAAGTT	
CREATE-84 (4.2kb)	SV40 NLS	ATGCCAAAGAAGAAGCGGAAGGTCTGGCGGCGGCAG C	128
	PARP1 ZnF1 and ZnF2 domain	ATGGCGGAGTCTTCGGATAAGCTCTATCGAGTCGAGT ACGCCAAGAGCGGGCGCGCCTCTTGCAAGAAATGCA GCGAGAGCATCCCCAAGGACTCGCTCCGGATGGCCA TCATGGTGCAGTCGCCCATGTTTGATGGAAAAGTCCC ACACTGGTACCACTTCTCCTGCTTCTGGAAGGTGGGC CACTCCATCCGGCACCCCTGACGTTGAGGTGGATGGGT TCTCTGAGCTTCGGTGGGATGACCAGCAGAAAGTCA AGAAGACAGCGGAAGCTGGAGGAGTGACAGGCAAA GGCCAGGATGGAATTGGTAGCAAGGCAGAGAAGACT CTGGGTGACTTTGCAGCAGAGTATGCCAAGTCCAAC AGAAGTACGTGCAAGGGGTGTATGGAGAAGATAGAA AAGGGCCAGGTGCGCCTGTCCAAGAAGATGGTGGAC CCGGAGAAGCCACAGCTAGGCATGATTGACCGCTGG TACCATCCAGGCTGCTTTGTCAAGAACAGGGAGGAG CTGGGTTTCCGGCCCCGAGTACAGTGCGAGTCAGCTCA AGGGCTTCAGCCTCCTTGCTACAGAGGATAAAGAAG CCCTGAAGAAGCAGCTCCCAGGAGTCAAGAGTGAAG GAAAGAGAAAAGGC	145
	33XTEN	TCTGGAGGATCTAGCGGAGGATCCTCTGGCAGCGAG ACACCAGGAACAAGCGAGTCAGCAACACCAGAGAG CAGTGGCGGCAGCAGCGGCGGCAGCAGC	142

Name (size)	Description	Sequence	SEQ ID NO:
	ORF2 (aa. 239-1275)	ATTAAGAATCTCACTCAAAGTAGATCCACGACATGG AAACCTGAACAATCTCCTCTTGAATGACTACTGGGTGC ATAACGAAATGAAGGCTGAAATAAAGATGTTCTTTG AGACCAACGAAAACAAAGACACCACGTACCAGAATC TCTGGGACGCTTTCAAAGCAGTGTGTCGAGGAAAATT TATTGCACTGAATGCTTACAAGCGGAAGCAGGAAAG ATCCAAAATAGACACCCTGACTAGCCAACTTAAAGA ACTGGAAAAGCAAGAGCAAACCTCATAGCAAAGCTAG CCGTCGCCAAGAAATTACGAAAATCAGAGCTGAACT GAAGGAAATTGAGACACAGAAAACCCTGCAAAAAGAT AAATGAAAGCCGCAGCTGGTTCTTTGAACGCATCAA CAAAATCGATAGGCCACTTGCTCGCCTTATCAAGAAG AAAAGGGAGAAGAATCAAATCGACACTATAAAGAAT GATAAAGGCGATATAACCACCGATCCACAGAAATT CAAACAACCATACGCGAATACTACAAACACCTCTAC GCCAATAAACTCGAAAATCTCGAGGAAATGGATACA TTCCTCGACACGTACACCCTTCCCAGGCTGAACCAGG AAGAAGTTGAATCACTGAATCGGCCTATCACGGGGA GTGAAATAGTAGCTATCATCAATTCCTCCCTACCAA GAAGTCACCCGGACCTGATGGATTACCCGCCGAATT CTACCAGAGATACATGGAAGAACTGGTGCCCTTCTTG CTGAAACTTTTCCAAAGTATTGAGAAAGAGGGAATA CTTCCAAACCTCATTTTATGAGGCATCCATCATTCTGA TCCCGAAGCCCGGCAGGGACACGACCAAGAAAGAG AATTTTCGACCAATCTCATTGATGAACATTGATGCAA AGATCCTCAATAAAATACTGGCAAATCGGATTACAGC AGCACATAAAGAAGCTGATCCACCATGATCAAGTAG GCTTCATCCCCGGTATGCAAGGTTGGTTCAATATACG AAAATCAATCAATGTTATCCAGCATATAAACCGGGC CAAAGACAAGAACCACATGATTATTAGTATCGATGC TGAGAAAGCCTTTGACAAAATACAACAACCCTTCAT GCTGAAAACATTGAATAAGCTGGGAATTGATGGCAC CTAATTCAAAATCATCAGAGCCATATATGACAAACCA	143

Name (size)	Description	Sequence	SEQ ID NO:
		ACAGCAAATATCATTCTGAATGGTCAGAAATTGGAA GCATTCCCCTTGAAAACCGGCACACGGCAGGGTTGC CCTCTGTCACCACTCCTCTTCAACATCGTGTGGAAG TTCTTGCCCGCGCAATCCGGCAGGAAAAGGAAATCA AGGGCATTCAACTGGGCAAAGAGGAAGTTAAATTGA GCCTGTTTGCAGACGACATGATCGTCTATTTGGAAAA CCCCATAGTTAGTGCACAAAATCTGCTGAAGTTGATC AGTAATTTCTCCAAAGTGAGTGGGTACAAAATCAAT GTGCAAAAGAGCCAAGCTTTCTTGTACACCAACAAC AGGCAAACCTGAGTCTCAAATCATGGGCGAACTCCCC TTCGTGATTGCATCCAAGCGGATCAAATACCTGGGGA TTCAATTGACTCGTGATGTGAAGGACCTCTTCAAGGA GAACTACAAACCCCTGCTCAAGGAAATCAAAGAGGA CACAAACAAATGGAAGAACATTCCATGCTCTTGGGT GGGAAGGATCAATATCGTCAAAATGGCCATCCTGCC CAAGGTAATTTACAGGTTCAATGCTATACCCATCAAG CTCCCCATGACATTCTTCACAGAACTTGAAAAGACGA CGCTGAAGTTCATTTGGAACCAGAAACGTGCCAGGA TTGCTAAATCTATTCTCTCCCAAAAGAACAAAGCTGG CGGAATCACACTCCCAGACTTCAAACCTTACTACAAG GCGACCGTGACGAAAACGGCTTGGTACTGGTACCAA AACAGGGATATAGATCAATGGAACCGAACGGAGCCC AGCGAAATTATGCCTCATATATACAACTATCTGATCT TTGACAAACCGGAGAAGAACAAGCAATGGGGAAAG GATAGTCTGTTTAATAAATGGTGCTGGGAAAACCTGGC TCGCAATCTGTAGGAAGCTGAAACTGGATCCATTCTT GACGCCTTATACAAAGATAAATTCCCGATGGATTAA AGATCTCAACGTGAAACCCAAAACAATTAAAACCCCT CGAGGAAAACCTGGGTATTACGATTCAGGACATTGG GGTGGGAAAGGACTTCATGTCCAAAACCCCAAAAGC GATGGCAACCAAAGACAAAATCGACAAATGGGATCT CATAAACTTAAGTCATTTTGCACAGCTAAAGAAAC GACAATTAGGGTGAACCGACAACCGACCACCTTGGGA	

Name (size)	Description	Sequence	SEQ ID NO:
		GAAAATCTTCGCAACATACAGTTCTGACAAAGGCCT GATTTCCAGGATCTACAATGAATTGAAACAAATTTAC AAGAAGAAGACGAACAACCCTATAAAGAAATGGGCC AAGGACATGAACAGACACTTCTCTAAGGAAGACATT TATGCAGCCAAGAAACACATGAAGAAATGCAGCTCT TCACTGGCAATCAGGGAAATGCAAATCAAAACAACA ATGAGATATCATCTCACACCCGTCAGAATGGCCATCA TTAAGAAGAGCGGAAACAACCGGTGCTGGCGTGGTT GCGGAGAAATCGGTACTCTCCTTCACTGTTGGTGGGA CTGTAACTCGTTCAACCACTGTGGAAGTCTGTGTGG CGGTTCCCTCAGAGATCTGGAACCTCGAAATCCCATTG ACCCAGCCATCCCTCTCCTGGGTATATACCCGAATGA GTATAAATCCTGCTGCTATAAAGACACCTGCACAAG GATGTTTATTGCAGCTCTCTTCACAATCGCGAAGACG TGGAACCAACCCAAATGTCCGACTATGATTGACTGG ATTAAGAAGATGTGGCACATATACACTATGGAATAC TATGCTGCGATCAAGAACGATGAGTTCATATCATTTG TGGGCACATGGATGAACTCGAAACCATCATACTCTC TAAATTGAGTCAAGAACAGAAAATAAACACCGTAT ATTTCCCTGATCGGTGGGAAT	
	Nucleoplasm in NLS	AAAAGGCCGGCGGCCACGAAAAAGGCCGGCCAGGC AAAAAAGAAAAAG	125
	FLAG	GACTACAAAGACGATGACGACAAG	138
	3' UTR	CTACAAAGACGATGACGACAAGGACCATGGAGACGG TGAGAGACACAAAAAATTCACACACTATTGCAAT GAAAATAAATTTCTTTATTAGCCAGAAGTCAGATGC TCAAGGGGCTTCATGATGTCCCCATAATTTTTGGCAG AGGGAAAAAGATCTCAGTGGTATTTGTGAGCCAGGG CATTGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGT GCGGCCGCTTAAGCTGCAATAAACAAGTT	139
	Full Length	ATGCCAAAGAAGAAGCGGAAGGTCGGCGGCGGCAG CATGGCGGAGTCTTCGGATAAGCTCTATCGAGTCGAG TACGCCAAGAGCGGGCGCGCCTCTTGCAAGAAATGC	146

Name (size)	Description	Sequence	SEQ ID NO:
		AGCGAGAGCATCCCCAAGGACTCGCTCCGGATGGCC ATCATGGTGCAGTCGCCCATGTTTGATGGAAAAGTCC CACACTGGTACCACTTCTCCTGCTTCTGGAAGGTGGG CCACTCCATCCGGCACCTGACGTTGAGGTGGATGG GTTCTCTGAGCTTCGGTGGGATGACCAGCAGAAAGTC AAGAAGACAGCGGAAGCTGGAGGAGTGACAGGCAA AGGCCAGGATGGAATTGGTAGCAAGGCAGAGAAGAC TCTGGGTGACTTTGCAGCAGAGTATGCCAAGTCCAAC AGAAGTACGTGCAAGGGGTGTATGGAGAAGATAGAA AAGGGCCAGGTGCGCCTGTCCAAGAAGATGGTGGAC CCGGAGAAGCCACAGCTAGGCATGATTGACCGCTGG TACCATCCAGGCTGCTTTGTCAAGAACAGGGAGGAG CTGGGTTTCCGGCCCGAGTACAGTGCGAGTCAGCTCA AGGGCTTCAGCCTCCTTGCTACAGAGGATAAAGAAG CCCTGAAGAAGCAGCTCCCAGGAGTCAAGAGTGAAG GAAAGAGAAAAGGCTCTGGAGGATCTAGCGGAGGAT CCTCTGGCAGCGAGACACCAGGAACAAGCGAGTCAG CAACACCAGAGAGCAGTGGCGGCAGCAGCGGCGGC AGCAGCATTAAAGAATCTCACTCAAAGTAGATCCACG ACATGGAACTGAACAATCTCCTCTTGAATGACTACT GGGTGCATAACGAAATGAAGGCTGAAATAAAGATGT TCTTTGAGACCAACGAAAACAAAGACACCACGTACC AGAATCTCTGGGACGCTTTCAAAGCAGTGTGTCGAG GAAAATTTATTGCACTGAATGCTTACAAGCGGAAGC AGGAAAGATCCAAAATAGACACCCTGACTAGCCAAC TTAAAGAACTGGAAAAGCAAGAGCAAATCATAGCA AAGCTAGCCGTCGCCAAGAAATTACGAAAATCAGAG CTGAAGTGAAGGAAATTGAGACACAGAAAACCTGC AAAAGATAAATGAAAGCCGCAGCTGGTTCTTTGAAC GCATCAACAAAATCGATAGGCCACTTGCTCGCCTTAT CAAGAAGAAAAGGGAGAAGAATCAAATCGACACTAT AAAGAATGATAAAGGCGATATAACCAACCGATCCCAC AGAAATTCAAACAACCATACGCGAATACTACAAACA	

Name (size)	Description	Sequence	SEQ ID NO:
		CCTCTACGCCAATAAACTCGAAAATCTCGAGGAAAT GGATACATTCTCGACACGTACACCCTTCCCAGGCTG AACCAGGAAGAAGTTGAATCACTGAATCGGCCTATC ACGGGGAGTGAAATAGTAGCTATCATCAATCACTCC CTACCAAGAAGTCACCCGGACCTGATGGATTCACCG CCGAATTCTACCAGAGATACATGGAAGAACTGGTGC CCTTCTTGCTGAAACTTTTCCAAAGTATTGAGAAAGA GGGAATACTTCCAAACTCATTTTATGAGGCATCCATC ATTCTGATCCCGAAGCCCGGCAGGGACACGACCAAG AAAGAGAATTTTCGACCAATCTCATTGATGAACATTG ATGCAAAGATCCTCAATAAAATACTGGCAAATCGGA TTCAGCAGCACATAAAGAAGCTGATCCACCATGATC AAGTAGGCTTCATCCCCGGTATGCAAGGTTGGTTCAA TATACGAAAATCAATCAATGTTATCCAGCATATAAAC CGGGCCAAAGACAAGAACCACATGATTATTAGTATC GATGCTGAGAAAGCCTTTGACAAAATACAACAACCC TTCATGCTGAAAACATTGAATAAGCTGGGAATTGATG GCACCTACTTCAAAATCATCAGAGCCATATATGACAA ACCAACAGCAAATATCATTCTGAATGGTCAGAAATT GGAAGCATTCCCCTTGAAAACCGGCACACGGCAGGG TTGCCCTCTGTCACCACTCCTCTTCAACATCGTGTTGG AAGTTCTTGCCCGCGCAATCCGGCAGGAAAAGGAAA TCAAGGGCATTCAACTGGGCAAAGAGGAAGTTAAAT TGAGCCTGTTTGCAGACGACATGATCGTCTATTTGGA AAACCCCATAGTTAGTGCACAAAATCTGCTGAAGTTG ATCAGTAATTTCTCCAAAGTGAGTGGGTACAAAATCA ATGTGCAAAGAGCCAAGCTTTCTTGTACACCAACA ACAGGCAAACCTGAGTCTCAAATCATGGGCGAACTCC CCTTCGTGATTGCATCCAAGCGGATCAAATACCTGGG GATTCAATTGACTCGTGATGTGAAGGACCTCTTCAAG GAGAACTACAAACCCCTGCTCAAGGAAATCAAAGAG GACACAAACAAATGGAAGAACATTCCATGCTCTTGG GTGGGAAGGATCAATATCGTCAAAATGGCCATCCTG	

Name (size)	Description	Sequence	SEQ ID NO:
		CCCAAGGTAATTTACAGGTTCAATGCTATACCCATCA AGCTCCCCATGACATTCTTCACAGAACTTGAAAAGAC GACGCTGAAGTTCATTTGGAACCAGAAACGTGCCAG GATTGCTAAATCTATTCTCTCCCAAAGAACAAGCT GGC GGAATCACACTCCCAGACTTCAAAC TTTACTACA AGGCGACCGTGACGAAAACGGCTTGGTACTGGTACC AAAACAGGGATATAGATCAATGGAACCGAACGGAGC CCAGCGAAATTATGCCTCATATATACTATCTGAT CTTTGACAAACCGGAGAAGAACAAGCAATGGGGAAA GGATAGTCTGTTTAATAAATGGTGCTGGGAAAAC TG GCTCGCAATCTGTAGGAAGCTGAAACTGGATCCATTC TTGACGCCTTATACAAAGATAAATTCCCGATGGATTA AAGATCTCAACGTGAAACCCAAAACAATTA AACCC TCGAGGAAAACCTGGGTATTACGATTCAGGACATTG GGGTGGGAAAGGACTTCATGTCCAAAACCCCAAAAG CGATGGCAACCAAAGACAAAATCGACAAATGGGATC TCATAAAACTTAAGTCATTTTGCACAGCTAAAGAAAC GACAATTAGGGTGAACCGACAACCGACCACTTGGGA GAAAATCTTCGCAACATACAGTTCTGACAAAGGCCT GATTTCCAGGATCTACAATGAATTGAAACAAATTTAC AAGAAGAAGACGAACAACCTATAAAGAAATGGGCC AAGGACATGAACAGACACTTCTCTAAGGAAGACATT TATGCAGCCAAGAAACACATGAAGAAATGCAGCTCT TCACTGGCAATCAGGGAAATGCAAATCAAAACAACA ATGAGATATCATCTCACACCCGTCAGAATGGCCATCA TTAAGAAGAGCGGAAACAACCGGTGCTGGCGTGGTT GCGGAGAAATCGGTACTCTCCTTCACTGTTGGTGGGA CTGTAAACTCGTTCAACCACTGTGGAAGTCTGTGTGG CGGTTCTCAGAGATCTGGAAC TCGAAATCCCATTG ACCCAGCCATCCCTCTCCTGGGTATATACCCGAATGA GTATAAATCCTGCTGCTATAAAGACACCTGCACAAG GATGTTTATTGCAGCTCTCTTCACAATCGCGAAGACG TGGAACCAACCCAAATGTCCGACTATGATTGACTGG	

