

## **R2 RETROTRANSPOSONS FOR GENE WRITING**

### **CROSS REFERENCE TO RELATED APPLICATIONS**

**[0001]** This application claims the benefit under 35 U.S.C. § 119(e) of the United States Provisional Application Serial No. 63/590,751 filed October 16, 2023, the content of which is hereby incorporated by reference in its entirety.

### **REFERENCE TO AN ELECTRONIC SEQUENCE LISTING**

**[0002]** The content of the electronic sequence listing (350720.xml; Size: 12,240 bytes; and Date of Creation: October 14, 2024) is herein incorporated by reference in its entirety.

### **BACKGROUND**

**[0003]** Unlike DNA transposons, which move via a “cut and paste” mechanism, retrotransposons move in a “copy and paste” manner, using RNA as an intermediate. Based on whether they are flanked by long terminal repeat (LTR), retrotransposons are classified as LTR type or non-LTR type. Non-LTR retrotransposons are sub-classified into two major groups, namely long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). LINEs are autonomous elements that encode proteins to mediate their own mobility, whereas SINEs are nonautonomous elements that do not encode protein and consequently require LINEs for their propagation.

**[0004]** Non-LTR retrotransposons are mobilized by a mechanism in which proteins translated from their open reading frame proteins combine with their own mRNA to form a ribonucleoprotein (RNP) complex. The RNP complex is inserted into the target site via a mechanism termed target-primed reverse transcription (TPRT). The TPRT process is initiated by an encoded EN domain that nicks one strand of DNA at a target site and creates a 3'-hydroxyl end, which is used as a primer for reverse transcription of the LINE mRNA onto the DNA target via RT.

**[0005]** R2 is a long interspersed element (LINE) found in a specific sequence of the 28S rDNA among a wide variety of animals. Known R2 retrotransposons include those from medaka fish *Oryzias latipes* (R2OI) and from *Bombyx mori* (R2Bm). Efforts have been made to use these R2 retrotransposons for targeted gene knock-in for practical application such as gene therapy.

## SUMMARY

[0006] Through bioinformatic approaches with confirmative testing in human cells, the instant inventors discovered highly efficient new R2 retrotransposons from bird species. Two of them, from *Corvus moneduloides* (New Caledonian crow) and *Accipiter gentilis* (Eurasian goshawk) and thus referred to as R2Cm and R2Ag, respectively, exhibited excellent retrotransposition activities in human cells. These new retrotransposons, as well as engineered versions thereof, can be suitably used for targeted gene delivery.

[0007] Accordingly, one embodiment of the present disclosure provides a method for inserting an exogenous sequence to a genomic DNA of a cell, comprising introducing to the cell (a) a polypeptide or a first polynucleotide encoding the polypeptide, wherein the polypeptide comprises (i) a DNA binding domain or a CRISPR Cas protein, (ii) a reverse-transcriptase comprising the amino acid sequence of SEQ ID NO:2 or 6, or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:2 or 6, and (iii) an endonuclease or a nickase, and (b) an RNA comprising (i) a template for the exogenous sequence and (ii) a 3' UTR and/or a 3' UTR, or a second polynucleotide encoding the RNA.

[0008] In some embodiments, the method comprises introducing the first polynucleotide and the second polynucleotide to the cell. In some embodiments the first polynucleotide and the second polynucleotide are on a same vector or on separate vectors.

[0009] In some embodiments the 3'UTR is an R2 retrotransposon 3' UTR. In some embodiments the 3'UTR comprise the sequence of SEQ ID NO:3 or 7. In some embodiments the reverse-transcriptase comprises the amino acid sequence of SEQ ID NO:2 or 6.

[0010] In some embodiments, the reverse-transcriptase comprises the amino acid sequence of SEQ ID NO:6 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:6. In some embodiments, the 3'UTR comprise nucleotides 302-371 of SEQ ID NO:7, or a nucleic acid sequence having at least 85% sequence identity to nucleotides 302-371 of SEQ ID NO:7. In some embodiments, the 3'UTR comprise SEQ ID NO:7, or a nucleic acid sequence having at least 85% sequence identity SEQ ID NO:7. In some embodiments, the RNA of (b) does not include a portion, such as at least 50 nucleotides, among nucleotides 1-301 of SEQ ID NO:7 upstream of nucleotides 302-371 of SEQ ID NO:7.

**[0011]** In some embodiments, the reverse-transcriptase comprises the amino acid sequence of SEQ ID NO:2 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:2. In some embodiments, the 3'UTR comprise nucleotides 194-264 of SEQ ID NO:3, or a nucleic acid sequence having at least 85% sequence identity to nucleotides 194-264 of SEQ ID NO:3. In some embodiments, the 3'UTR comprise SEQ ID NO:3, or a nucleic acid sequence having at least 85% sequence identity SEQ ID NO:3. In some embodiments, the RNA of (b) does not include at least 50 nucleotides among nucleotides 1-193 of SEQ ID NO:3 upstream of nucleotides 194-264 of SEQ ID NO:3.

**[0012]** In some embodiments the polypeptide comprises the endonuclease. In some embodiments the endonuclease comprises amino acid residues 1208-1364 of SEQ ID NO:1 or amino acid residues 1175-1261 of SEQ ID NO:5 or an amino acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to amino acid residues 1208-1364 of SEQ ID NO:1 or amino acid residues 1175-1261 of SEQ ID NO:5.

**[0013]** In some embodiments the polypeptide comprises the nickase, which optionally is Cas nickase. In some embodiments the CRISPR Cas protein is selected from the group consisting of Cas9, Cas12 and Cas13.

**[0014]** In some embodiments the DNA binding domain is selected from the group consisting of a transcription activator-like effector (TALE) DNA binding domain, a homeodomain protein, a zinc finger, a helix-turn-helix, a leucine zipper, and a DNA-binding domain of a homing endonuclease, a transposon or retro-transposon.

**[0015]** In some embodiments the DNA binding domain is a TALE. In some embodiments the homing endonuclease is selected from the group consisting of I-SceI, I-CreI and I-DmoI.

**[0016]** In some embodiments the retro-transposon is a long interspersed nuclear element (LINE), optionally selected from the group consisting of R1, R2, R4, R5, R6, R7, R8, and R9.

**[0017]** In some embodiments the polypeptide comprises the amino acid sequence of SEQ ID NO:1 or 5, or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:1 or 5. In some embodiments the polypeptide comprises the amino acid sequence of SEQ ID NO:1 or 5.

**[0018]** In some embodiments the RNA further comprises a 5'UTR. In some embodiments the 5'UTR comprises the sequence of SEQ ID NO:4 or 8.

**[0019]** In some embodiments, the RNA further comprises one or two flanking sequences that flank a target site on a target genome at which the exogenous sequence is expected to be inserted. In some embodiments, the target site is a site within a 28S rDNA gene.

**[0020]** In some embodiments the cell is a mammalian cell, preferably a human cell. In some embodiments the exogenous sequence encodes a protein, preferably a therapeutic protein.

**[0021]** Also provided, in one embodiment, is a chimeric polypeptide, comprising (i) a DNA binding domain or a CRISPR Cas protein, (ii) a reverse-transcriptase comprising the amino acid sequence of SEQ ID NO:2 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:2, and (iii) an endonuclease or a nickase, wherein the chimeric protein does not include any one of (a) amino acid residues 205-226 of SEQ ID NO:1, (b) amino acid residues 232-255 of SEQ ID NO:1, (c) amino acid residues 269-290 of SEQ ID NO:1, and (d) amino acid residues 317-357 of SEQ ID NO:1, wherein at least (i) or (ii) is heterologous to SEQ ID NO:1.

**[0022]** Also provided is a chimeric polypeptide, comprising (i) a DNA binding domain or a CRISPR Cas protein, (ii) a reverse-transcriptase comprising the amino acid sequence of SEQ ID NO:6 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:6, and (iii) an endonuclease or a nickase, wherein the chimeric protein does not include any one of (a) amino acid residues 145-166 of SEQ ID NO:1, (b) amino acid residues 175-195 of SEQ ID NO:1, (c) amino acid residues 209-230 of SEQ ID NO:1, and (d) amino acid residues 257-297 of SEQ ID NO:5, wherein at least (i) or (ii) is heterologous to SEQ ID NO:5.