Name (size)	Description	Sequence	SEQ ID NO:
		ATTAAGAAGATGTGGCACATATACACTATGGAATAC TATGCTGCGATCAAGAACGATGAGTTCATATCATTG TGGGCACATGGATGAACTCGAAACCATCATACTCTC TAAATTGAGTCAAGAACAGAAAATAAACACCGTAT ATTTTCCCTGATCGGTGGGAATGGAGGCAAAAGGCC GGC GCCACGAAAAAGGCCGCCAGGCAAAAAAGA AAAAGGACTACAAAGACGATGACGACAAGTA ACTAC AAAGACGATGACGACAAGGACCATGGAGACGGTGA GAGACACAAAAAATTCCAACACACTATTGCAATGAA AATAAATTTCTTTATTAGCCAGAAGTCAGATGCTCA AGGGGCTTCATGATGTCCCATAATTTTTGGCAGAGG GAAAAAGATCTCAGTGGTATTTGTGAGCCAGGGCAT TGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGTGCG GCCGCTTAAGCTGCAATAAACAAGTT	
CREATE-25 (6.4kb)	ORF1	ATGGGCAAGAAGCAAAATCGCAAGACGGGGAATTCC AAGACACAATCCGCTAGCCCACCACCTAAAGAGCGT TCTAGCTCCCCTGCTACTGAGCAGTCCTGGATGGAAA ACGACTTCGATGAACTCCGGAAGAGGGATTTAGGC GATCCA ACTATT CAGAACTCCGCGAAGATATCCAGA CAAAGGGGAAGGAAGTCGAGAATTTGAGAAGAACC TCGAGGAGTGCATCACCCGTATCACAAACACTGAGA AATGTCTCAAAGAACTCATGGA ACTTAAGACAAAAG CCAGGGAGCTTCGAGAGGAGTGTGCGAGTCTGAGAT CCAGGTGTGACCAGCTCGAGGAGCGCGTGAGCGCGA TGGAAGACGAGATGAACGAGATGAAAAGAGAGGGC AAATTCAGGGAGAAGCGCATTAAAGAGGAACGAACAG AGTCTGCAGGAGATTTGGGATTACGTCAAGAGGCCT AACCTGCGGTTGATCGGCGTCCCCGAGAGCGACGTA GAAAACGGGACTAACTGGAGAATACACTTCAAGAC ATCATTCAAGAAAATTTTCCAAACCTGGCTCGGCAAG CTAATGTGCAAATCCAAGAGATCCAACGCACACCCC AGCGGTATAGCTCTCGGCGTGCCACCCCTAGGCATAT TATCGTGCGCTTTACTAAGGTGGAGATGAAAGAGAA	10

Name (size)	Description	Sequence	SEQ ID NO:
		GATGCTGCGAGCCGCTCGGGAAAAGGGAAGGGTGAC TTTGAAGGGCAAACCTATTTCGGCTGACGGTTGACCTT AGCGCCGAGACACTCCAGGCACGCCGGGAATGGGGC CCCATCTTTAATATCCTGAAGGAGAAGAAGTTCCAGC CACGAATCTCTTACCCTGCAAAGTTGAGTTTTATCTC CGAGGGTGAGATTAAGTATTTTCATCGATAAACAGAT GCTGCGGAGACTTCGTGACAACTCGCCCAGCTCTCAAG GAACTGCTCAAAGAGGCTCTTAATATGGAGCGCAAT AATAGATATCAACCCTTGCAGAACCACGCAAAGATG TGA	
	InterORF Sequence	AGACCATCAAGACTAGGAAGAACTGCATCAACTAA TGAGCAAAATCACCAGCTAACATCATA	127
	SV40 NLS	ATGCCAAAGAAGAAGCGGAAGGTCGGCGGCGGCAG C	128
	ORF2 D205A	ACCGGCTCTAACTCACATATCACCATCCTTACACTTA ACATTAACGGCCTCAACTCAGCTATCAAGCGCCATCG GCTGGCCAGCTGGATCAAATCACAGGATCCAAGCGT TTGTTGCATCCAAGAGACCCACCTGACGTGCAGGGA CACCCATCGGCTGAAAATAAAAGGCTGGCGGAAGAT CTACCAAGCCAATGGAAAACAAAAGAAGGCTGGGGT GGCGATACTTGTAAGCGATAAAACAGACTTTAAACC AACTAAGATCAAACGGGACAAAGAGGGCCATTACAT CATGGTAAAGGGTAGTATTCAACAAGAGGAGCTGAC TATCCTGAATATTTATGCACCTAATACTGGAGCCCC AGATTCATAAAGCAAGTGTTGAGTGACCTTCAACGC GACCTCGACTCCCACACTCTGATCATGGGAGACTTTA ACACCCCGCTGTCCACTCTCGACAGATCTACTAGACA GAAAGTCAACAAGGATACACAGGAAGTGAACAGTGC TCTCCACCAAGCGGACCTTATCGACATCTACAGAACA CTCCACCCCAAAAGCACAGAATATACCTTCTTTTCAG CCCCTCACCACACCTATTCCAAAATTGCCACATTGT GGGGAGTAAAGCCCTTCTCTCCAAATGTAAACGGAC CGAAATTATCACTAACTATCTCTCCGACCACAGTGCA	129

Name (size)	Description	Sequence	SEQ ID NO:
		ATAAAACTTGAATTGCGAATTAAGAATCTCACTCAA GTAGATCCACGACATGGAACTGAACAATCTCCTCTT GAATGACTACTGGGTGCATAACGAAATGAAGGCTGA AATAAAGATGTTCTTTGAGACCAACGAAAACAAAGA CACCACGTACCAGAATCTCTGGGACGCTTTCAAAGC AGTGTGTCGAGGAAAATTTATTGCACTGAATGCTTAC AAGCGGAAGCAGGAAAGATCCAAAATAGACACCCTG ACTAGCCAACCTTAAAGAACTGGAAAAGCAAGAGCAA ACTCATAGCAAAGCTAGCCGTCGCCAAGAAATTACG AAAATCAGAGCTGAAGTGAAGGAAATTGAGACACAG AAAACCCTGCAAAAGATAAATGAAAGCCGCAGCTGG TTCTTTGAACGCATCAACAAAATCGATAGGCCACTTG CTCGCCTTATCAAGAAGAAAAGGGAGAAGAATCAAA TCGACACTATAAAGAATGATAAAGGCGATATAACCA CCGATCCCACAGAAATTCAAACAACCATACGCGAAT ACTACAAACACCTCTACGCCAATAAACTCGAAAATC TCGAGGAAATGGATACATTCCTCGACACGTACACCCT TCCCAGGCTGAACCAGGAAGAAGTTGAATCACTGAA TCGGCCTATCACGGGGAGTGAAATAGTAGCTATCATC AATTCACTCCCTACCAAGAAGTCACCCGGACCTGATG GATTCACCGCCGAATTCTACCAGAGATACATGGAAG AACTGGTGCCCTTCTTGCTGAACTTTTCCAAAGTAT TGAGAAAGAGGGAATACTTCCAACTCATTTTATGA GGCATCCATCATTCTGATCCCGAAGCCCGGCAGGGA CACGACCAAGAAAGAGAATTTTCGACCAATCTCATT GATGAACATTGATGCAAAGATCCTCAATAAAATACT GGCAAATCGGATTCAGCAGCACATAAAGAAGCTGAT CCACCATGATCAAGTAGGCTTCATCCCCGGTATGCAA GGTTGGTTCAATATACGAAAATCAATCAATGTTATCC AGCATATAAACCGGGCCAAAGACAAGAACCACATGA TTATTAGTATCGATGCTGAGAAAGCCTTTGACAAAAT ACAACAACCCTTCATGCTGAAAACATTGAATAAGCT GGGAATTGATGGCACCTACTTCAAATCATCAGAGC	

Name (size)	Description	Sequence	SEQ ID NO:
		CATATATGACAAACCAACAGCAAATATCATTCTGAAT GGTCAGAAATTGGAAGCATTCCCCTTGAAAACCGGC ACACGGCAGGGTTGCCCTCTGTCACCACTCCTCTTCA ACATCGTGTTGGAAGTTCTTGCCCGCGCAATCCGGCA GGAAAAGGAAATCAAGGGCATTCAACTGGGCAAAGA GGAAGTTAAATTGAGCCTGTTTGCAGACGACATGATC GTCTATTTGGAAAACCCCATAGTTAGTGCACAAAATC TGCTGAAGTTGATCAGTAATTTCTCCAAAGTGAGTGG GTACAAAATCAATGTGCAAAAGAGCCAAGCTTTCTT GTACACCAACAACAGGCCAACTGAGTCTCAAATCAT GGGCGAACTCCCCTTCGTGATTGCATCCAAGCGGATC AAATACCTGGGGATTCAATTGACTCGTGATGTGAAG GACCTCTTCAAGGAGAACTACAAACCCCTGCTCAAG GAAATCAAAGAGGACACAAACAAATGGAAGAACATT CCATGCTCTTGGGTGGGAAGGATCAATATCGTCAAA ATGGCCATCCTGCCCAAGGTAATTTACAGGTTCAATG CTATACCCATCAAGCTCCCCATGACATTCTTCACAGA ACTTGAAAAGACGACGCTGAAGTTCATTTGGAACCA GAAACGTGCCAGGATTGCTAAATCTATTCTCTCCCAA AAGAACAAAGCTGGCGGAATCACACTCCCAGACTTC AAACCTTACTACAAGGCGACCGTGACGAAAACGGCT TGGTACTGGTACCAAAACAGGGATATAGATCAATGG AACCGAACGGAGCCCAGCGAAATTATGCCTCATATA TACAACATCTGATCTTTGACAAACCGGAGAAGAAC AAGCAATGGGGAAAGGATAGTCTGTTTAATAAATGG TGCTGGGAAAACCTGGCTCGCAATCTGTAGGAAGCTG AAACCTGGATCCATTCTTGACGCCTTATACAAAGATAA ATTCCCGATGGATTAAAGATCTCAACGTGAAACCCA AAACAATTAAAACCCTCGAGGAAAACCTGGGTATTA CGATTCAGGACATTGGGGTGGGAAAGGACTTCATGT CCAAAACCCCAAAAGCGATGGCAACCAAAGACAAA ATCGACAAATGGGATCTCATAAACTTAAGTCATTTT GCACAGCTAAAGAAACGACAATTAGGGTGAACCGAC	

Name (size)	Description	Sequence	SEQ ID NO:
		AACCGACCACTTGGGAGAAAATCTTCGCAACATACA GTTCTGACAAAGGCCTGATTTCCAGGATCTACAATGA ATTGAAACAAATTTACAAGAAGAAGACGAACAACCC TATAAAGAAATGGGCCAAGGACATGAACAGACACTT CTCTAAGGAAGACATTTATGCAGCCAAGAAACACAT GAAGAAATGCAGCTCTTCACTGGCAATCAGGGAAAT GCAAATCAAAACAACAATGAGATATCATCTCACACC CGTCAGAATGGCCATCATTAAAGAAGAGCGGAAACAA CCGGTGCTGGCGTGGTTGCGGAGAAATCGGTACTCTC CTTCACTGTTGGTGGGACTGTAACTCGTTCAACCAC TGTGGAAGTCTGTGTGGCGGTTCCCTCAGAGATCTGGA ACTCGAAATCCCATTGACCCAGCCATCCCTCTCCTG GGTATATACCCGAATGAGTATAAATCCTGCTGCTATA AAGACACCTGCACAAGGATGTTTATTGCAGCTCTCTT CACAATCGCGAAGACGTGGAACCAACCCAAATGTCC GACTATGATTGACTGGATTAAGAAGATGTGGCACAT ATACACTATGGAATACTATGCTGCGATCAAGAACGA TGAGTTCATATCATTGTGGGCACATGGATGAAACTC GAAACCATCATACTCTCTAAATTGAGTCAAGAACAG AAAATAAACACCGTATATTTTCCCTGATCGGTGGGA AT	
	Nucleoplasm in NLS	AAAAGGCCGCGGCCACGAAAAAGGCCGGCCAGGC AAAAAAGAAAAAG	125
	3' UTR	CTACAAAGACGATGACGACAAGGACCATGGAGACGG TGAGAGACACAAAAAATCCAACACACTATTGCAAT GAAAATAAATTTCTTTATTAGCCAGAAGTCAGATGC TCAAGGGGCTTCATGATGTCCCCATAATTTTTGGCAG AGGGAAAAAGATCTCAGTGGTATTTGTGAGCCAGGG CATTGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGT GCGGCCGCT	139
	PBS2 (30nt)	CAGCCGCGTCAGAGCAGCTCAGGTTCTGGG	51
	SV40 PolyA	TAAGATACATTGATGAGTTTGGACAAACCACAATA GAATGCAGTGAAAAAATGCTTTATTTGTGAAATTTG	131

Name (size)	Description	Sequence	SEQ ID NO:
		TGATGCTATTGCTTTATTTGTAACCATTATAAGCTGC AATAAACAAGTT	
	GFP	TTACTTGTACAGCTCGTCCATGCCGAGAGTGATCCCG GCGGCGGTACGAACTCCAGCAGGACCATGTGATCG CGCTTCTCGTTGGGGTCTTTGCTCAGGGCGGACTGGG TGCTCAGGTAGTGGTTGTCGGGCAGCAGCACGGGGC CGTCGCCGATGGGGGTGTTCTGCTGGTAGTGGTCGGC GAGCTGCACGCTGCCGTCTCGATGTTGTGGCGGATC TTGAAGTTCACCTTGATGCCGTTCTTCTGCTTGTGGC CATGATATAGACGTTGTGGCTGTTGTAGTTGTACTCC AGCTTGTGCCCCAGGATGTTGCCGTCTCTCTGAAGT CGATGCCCTTCAGCTCGATGCGGTTACCAGGGTGTC GCCCTCGAACTTCACCTCGGCGCGGGTCTTGTAGTTG CCGTCGTCCTTGAAGAAGATGGTGCGCTCCTGGACGT AGCCTTCGGGCATGGCGGACTTGAAGAAGTCGTGCT GCTTCATGTGGTCGGGGTAGCGGCTGAAGCACTGCA CGCCGTAGGTACAGGGTGGTCACGAGGGTGGGCCAGG GCACGGGCAGCTTGCCGGTGGTGCAGATGAACTTCA GGGTCAGCTTGCCGTAGGTGGCATCGCCCTCGCCCTC GCCGGACACGCTGAACTTGTGGCCGTTTACGTCGCCG TCCAGCTCGACCAGGATGGGCACCACCCCGGTGAAC AGCTCCTCGCCCTTGCTCACCAT	132
	EF-1a Promoter	CTGTGTTCTGGCGGCAAACCCGTTGCGAAAAAGAAC GTTACGGCGACTACTGCACTTATATACGGTTCTCCC CCACCCTCGGGAAAAAGGCGGAGCCAGTACACGACA TCACTTTCCCAGTTTACCCCGCGCCACCTTCTCTAGG CACCGGATCAATTGCCGACCCCTCCCCCAACTTCTC GGGGACTGTGGGCGATGTGCGCTCTGCCC	133
	PBS1 (30nt)	TCCTCTCTGGCTCCATCGTAAGCAAACCTT	50
	Full Length	ATGGGCAAGAAGCAAAATCGCAAGACGGGGAATTCC AAGACACAATCCGCTAGCCCACCACCTAAAGAGCGT TCTAGCTCCCCTGCTACTGAGCAGTCCTGGATGGAAA ACGACTTCGATGAACTCCGGGAAGAGGGATTTAGGC	147

Name (size)	Description	Sequence	SEQ ID NO:
		GATCCAAC TATTCAGAACTCCGCGAAGATATCCAGA CAAAGGGGAAGGAAGTCGAGAATTTTCGAGAAGAACC TCGAGGAGTGCATCACCCGTATCACAAACACTGAGA AATGTCTCAAAGAACTCATGGAAC TTAAGACAAAAG CCAGGGAGCTTCGAGAGGAGTGTCTGGAGTCTGAGAT CCAGGTGTGACCAGCTCGAGGAGCGCGTGAGCGCGA TGGAAGACGAGATGAACGAGATGAAAAGAGAGGGC AAATTCAGGGAGAAGCGCATTAAAGAGGAACGAACAG AGTCTGCAGGAGATTTGGGATTACGTCAAGAGGCCT AACCTGCGGTTGATCGGCGTCCCCGAGAGCGACGTA GAAAACGGGACTAAACTGGAGAATACACTTCAAGAC ATCATTCAAGAAAATTTTCCAAACCTGGCTCGGCAAG CTAATGTGCAAATCCAAGAGATCCAACGCACACCCC AGCGGTATAGCTCTCGGCGTGCCACCCCTAGGCATAT TATCGTGCGCTTTACTAAGGTGGAGATGAAAGAGAA GATGCTGCGAGCCGCTCGGGAAAAGGGAAGGGTGAC TTTGAAGGGCAAACCTATTTCGGCTGACGGTTGACCTT AGCGCCGAGACACTCCAGGCACGCCGGGAATGGGGC CCCATCTTTAATATCCTGAAGGAGAAGAACTTCCAGC CACGAATCTCTTACCCTGCAAAGTTGAGTTTTATCTC CGAGGGTGAGATTAAGTATTTTCATCGATAAACAGAT GCTGCGAGACTTCGTGACAACTCGCCCAGCTCTCAAG GAACTGCTCAAAGAGGCTCTTAATATGGAGCGCAAT AATAGATATCAACCCTTGCAAGAACACGCAAAGATG TGAGACAGCCGTCAGACCATCAAGACTAGGAAGAAA CTGCATCAACTAATGAGCAAAATCACCAGCTAACAT CATAGTATACATGCCAAAGAAGAAGCGGAAGGTCGG CGGCGGCAGCACCGGCTCTAACTCACATATCACCATC CTTACACTTAACATTAACGGCCTCAACTCAGCTATCA AGCGCCATCGGCTGGCCAGCTGGATCAAATCACAGG ATCCAAGCGTTTGTTGCATCCAAGAGACCCACCTGAC GTGCAGGGACACCCATCGGCTGAAAATAAAAGGCTG GCGGAAGATCTACCAAGCCAATGGAAAACAAAAGAA	