**[0023]** Further provided is an RNA or a DNA vector encoding the RNA, wherein the RNA comprises: (a) a first fragment encoding the sequence of SEQ ID NO:5 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:5; (b) a 3'UTR comprising nucleotides 302-371 of SEQ ID NO:7 or a nucleic acid sequence having at least 85% sequence identity to 302-371 of SEQ ID NO:7; and (c) a second fragment disposed between (a) and (b), which encodes a heterologous protein.

**[0024]** In some embodiments, the RNA does not include at least 50 nucleotides among nucleotides 1-301 of SEQ ID NO:7 upstream of nucleotides 302-371 of SEQ ID NO:7. In

some embodiments, the RNA does not include at least 50 nucleotides of SEQ ID NO:8 upstream of the first fragment.

**[0025]** Also provided is an RNA or a DNA vector encoding the RNA, wherein the RNA comprises: (a) a first fragment encoding the sequence of SEQ ID NO:1 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:1; (b) a 3'UTR comprising nucleotides 194-264 of SEQ ID NO:3 or a nucleic acid sequence having at least 85% sequence identity to 194-264 of SEQ ID NO:3; and (c) a second fragment disposed between (a) and (b), which encodes a heterologous protein.

**[0026]** In some embodiments, the RNA does not include at least 50 nucleotides among nucleotides 1-193 of SEQ ID NO:3 upstream of nucleotides 194-264 of SEQ ID NO:3. In some embodiments, the RNA does not include at least 50 nucleotides of SEQ ID NO:4 upstream of the first fragment.

**[0027]** In some embodiments, the RNA further comprises one or two flanking sequences that flank a target site on a target genome at which the heterologous sequence is expected to be inserted. In some embodiments, the target site is a site within a 28S rDNA gene.

**[0028]** Still further, in one embodiment, is an RNA or a DNA vector encoding the RNA, wherein the RNA comprises: (a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:5 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:5; and (b) a 3'UTR comprising nucleotides 302-371 of SEQ ID NO:7 or a nucleic acid sequence having at least 85% sequence identity to 302-371 of SEQ ID NO:7, wherein the RNA does not include: (1) at least 50 nucleotides among nucleotides 1-301 of SEQ ID NO:7 upstream of nucleotides 302-371 of SEQ ID NO:7; (2) at least 50 nucleotides of SEQ ID NO:8 upstream of the ORF; or (3) either or both sequences upstream of the ORF and downstream of the 3'UTR complementary to a 28S rDNA gene.

**[0029]** Also provided is an RNA or a DNA vector encoding the RNA, wherein the RNA comprises: (a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:1 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:1; and (b) a 3'UTR comprising nucleotides 194-264 of SEQ ID NO:3 or a nucleic acid sequence having at least 85% sequence identity to 194-264 of SEQ ID NO:3, wherein the RNA does not include: (1) at least 50 nucleotides among nucleotides 1-193 of SEQ ID NO:3 upstream of nucleotides 194-264 of SEQ ID NO:3; (2) at least 50 nucleotides of SEQ ID NO:4 upstream

of the ORF; or (3) either or both sequences upstream of the ORF and downstream of the 3'UTR complementary to a 28S rDNA gene.

[0030] Still further provided is a method for inserting an exogenous sequence to a genomic DNA of a cell, comprising introducing to the cell (a) the polypeptide of any of the above embodiments or a first polynucleotide encoding the polypeptide, and (b) an RNA comprising (i) a template for the exogenous sequence and (ii) a 5' UTR and/or a 3' UTR, or a second polynucleotide encoding the RNA.

### BRIEF DESCRIPTION OF THE DRAWINGS

[0031] FIG. 1 shows that the newly discovered R2Ag had higher transfection efficiency to deliver a GFP cargo to target HEK293T cells than a published R2 retrotransposon system.