Name (size)	Description	Sequence	SEQ ID NO:
		GGCTGGGGTGGCGATACTTGTAAGCGATAAAACAGA CTTTAAACCAACTAAGATCAAACGGGACAAAGAGGG CCATTACATCATGGTAAAGGGTAGTATTCAACAAGA GGAGCTGACTATCCTGAATATTTATGCACCTAATACT GGAGCCCCCAGATTCATAAAGCAAGTGTTGAGTGAC CTTCAACGCGACCTCGACTCCCACACTCTGATCATGG GAGACTTTAACACCCCCTGTCCACTCTCGACAGATC TACTAGACAGAAAGTCAACAAGGATACACAGGAACT GAACAGTGCTCTCCACCAAGCGGACCTTATCGACATC TACAGAACTCCACCCCAAAGCACAGAATATACC TTCTTTTCAGCCCCTCACCACACCTATTCCAAAATTG CCCACATTGTGGGGAGTAAAGCCCTTCTCTCCAAATG TAAACGGACCGAAATTATCACTAACTATCTCTCCGAC CACAGTGCAATAAACTTGAATTGCGAATTAAGAAT CTCACTCAAAGTAGATCCACGACATGGAACTGAAC AATCTCCTCTTGAATGACTACTGGGTGCATAACGAAA TGAAGGCTGAAATAAAGATGTTCTTTGAGACCAACG AAAACAAAGACACCACGTACCAGAATCTCTGGGACG CTTTCAAAGCAGTGTGTGCGAGGAAAATTTATTGCACT GAATGCTTACAAGCGGAAGCAGGAAAGATCCAAAAT AGACACCCTGACTAGCCAACCTTAAAGAAGTGGAAAA GCAAGAGCAAACCTCATAGCAAAGCTAGCCGTCGCCA AGAAATTACGAAAATCAGAGCTGAACTGAAGGAAAT TGAGACACAGAAAACCCTGCAAAAGATAAATGAAAG CCGCAGCTGGTTCTTTGAACGCATCAACAAAATCGAT AGGCCACTTGCTCGCCTTATCAAGAAGAAAAGGGAG AAGAATCAAATCGACACTATAAAGAATGATAAAGGC GATATAACCACCGATCCCACAGAAATTCAAACAACC ATACGCGAATACTACAAACACCTCTACGCCAATAAA CTCGAAAATCTCGAGGAAATGGATACATTCTCGAC ACGTACACCCTTCCCAGGCTGAACCAGGAAGAAGTT GAATCACTGAATCGGCCTATCACGGGGAGTGAAATA GTAGCTATCATCAATCACTCCCTACCAAGAAGTCAC	

Name (size)	Description	Sequence	SEQ ID NO:
		CCGGACCTGATGGATTACCGCCGAATTCTACCAGA GATACATGGAAGAACTGGTGCCCTTCTTGCTGAAACT TTTCCAAAGTATTGAGAAAGAGGGAATACTTCCAAA CTCATTTTATGAGGCATCCATCATTCTGATCCCGAAG CCCGGCAGGGACACGACCAAGAAAGAGAATTTTCGA CCAATCTCATTGATGAACATTGATGCAAAGATCCTCA ATAAAATACTGGCAAATCGGATTCAGCAGCACATAA AGAAGCTGATCCACCATGATCAAGTAGGCTTCATCCC CGGTATGCAAGGTTGGTTCAATATACGAAAATCAATC AATGTTATCCAGCATATAAACCGGGCCAAAGACAAG AACCACATGATTATTAGTATCGATGCTGAGAAAGCCT TTGACAAAATACAACAACCCTTCATGCTGAAAACATT GAATAAGCTGGGAATTGATGGCACCTACTTCAAAAT CATCAGAGCCATATATGACAAACCAACAGCAAATAT CATTCTGAATGGTCAGAAATTGGAAGCATTCCCCTTG AAAACCGGCACACGGCAGGGTTGCCCTCTGTCACCA CTCCTCTTCAACATCGTGTGGAAGTTCTTGCCCGCG CAATCCGGCAGGAAAAGGAAATCAAGGGCATTCAAC TGGGCAAAGAGGAAGTTAAATTGAGCCTGTTTGCAG ACGACATGATCGTCTATTTGGAAAACCCCATAGTTAG TGCACAAAATCTGCTGAAGTTGATCAGTAATTTCTCC AAAGTGAGTGGGTACAAAATCAATGTGCAAAAGAGC CAAGCTTTCTTGACACCAACAACAGGCAAACCTGAGT CTCAAATCATGGGCGAACTCCCCTTCGTGATTGCATC CAAGCGGATCAAATACCTGGGGATTCAATTGACTCGT GATGTGAAGGACCTCTTCAAGGAGAACTACAAACCC CTGCTCAAGGAAATCAAAGAGGACACAAACAAATGG AAGAACATTCCATGCTCTTGGGTGGGAAGGATCAAT ATCGTCAAAATGGCCATCCTGCCCAAGGTAATTTACA GGTTCAATGCTATACCCATCAAGCTCCCCATGACATT CTTCACAGAACTTGAAAAGACGACGCTGAAGTTCATT TGGAACCAGAAACGTGCCAGGATTGCTAAATCTATTC TCTCCCAAAGAACAAGCTGGCGGAATCACACTCC	

Name (size)	Description	Sequence	SEQ ID NO:
		CAGACTTCAAACCTTTACTACAAGGCGACCGTGACGA AAACGGCTTGGTACTGGTACCAAAACAGGGATATAG ATCAATGGAACCGAACGGAGCCCAGCGAAATTATGC CTCATATATACAACCTATCTGATCTTTGACAAACCGGA GAAGAACAAGCAATGGGGAAAGGATAGTCTGTTTAA TAAATGGTGCTGGGAAAACCTGGCTCGCAATCTGTAG GAAGCTGAAACTGGATCCATTCTTGACGCCTTATACA AAGATAAATTCCCGATGGATTAAAGATCTCAACGTG AAACCCAAAACAATTAACCCCTCGAGGAAAACCTG GGTATTACGATTACAGGACATTGGGGTGGGAAAGGAC TTCATGTCCAAAACCCCAAAGCGATGGCAACCAAA GACAAAATCGACAAATGGGATCTCATAAACTTAAG TCATTTTGCACAGCTAAAGAAACGACAATTAGGGTG AACCGACAACCGACCACTTGGGAGAAAATCTTCGCA ACATACAGTTCTGACAAAGGCCTGATTTCCAGGATCT ACAATGAATTGAAACAAATTTACAAGAAGAAGACGA ACAACCCTATAAAGAAATGGGCCAAGGACATGAACA GACACTTCTCTAAGGAAGACATTTATGCAGCCAAGA AACACATGAAGAAATGCAGCTCTTCACTGGCAATCA GGGAAATGCAAATCAAAACAACAATGAGATATCATC TCACACCCGTCAGAATGGCCATCATTAAAGAAGAGCG GAAACAACCGGTGCTGGCGTGGTTGCGGAGAAATCG GTACTCTCCTTCACTGTTGGTGGGACTGTAAACTCGT TCAACCACTGTGGAAGTCTGTGTGGCGGTTCTCAGA GATCTGGAACCTCGAAATCCCATTGACCCAGCCATCC CTCTCCTGGGTATATACCCGAATGAGTATAAATCCTG CTGCTATAAAGACACCTGCACAAGGATGTTTATTGCA GCTCTCTTCACAATCGCGAAGACGTGGAACCAACCC AAATGTCCGACTATGATTGACTGGATTAAGAAGATGT GGCACATATACACTATGGAATACTATGCTGCGATCAA GAACGATGAGTTCATATCATTTGTGGGCACATGGATG AAACTCGAAACCATCATACTCTCTAAATTGAGTCAAG AACAGAAAACCTAAACACCGTATATTTCCCTGATCGG	

Name (size)	Description	Sequence	SEQ ID NO:
		TGGGAATGGAGGCAAAAGGCCGGCGGCCACGAAAA AGGCCGGCCAGGCAAAAAAGAAAAAGTAGCTACAA AGACGATGACGACAAGGACCATGGAGACGGTGAGA GACACAAAAAATTCCAACACACTATTGCAATGAAAA TAAATTTCTTTATTAGCCAGAAGTCAGATGCTCAAG GGGCTTCATGATGTCCCCATAATTTTTGGCAGAGGGA AAAAGATCTCAGTGGTATTTGTGAGCCAGGGCATTG GCCTTCTGATAGGCAGCCTGCACCTGAGGAGTGCGG CCGCTCAGCCGCGTCAGAGCAGCTCAGGTTCTGGGT AAGATACATTGATGAGTTTGGACAAACCACAACCTAG AATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGT GATGCTATTGCTTTATTTGTAACCATTATAAGCTGCA ATAAACAAGTTTTACTTGTACAGCTCGTCCATGCCGA GAGTGATCCCGGCGGCGGTCACGAACTCCAGCAGGA CCATGTGATCGCGCTTCTCGTTGGGGTCTTTGCTCAG GGCGGACTGGGTGCTCAGGTAGTGGTTGTCGGGCAG CAGCACGGGGCCGTCGCCGATGGGGGTGTTCTGCTG GTAGTGGTCGGCGAGCTGCACGCTGCCGTCCTCGATG TTGTGGCGGATCTTGAAGTTCACCTTGATGCCGTTCT TCTGCTTGTCGGCCATGATATAGACGTTGTGGCTGTT GTAGTTGTACTCCAGCTTGTGCCCCAGGATGTTGCCG TCCTCCTTGAAGTCGATGCCCTTCAGCTCGATGCGGT TCACCAGGGTGTCGCCCTCGAACTTCACCTCGGCGCG GGTCTTGTAGTTGCCGTCGTCCTTGAAGAAGATGGTG CGCTCCTGGACGTAGCCTTCGGGCATGGCGGACTTGA AGAAGTCGTGCTGCTTCATGTGGTCGGGGTAGCGGCT GAAGCACTGCACGCCGTAGGTCAGGGTGGTCACGAG GGTGGGCCAGGGCACGGGCAGCTTGCCGGTGGTGCA GATGAACTTCAGGGTCAGCTTGCCGTAGGTGGCATCG CCCTCGCCCTCGCCGGACACGCTGAACTTGTGGCCGT TTACGTCGCCGTCCAGCTCGACCAGGATGGGCACCA CCCCGGTGAACAGCTCCTCGCCCTTGCTCACCATGGT GGCGGGATCTGACGGTTCATAAACCTGTGTTCTGG	

Name (size)	Description	Sequence	SEQ ID NO:
		CGGCAAACCCGTTGCGAAAAAGAACGTTACGGCGA CTACTGCACTTATATACGGTTCTCCCCACCCTCGGG AAAAAGGCGGAGCCAGTACACGACATCACTTTCCCA GTTTACCCCGCGCCACCTTCTCTAGGCACCGGATCAA TTGCCGACCCCTCCCCCAACTTCTCGGGGACTGTGG GCGATGTGCGCTCTGCCCTCCTCTCTGGCTCCATCGT AAGCAAACCTT	
CREATE-37 (6.4kb)	ORF1	ATGGGCAAGAAGCAAAATCGCAAGACGGGGAATTCC AAGACACAATCCGCTAGCCCACCACCTAAAGAGCGT TCTAGCTCCCCTGCTACTGAGCAGTCCTGGATGGAAA ACGACTTCGATGAACTCCGGGAAGAGGGATTTAGGC GATCCAATATTCAGAACTCCGCGAAGATATCCAGA CAAAGGGGAAGGAAGTCGAGAATTTGAGAAGAACC TCGAGGAGTGCATCACCCGTATCACAAACACTGAGA AATGTCTCAAAGAACTCATGGAACCTAAGACAAAAG CCAGGGAGCTTCGAGAGGAGTGTCTGGAGTCTGAGAT CCAGGTGTGACCAGCTCGAGGAGCGCGTGAGCGCGA TGGAAGACGAGATGAACGAGATGAAAAGAGAGGGC AAATTCAGGGAGAAGCGCATTAAAGAGGAACGAACAG AGTCTGCAGGAGATTTGGGATTACGTCAAGAGGCCT AACCTGCGGTTGATCGGCGTCCCCGAGAGCGACGTA GAAAACGGGACTAACTGGAGAATACACTTCAAGAC ATCATTCAAGAAAATTTTCCAAACCTGGCTCGGCAAG CTAATGTGCAAATCCAAGAGATCCAACGCACACCCC AGCGGTATAGCTCTCGGCGTGCCACCCCTAGGCATAT TATCGTGCGCTTTACTAAGGTGGAGATGAAAGAGAA GATGCTGCGAGCCGCTCGGGAAAAGGGAAGGGTGAC TTTGAAGGGCAAACCTATTCTGGCTGACGGTTGACCTT AGCGCCGAGACACTCCAGGCACGCCGGGAATGGGGC CCCATCTTTAATATCCTGAAGGAGAAGAACTTCCAGC CACGAATCTCTTACCCTGCAAAGTTGAGTTTATCTC CGAGGGTGAGATTAAGTATTTTCATCGATAAACAGAT GCTGCGAGACTTCGTGACAACTCGCCCAGCTCTCAAG	10

Name (size)	Description	Sequence	SEQ ID NO:
		GAAGTCTCAAAGAGGCTCTTAATATGGAGCGCAAT AATAGATATCAACCCTTGCAGAACCACGCAAAGATG TGA	
	InterORF Sequence	AGACCATCAAGACTAGGAAGAACTGCATCAACTAA TGAGCAAAATCACCAGCTAACATCATA	148
	SV40 NLS	ATGCCAAAGAAGAAGCGGAAGGTCGGCGGCGGCAG C	128
	ORF2 D205A	ACCGGCTCTAACTCACATATCACCATCCTTACACTTA ACATTAACGGCCTCAACTCAGCTATCAAGCGCCATCG GCTGGCCAGCTGGATCAAATCACAGGATCCAAGCGT TTGTTGCATCCAAGAGACCCACCTGACGTGCAGGGA CACCCATCGGCTGAAAATAAAAGGCTGGCGGAAGAT CTACCAAGCCAATGGAAAACAAAAGAAGGCTGGGGT GGCGATACTTGTAAGCGATAAAACAGACTTTAAACC AACTAAGATCAAACGGGACAAAGAGGGCCATTACAT CATGGTAAAGGGTAGTATTCAACAAGAGGAGCTGAC TATCCTGAATATTTATGCACCTAATACTGGAGCCCC AGATTCATAAAGCAAGTGTTGAGTGACCTTCAACGC GACCTCGACTCCCACACTCTGATCATGGGAGACTTTA ACACCCCGCTGTCCACTCTCGACAGATCTACTAGACA GAAAGTCAACAAGGATACACAGGAAGTGAACAGTGC TCTCCACCAAGCGGACCTTATCGACATCTACAGAACA CTCCACCCCAAAAGCACAGAATATACCTTCTTTTCAG CCCCTCACCACACCTATTCCAAAATTGCCACATTGT GGGGAGTAAAGCCCTTCTCTCCAAATGTAAACGGAC CGAAATTATCACTAACTATCTCTCCGACCACAGTGCA ATAAACTTGAATTGCGAATTAAGAATCTCACTCAA GTAGATCCACGACATGGAAACTGAACAATCTCCTCTT GAATGACTACTGGGTGCATAACGAAATGAAGGCTGA AATAAAGATGTTCTTTGAGACCAACGAAAACAAAGA CACCACGTACCAGAATCTCTGGGACGCTTCAAAGC AGTGTGTCGAGGAAAATTTATTGCACTGAATGCTTAC AAGCGGAAGCAGGAAAGATCCAAAATAGACACCCTG	129

Name (size)	Description	Sequence	SEQ ID NO:
		ACTAGCCAACTTAAAGAACTGGAAAAGCAAGAGCAA ACTCATAGCAAAGCTAGCCGTCGCCAAGAAATTACG AAAATCAGAGCTGAACTGAAGGAAATTGAGACACAG AAAACCCTGCAAAAGATAAATGAAAGCCGCAGCTGG TTCTTTGAACGCATCAACAAAATCGATAGGCCACTTG CTCGCCTTATCAAGAAGAAAAGGGAGAAGAATCAAA TCGACACTATAAAGAATGATAAAGGCGATATAACCA CCGATCCCACAGAAATTCAAACAACCATACGCGAAT ACTACAAACACCTCTACGCCAATAAACTCGAAAATC TCGAGGAAATGGATACATTCCCTCGACACGTACACCCT TCCCAGGCTGAACCAGGAAGAAGTTGAATCACTGAA TCGGCCTATCACGGGGAGTGAAATAGTAGCTATCATC AATTCACCTCCCTACCAAGAAGTCACCCGGACCTGATG GATTCACCGCCGAATTCTACCAGAGATACATGGAAG AACTGGTGCCCTTCTTGCTGAAACTTTTCCAAAGTAT TGAGAAAGAGGGAATACTTCCAAACTCATTTTATGA GGCATCCATCATTCTGATCCCGAAGCCCGGCAGGGA CACGACCAAGAAAGAGAATTTTCGACCAATCTCATT GATGAACATTGATGCAAAGATCCTCAATAAAATACT GGCAAATCGGATTCAGCAGCACATAAAGAAGCTGAT CCACCATGATCAAGTAGGCTTCATCCCCGGTATGCAA GGTTGGTTCAATATACGAAAATCAATCAATGTTATCC AGCATATAAACCGGGCCAAAGACAAGAACCACATGA TTATTAGTATCGATGCTGAGAAAGCCTTTGACAAAAT ACAACAACCCTTCATGCTGAAAACATTGAATAAGCT GGGAATTGATGGCACCTACTTCAAAATCATCAGAGC CATATATGACAAACCAACAGCAAATATCATTCTGAAT GGTCAGAAATTGGAAGCATTCCCCTTGAAAACCGGC ACACGGCAGGGTTGCCCTCTGTCACCACTCCTCTTCA ACATCGTGTTGGAAGTTCTTGCCCGCGCAATCCGGCA GGAAAAGGAAATCAAGGGCATTCAACTGGGCAAAGA GGAAGTTAAATTGAGCCTGTTTGCAGACGACATGATC GTCTATTTGAAAACCCCATAGTTAGTGCACAAAATC	