[0032] FIG. 2 shows that the newly discovered R2Ag had higher transfection efficiency to deliver a GFP cargo to human primary T cells than a published R2 retrotransposon system.

[0033] FIG. 3 shows R2Ag-mediated gene delivery efficiencies to primary human T cells and hepatocytes.

[0034] FIG. 4 illustrates a *trans* and a *cis* configuration for R2 retrotransposon systems for delivering a transgene into a target cell.

[0035] FIG. 5 shows that, in the *cis* configuration, both forward and reverse orientations of the transgene achieved efficiency gene delivery to target HEK293T cells.

[0036] FIG. 6 shows that a majority of the 3'UTR sequence, except the C-terminal 70 nt core, was not required for the function of the R2Ag.

[0037] FIG. 7 shows that R2Ag can highly efficiently integrate full-length cargos as confirmed by long-read sequencing.

[0038] FIG. 8 shows that R2Ag had a >99.5% on-target site-specific integration profile in HEK cells.

[0039] FIG. 9 shows that R2Ag had greatly higher integration efficiency than reference R2OI.

## DETAILED DESCRIPTION

[0040] The following description sets forth exemplary embodiments of the present technology. It should be recognized, however, that such description is not intended as a limitation on the scope of the present disclosure but is instead provided as a description of exemplary embodiments.

### Newly Discovered R2 Retrotransposons

[0041] Through bioinformatic approaches with confirmative testing in human cells, the instant inventors discovered highly efficient new R2 retrotransposons from bird species. Two of them were further investigated, R2Cm from *Corvus moneduloides* (New Caledonian crow) and R2Ag from *Accipiter gentilis* (Eurasian goshawk). Both exhibited excellent retrotransposition activities in human cells, higher than known R2 retrotransposons.

[0042] Each of the R2Cm and R2Ag mRNA includes a 5' untranslated region (UTR), a single open reading frame (ORF) encoding the endonuclease (EN) and reverse-transcriptase (RT) domains, and a 3' UTR (**Table 1**). The R2 mRNA can anneal with the target site (*i.e.*, the 28S rDNA in the human genome) and initiate the reverse transcription mechanism.

[0043] More specifically, each of the R2Cm and R2Ag ORFs includes, from the N- to the C-terminus, three N-terminal zinc finger domains (N-ZnF1, N-ZnF2 and N-ZnF3), a SANT DNA-binding domain, a 3' UTR binding domain, a template switching domain, the reverse-transcriptase (RT), a C-terminal zinc finger domain (C-ZnF), and the endonuclease (EN).

[0044] With reference to SEQ ID NO:1, the R2Cm ORF includes N-ZnF1 (residues 205-226), N-ZnF2 (residues 232-255), N-ZnF3 (residues 269-290), the SANT DNA-binding domain (317-357), the 3' UTR binding domain (residues 528-546), the template switching domain (residues 601-641), the reverse-transcriptase (residues 662-928), C-ZnF (residues 1186-1203), and the endonuclease (residues 1208-1364). The sequence of the R2Cm reverse-transcriptase is also provided in SEQ ID NO:2.

[0045] With reference to SEQ ID NO:5, the R2Ag ORF includes N-ZnF1 (residues 145-166), N-ZnF2 (residues 175-195), N-ZnF3 (residues 209-230), the SANT DNA-binding domain (257-297), the 3' UTR binding domain (residues 425-452), the template switching domain (residues 497-537), the reverse-transcriptase (residues 558-824), C-ZnF (residues

1083-1100), and the endonuclease (residues 1105-1261). The sequence of the R2Ag reverse-transcriptase is also provided in SEQ ID NO:6.

**[0046]** As demonstrated in **Example 4**, the entire 5'UTR was not required for efficient transfection or integration of the R2 retrotransposons and thus can be truncated or replaced by alternative 5'UTR sequence of fragments thereof. More interestingly, within the R2Ag 3'UTR, only the last 70nt portion (referred to as "70-nt core", nucleotides 302-371 of SEQ ID NO:7) was required. Sequence alignment indicates that the R2Cm 3'UTR has a highly homologous 71-nt core (nucleotides 194-264 of SEQ ID NO:3; sequence identity to the 70-nt core of about 83%, with 59 nt being identical out of 70/71 nt). In other words, a highly active R2Cm or R2Ag can function without the 5'UTR or the 5' majority of the 3'UTR.

### **Engineered R2Cm and R2Ag Retrotransposons**

**[0047]** The retrotransposition of R2 retrotransposons does not require homology-directed repair (HDR) or donor DNA. Compared to HDR, which is a passive process in cells, reverse transcription of mRNA into cDNA via TPRT using the RT domain shows relatively higher processivity and fidelity, which is advantageous for synthesizing long cDNAs at the target site rather than recombination at the target site as donor DNA.

**[0048]** It has been demonstrated herein that the R2Cm and R2Ag retrotransposons specifically target the 28S gene in the human genome. Based on sequence analysis, it is contemplated that the specific 28S binding is attributed to the three N-terminal ZnF domains, likely assisted by the SANT DNA-binding domain. Replacement of these N-terminal regions of the R2 proteins, therefore, may alternate the target specificity of the R2Cm and R2Ag retrotransposons.

**[0049]** Accordingly, in some embodiments, the present disclosure provides an engineered R2Cm, in the form of a fusion protein, that includes a truncated version of the R2Cm ORF and one or more other heterologous elements. The truncated R2Cm ORF, in some embodiments, includes the reverse-transcriptase (or a biological variant thereof) of R2Cm, but does not include portion of the N-terminal region of R2Cm, such as one or more of the N-Znf1, N-Znf2, N-Znf3, and SANT DNA-Binding Domain.

**[0050]** In some embodiments, the heterologous elements may be a DNA binding domain or a CRISPR Cas protein and/or an endonuclease or a nickase. It is appreciated that it is not

necessary that both elements are heterologous. For instance, the fusion protein can include one or more heterologous DNA binding domains but the native endonuclease. In another example, the fusion protein can include the native DNA binding domains and a heterologous endonuclease.

**[0051]** Likewise, another embodiment of the present disclosure provides an engineered R2Ag, in the form of a fusion protein, that includes a truncated version of the R2Ag ORF and one or more other heterologous elements. The truncated R2Ag ORF, in some embodiments, includes the reverse-transcriptase (or a biological variant thereof) of R2Cm, but does not include portion of the N-terminal region of R2Ag, such as one or more of the N-Znf1, N-Znf2, N-Znf3, and SANT DNA-Binding Domain.

**[0052]** As used herein, a “biological variant” of a reference protein refers to a protein or fragment that has at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98% or 99% sequence identity to the reference protein. In some embodiments, the biological variant retains the desired activity (*e.g.*, reverse-transcriptase) of the reference protein.

#### ***A. Truncated R2Cm or R2Ag ORF***

**[0053]** As provided, an engineered R2Cm or R2Ag protein can include a truncated version of the R2 ORF fused to at least one heterologous element.

**[0054]** In one embodiment, the truncated R2Cm ORF includes a reverse-transcriptase comprising the amino acid sequence of SEQ ID NO:2 or an amino acid sequence having at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98% or 99% sequence identity to SEQ ID NO:2. In some embodiments, the truncated R2Cm ORF further includes the 3' UTR binding domain (residues 528-546 of SEQ ID NO:1). In some embodiments, the truncated R2Cm ORF further includes the template switching domain (residues 601-641 of SEQ ID NO:1). In some embodiments, the truncated R2Cm ORF further includes the C-ZnF (residues 1186-1203 of SEQ ID NO:1). In some embodiments, the truncated R2Cm ORF further includes the endonuclease (residues 1208-1364 of SEQ ID NO:1). In some embodiments, the truncated R2Cm ORF includes all elements of the wild-type R2Cm ORF, except those that are specifically substituted with a corresponding heterologous element (*e.g.*, a heterologous endonuclease).