Name (size)	Description	Sequence	SEQ ID NO:
		TGCTGAAGTTGATCAGTAATTTCTCCAAAGTGAGTGG GTACAAAATCAATGTGCAAAAGAGCCAAGCTTTCTT GTACACCAACAACAGGCAAAGTCTCAAATCAT GGGCGAACTCCCCTTCGTGATTGCATCCAAGCGGATC AAATACCTGGGGATTCAATTGACTCGTGATGTGAAG GACCTCTTCAAGGAGAACTACAAACCCCTGCTCAAG GAAATCAAAGAGGACACAAACAAATGGAAGAACATT CCATGCTCTTGGGTGGGAAGGATCAATATCGTCAAA ATGGCCATCCTGCCCAAGGTAATTTACAGGTTCAATG CTATACCCATCAAGCTCCCCATGACATTCTTCACAGA ACTTGAAAAGACGACGCTGAAGTTCATTTGGAACCA GAAACGTGCCAGGATTGCTAAATCTATTCTCTCCCAA AAGAACAAAGCTGGCGGAATCACACTCCAGACTTC AACTTTACTACAAGGCGACCGTGACGAAAACGGCT TGGTACTGGTACCAAAACAGGGATATAGATCAATGG AACCGAACGGAGCCCAGCGAAATTATGCCTCATATA TACAACTATCTGATCTTTGACAAACCGGAGAAGAAC AAGCAATGGGGAAAGGATAGTCTGTTTAATAAATGG TGCTGGGAAAACCTGGCTCGCAATCTGTAGGAAGCTG AACTGGATCCATTCTTGACGCCTTATACAAAGATAA ATTCCCGATGGATTAAAGATCTCAACGTGAAAACCCA AAACAATTAAAACCCCTCGAGGAAAACCTGGGTATTA CGATTACAGGACATTGGGGTGGGAAAGGACTTCATGT CCAAAACCCCAAAAGCGATGGCAACCAAAGACAAA ATCGACAAATGGGATCTCATAAACTTAAGTCATTTT GCACAGCTAAAGAAACGACAATTAGGGTGAACCGAC AACCGACCACTTGGGAGAAAATCTTCGCAACATACA GTTCTGACAAAGGCCTGATTTCAGGATCTACAATGA ATTGAAACAAATTTACAAGAAGAAGACGAACAACCC TATAAAGAAATGGGCCAAGGACATGAACAGACACTT CTCTAAGGAAGACATTTATGCAGCCAAGAAACACAT GAAGAAATGCAGCTCTTCACTGGCAATCAGGGAAAT GCAAATCAAAACAACAATGAGATATCATCTCACACC	

Name (size)	Description	Sequence	SEQ ID NO:
		CGTCAGAATGGCCATCATTAAAGAAGAGCGGAAACAA CCGGTGCTGGCGTGGTTGCGGAGAAATCGGTACTCTC CTTCACTGTTGGTGGGACTGTAAACTCGTTCAACCAC TGTGGAAGTCTGTGTGGCGGTTCCCTCAGAGATCTGGA ACTCGAAATCCCATTGACCCAGCCATCCCTCTCCTG GGTATATACCCGAATGAGTATAAATCCTGCTGCTATA AAGACACCTGCACAAGGATGTTTATTGCAGCTCTCTT CACAATCGCGAAGACGTGGAACCAACCCAAATGTCC GACTATGATTGACTGGATTAAGAAGATGTGGCACAT ATACACTATGGAATACTATGCTGCGATCAAGAACGA TGAGTTCATATCATTTGTGGGCACATGGATGAAACTC GAAACCATCATACTCTCTAAATTGAGTCAAGAACAG AAAATAAACACCGTATATTTTCCCTGATCGGTGGGA AT	
	Nucleoplasm in NLS	AAAAGGCCGCGGCCACGAAAAAGGCCGCGCCAGGC AAAAAAGAAAAAG	125
	3' UTR	CTACAAAGACGATGACGACAAGGACCATGGAGACGG TGAGAGACACAAAAAATTCCAACACACTATTGCAAT GAAAATAAATTTCTTTATTAGCCAGAAGTCAGATGC TCAAGGGGCTTCATGATGTCCCCATAATTTTGGCAG AGGGAAAAAGATCTCAGTGGTATTTGTGAGCCAGGG CATTGGCCTTCTGATAGGCAGCCTGCACCTGAGGAGT GCGGCCGCT	139
	PBS2 (50nt)	GATCAGTGAAACGCACCAGACAGCCGCGTCAGAGCA GCTCAGGTTCTGGG	53
	SV40 PolyA	TAAGATACATTGATGAGTTTGGACAAACCACAATA GAATGCAGTGAAAAAATGCTTTATTTGTGAAATTTG TGATGCTATTGCTTTATTTGTAACCATTATAAGCTGC AATAACAAGTT	131
	GFP	TTACTTGTACAGCTCGTCCATGCCGAGAGTGATCCCG GCGGCGGTACGAACTCCAGCAGGACCATGTGATCG CGTTCTCGTTGGGGTCTTTGCTCAGGGCGGACTGGG TGCTCAGGTAGTGGTTGTCTGGGCAGCAGCACGGGGC	132

Name (size)	Description	Sequence	SEQ ID NO:
		CGTCGCCGATGGGGGTGTTCTGCTGGTAGTGGTCGGC GAGCTGCACGCTGCCGTCCTCGATGTTGTGGCGGATC TTGAAGTTCACCTTGATGCCGTTCTTCTGCTTGTCCGGC CATGATATAGACGTTGTGGCTGTTGTAGTTGTACTCC AGCTTGTGCCCCAGGATGTTGCCGTCCTCCTTGAAGT CGATGCCCTTCAGCTCGATGCGGTTACCAGGGTGTG GCCCTCGAACTTCACCTCGGCGCGGGTCTTGTAGTTG CCGTCGTCCTTGAAGAAGATGGTGCGCTCCTGGACGT AGCCTTCGGGCATGGCGGACTTGAAGAAGTCGTGCT GCTTCATGTGGTCGGGGTAGCGGCTGAAGCACTGCA CGCCGTAGGTACAGGGTGGTCACGAGGGTGGGCCAGG GCACGGGCAGCTTGCCGGTGGTGCAGATGAACTTCA GGGTACAGTTGCCGTAGGTGGCATCGCCCTCGCCCTC GCCGGACACGCTGAACTTGTGGCCGTTTACGTCGCCG TCCAGCTCGACCAGGATGGGCACCACCCCGGTGAAC AGCTCCTCGCCCTTGCTACCAT	
	EF-1a Promoter	CTGTGTTCTGGCGCAAACCCGTTGCGAAAAAGAAC GTTACGGCGACTACTGCACTTATATACGGTTCTCCC CCACCCTCGGGAAAAAGGCGGAGCCAGTACACGACA TCACTTTCCCAGTTTACCCCGCGCCACCTTCTCTAGG CACCGGATCAATTGCCGACCCCTCCCCCAACTTCTC GGGGACTGTGGGCGATGTGCGCTCTGCCC	133
	PBS1 (50nt)	TCCTCTCTGGCTCCATCGTAAGCAAACCTTAGAGGTT CTGGCAAGGAGAG	52
	Full Length	ATGGGCAAGAAGCAAAATCGCAAGACGGGGAATTCC AAGACACAATCCGCTAGCCCACCACCTAAAGAGCGT TCTAGCTCCCCTGCTACTGAGCAGTCCTGGATGGAAA ACGACTTCGATGAACTCCGGGAAGAGGGATTTAGGC GATCCAATAATTCAGAACTCCGCGAAGATATCCAGA CAAAGGGGAAGGAAGTCGAGAATTTGAGAAGAACC TCGAGGAGTGCATACCCGTATCACAACACTGAGA AATGTCTCAAAGAACTCATGGAACCTTAAGACAAAAG CCAGGGAGCTTCGAGAGGAGTGTGCGAGTCTGAGAT	149

Name (size)	Description	Sequence	SEQ ID NO:
		CCAGGTGTGACCAGCTCGAGGAGCGCGTGAGCGCGA TGGAAGACGAGATGAACGAGATGAAAAGAGAGGGC AAATTCAGGGAGAAGCGCATTAAAGAGGAACGAACAG AGTCTGCAGGAGATTTGGGATTACGTCAAGAGGCCT AACCTGCGGTTGATCGGCGTCCCCGAGAGCGACGTA GAAAACGGGACTAACTGGAGAATACACTTCAAGAC ATCATTCAAGAAAATTTTCCAAACCTGGCTCGGCAAG CTAATGTGCAAATCCAAGAGATCCAACGCACACCCC AGCGGTATAGCTCTCGGCGTGCCACCCCTAGGCATAT TATCGTGCGCTTTACTAAGGTGGAGATGAAAGAGAA GATGCTGCGAGCCGCTCGGGAAAAGGGAAGGGTGAC TTTGAAGGGCAAACCTATTCGGCTGACGGTTGACCTT AGCGCCGAGACACTCCAGGCACGCCGGGAATGGGGC CCCATCTTTAATATCCTGAAGGAGAAGAACTTCCAGC CACGAATCTCTTACCCTGCAAAGTTGAGTTTTATCTC CGAGGGTGAGATTAAGTATTTTCATCGATAAACAGAT GCTGCGAGACTTCGTGACAACTCGCCCAGCTCTCAAG GAACTGCTCAAAGAGGCTCTTAATATGGAGCGCAAT AATAGATATCAACCCTTGCAAGAACCACGCAAAGATG TGAGACAGCCGTCAGACCATCAAGACTAGGAAGAAA CTGCATCAACTAATGAGCAAAATCACCAGCTAACAT CATAGTATACATGCCAAAGAAGAAGCGGAAGGTCGG CGGCGGCAGCACCGGCTCTAACTCACATATCACCATC CTTACACTTAACATTAACGGCCTCAACTCAGCTATCA AGCGCCATCGGCTGGCCAGCTGGATCAAATCACAGG ATCCAAGCGTTTGTGTCATCCAAGAGACCCACCTGAC GTGCAGGGACACCCATCGGCTGAAAATAAAAGGCTG GCGGAAGATCTACCAAGCCAATGGAACAAAAGAA GGCTGGGGTGGCGATACTTGTAAGCGATAAAACAGA CTTTAAACCAACTAAGATCAAACGGGACAAAGAGGG CCATTACATCATGGTAAAGGGTAGTATTCAACAAGA GGAGCTGACTATCCTGAATATTTATGCACCTAATACT GGAGCCCCCAGATTCATAAAGCAAGTGTTGAGTGAC	

Name (size)	Description	Sequence	SEQ ID NO:
		CTTCAACGCGACCTCGACTCCCACACTCTGATCATGG GAGACTTTAACACCCCGCTGTCCACTCTCGACAGATC TACTAGACAGAAAGTCAACAAGGATACACAGGAACT GAACAGTGCTCTCCACCAAGCGGACCTTATCGACATC TACAGAACACTCCACCCCAAAAGCACAGAATATACC TTCTTTTCAGCCCCTCACCACACCTATTCCAAAATTG CCCACATTGTGGGGAGTAAAGCCCTTCTCTCCAAATG TAAACGGACCGAAATTATCACTAACTATCTCTCCGAC CACAGTGCAATAAACTTGAATTGCGAATTAAGAAT CTCACTCAAAGTAGATCCACGACATGGAACTGAAC AATCTCCTCTTGAATGACTACTGGGTGCATAACGAAA TGAAGGCTGAAATAAAGATGTTCTTTGAGACCAACG AAAACAAAGACACCACGTACCAGAATCTCTGGGACG CTTTCAAAGCAGTGTGTCGAGGAAAATTTATTGCACT GAATGCTTACAAGCGGAAGCAGGAAAGATCCAAAAT AGACACCCTGACTAGCCAACCTTAAAGAACTGGAAAA GCAAGAGCAAACCTCATAGCAAAGCTAGCCGTCGCCA AGAAATTACGAAAATCAGAGCTGAACTGAAGGAAAT TGAGACACAGAAAACCCTGCAAAAGATAAATGAAAG CCGCAGCTGGTTCTTTGAACGCATCAACAAAATCGAT AGGCCACTTGCTCGCCTTATCAAGAAGAAAAGGGAG AAGAATCAAATCGACACTATAAAGAATGATAAAGGC GATATAACCACCGATCCACAGAAATTCAAACAACC ATACGCGAATACTACAAACACCTCTACGCCAATAAA CTCGAAAATCTCGAGGAAATGGATACATTCTCGAC ACGTACACCCTTCCCAGGCTGAACCAGGAAGAAGTT GAATCACTGAATCGGCCTATCACGGGGAGTGAAATA GTAGCTATCATCAATCACTCCCTACCAAGAAGTCAC CCGGACCTGATGGATTCAACGCCGAATTCTACCAGA GATACATGGAAGAACTGGTGCCCTTCTTGCTGAAACT TTTCCAAAGTATTGAGAAAGAGGGAATACTTCCAAA CTCATTTTATGAGGCATCCATCATTCTGATCCCGAAG CCCGGCAGGGACACGACCAAGAAAGAGAATTTTCGA	

Name (size)	Description	Sequence	SEQ ID NO:
		CCAATCTCATTGATGAACATTGATGCAAAGATCCTCA ATAAAATACTGGCAAATCGGATTCAGCAGCACATAA AGAAGCTGATCCACCATGATCAAGTAGGCTTCATCCC CGGTATGCAAGGTTGGTTCAATATACGAAAATCAATC AATGTTATCCAGCATATAAACC GGGCCAAAGACAAG AACCACATGATTATTAGTATCGATGCTGAGAAAGCCT TTGACAAAATACAACAACCCTTCATGCTGAAAACATT GAATAAGCTGGGAATTGATGGCACCTACTTCAAAAT CATCAGAGCCATATATGACAAACCAACAGCAAATAT CATTCTGAATGGTCAGAAATTGGAAGCATTCCCCTTG AAAACCGGCACACGGCAGGGTTGCCCTCTGTCACCA CTCCTCTTCAACATCGTGTGGAAGTTCTTGCCCGCG CAATCCGGCAGGAAAAGGAAATCAAGGGCATTCAAC TGGGCAAAGAGGAAGTTAAATTGAGCCTGTTTGCAG ACGACATGATCGTCTATTTGGA AAAACCCCATAGTTAG TGCACAAAATCTGCTGAAGTTGATCAGTAATTTCTCC AAAGTGAGTGGGTACAAAATCAATGTGCAAAGAGC CAAGCTTTCTTGTACACCAACAACAGGCAAACCTGAGT CTCAAATCATGGGCGAACTCCCCTTCGTGATTGCATC CAAGCGGATCAAATACCTGGGGATTCAATTGACTCGT GATGTGAAGGACCTCTTCAAGGAGAACTACAAACCC CTGCTCAAGGAAATCAAAGAGGACACAAACAAATGG AAGAACATTCCATGCTCTTGGGTGGGAAGGATCAAT ATCGTCAAATGGCCATCCTGCCCCAAGGTAATTTACA GGTTCAATGCTATACCCATCAAGCTCCCCATGACATT CTTACAGAACTTGAAAAGACGACGCTGAAGTTCATT TGGAACCAGAAACGTGCCAGGATTGCTAAATCTATTC TCTCCCAAAGAACAAGCTGGCGGAATCACACTCC CAGACTTCAAACCTTACTACAAGGCGACCGTGACGA AAACGGCTTGGTACTGGTACCAAAACAGGGATATAG ATCAATGGAACCGAACGGAGCCCAGCGAAATTATGC CTCATATATACAACTATCTGATCTTTGACAAACCGGA GAAGAACAAGCAATGGGGAAAGGATAGTCTGTTTAA	

Name (size)	Description	Sequence	SEQ ID NO:
		TAAATGGTGCTGGGAAAACCTGGCTCGCAATCTGTAG GAAGCTGAAACTGGATCCATTCTTGACGCCTTATACA AAGATAAATTCCCGATGGATTAAAGATCTCAACGTG AAACCCAAAACAATTAAAACCCCTCGAGGAAAACCTG GGTATTACGATTACAGGACATTGGGGTGGGAAAGGAC TTCATGTCCAAAACCCCAAAAGCGATGGCAACCAAA GACAAAATCGACAAATGGGATCTCATAAACTTAAG TCATTTTGCACAGCTAAAGAAACGACAATTAGGGTG AACCGACAACCGACCACTTGGGAGAAAATCTTCGCA ACATACAGTTCTGACAAAGGCCTGATTTCAGGATCT ACAATGAATTGAAACAAATTTACAAGAAGAAGACGA ACAACCCTATAAAGAAATGGGCCAAGGACATGAACA GACACTTCTCTAAGGAAGACATTTATGCAGCCAAGA AACACATGAAGAAATGCAGCTCTTCACTGGCAATCA GGGAAATGCAAATCAAAACAACAATGAGATATCATC TCACACCCGTCAGAATGGCCATCATTAAAGAAGAGCG GAAACAACCGGTGCTGGCGTGGTTGCGGAGAAATCG GTACTCTCCTTCACTGTTGGTGGGACTGTAACTCGT TCAACCACTGTGGAAGTCTGTGTGGCGGTTCTCAGA GATCTGGAACCTCGAAATCCCATTTGACCCAGCCATCC CTCTCCTGGGTATATACCCGAATGAGTATAAATCCTG CTGCTATAAAGACACCTGCACAAGGATGTTTATTGCA GCTCTCTTCACAATCGCGAAGACGTGGAACCAACCC AAATGTCCGACTATGATTGACTGGATTAAGAAGATGT GGCACATATACACTATGGAATACTATGCTGCGATCAA GAACGATGAGTTCATATCATTTGTGGGCACATGGATG AAACTCGAAACCATCATACTCTCTAAATTGAGTCAAG AACAGAAAACATAACACCGTATATTTCCCTGATCGG TGGAATGGAGGCAAAAGGCCGCGGCCACGAAAA AGGCCGGCCAGGCAAAAAAGAAAAAGTAGCTACAA AGACGATGACGACAAGGACCATGGAGACGGTGAGA GACACAAAAAATCCAACACACTATTGCAATGAAAA TAAATTTCTTTATTAGCCAGAAGTCAGATGCTCAAG	