[0055] In one embodiment, the truncated R2Ag ORF includes a reverse-transcriptase comprising the amino acid sequence of SEQ ID NO:6 or an amino acid sequence having at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98% or 99% sequence identity to SEQ ID NO:6. In some embodiments, the truncated R2Ag ORF further includes the 3' UTR binding domain (residues 425-452 of SEQ ID NO:5). In some embodiments, the truncated R2Ag ORF further includes the template switching domain (residues 497-537 of SEQ ID NO:5). In some embodiments, the truncated R2Ag ORF further includes the C-ZnF (residues 1083-1100 of SEQ ID NO:5). In some embodiments, the truncated R2Ag ORF further includes the endonuclease (residues 1105-1261 of SEQ ID NO:5). In some embodiments, the truncated R2Ag ORF includes all elements of the wild-type R2Ag ORF, except those that are specifically substituted with a corresponding heterologous element (*e.g.*, a heterologous endonuclease).

### ***B. Heterologous Targeting Domain***

[0056] In some embodiments, one or more of the N-terminal ZnF domains is replaced with a heterologous targeting domain (*e.g.*, a DNA-binding domain or a CRISPR Cas protein).

[0057] A “DNA-binding domain” refers to a protein domain capable of binding a nucleic acid sequence of interest, preferably a double strand nucleic acid molecule. The DNA binding domain recognizes and binds nucleic acid at specific polynucleotide sequences, further referred to as “nucleic acid target sequence.” The DNA-binding domain comprises any of the DNA-binding domain known in the art, such as transcription activator-like effector (TALE), homeodomain protein, a zinc finger, a helix-turn-helix, a leucine zipper, and a DNA-binding domain of a homing endonuclease, a transposon or retro-transposon. In certain embodiments, the DNA-binding domain recognizes a target DNA sequence in the genome.

[0058] A “homeodomain protein” consists of three linked alpha helices (helices 1, 2 and 3). Helices 2 and 3 are arranged in a conspicuous helix turn helix motif. A 60 amino acid long region (homeodomain) within helix 3 binds specifically to DNA segments that contain the sequence 5'ATTA3'. In animals, there are 16 major classes of homeodomain protein, ANTP, PRD, PRD-LIKE, POU, HNF, CUT (with four subclasses: ONECUT, CUX, SATB, and CMP), LIM, ZF, CERS, PROS, SIX/SO, plus the TALE superclass with the classes IRO, MKX, TGIF, PBC, and MEIS. Further homeodomain proteins include but not limited to HEX motif, EH1 motif, Octapeptide/Hep/EH1/TN/GEH motif, WRPW motif, TUP1, OAR motif,

CUT domain, such as ONECUT, CUX, SATB, CMP domains, COMPASS domain, HNF1A (LFB1) transcription factor, POU, OCT-1, OCT-2, Pit 1, LIM domain, PROSPERO (PROS), SIX/SO and CERS (LASS).

**[0059]** The “zinc finger” (ZF) includes at least nine types, C2H2, C2HC, C3H, C4, C6, C3HC4, C2HC5, C4HC3, and C8, in which C and H represent cysteine and histidine, respectively. C2H2 zinc finger is a loop of 12 amino acids with two cysteines and two histidines at the base of the loop (Cys2His2 zinc finger motif consisting of an  $\alpha$  helix and an antiparallel  $\beta$  sheet) that tetrahedrally coordinate a zinc ion. C2H2 zinc finger is represented by TFIIIA and further comprises Adnp, Tshz, Zeb, Zfhx and Zhx. C2HC-type zinc finger domains, also referred to as retroviral-type (RT) zinc finger sequences, are required for viral genome packaging and RNA or single-stranded DNA binding in eukaryotes. The C3H family proteins are divided into 18 groups based on the different amino acid spacing numbers between C and H in zinc finger motif. C4 finger has one alpha helix and contains a zinc atom bound to four cysteine amino acids. The C4 zinc finger contains a 70 amino acid long region near the zinc atom that binds specifically to DNA segments. The C4 family includes seven types: GATA, FYVE, Tim10/DDP, LSD1, A20, TFIIIB, and Zn-finger in Ran-binding protein. C3HC4-type finger, also termed RING finger, can be categorized into seven types with different conserved motifs, such as RING-H2, RING-HC, RING-v, RING-D, RING-S/T, RING-G and RING-C2. The C2HC5 motif, also referred to as LIM. More information on the zinc finger can be found in the prior art publication, such as Li et al., International Journal of Molecular Sciences, (2020) 21(4):1361. Further examples of the zinc finger proteins include Sp1, Glucocorticoid receptor, estrogen receptor, progesterone receptor, thyroid hormone receptor (erbA), retinoid acid receptor, and the vitamin D3 receptor.

**[0060]** A “leucine zipper” consists of an alpha helix that contains a region in which every seventh amino acid is leucine, which has the effect of lining up all the leucine residues on one side of the alpha helix. The leucine residues allow for the dimerization of the two leucine zipper proteins and formation of Y shaped dimer. Dimerization may occur between two of the same proteins (homodimers, e.g., Jun-Jun) or two different proteins (heterodimers, e.g., Fos-Jun). A leucine zipper contains a 20 amino acid long region that binds specifically to DNA segments. Examples of the Leucine Zipper proteins include but not limited to CCAAT/enhancer binding protein (C/EBP), Cyclic AMP response element binding protein (CREB), Finkel osteogenic sarcoma virus (Fos) protein, Jun, GCN4, and HSF.

**[0061]** A “helix loop helix” (HLH) or “helix-turn-helix” consists of a short alpha helix connected by a loop to a longer alpha helix. The loop allows for dimerization of two HLH proteins and formation of Y shaped dimer. The dimerization may occur between two of the same proteins (homo dimers) or two different proteins (heterodimers). Example of HLH domains include but not limited to MyoD, Myc, Pho4, SREBP-1a, and Max-Mad.

**[0062]** A “Transcription Activator like Effector” (TALE) are proteins that are encoded by phytopathogenic bacteria of the genus *Xanthomonas* and *Ralstonia* to influence the gene expression of host plant cells during bacterial infection. These proteins are helix-loop-helix-containing transcriptional factors that comprise a DNA binding region and an N-terminal domain that appears to interact with the bacterial transport machinery for introducing the protein into the plant cell. The C-terminal domain of the TALE protein seems to interact with the plant host's transcriptional machinery to induce expression of sets of plant genes that are beneficial to the invading bacteria. The DNA binding portion of the proteins is found in the middle section of the protein and is made of an array of repeat units, each approximately 33-35 amino acids in length, which have been shown to be responsible for interacting with the target DNA. In a preferred embodiment, said DNA-binding domain is derived from a TALE engineered to bind a specific nucleic acid target sequence. Unlimited examples of the TALE include KNOX and BEL, such as PBC, MEIS, PREP, IRO, MKX and TGIF.

**[0063]** A “homing endonuclease (HEs)” or meganucleases (MNs), are sequence-specific endonucleases with large cleavage sites (14–25 bp) that can create double-stranded breaks at specific locations. The homing endonucleases are encoded by mobile genetic elements that induce recombination in a process called homing. The homing endonucleases genes (HEGs) that have invaded many genomic niches including group I and group II introns. Some HEGs can move from an ORF-containing intron to an “ORF-less” intron. There are four major families of homing endonucleases (HEs), naming is based on conserved amino acid motifs: the H-N-H, HIS-CYS, LAGLIDADG, and GIY-YIG families of HEs. The LAGLIDADG family of HEs is the most frequently encountered among group I introns. However GIY-YIG endonucleases have been identified within numerous group I introns, and LAGLIDADG HEs and the H-N-H domain are present within the ORFs of some group II introns. Additional HE-like proteins have been described, the PD-(D/E)XK HEs are found in bacterial tRNA group I introns, the very-short patch repair (Vsr) endonucleases (a predicted family of phage HEs based on metagenomic), and the Holliday junction resolvase-like HEs found in some phage

introns. In certain embodiments of the present disclosure, the homing endonuclease comprises PI-SceI, PI-PfuI, I-CreI, I-CeuI, I-DmoI, F-TevI, F-TevII, I-TevI, I-TevII, I-PpoI, I-DirI, I-NjaI, I-NanI, I-NitI, I-SceV, I-SceVI and I-LlaI, I-HmuI, I-HmuII, I-TevIII and I-CmoI. In a particular embodiment, the homing endonuclease is I-PpoI (Intron-Encoded Endonuclease).