Name (size)	Description	Sequence	SEQ ID NO:
		GGGCTTCATGATGTCCCCATAATTTTTGGCAGAGGGA AAAAGATCTCAGTGGTATTTGTGAGCCAGGGCATTG GCCTTCTGATAGGCAGCCTGCACCTGAGGAGTGCGG CCGCTGATCAGTGAAACGCACCAGACAGCCGCGTCA GAGCAGCTCAGGTTCTGGGTAAGATACATTGATGAG TTTGGACAAACCACAAC TAGAATGCAGTGAAAAAA TGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATT TGTAACCATTATAAGCTGCAATAAACAAGTTTTACTT GTACAGCTCGTCCATGCCGAGAGTGATCCCGGCGGC GGTCACGAACTCCAGCAGGACCATGTGATCGCGCTT CTCGTTGGGGTCTTTGCTCAGGGCGGACTGGGTGCTC AGGTAGTGGTTGTCGGGCAGCAGCACGGGGCCGTCG CCGATGGGGGTGTTCTGCTGGTAGTGGTCGGCGAGCT GCACGCTGCCGTCCTCGATGTTGTGGCGGATCTTGAA GTTACCTTGATGCCGTTCTTCTGCTTGTCGGCCATGA TATAGACGTTGTGGCTGTTGTAGTTGTACTCCAGCTT GTGCCCCAGGATGTTGCCGTCCTCCTTGAAGTCGATG CCCTTCAGCTCGATGCGGTTACACAGGGTGTCGCCCT CGAACTTCACCTCGGCGCGGGTCTTG TAGTTGCCGTC GTCCTTGAAGAAGATGGTGCGCTCCTGGACGTAGCCT TCGGGCATGGCGGACTTGAAGAAGTCGTGCTGCTTCA TGTTGGTCGGGGTAGCGGCTGAAGCACTGCACGCCGT AGGTCAGGGTGGTCACGAGGGTGGGCCAGGGCACGG GCAGCTTGCCGGTGGTGACAGATGAACTTCAGGGTCA GCTTGCCGTAGGTGGCATCGCCCTCGCCCTCGCCGGA CACGCTGAACTTGTGGCCGTTTACGTCGCCGTCCAGC TCGACCAGGATGGGCACCACCCCGGTGAACAGCTCC TCGCCCTTGCTCACCATGGTGGCGGGATCTGACGGTT CACTAAACCCTGTGTTCTGGCGGCAAACCCGTTGCGA AAAAGAACGTTACGGCGACTACTGCACTTATATAC GGTTCTCCCCACCCTCGGGAAAAAGGCGGAGCCAG TACACGACATCACTTTCCAGTTTACCCCGCGCCACC TTCTCTAGGCACCGGATCAATTGCCGACCCCTCCCC	

Name (size)	Description	Sequence	SEQ ID NO:
		CAACTTCTCGGGGACTGTGGGCGATGTGCGCTCTGCC CTCCTCTCTGGCTCCATCGTAAGCAAACCTTAGAGGT TCTGGCAAGGAGAG	

Table 12. CREATE Construct Amino Acid Sequences

Name	Description	Sequence	SEQ ID NO:
CREATE-29	3X FLAG	DYKDHDGDYKDHDIDYKDDDDK	156
	SV40 NLS	PKKKRKV	157
	Cas9 H840A Nickase (H840A mutation)	DKKYSIGLDIGTNSVGWAVITDEYKVPS KKFKVLGNTDRHSIKKNLIGALLFDSGE TAEATRLKRTARRRYTRRKNRICYLQEI FSNEMAKVDDSFHRLSEESFLVEEDKK HERHPIFGNIVDEVAYHEKYPTIYHLRK KLVDSTDKADLRLIYLALAHMIKFRGH FLIEGDLNPDNSDVKLFIQLVQTYNQL FEENPINASGVDAKILSARLSKSRRL NLIAQLPGEKKNLFGNLIALSLGLTPN FKSNFDLAEDAKLQLSKDTYDDDLN LAQIGDQYADLFLAAKNLSDAILSDIL RVNTEITKAPLSASMIKRYDEHHQDLTL LKALVRQQLPEKYKEIFFDQSKNGYAG YIDGGASQEEFYKFIKPILEKMDGTEELL VKLNREDLLRKQRTFDNGSIPHQIHLGE LHAILRRQEDFYPLKDNREKIEKILTR IPYYVGPLARGNSRFAWMTRKSEETITP WNFEEVVDKGASAQSFIERMTNFDKNL PNEKVLPKHSLLYEYFTVYNELTKVKY VTEGMRKPAFLSGEQKKAIVDLLFKTN RKVTVKQLKEDYFKKIECFDSVEISGVE DRFNASLGTYHDLLKIKDKDFLDNEEN EDILEDIVLTLTLFEDREMIEERLKTYAH LFDDKVMKQLKRRRYTGWGRLSRKLI NGIRDKQSGKTILDFLKSDGFANRNF	150

Name	Description	Sequence	SEQ ID NO:
		QLIHDDSLTFKEDIQKAQVSGQGDSLHE HIANLAGSPAIKKGILQTVKVVDDELVKV MGRHKPENIVIAMARENQTTQKGQKNS RERMKRIEEGIKELGSQILKEHPVENTQ LQNEKLYLYYLQNGRDMYVDQELDIN RLSDYDVDAIVPQSFLKDDSIDNKVLTR SDKNRGKSDNVPSEEVVKKMKNYWRQ LLNAKLITQRKFDNLTKAERGGLSELDK AGFIKRQLVETRQITKHVAQILDSRMNT KYDENDKLIREVKVITLKSCLVSDFRKD FQFYKVVREINNYHHAHDAYLNAVVG ALIKKYPKLESEFVYGDYKVYDVRKMI AKSEQEIGKATAKYFFYSNIMNFFKTEIT LANGEIRKRPLIETNGETGEIVWDKGRD FATVRKVLSPQVNIVKKTEVQTGGFS KESILPKRNSDKLIARKKDWDPKKYGG FDSPTVAYSVLVVAKEKGKSKKLKSV KELLGITIMERSSEKPNIDFLEAKGYKE VKKDLIKLPKYSLELENGRKRMLASA GELQKGNELALPSKYVNFLYLASHYEK LKGSPEDNEQKQLFVEQHKHYLDEIIEQ ISEFSKRVLADANLDKVL SAYNKHARDK PIREQAENIIHLFTLTNLGAPAAFKYFDT TIDRKRYTSTKEVLDTLIHQSIITGLYET RIDLSQLGGD	
	Nucleoplasm in NLS	KRPAATKKAGQAKKKK	158
CREATE- 24	ORF1	MGKKQNRKTGNSKTQSASPPPKERSSSP ATEQSWMENDFDELREEGFRRSNYSEL REDIQTKGKEVENFEKNLEECITRITNTE KCLKELMELKTKARELREECRSLRSRC DQLEERVSAEMEDEMNMKREGKFREK RIKRNEQSLQEIWYVYKRPNLRLIGVPE SDVENGTKLENTLQDIIQENFPNLARQA NVQIQEIQRTPQRYSSRRATPRHIIVRFT	9

Name	Description	Sequence	SEQ ID NO:
		KVEMKEKMLRAAREKGRVTLKGKPIRL TVDLAETLQARREWGPIFNILKEKNFQ PRISYPAKLSFISEGEIKYFIDKQMLRDF VTRPALKELLKEALNMERNNRYQPLQ NHAKM	
	SV40 NLS	MPKKKRKVGGGS	159
	ORF2 (EN dead, D205A mutation)	TGSNSHITILTLNINGLNSAIKRHRLASW IKSQDPSVCCIQETHLTRDTHRLKIKG WRKIYQANGKQKKAGVAILVSDKTDF KPTKIKRDKEGHYIMVKGSIQQEELTIL NIYAPNTGAPRFIKQVLSDLQRDLDSHT LIMGDFNTPLSTLDRSTRQKVNKDTQEL NSALHQADLIDIYRTLHPKSTEYTFFSAP HHTYSKIAHIVGSKALLSKCKRTEIITNY LSDHSAIKLELRIKNLTQSRSTTWKLNN LLLNDYWVHNEMKAEIKMFFETNENK DTTYQNLWDAFKAVCRGKFIALNAYK RKQERSKIDTLTSQLKELEKQEQTTHSKA SRRQEITKIRAELEIETQKTLQKINESRS WFFERINKIDRPLARLIKKKREKNQIDTI KNDKGDITDPTEIQTIREYYKHLNAN KLENLEEMDTFLDTYTLPRLNQEEVESL NRPITGSEIVAIINSLPTKKSPGPDGFTA FYQRYMEELVPFLLKLFQSIEKEGILPNS FYEASIILIPKGRDTTKKENFRPISLMNI DAKILNKILANRIQQHIKKLIHHDQVGFI PGMQGWFNIRKSINVIQHINRAKDKNH MIISIDAEKAFDKIQPFMLKTLNKLIGID GTYFKIIRAIYDKPTANIILNGQKLEAFP LKTGTRQGCPSPLLFNIVLEVLARAIQ EKEIKGIQLGKEEVKLSLFADDMIVYLE NPIVSAQNLLKLISNFSKVSGYKINVQKS QAFLYTNNRQTESQIMGELPFVIASKRI KYLGIQLTRDVKDLFKENYKPLLKEIKE DTNKWKNIPCSWVGRINIVKMAILPKVI	160

Name	Description	Sequence	SEQ ID NO:
		YRFNAIPIKLPMTEFFTELEKTTLKFIWNQ KRARIAKSILSQKNKAGGITLPDFKLYY KATVTKTAWYWYQNRDIDQWNRTEPS EIMPHIYNYLIFDKPEKNKQWGKDSLFN KWCWENWLAICRKLKLDPFLTPYTKIN SRWIKDLNVKPKTIKTLEENLGITIQDIG VGKDFMSKTPKAMATKDKIDKWDLIK LKSFCTAKETTIRVNRQPTTWEEKIFATY SSDKGLISRIYNELKQIYKKKTNNPIKK WAKDMNRHFSKEDIYAAKKHMKKCSS SLAIREMQIKTTMRYHLTPVRMAIKKS GNNRCWRGCGEIGTLLHCWWDCCLVQ PLWKS VWRFLRDLELEIPFDPAIPLLG IY PNEYKSCCYKDTCTRMFIAALFTIAKTW NQPKCPTMIDWIKKMWHIYTMEEYAAI KNDEFISFVGTWMKLETHLSKLSQEQK TKHRIFSLIGGN	
	Nucleoplasm in NLS	KRPAATKKAGQAKKKK	158
	GFP	LLVQLVHAESDPGGGHELQQDHVIAL VGVFAQGGGLGAQVVVVGGQHGAVAD GGVLLVVVGELHAAVLDVVADLEVHL DAVLLLVGHDIDVVAVVVVLQLVPQD VAVLLEVDALQLDAVHQGVALELHLG AGLVVAVVLEEDGALLDVAFGHGGL EVVLLHVVGVAEALHAVGQGGHEGG PGHGQLAGGADELQGQLAVGGIALALA GHAELVAVYVAVQLDQDGHHPGEQL ALAHH	161
CREATE- 81	ORF1	MGKKQNRKTGNSKTQSASPPPKERSSSP ATEQSWMENDFDELREEGFRRSNYSEL REDIQTGKKEVENFEKNLEECITRITNTE KCLKELMELKTKARELREECRLRSRC DQLEERVSAEMEDEMNMKREGKFREK RIKRNEQSLQEIWVYVVRPNLRLIGVPE	9

Name	Description	Sequence	SEQ ID NO:
		SDVENGTKLENTLQDIIQENFPNLARQA NVQIQEIQRTPQRYSSRRATPRHIIVRFT KVEMKEKMLRAAREKGRVTLKGKPIRL TVDLAETLQARREWGPIFNILKEKNFQ PRISYPAKLSFISEGEIKYFIDKQMLRDF VTTRPALKELLKEALNMERNNRYQPLQ NHAKM	
	GSG Linker	GSG	162
	T2A Sequence	EGRGSLTTCGDVEENPGP	163
	ORF2 (EN dead, D205A)	TGSNSHITILTLNINGLNSAIKRHRLASW IKSQDPSVCCIQETHLTRDTHRLKIKG WRKIYQANGKQKKAGVAILVSDKTDF KPTKIKRDKEGHYIMVKGSIQQEELTIL NIYAPNTGAPRFIKQVLSDLQRDLDSHT LIMGDFNTPLSTLDRSTRQKVNKDTQEL NSALHQADLIDIYRTLHPKSTEYTFFSAP HHTYSKIAHIVGSKALLSKCKRTEITNY LSDHSAIKLELRIKNLTQSRSTTWKLNN LLLNDYWVHNEMKAEIKMFFETNENK DTTYQNLWDAFKAVCRGKFIALNAYK RKQERSKIDTLTSQLKELEKQEQTHTSKA SRRQEITKIRAELEIETQKTLQKINESRS WFFERINKIDRPLARLIKKKREKNQIDTI KNDKGDITDPTEIQTIREYYKHLNAN KLENLEEMDTFLDTYTLPRLNQEEVESL NRPITGSEIVAIINSLPTKKSPGPDGFTA FYQRYMEELVPFLLKLFQSIEKEGILPNS FYEASIILIPKPGRDTTKKENFRPISLMNI DAKILNKILANRIQQHIKKLIHHDQVGFI PGMQGWFNIRKSINVIQHINRAKDKNH MIISIDAEKAFDKIQPFMLKTLNKLID GTYFKIIRAIYDKPTANIILNGQKLEAFP LKTGTRQGCPSPLLFNIVLEVLAIRQ EKEIKGIQLGKEEVKLSLFADDMIVYLE	160

Name	Description	Sequence	SEQ ID NO:
		NPIVSAQNLLKLISNFSKVSGYKINVQKS QAFLYTNNRQTESQIMGELPFVIASKRI KYLGIQLTRDVKDLFKENYKPLLKEIKE DTNKWKNIPCSWVGRINIVKMAILPKVI YRFNAIPIKLPMTFFTELEKTTLKFIWNQ KRARIAKSILSQKNKAGGITLPDFKLYY KATVTKTAWYWYQNRDIDQWNRTEPS EIMPHIYNYLIFDKPEKNKQWGKDSL FN KWCWENWLAICRKLKLDPFLTPYTKIN SRWIKDLNVKPKTIKTLEENLGITIQDIG VGKDFMSKTPKAMATKDKIDKWDLIK LKSFCTAKETTIRVNRQPTTWEEKIFATY SSDKGLISRIYNELKQIYKKKTNNPIKK WAKDMNRHFSKEDIYAAKKHMKKCSS SLAIREMQIKTTMRYHLTPVRMAIKKS GNNRCWRGCGEIGTLLHCWWDCCLVQ PLWKS VWRFLRDLELEIPFDPAIPLGIY PNEYKSCCYKDTCTRMFIAALFTIAKTW NQPKCPTMIDWIKKMWHIYTM EYYAAI KNDEFISFVGTWMKLETILSKLSQE QK TKHRIFSLIGGN	
	Nucleoplasm in NLS	KRPAATKKAGQAKKKK	158
	GFP	LLVQLVHAESDPGGGHELQQDHVIAL L VGVFAQGGLGAQVVVVGQQHGAVAD GGVLLVVVGELHAAVL DVVADLEVHL DAVLLL VGHDIDVAVVVVLQLVPQD VAVLLEVDALQLDAVHQGV ALELHLG AGLVVAVVLEEDGALLDVAFGHG GLE EVL LHVVGVA AEALHAVGQGGHEGG PGHGQLAGGADELQGQLAVGGIALALA GHAELVAVYVAVQLDQDGHHPGEQLL ALAHH	161
	SV 40 NLS	MPKKKRKVGGGS	159

Name	Description	Sequence	SEQ ID NO:
CREATE-80	ORF2 (EN dead, D205A)	TGSNSHITILTLNINGLNSAIKRHRLASW IKSQDPSVCCIQETHLTRDTHRLKIKG WRKIYQANGKQKKAGVAILVSDKTDF KPTKIKRDKEGHYIMVKGSIQQEELTIL NIYAPNTGAPRFIKQVLSDLQRDLDSHT LIMGDFNTPLSTLDRSTRQKVNKDTQEL NSALHQADLIDIYRTLHPKSTEYTFFSAP HHTYSKIAHIVGSKALLSKCKRTEITNY LSDHSAIKLELRIKNLTQSRSTTWKLNN LLLNDYWVHNEMKAEIKMFFETNENK DDTYQNLWDAFKAVCRGKFIALNAYK RKQERSKIDTLTSQLKELEKQEQTHSKA SRRQEITKIRAELEIETQKTLQKINESRS WFFERINKIDRPLARLIKKKREKNQIDTI KNDKGDITTDPTIEIQTIREYYKHLNAN KLENLEEMDTFLDYTLPRLNQEEVESL NRPITGSEIVAIINSLPTKKSPGPDGFTA FYQRYMEELVPFLLKLFQSIEKEGILPNS FYEASIIIPKPGRDTTKENFRPISLMNI DAKILNKILANRIQQHIKKLIHHDQVGFI PGMQGWFNIRKSINVIQHINRAKDKNH MIISIDAEKAFDKIQPFMLKTLNKLGD GTYFKIIRAIYDKPTANIILNGQKLEAFP LKTGTRQGCPSPLLFNIVLEVLARAIQ EKEIKGIQLGKEEVKLSLFADDMIVYLE NPIVSAQNLLKLISNFSKVSGYKINVQKS QAFLYTNNRQTESQIMGELPFVIASKRI KYLGIQLTRDVKDLFKENYKPLLKEIKE DTNKWKNIPCSWVGRINIVKMAILPKVI YRFNAIPIKLPMTEFFTELEKTTLKFIWNQ KRARIAKSILSQKNKAGGITLPDFKLYY KATVTKTAWYWYQNRDIDQWNRTEPS EIMPHIYNYLIFDKPEKNKQWGKDSL FN KWCWENWLAICRKLKLDPFLTPYTKIN SRWIKDLNVKPKTIKTLEENLGITQDIG	160

Name	Description	Sequence	SEQ ID NO:
		VGKDFMSKTPKAMATKDKIDKWDLIK LKSFCTAKETTIRVNRQPTTWEEKIFATY SSDKGLISRIYNELKQIYKKKTNNPIKK WAKDMNRHFSKEDIYAAKKHMKKCSS SLAIREMQIKTTMRYHLTPVRMAIKKS GNNRCWRGCGEIGTLLHCWWDCCLVQ PLWKS VWRFLRDLELEIPFDPAIPLLG IY PNEYKSCCYKDTCTRMFIAALFTIAKTW NQPKCPTMIDWIKKMWHIYTMEYYAAI KNDEFISFVGTWMKLETILSKLSQEQK TKHRIFSLIGGN	
	Nucleoplasm in NLS	KRPAATKKAGQAKKKK	158
	FLAG	DYKDDDDK	16
CREATE- 82	SV40 NLS	MPKKKRKVGGGS	159
	Sso7d	MATVKFKYKGEEKEVDISKIKKVWRVG KMISFTYDEGGGKTGRGAVSEKDAPKE LLQMLEKQKK	152
	33XTEN	SGGSSGGSSGSETPGTSESATPESSGGSS GGSS	153
	ORF2 (aa. 239-1275)	IKNLTQSRSTTWKLNLLNDYWVHNE MKAEIKMFFETNENKDTTYQNLWDAF KAVCRGKFIALNAYKRKQERSKIDTLTS QLKELEKQEQTTHSKASRRQEITKIRAE KEIETQKTLQKINESRSWFFERINKIDRP LARLIKKKREKNQIDTIKNDKGDITDP TEIQTIREYYKHL YANKLENLEEMDTF LDTYTL PRLNQEEVESLNR PITGSEIVAI NSLPTKKSPGPDGFTA EFYQRYMEELVP FLLKLFQSIEKEGILPNSFYEASIILIPKPG RDTTKKENFRPISLMNIDAKILNKILANR IQQHIKKLIHHDQVGFIPGMQGWFNIRK SINVIQHINRAKDKNHMIISIDAEKAFDK IQQPFMLKTLNKL GIDGTYFKIIRAIYDK PTANIILNGQKLEAFPLKTGTRQGCPLSP	151

Name	Description	Sequence	SEQ ID NO:
		LLFNIVLEVLARAIRQEKEIKGIQLGKEE VKLSLFADDMIVYLENPIVSAQNLLKLI SNFSKVSGYKINVQKSQAFLYTNNRQT ESQIMGELPFVIAASKRIKYLGIQLTRDVK DLFKENYKPLLKEIKEDTNKWKNIPCS WVGRINIVKMAILPKVIYRFNAIPIKLPM TFFTELEKTTLKFIWNQKRARIAKSILSQ KNKAGGITLPDFKLYYKATVTKTAWY WYQNRDIDQWNRTEPSEIMPHIYNLIF DKPEKNKQWGKDSL FNKWCWENWLAI CRKLKLDPFLTPYTKINSRWIKDLNVKP KTIKTLEENLGITIQDIGVGKDFMSKTPK AMATKDKIDKWDLIKLKSFCTAKETTIR VNRQPTTWEKIFATYSSDKGLISRIYNEL KQIYKKKTNNPIKKWAKDMNRHFSKE DIYAAKKHMKKCSSLAIEMQIKTTM RYHLTPVRMAIIKSGNNRCWRGCGEI GTLLHCWWDCCLVQPLWKS VWRFLRD LELEIPFDPAIPLLGIYPNEYKSCCYKDT CTRMFIAALFTIAKTWNQPKCPTMIDWI KKMWHIYTMEYYAAIKNDEFISFVGTW MKLETIILSKLSQEQKTKHRIFSLIGGN	
	Nucleoplasm in NLS	KRPAATKKAGQAKKKK	158
	FLAG	DYKDDDDK	16
CREATE- 83	SV40 NLS	MPKKKRKVGGGS	159
	HMGD	MSDKPKRPLSAYMLWLNSARESIKREN PGIKVTEVAKRGGELWRAMKDKSEWE AKAAKAKDDYDRAVKEFEANGGSSAA NGGGAKKRAKPAKKVAKKSKKEESDE DDDESE	154
	33XTEN	SGGSSGGSSGSETPGTSESATPESSGGSS GGSS	153
	ORF2 (aa. 239-1275)	IKNLTQSRSTTWKLNNLLNDYWVHNE MKAEIKMFFETNENKDTTYQNLWDAF	151

Name	Description	Sequence	SEQ ID NO:
		KAVCRGKFIALNAYKRKQERSKIDTLTS QLKELEKQEQTHSKASRRQEITKIRAE KEIETQKTLQKINESRSWFFERINKIDRP LARLIKKKREKNQIDTIKNDKGDITDP TEIQTIREYYKHL YANKLENLEEMDTF LDTYTLPRLNQEEVESLNR PITGSEIVAH NSLPTKKSPGPDGFTA EFYQRYMEELVP FLLKLFQSIEKEGILPNSFYEASIIIPKPG RDTTKKENFRPISLMNIDAKILNKILANR IQQHIKKLIHHDQVGFIPGMQGWFNIRK SINVIQHINRAKDKNHMIISIDAEKAFDK IQQPFMLKTLNKL GIDGTYFKIIRAIYDK PTANIILNGQKLEAFPLKTGTRQGCPLSP LLFNIVLEV LARAIRQEKEIKGIQLGKEE VKLSLFADDMIVYLENPIVSAQNLLKLI SNFSKVSGYKINVQKSQAFLYTNNRQT ESQIMGELPFVIA SKRIKYLGIQLTRDVK DLFKENYKPLLKEIKEDTNKWKNIPCS WVGRINIVKMAILPKVIYRFNAIPIKLPM TFFTELEKTTLKFIWNQKRARIAKSILSQ KNKAGGITLPDFKLYYKATVTKTAWY WYQNRDIDQWNRTEPSEIMPHIYNLIF DKPEKNKQWGKDSL FNKWCWENWLAI CRKLKLDPFLTPYTKINSRWIKDLNVKP KTIKTLEENLGITIQDIGVGKDFMSKTPK AMATKDKIDKWDLIKLSFCTAKETTIR VNRQPTTWEKIFATYSSDKGLISRIYNEL KQIYKKKTNNPIKKWAKDMNRHFSKE DIYAAKHKMKCSSSLAIREMQIKTTM RYHLTPVRMAIHKSGNNRCWRGCGEI GTLLHCWWDCKLVQPLWKS VWRFLRD LELEIPFDPAIPLLG IYPNEYKSCCYKDT CTRMFIAALFTIAKTWNQPKCPTMIDWI KKMWHIYTMEYYAAIKNDEFISFVGW MKLETHLSKLSQEQKTKHRIFSLIGN	

Name	Description	Sequence	SEQ ID NO:
	Nucleoplasm in NLS	KRPAATKKAGQAKKKK	158
	FLAG	DYKDDDDK	16
	SV40 NLS	MPKKKRKVGGGS	159
	PARP1 ZnF1 and ZnF2 domain	MAESSDKLYRVEYAKSGRASCKKCSSES IPKDSLRMAIMVQSPMFDGKVPHWYHF SCFWKVGHSIRHPDVEVDGFSELRWDD QQKVKKTAEAGGVTGKGQDGIGSKAE KTLGDFAAEYAKSNRSTCKGCMEKIEK GQVRLSKKMVDPEKPQLGMIDRWYHP GCFVKNREELGFRPEYSASQLKGFSLLA TEDKEALKKQLPGVKSEGKRKG	155
	33XTEN	SGGSSGGSSGSETPGTSESATPESSGGSS GGSS	153
CREATE- 84	ORF2 (aa. 239-1275)	IKNLTQSRSTTWKLNLLNDYWVHNE MKAIEIKMFFETNENKDDTYQNLWDAF KAVCRGKFIALNAYKRKQERSKIDTLTS QLKELEKQEQTHSKASRRQEITKIRAE KEIETQKTLQKINESRSWFFERINKIDRP LARLIKKKREKNQIDTIKNDKGDITDP TEIQTIREYYKHL YANKLENLEEMDTF LDITYTLPRLNQEEVESLNR PITGSEIVAI NSLPTKKSPGPDGFTA EFYQRYMEELVP FLLKLFQSIEKEGILPNSFYEASIIIPKPG RDTTKKENFRPISLMNIDAKILNKILANR IQQHIKKLIHHDQVGFIPGMQGWFNIRK SINV IQHINRAKDKNHMIISIDAEKAFDK IQQPFMLKTLNKL GIDGTYFKIIRAIYDK PTANIILNGQKLEAFPLKTGTRQGCPLSP LLFNIVLEV LARAIRQEKEIKGIQLGKEE VKLSLFADDMIVYLENPIVSAQNLLKLI SNFSKVSGYKINVQKSQAFLYTNNRQT ESQIMGELPFVIA SKRIKYLGIQLTRDVK DLFKENYKPLLKEIKEDTNKWKNIPCS WVGRINIVKMAILPKVIYRFNAIPIKLPM	151

Name	Description	Sequence	SEQ ID NO:
		TFFTELEKTTLKFIWNQKRARIAKSILSQ KNKAGGITLPDFKLYYKATVTKTAWY WYQNRDIDQWNRTEPSEIMPHIYNLIF DKPEKNKQWGKDSL FNKWCWENWLAI CRKLKLDPFLTPYTKINSRWIKDLNVKP KTIKTLEENLGITIQDIGVGKDFMSKTPK AMATKDKIDKWDLIKLSFCTAKETTIR VNRQPTTWEEKIFATYSSDKGLISRIYNEL KQIYKKKTNNPIKKWAKDMNRHFSKE DIYAAKKHMKKCSSSLAIREMQIKTTM RYHLTPVRMAIIKSGNNRCWRGCGEI GTLLHCWWDCCLVQPLWKS VWRFLRD LELEIPFDPAIPLLGIYPNEYKSCCYKDT CTRMFIAALFTIAKTWNQPKCPTMIDWI KKMWHIYTMEYYAAIKNDEFISFVG TW MKLETHLSKLSQEQKTKHRIFSLIGN	
	Nucleoplasm in NLS	KRPAATKKAGQAKKKK	158
	FLAG	DYKDDDDK	16
CREATE- 25	ORF1	MGKKQNRKTGNSKTQSASPPPKERSSSP ATEQSWMENDFDELREEGFRRSNYSEL REDIQTGKEVENFEKNLEECITRITNTE KCLKELMELKTKARELREECRSLRSRC DQLEERVSAMEDEMNMKREGKFREK RIKRNEQSLQEIW DYVKRPNLRLIGVPE SDVENGT KLENTLQDIIQENFPNLARQA NVQIQEIQRTPQRYSSRRATPRHIIVRFT KVEMKEKMLRAAREKGRVTLKGKPIRL TVDL SAETLQARREWGPIFNILKEKNFQ PRISYPAKLSFISEGEIKYFIDKQMLRDF VTTRPALKELLKEALNMERNNRYQPLQ NHAKM	9
	SV40 NLS	MPKKKRKVGGGS	159
	ORF2 D205A	TGSNSHITILTLNINGLNSAIKRHRLASW IKSQDPSVCCIQETHLTCRDTHRLKIKG	160

Name	Description	Sequence	SEQ ID NO:
		WRKIYQANGKQKKAGVAILVSDKTD KPTKIKRDKEGHYIMVKGSIQQEELTIL NIYAPNTGAPRFIKQVLSDLQRDLDSHT LIMGDFNTPLSTLDRSTRQKVNKDTQEL NSALHQADLIDIYRTLHPKSTEYTFFSAP HHTYSKIAHIVGSKALLSKCKRTEITNY LSDHSAIKLELRIKNLTQSRSTTWKLNN LLLNDYWVHNEMKAEIKMFFETNENK DTTYQNLWDAFKAVCRGKFIALNAYK RKQERSKIDTLTSQLKELEKQEETHSKA SRRQEITKIRAELEKETQKTLQKINESRS WFFERINKIDRPLARLIKKKREKNQIDTI KNDKGDITTDPTIEIQTIREYYKHLNAN KLENLEEMDTFLDTYTLPRLNQEEVESL NRPITGSEIVAIINSLPTKKSPGPDGFTA FYQRYMEELVPFLLKLFQSIEKEGILPNS FYEASIIIPKPGRDTTKKENFRPISLMNI DAKILNKILANRIQQHIKKLIHHDQVGFI PGMQGWFNIRKSINVIQHINRAKDKNH MIISIDAEKAFDKIQQPFMLKTLNKLGD GTYFKIIRAIYDKPTANIILNGQKLEAFP LKTGTRQGCPLSPLLFNIVLEVLARAIQ EKEIKGIQLGKEEVKLSLFADDMIVYLE NPIVSAQNLLKLISNFSKVSGYKINVQKS QAFLYTNNRQTESQIMGELPFVIASKRI KYLGIQLTRDVKDLFKENYKPLLKEIKE DTNKWKNIPCSWVGRINIVKMAILPKVI YRFNAIPIKLPMTEFFTELEKTTLKFIWNQ KRARIAKSILSQKNKAGGITLPDFKLYY KATVTKTAWYWYQNRDIDQWNRTEPS EIMPHIYNYLIFDKPEKNKQWGKDSLFN KWCWENWLAICRKLKLDPFLTPYTKIN SRWIKDLNVKPKTIKTLEENLGITIQDIG VGKDFMSKTPKAMATKDKIDKWDLIK LKSFCTAKETTIRVNRQPTTWEKIFATY	

Name	Description	Sequence	SEQ ID NO:
CREATE-37		SSDKGLISRIYNELKQIYKKKTNNPIKK WAKDMNRHFSKEDIYAAKKHMKKCSS SLAIREMQIKTTMRYHLTPVRMAIHKKS GNNRCWRGCGEIGTLLHCWWDCCLVQ PLWKS VWRFLRDLELEIPFDPAIPLLG IY PNEYKSCCYKDTCTRMFIAALFTIAKTW NQPKCPTMIDWIKKMWHIYTMEYYAAI KNDEFISFVGTWMKLETIILSKLSQEQK TKHRIFSLIGGN	
	Nucleoplasm in NLS	KRPAATKKAGQAKKKK	158
	GFP	LLVQLVHAESDPGGGHELQQDHVIAL VGVFAQGGGLGAQVVVVGQQHGAVAD GGVLLVVVGELHAAVLDDVADLEVHL DAVLLLVGHDIDVVAVVVVLQLVPQD VAVLLEVDALQLDAVHQGV ALELHLG AGLVVAVVLEEDGALLDVAFGHGGL EVLVLLHVVGVA AEALHAVGQGGHEGG PGHGQLAGGADELQGQLAVGGIALALA GHAELVAVYVAVQLDQDGHHPGEQLL ALAHH	161
CREATE-37	ORF1	MGKKQNRKTGNSKTQSASPPPKERSSSP ATEQSWMENDFDELREEGFRRSNYSEL REDIQTGKKEVENFEKNLEECITRITNTE KCLKELMELKTKARELREECRSLR SRC DQLEERV SAMEDEMNMKREGKFREK RIKRNEQSLQEIW DYVKRPNLRLIGVPE SDVENGTKLENTLQDIIQENFPNLARQA NVQIQEIQRTPQRYSSRRATPRHIIVRFT KVEMKEKMLRAAREKGRVTLKGKPIRL TVDL SAETLQARREWGPIFNILKEKNFQ PRISYPAKLSFISEGEIKYFIDKQMLRDF VTTRPALKELLKEALNMERNNRYQPLQ NHAKM	9
	SV40 NLS	MPKKKRKVGGGS	159

Name	Description	Sequence	SEQ ID NO:
	ORF2 D205A	TGSNSHITILTLNINGLNSAIKRHRLASW IKSQDPSVCCIQETHLTRDTHRLKIKG WRKIYQANGKQKKAGVAILVSDKTDF KPTKIKRDKEGHYIMVKGSIQQEELTIL NIYAPNTGAPRFIKQVLSDLQRDLDSHT LIMGDFNTPLSTLDRSTRQKVNKDTQEL NSALHQADLIDIYRTLHPKSTEYTFPSAP HHTYSKIAHIVGSKALLSKCKRTEITNY LSDHSAIKLELRIKNLTQSRSTTWKLNN LLLNDYWVHNEMKAEIKMFFETNENK DDTYQNLWDAFKAVCRGKFIALNAYK RKQERSKIDTLTSQLKELEKQEQTHSKA SRRQEITKIRAELEIETQKTLQKINESRS WFFERINKIDRPLARLIKKREKNQIDTI KNDKGDITTDPTIEIQTIREYYKHLNAN KLENLEEMDTFLDYTLPRLNQEEVESL NRPITGSEIVAIINSLPTKKSPGPDGFTA FYQRYMEELVPFLLKLFQSIEKEGILPNS FYEASIIIPKPGRDTTKKENFRPISLMNI DAKILNKILANRIQQHIKKLIHHDQVGFI PGMQGWFNIRKSINVIQHINRAKDKNH MIISIDAEKAFDKIQPFMLKTLNKLGD GTYFKIIRAIYDKPTANIILNGQKLEAFP LKTGTRQGCPSPLLFNIVLEVLARAIQ EKEIKGIQLGKEEVKLSLFADDMIVYLE NPIVSAQNLLKLISNFSKVSGYKINVQKS QAFLYTNNRQTESQIMGELPFVIASKRI KYLGIQLTRDVKDLFKENYKPLLKEIKE DTNKWKNIPCSWVGRINIVKMAILPKVI YRFNAIPIKLPMTEFFTELEKTTLKFIWNQ KRARIAKSILSQKNKAGGITLPDFKLYY KATVTKTAWYWYQNRDIDQWNRTEPS EIMPHIYNYLIFDKPEKNKQWGKDSLFN KWCWENWLAICRKLKLDPFLTPYTKIN SRWIKDLNVKPKTIKTLEENLGITIQDIG	160

Name	Description	Sequence	SEQ ID NO:
		VGKDFMSKTPKAMATKDKIDKWDLIK LKSFCTAKETTIRVNRQPTTWEEKIFATY SSDKGLISRIYNELKQIYKKKTNNPIKK WAKDMNRHFSKEDIYAAKKHMKKCSS SLAIREMQIKTTMRYHLTPVRMAIKKS GNNRCWRGCGEIGTLLHCWWDCCLVQ PLWKS VWRFLRDLELEIPFDPAIPLLG IY PNEYKSCCYKDTCTRMFIAALFTIAKTW NQPKCPTMIDWIKKMWHIYTMEYYAAI KNDEFISFVGTWMKLETILSKLSQEQK TKHRIFSLIGGN	
	Nucleoplasm in NLS	KRPAATKKAGQAKKKK	158
	GFP	LLVQLVHAESDPGGGHELQQDHVIAL VGVFAQGGLGAQVVVVGQQHGAVAD GGVLLVVVGELHAAVL DVVADLEVHL DAVLLL VGHDIDVAVVVVLQLVPQD VAVLLEVDALQLDAVHQGV ALELHLG AGLVVAVVLEEDGALLDVAFGHGGL EVLVLLHVVGVA AEALHAVGQGGHEGG PGHGQLAGGADELQGGQLAVGGIALALA GHAELVAVYVAVQLDQDGHHPGEQLL ALAHH	161

CLAIMS

1. A composition comprising:
 - (a) an RNA molecule comprising:
 - (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide;
 - (ii) a 5' homology arm and a 3' homology arm; and
 - (iii) a sequence encoding a human ORF2p polypeptide, wherein the sequence encoding a human ORF2p polypeptide is upstream of the 5' homology arm or downstream of the 3' homology arm;
 - (b) an endonuclease or a polynucleic acid sequence encoding the endonuclease; and
 - (c) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.
2. A composition comprising:
 - (a) an RNA molecule comprising:
 - (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide;
 - (ii) a 5' homology arm and a 3' homology arm; and
 - (iii) a sequence encoding a polypeptide with reverse transcriptase activity, wherein the sequence encoding a polypeptide with reverse transcriptase

activity is upstream of the 5' homology arm or downstream of the 3' homology arm;

- (b) an endonuclease or a polynucleic acid sequence encoding the endonuclease; and
- (c) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.

3. A composition comprising:

- (a) an RNA molecule comprising:
 - (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide;
 - (ii) a 5' homology arm that is upstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide and a 3' homology arm that is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide; and
 - (iii) a sequence encoding a polypeptide with reverse transcriptase activity, wherein the sequence encoding a polypeptide with reverse transcriptase activity is upstream of the 5' homology arm or downstream of the 3' homology arm;
- (b) an endonuclease or a polynucleic acid sequence encoding the endonuclease; and
- (c) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.

4. The composition of any one of claims 1-3, wherein the RNA molecule of (a) comprises the polynucleic acid sequence encoding the endonuclease of (b).
5. The composition of any one of claims 1-4, wherein the sequence of the 5' homology arm or the 3' homology arm is about 50 nucleotides in length.
6. The composition of any one of claims 1-5, wherein the polynucleic acid sequence encoding the endonuclease is upstream of the 5' homology arm sequence or downstream of the 3' homology arm sequence.
7. The composition of any one of claims 1-6, wherein the polynucleic acid sequence encoding the endonuclease is present in a polynucleic acid molecule that is different than the RNA molecule of (a).
8. The composition of any one of claims 1-7, wherein the RNA molecule of (a) comprises the one or more polynucleic acids encoding the one or more guide RNA molecules of (c), and wherein the one or more polynucleic acids encoding the one or more guide RNA molecules are upstream of the 5' homology arm sequence or downstream of the 3' homology arm sequence.
9. The composition of any one of claims 1-8, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of a second strand of the genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of a first strand of the genomic DNA of the target cell.
10. The composition of claim 9, wherein the number of nucleobases between first target sequence of the second strand of the genomic DNA of a target cell and the second target sequence of the first strand of the of the genomic DNA of the target cell is about 10-1000 nucleobases, about 50-500 nucleobases, or about 100-300 nucleobases.
11. The composition of any one of claims 1-10, wherein the 5' homology arm sequence comprises a sequence capable of hybridizing to a target sequence of the second strand of the genomic DNA of the target cell and the 3' homology arm sequence comprises a sequence capable of hybridizing to a target sequence of the first strand of the genomic DNA of the target cell.
12. The composition of claim 11, wherein the sequence of the 3' homology arm comprises a sequence that primes with the 3'-flap on the first strand created by Cas 9 nicking.
13. The composition of claim 11 or 12, wherein the 3' end of the target sequence of the second strand of the genomic DNA to which the 5' homology arm hybridizes is at a distance of at most 50, 40, 30, 20, 10, 5, 4, 3, 2, or 1 bases from the 3' end of the target sequence of the first strand to which the 3' homology arm hybridizes.

14. The composition of any one of claims 11-13, wherein the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA overlaps with the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence.
15. The composition of any one of claims 11-14, wherein the sequence of the 3' homology arm sequence capable of hybridizing to a target sequence of a first strand of the genomic DNA is at most 50, 40, 30, 20, 10, 5, 4, 3, 2, or 1 bases downstream (3') of the second target sequence of a first strand of the genomic DNA of the target cell capable of hybridizing to the second guide RNA sequence.
16. The composition of any one of claims 1-15, wherein the exogenous sequence encoding a polypeptide is flanked by the 5' homology arm and the 3' homology arm.
17. The composition of any one of claims 1-16, wherein the endonuclease is a nickase.
18. The composition of claim 17, wherein the nickase is a Cas9 nickase.
19. The composition of claim 18, wherein the Cas9 nickase comprises an H840A mutation relative to the wild type Cas9.
20. The composition of claim 19, wherein the Cas9 nickase is encoded by a nucleic acid sequence with at least 90% sequence identity to SEQ ID NO: 124.
21. The composition of any one of claims 1-20, wherein the polynucleic acid sequence encoding the endonuclease is an RNA sequence.
22. The composition of any one of claims 1-21, wherein the one or more guide RNA molecules comprises the first guide RNA molecule comprising a first guide RNA sequence and the second guide RNA molecule comprising a second guide RNA sequence, and wherein the first guide RNA molecule forms a complex with a first endonuclease molecule and the second guide RNA molecule forms a complex with a second endonuclease molecule.
23. The composition of any one of claims 1-22, wherein the one or more guide RNA molecules direct the endonuclease to cut (i) a first strand of the genome DNA of the target cell and (ii) a second strand of the genomic DNA of the target cell.
24. The composition of any one of claims 1-23, wherein the one or more guide RNA molecules do not direct the endonuclease to create a double strand break (DSB) in the genome DNA of the target cell.
25. The composition of any one of claims 1-24, wherein the first guide RNA molecule directs the endonuclease to cut a first strand of the genome DNA of the target cell and the second guide RNA molecule directs the endonuclease to cut a second strand of the genomic DNA of the target cell.

26. The composition of any one of claims 23-25, wherein the number of nucleobases between the first cut of the first strand of the genomic DNA and the second cut of the second strand of the genomic DNA is about 10-1000 bases, about 50-500 bases, or about 100-300 bases.
27. The composition of any one of claims 1-26, wherein the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell are within a same region of the genomic DNA of the target cell.
28. The composition of any one of claims 1-27, wherein the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell are within a same locus of the genomic DNA of the target cell.
29. The composition of claim 28, wherein the locus is a genomic safe harbor locus.
30. The composition of claim 28 or 29, wherein the locus is non-ribosomal DNA.
31. The composition of any one of claims 1-30, wherein the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell do not comprise a sequence of ribosomal DNA.
32. The composition of any one of claims 28-30, wherein the locus is a human ortholog of the mouse Rosa26 locus, adeno-associated virus site 1 (AAVS1), CCR5 gene, HEK3, PRNP, or IDS loci.
33. The composition of any one of claims 1-32, wherein the first target sequence of the genomic DNA of the target cell and the second target sequence of the genomic DNA of the target cell comprise sequences from a single gene within the genomic DNA of the target cell.
34. The composition of claim 33, wherein the single gene is a non-essential gene.
35. The composition of any one of claims 2-34, wherein the polypeptide with reverse transcriptase activity promotes (i) reverse transcription of the reverse complement sequence of the insert sequence thereby producing the insert sequence, (ii) integration of the insert sequence into the genomic DNA of the target cell, or (iii) integration of the insert sequence into the genomic DNA of the target cell via target primed reverse transcription (TPRT).
36. The composition of any one of claims 2-35, wherein the polypeptide with reverse transcriptase activity promotes (i) integration of the insert sequence into the genomic DNA of the target cell at the first target sequence of the genomic DNA of the target cell, (ii) integration of the insert sequence into the genomic DNA of the target cell at the second target sequence of the genomic DNA of the target cell, or (iii) both.

37. The composition of any one of claims 2-36, wherein the sequence encoding a polypeptide with reverse transcriptase activity comprises a mobile genetic element sequence.
38. The composition of any one of claims 2-37, wherein the sequence encoding a polypeptide with reverse transcriptase activity comprises a human mobile genetic element sequence.
39. The composition of any one of claims 2-38, wherein the mobile human mobile genetic element comprises LINE-1 or a fragment thereof.
40. The composition of any one of claims 2-39, wherein the polypeptide with reverse transcriptase activity comprises a human ORF2p polypeptide or a functional fragment thereof.
41. The composition of claim 1 or 40, wherein the human ORF2p polypeptide or functional fragment thereof lacks endonuclease (EN) activity.
42. The composition of claim 41, wherein the human ORF2p polypeptide or functional fragment thereof comprises a mutation in the endonuclease domain that results in a lack of endonuclease activity.
43. The composition of any one of claims 40-42, wherein the human ORF2p polypeptide or functional fragment thereof comprises a mutation at D205 relative to wild type human ORF2p.
44. The composition of claim 43, wherein the mutation at D205 comprises a D205A mutation relative to wild type human ORF2p.
45. The composition of any one of claims 40-44, wherein the human ORF2p polypeptide or functional fragment thereof is a functional fragment of a human ORF2p polypeptide that lacks an endonuclease domain.
46. The composition of any one of claim 1-45, wherein the composition further comprises (d) an RNA molecule comprising a sequence encoding a human ORF2p polypeptide that lacks a functional endonuclease domain, wherein the RNA molecule of (d) lacks sequences encoding a human ORF1p polypeptide or the exogenous sequence encoding a polypeptide.
47. The composition of claim 46, wherein the human ORF2p polypeptide that lacks a functional endonuclease domain comprises an exogenous DNA binding domain.
48. The composition of claim 47, wherein the exogenous DNA binding domain increases DNA binding activity of the human ORF2p polypeptide or functional fragment thereof.

49. The composition of claim 47 or 48, wherein the exogenous DNA binding domain increases TPRT processivity of the human ORF2p polypeptide or functional fragment thereof.
50. The composition of any one of claims 46-49, wherein the exogenous DNA binding domain comprises a Sso7d DNA binding domain.
51. The composition of claim 50, wherein the Sso7d DNA binding domain comprises an amino acid sequence with at least 90% sequence identity to SEQ ID NO: 152.
52. The composition of any one of claims 49-51, wherein the RNA molecule of (d) comprises a sequence with at least 90% sequence identity to SEQ ID NO: 144.
53. The composition of any one of claims 46-49, wherein the exogenous DNA binding domain is derived from an HMGD protein.
54. The composition of claim 53, wherein the exogenous DNA binding domain comprises an amino acid sequence with at least 90% sequence identity to SEQ ID NO: 154.
55. The composition of claim 53 or 54, wherein the RNA molecule of (d) comprises a sequence with at least 90% sequence identity to SEQ ID NO: 164.
56. The composition of any one of claims 46-49, wherein the exogenous DNA binding domain increases direct recruitment of the endonuclease to the target sequence in the genomic DNA of the target cell.
57. The composition of claim 46-49, wherein the exogenous DNA binding domain is derived from an hPARP1 protein.
58. The composition of claim 57, wherein the exogenous DNA binding domain comprises ZnF1 domain and ZnF2 domain of the hPARP1 protein.
59. The composition of claim 58, wherein the exogenous DNA binding domain comprises an amino acid sequence with at least 90% sequence identity to SEQ ID NO: 155.
60. The composition of any one of claims 56-59, wherein the RNA molecule of (d) comprises a sequence with at least 90% sequence identity to SEQ ID NO: 146.
61. The composition of any one of claims 46-60, wherein the additional DNA binding domain is operably linked to the functional fragment of a human ORF2p polypeptide via a linker.
62. The composition of claim 61, wherein the linker is a flexible linker.
63. The composition of claim 62, wherein the flexible linker comprises a 33XTEN linker.
64. The composition of any one of claims 46-63, wherein the human ORF2p polypeptide that lacks a functional endonuclease domain is a human ORF2p polypeptide that lacks an endonuclease domain.

65. The composition of any one of claims 46-64, wherein a ratio of an amount of the RNA molecule of (a) to that of the RNA molecule of (d) is about 5:1.
66. The composition of any one of claims 1-65, wherein the RNA molecule of (a) further comprises a sequence encoding a human ORF1p polypeptide, wherein the sequence encoding the human ORF1p polypeptide is upstream of the 5' homology arm or downstream of the 3' homology arm.
67. The composition of claim 66, wherein the sequence encoding a human ORF1p polypeptide and the sequence encoding a human ORF2p polypeptide are separated by an interORF sequence.
68. The composition of claim 66, wherein the sequence encoding a human ORF1p polypeptide and the sequence encoding a human ORF2p polypeptide are separated by (a) a sequence encoding a GSS linker and (b) a sequence encoding a T2A cleavage sequence.
69. The composition of any one of claims 66-68, wherein the sequence encoding the human ORF1 polypeptide comprises a sequence with at least 90% sequence identity to SEQ ID NO: 10.
70. The composition of any one of claims 1-69, wherein the polypeptide is a therapeutic polypeptide.
71. The composition of any one of claims 1-70, wherein the polypeptide is selected from the group consisting of a ligand, an antibody, a receptor, an enzyme, a transport protein, a structural protein, a hormone, a contractile protein, a storage protein and a transcription factor.
72. The composition of any one of claims 1-71, wherein the polypeptide is a receptor selected from the group consisting of a chimeric antigen receptor (CAR) and a T cell receptor (TCR).
73. The composition of any one of claims 1-72, wherein the target cell is a mammalian cell, a human cell, a primary cell, an ex vivo cell, or an in vivo cell.
74. The composition of any one of claims 1-73, wherein the target cell is an immune cell selected from the group consisting of a T cell, a B cell, a myeloid cell, a monocyte, a macrophage, and a dendritic cell.
75. The composition of any one of claims 1-74, wherein the composition further comprises one or more delivery vehicles for delivery of the RNA molecule, the endonuclease or the polynucleic acid sequence encoding the endonuclease, and the one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules into the target cell.

76. The composition of claim 75, wherein the one or more delivery vehicles is selected from the group consisting of: a nanoparticle delivery vehicle, a plasmid vector, and a viral vector.
77. The composition of claim 76, wherein the viral vector is an adenoviral vector, an adeno-associated viral vector, a lentiviral vector, or a retroviral vector.
78. The composition of claim 76, wherein the nanoparticle delivery vehicle is selected from the group consisting of a lipid nanoparticle and a polymeric nanoparticle.
79. The composition of any one of claims 1-78, wherein the 3' homology arm comprises a nucleic acid sequence of any of SEQ ID NOs: 49, 51, 53, 55, 57, 59, 61, or 63.
80. The composition of any one of claims 1-79, wherein the 5' homology arm comprises a nucleic acid sequence of any of SEQ ID NOs: 48, 50, 52, 54, 56, 58, 60, or 62.
81. The composition of any one of claims 1-80, wherein the RNA molecule of (a) comprises a nucleic acid sequence encoding a human ORF2p with D205A mutation, and wherein the nucleic acid sequence comprises a sequence with at least 90% sequence identity to SEQ ID NO: 129.
82. A pharmaceutical composition comprising (a) the composition of any one of claims 1-81; and (b) a pharmaceutically acceptable excipient.
83. A gene editing system for use in incorporating an exogenous sequence encoding an exogenous therapeutic polypeptide in a genome of a target cell, the system comprising a composition comprising:
- (a) an endonuclease or a polynucleic acid encoding the same;
 - (b) a first guide RNA sequence or a polynucleic acid encoding the same;
 - (c) a second guide RNA sequence or polynucleic acid encoding the same; and
 - (d) an RNA comprising a sequence encoding a retrotransposon machinery, the RNA comprising:
 - (i) a sequence that is a reverse complement of the sequence encoding the exogenous therapeutic polypeptide; and
 - (ii) a 5' homology arm and a 3' homology arm flanking the sequence encoding the exogenous therapeutic polypeptide, and wherein each homology arm is complementary to a sequence of a target site in the genome; and
 - (iii) a sequence encoding a human LINE1 ORF1p or functional fragment thereof; and

- (iv) a sequence encoding a human ORF2p comprising a reverse transcriptase (RT), wherein the endonuclease activity of ORF2p is inactivated.
84. The system of claim 83, wherein the endonuclease is a nickase, and wherein the nickase is (i) a Cas9 nickase, or (ii) a Cas 9 nickase comprising an H840A mutation relative to the wild type Cas9.
85. The system of claim 83 or 84, wherein the nucleic acid sequence flanked by the two homology arms does not comprise or overlap with the sequences encoding the endonuclease, the LINE1 ORF1p, or the ORF2p, or a fragment thereof.
86. The system of any one of claims 83-85, wherein the system further comprises (e) an RNA molecule comprising a sequence encoding a human ORF2p polypeptide that lacks a functional endonuclease domain, wherein the RNA molecule of (d) lacks sequences encoding a human ORF1p polypeptide or the exogenous sequence encoding a polypeptide, wherein the human ORF2p polypeptide that lacks a functional endonuclease domain comprises an exogenous DNA binding domain, wherein the exogenous DNA binding domain increases (i) DNA binding activity of the human ORF2p polypeptide or functional fragment thereof, (ii) TPRT processivity of the human ORF2p polypeptide or functional fragment thereof, or (iii) direct recruitment of the endonuclease to the target sequence in the genomic DNA of the target cell.
87. A method of incorporating an exogenous sequence encoding a therapeutic polypeptide at a specific site in a genome of a target mammalian cell, the method comprising contacting the target mammalian cell with the composition of any one of claims 1-81, wherein only the exogenous sequence encoding a therapeutic polypeptide is integrated into the genome of the target mammalian cell.
88. A method of incorporating an exogenous sequence encoding a therapeutic polypeptide at a specific site in a genome of a target mammalian cell, the method comprising:
- (a) contacting the target mammalian cell with a composition comprising:
 - (i) an endonuclease or a polynucleic acid encoding the same;
 - (ii) a first guide RNA sequence or a polynucleic acid encoding the same;
 - (iii) a second guide RNA sequence or a polynucleic acid sequence encoding the same; and
 - (iv) an RNA molecule comprising:
 - (A) an RNA sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide;

- (B) a 5' homology arm and a 3' homology arm; and
- (C) a human mobile genetic element comprising a sequence encoding a polypeptide that promotes integration of the exogenous sequence encoding the therapeutic polypeptide into the genome of the target mammalian cell via target primed reverse transcription (TPRT); and

(b) expressing the exogenous sequence encoding the therapeutic polypeptide from a genomically incorporated sequence of the target mammalian cell;

wherein, only the exogenous sequence encoding the therapeutic polypeptide is the genomically incorporated sequence that is incorporated at the specific site in the genome of the target mammalian cell.

- 89. The method of claim 88, wherein the nickase is a Cas9 nickase.
- 90. The method of claim 88 or 89, wherein the 5' homology arm and 3' homology arm flank the RNA sequence that is a reverse complement of an exogenous sequence encoding the therapeutic polypeptide.
- 91. The method of claim 90, wherein the nucleic acid sequence flanked by the two homology arms does not comprise or overlap with the sequences encoding the endonuclease or the human mobile genetic element, or a fragment thereof.
- 92. The method of any one of claims 88-91, wherein the human mobile genetic element comprises one or more elements of human LINE1.
- 93. The method of claim 92, wherein the one or more elements of human LINE1 comprise ORF1p or a functional fragment thereof.
- 94. The method of claim 93, wherein the sequence encoding a human ORF1p polypeptide and the sequence encoding a human ORF2p polypeptide are separated by (a) a sequence encoding a GSS linker and (b) a sequence encoding a T2A cleavage sequence.
- 95. The method of any one of claims 92-94, wherein the one or more elements of human LINE 1 comprise an ORF2p reverse transcriptase.
- 96. The method of any one of claims 92-95, wherein the one of more elements of human LINE 1 comprises an ORF2p endonuclease domain (EN), and wherein the endonuclease activity of ORF2p is inactivated.
- 97. The method of claim 96, wherein the endonuclease activity of ORF2p is inactivated by introduction of one or more mutations, selected from the group consisting of S228P, Y1180A and D205A.
- 98. The method of claim 96, wherein the endonuclease activity of ORF2p is inactivated by truncation.

99. The method of any one of claims 88-98, wherein the composition in step (a) further comprises (v) an RNA molecule comprising a sequence encoding a human ORF2p polypeptide that lacks a functional endonuclease domain, wherein the RNA molecule of (d) lacks sequences encoding a human ORF1p polypeptide or the exogenous sequence encoding a polypeptide.
100. The method of claim 99, wherein the human ORF2p polypeptide that lacks a functional endonuclease domain lacks an endonuclease domain.
101. The method of claim 99, wherein the human ORF2p polypeptide comprises an exogenous DNA binding domain.
102. The method of claim 101, wherein the exogenous DNA binding domain is derived from (i) a Sso7d protein, (ii) an HMGD protein, or (iii) an hPARP1 protein.
103. The method of claim 101 or 102, wherein the exogenous DNA binding domain is operably linked to the ORF2p via a linker.
104. The method of claim 103, wherein the linker is a 33XTEN linker.
105. The method of any one of claims 99-104, wherein a ratio of an amount of the RNA molecule of (iv) to that of the RNA molecule of (v) is about 5:1.
106. The method of any one of claims 89-105, wherein, the Cas9 nickase cuts the genomic DNA at a site that is within 1000, 800, 600, 400, 200, or 100 nucleotides from the specific site in the genome of a target mammalian cell.
107. The method of any one of claims 88-105, wherein the method prevents double stranded breaks on the genomic DNA.
108. The method of any one of claims 88-107, wherein the first guide RNA sequence comprises a sequence that has Watson-Crick pairing with a plurality of contiguous nucleotides on a first strand of the genomic DNA, and the second guide RNA sequence comprises a sequence that has Watson-Crick pairing with a plurality of contiguous nucleotides on the second and opposite strand of the genomic DNA.
109. The method of claim 108, wherein the first guide RNA sequence and the second guide RNA sequence do not comprise sequences that have Watson-Crick pairing with a plurality of contiguous nucleotides of the genomic DNA that are themselves complementary to each other.
110. The method of any one of claims 88-109, wherein contacting the target mammalian cell with a composition comprises delivering the composition into the target mammalian cell.
111. The method of any one of claims 88-110, wherein the composition is delivered to the target mammalian cell using a nanoparticle selected from the group consisting of a lipid nanoparticle and a polymeric nanoparticle.

112. The method of any one of claims 88-111, wherein the therapeutic polypeptide is selected from the group consisting of a ligand, an antibody, a receptor, an enzyme, a transport protein, a structural protein, a hormone, a contractile protein, a storage protein and a transcription factor.
113. The method of claim 112, wherein the therapeutic polypeptide is a receptor selected from the group consisting of a chimeric antigen receptor (CAR) and a T cell receptor (TCR).
114. The method of any one of claims 88-113, wherein the target mammalian cell is an immune cell selected from the group consisting of a T cell, a B cell, a myeloid cell, a monocyte, a macrophage, and a dendritic cell.
115. The method of any one of claims 88-114, wherein the human mobile genetic element comprising a sequence encoding a polypeptide that promotes integration of the exogenous sequence encoding the therapeutic polypeptide into the genome of the target mammalian cell via target primed reverse transcription (TPRT) comprises (i) a sequence encoding the ORF1p and (ii) a sequence encoding the ORF2p, wherein both (i) and (ii) are on the same polynucleic acid molecule.
116. The method of any one of claims 96-115, wherein the ORF2p endonuclease comprises a sequence having a D205A mutation.
117. The method of any one of claims 88-116, wherein contacting the target mammalian cell with (i) a first composition comprising the RNA molecule comprising the mRNA sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide; the 5' homology arm and the 3' homology arm; and the human mobile genetic element is at a time separate from contacting the target mammalian cell with (ii) a second composition comprising the first guide RNA sequence and the second guide RNA sequence or one or more polynucleic acid sequences encoding the same.
118. The method of claim 117, wherein contacting comprises contacting the target mammalian cell with the first composition and the second composition at an interval of at least 2 hours.
119. The method of claim 118, wherein the target mammalian cell is contacted with the second composition at 3, 4, 5, 6, 7 or 8 hours after contacting with the first composition.
120. The method of any one of claims 117-119, wherein the endonuclease is Cas9 nickase, wherein the Cas 9 nickase is contacted to the target mammalian cell at the same time as contacting with the first composition.
121. The method of any one of claims 117-119, wherein the target mammalian cell is contacted with a composition comprising Cas9 nickase or a polynucleic acid encoding the same (i) prior to contacting with a first composition comprising the RNA molecule

- comprising the mRNA sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide; the 5' homology arm and the 3' homology arm; and the human mobile genetic element; or (ii) prior to contacting with a second composition comprising the first guide RNA sequence and the second guide RNA sequence or one or more polynucleic acid sequences encoding the same; or (iii) after contacting with the second composition.
122. The method of any one of claims 88-121, wherein the target mammalian cell is contacted with an RNA encoding Cas9 nickase.
123. The method of any one of claims 88-122, wherein the orientation of the 5' homology arm and the 3' homology arm with respect to the reverse complement of the exogenous sequence encoding the therapeutic polypeptide is reversed for insertion of the exogenous sequence in the opposite orientation.
124. The method of any one of claims 88-123, wherein the sequence of the 5'-homology arm or the 3' homology arm between 10-80 nucleotides, between 20-70 nucleotides, between 30-60 nucleotide, between 40-60 nucleotides, between 15-20 nucleotides, less than 20 nucleotides, about 18 nucleotides, or about 17 nucleotides in length.
125. The method of any one of claims 88-123, wherein the sequence of the 5'-homology arm or the 3' homology arm is about 30 nucleotides in length.
126. The method of any one of claims 88-123, wherein the sequence of the 5'-homology arm or the 3' homology arm is about 50 nucleotides in length.
127. A single mRNA molecule comprising a genome editing system for incorporating an exogenous sequence encoding a therapeutic polypeptide in a genome of a target cell, the genomic editing system comprising:
- (a) a Cas endonuclease or a functional fragment thereof;
 - (b) a human LINE1 retrotransposon reverse transcriptase or a fragment thereof,
 - and
 - (c) the exogenous sequence encoding the therapeutic polypeptide flanked by one or more homology arms comprising non-ribosomal genomic sequence; or
 - (d) a sequence comprising reverse complement of (c).
128. The single mRNA molecule of claim 127, wherein the Cas endonuclease is a Cas 9 endonuclease, and the Cas endonuclease is a Cas 9 nickase.
129. The single mRNA molecule of claim 127 or 128, wherein the exogenous sequence encoding the therapeutic polypeptide is greater than 1kb, 1.2kb, 1.5kb, 1.7kb, 1.8 kb, 1.9kb, 2 kb, 2.1kb, 2.5kb, 2.7kb, 2.8 kb, 2.9kb, 3 kb, 3.1kb, 4kb, or 5 kb in length.

130. A composition comprising the single mRNA molecule of any one of claims 127-129 and a nucleic acid delivery vehicle, and wherein the nucleic acid delivery vehicle comprises one or more lipids.
131. The composition of claim 130, wherein the composition further comprises one or more guide RNAs.
132. The composition of any one of claims 130-131, wherein the composition further comprises one or more siRNAs.
133. The composition of any one of claims 130-132, wherein the composition further comprises one or more nuclear localization sequences (NLS), and wherein the NLS helps transport of one or more ORF polypeptides into the nucleus.
134. The composition of claim 133, wherein the NLS is SV40 NLS.
135. A pharmaceutical composition for treating a disease in a subject, comprising: (a) an RNA comprising a sequence encoding Cas9 nickase; (b) an RNA comprising a sequence that is a reverse complement of the exogenous sequence encoding the therapeutic polypeptide; a 5' homology arm and a 3' homology arm, wherein the 5' homology arm or the 3' homology arm is in reverse complement orientation with respect to a sequence on one strand of the genomic DNA; and a sequence encoding a human mobile genetic element; (c) a composition comprising a first guide RNA sequence and a second guide RNA sequence or one or more polynucleic acid sequences encoding the same; and a therapeutically acceptable excipient, wherein each of the compositions of (a), (b) and (c) are comprised in a separate delivery vehicle.
136. A composition comprising:
- (a) an RNA molecule comprising:
 - (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide;
 - (ii) a 5' homology arm and a 3' homology arm; and
 - (iii) a sequence encoding a modified human LINE1 sequence or fragment thereof, wherein the sequence encoding the modified human LINE1

sequence or fragment thereof is upstream of the 5' homology arm or downstream of the 3' homology arm;

- (b) an RNA molecule encoding an endonuclease, wherein the endonuclease is a nickase;
 - (c) an RNA molecule comprising a sequence encoding a human ORF2p polypeptide comprising an endonuclease wherein the endonuclease is functionally deficient; and
 - (d) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.
137. The composition of claim 136, wherein the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF2p polypeptide, wherein the human ORF2p polypeptide comprises an endonuclease and a reverse transcriptase, wherein the endonuclease comprises a mutation.
138. The composition of claim 136 or 137, wherein the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF1p polypeptide upstream of the sequence encoding the human ORF2p polypeptide, and is separated by a LINE1 interORF sequence.
139. The composition of claim 136 or 137, wherein the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF1p polypeptide upstream of the sequence encoding the human ORF2p polypeptide, and is separated by an IRES sequence or a cleavage site.
140. The composition of claim 136 or 137, wherein the sequence encoding the modified human LINE1 sequence or fragment thereof comprises a sequence encoding a human ORF1p polypeptide upstream of the sequence encoding the human ORF2p polypeptide, and is separated by (a) a sequence encoding a GSS linker and (b) a T2A cleavage sequence.
141. The composition of any one of claims 136-140, wherein the sequence encoding the modified human LINE1 sequence as well as the sequence encoding the human ORF2p

- polypeptide comprise a mutation in ORF2p endonuclease domains, wherein the mutation in the ORF2p endonuclease domains render the endonucleases functionally deficient.
142. The composition of claim 141, wherein none of the ORF2p endonucleases have endonuclease activity.
143. The composition of any one of claims 136-142, wherein the human ORF2p polypeptide lacks an endonuclease domain, and wherein the human ORF2p polypeptide comprises an exogenous DNA binding domain.
144. The composition of claim of claim 143, wherein the exogenous DNA binding domain is derived from (i) a Sso7d protein, (ii) an HMGD protein, or (iii) an hPARP1 protein.
145. The composition of any one of claims 136-144, wherein the RNA molecule comprising a sequence encoding a human ORF2p polypeptide comprises a 5'-methylated guanosyl cap structure (m7G cap).
146. The composition of any one of claims 136-145, wherein the RNA molecule comprising a sequence encoding a human ORF2p polypeptide is translated via a canonical cap-dependent mechanism.
147. The composition of any one of claims 136-146, wherein the sequence encoding a human ORF2p polypeptide in the RNA molecule (c) is located at the 5' end of the RNA molecule (c).
148. The composition of any one of claims 136-147, wherein the sequence encoding the human ORF2p polypeptide in the modified human LINE1 sequence comprises an N-terminal NLS and a C-terminal NLS.
149. The composition of any one of claims 136-148, wherein the nickase is a functional nickase, and wherein the nickase comprises an N-terminal NLS, a C-terminal NLS or both N-terminal and C-terminal NLS.
150. The composition of any one of claims 136-149, wherein the sequence encoding the modified human LINE1 sequence or fragment thereof comprises one or more STOP codons at the end of the sequence encoding the human ORF2p polypeptide, the sequence encoding the human ORF1p polypeptide or both.
151. The composition of any one of claims 136-150, wherein any of the RNA molecules comprise one or more ribosomal entry sites, one or more cleavage sites, or one or more oligomerization domains.
152. The composition of any one of claims 136-151, wherein each of the RNA molecules comprise a poly A sequence at the 3' end.
153. The composition of any one of claims 136-152, wherein the RNA molecule of (c) further comprises a reverse complement sequence of an insert sequence, wherein the reverse

complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide.

154. The composition of claim 153, wherein the RNA molecule of (c) further comprises a 5' homology arm and a 3' homology arm.
155. The composition of any one of claims 136-154, wherein the RNA molecule of (c) does not comprise a sequence encoding ORF1p.
156. The composition of any one of claims 136-155, wherein the RNA molecule of (c) does not comprise a reverse complement sequence of an insert sequence.
157. The composition of any one of claims 136-156, wherein the RNA molecule of (c) consists of the sequence encoding the human ORF2p polypeptide with at least one NLS.
158. The composition of any one of claims 136-157, wherein the promoter comprises a EF1alpha promoter or a functional variant.
159. The composition of claim 158, wherein the EF1alpha promoter is a full-length EF1alpha promoter.
160. The composition of claim 159, wherein the EF1alpha promoter is a short EF1alpha promoter.
161. A composition comprising:
 - (a) an RNA molecule comprising:
 - (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide;
 - (ii) a 5' homology arm and a 3' homology arm; and
 - (iii) a sequence encoding a human OFR2p polypeptide comprising an endonuclease wherein the endonuclease is functionally deficient;
 - (b) an RNA molecule encoding an endonuclease, wherein the endonuclease is a nickase;
 - (c) an RNA molecule comprising a sequence encoding a human ORF1p, wherein the RNA molecule does not comprise a sequence encoding a human ORF2p;

- (d) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell.
162. The composition of claim 161, wherein the RNA molecule of (c) consists of a sequence encoding the human ORF1p.
163. A method for incorporating an exogenous sequence encoding a therapeutic polypeptide at a specific site in a genome of a target mammalian cell, the method comprising contacting the mammalian target cell with the composition of any one of claims 136-162, wherein only the exogenous sequence encoding a therapeutic polypeptide is integrated into the genome of the target mammalian cell.
164. The method of claim 163, wherein the integration efficiency is increased at least by 0.1%, 0.2%, or 0.5% compared to a system where the sequence that is a reverse complement of an exogenous sequence encoding the polypeptide is not operably linked to the sequence that is a reverse complement of the promoter sequence or does not comprise a promoter.
165. A composition comprising:
- (a) an RNA molecule comprising:
 - (i) a reverse complement sequence of an insert sequence, wherein the reverse complement sequence comprises (A) a sequence that is a reverse complement of an exogenous sequence encoding a polypeptide operably linked to (B) a sequence that is a reverse complement of a promoter sequence, wherein the sequence that is a reverse complement of a promoter sequence is downstream of the sequence that is a reverse complement of an exogenous sequence encoding a polypeptide;
 - (ii) a 5' homology arm and a 3' homology arm; and
 - (iii) a sequence encoding a human ORF1p polypeptide and a human OFR2p polypeptide, the human ORF2p polypeptide comprising an endonuclease wherein the endonuclease is functionally deficient;
 - (b) an RNA molecule encoding an endonuclease, wherein the endonuclease is a nickase;

- (c) an RNA molecule comprising a sequence encoding a human ORF2p, wherein the RNA molecule does not comprise a sequence encoding a human ORF1p or the reverse complement sequence of an insert sequence;
 - (d) one or more guide RNA molecules or one or more polynucleic acids encoding the one or more guide RNA molecules, wherein the one or more guide RNA molecules comprise a first guide RNA molecule and a second guide RNA molecule, wherein the first guide RNA molecule comprises a sequence capable of hybridizing to a first target sequence of genomic DNA of a target cell and the second guide RNA molecule comprises a sequence capable of hybridizing to a second target sequence of the genomic DNA of the target cell, wherein the integration efficiency of the insert sequence is increased by at least 0.05% compared to a composition that lacks (c).
166. A method of treating a disease in a subject, comprising administering to the subject (a) the composition of any one of claims 1-81, 130-134, 136-162, and 165 or (b) the pharmaceutical composition of claim 82 or 135.
167. The method of claim 166, wherein the disease is cancer.