**[0064]** Retrotransposon is a class of eucaryotic genes capable of replicating to new locations within their own genome through an RNA intermediate. Retrotransposons are widespread in metazoan genomes. The proportion of the genome that contains retrotransposons often exceeds that containing DNA transposons, particularly in higher vertebrates such as humans. Most retrotransposons integrate into random sites of the host genome, but some have a sequence preference. In particular, a few non-LTR subclades integrate into the genome in a highly sequence-specific manner. By leaving their 5' and/or 3' ends of genomic copies (usually with a poly (A) stretch and target-site duplication (TSD)), the site-specific non-LTR retrotransposons can be searched throughout the DNA database.

**[0065]** There are five clades of restriction enzyme-like endonuclease (RLE)-encoding elements based on their RT sequence similarity. Most of three clades (NeSL, R2, and R4) are target specific and two clades (HERO and CRE) have some site-specific elements. R2 (R2 clade) was found at a specific sequence in the 28S rDNA of many invertebrate and vertebrate species. R4 in *Ascaris* and Dong (R4 clade) were found at another site of the 28S rDNA and microsatellite TAA repeats, respectively.

**[0066]** In contrast to RLE-encoding elements, most of the apurinic/apyrimidinic endonuclease (APE)-encoding non-LTR elements do not insert themselves in a sequence-specific manner, but do have weak target-site specificity, *e.g.*, human L1 for TAAA repeats. However, two clades of APE-encoding non-LTR elements, Tx1 and R1, are known to be sequence-specific.

**[0067]** Among the site-specific non-LTR retrotransposons, R1 clade (R1/R6/R7/RT), R2 clade (R2/R8/R9), R4 clade (R4) and NeSl clade (R5) target the rRNA genes, whereas R7 and R8 target the 18S rDNA, and R1, R2, R4, R5, R6, R9, and RT target the 28S rDNA. All R-element target sites within 28S and 18S rDNA are highly conserved among organisms (see Fujiwara et al., Microbiology Spectrum, (2015) Vol. 3, Issue 2).

**[0068]** In a preferred embodiment, the retrotransposon is rDNA specific and comprises R1, R2, R4, R5, R6, R7, R8, R9 and RT. In certain embodiments, the heterologous DNA-binding domain provided herein comprises the DNA-binding domain of any one of the R1, R2, R4, R5, R6, R7, R8, R9 and RT.

**[0069]** In certain embodiments, the one or more N-terminal ZnF domains is replaced with a CRISPR Cas protein. In certain embodiments, the CRISPR Cas protein comprises Cas9, Cas12a/Cpf1 and Cas13. The clustered regularly interspaced short palindromic repeat (CRISPR)/Cas mediated genome editing technology has seen wild use since its invention due to its simplicity and efficiency, leading to a new era of genetic engineering and gene therapy. The targeting and cleavage are achieved by a single CRISPR RNA ("crRNA")-bound Cas protein in Class 2 CRISPR-Cas systems. Class 2 systems can be subdivided into Type II Cas9 and Type V Cas12 that target DNA, as well as Type VI Cas13 that target RNA. Cas9 and Cas12a/Cpf1 have been well studied in the past and broadly harnessed for gene editing in various cell types and organisms in prokaryotes and eukaryotes. Both Cas9 and Cas12 systems utilize guide RNA to recognize the target site and protospacer adjacent motifs (PAMs) to determine the cleavage site and generate double strand breaks. However, the binding and cleavage of DNA by Cas9 and Cas12 are quite different. Cas9 recognizes a 3'-G-rich PAM and produces blunt ends cleaved by the RuvC and HNH domains, whereas Cas12 recognizes a 5' T-rich PAM and produces staggered ends cleaved solely by the RuvC domain. Pre-assembled Cas 13 and crRNA recognizes target RNAs. Upon RNA-binding, Cas13 will undergo a conformational change and induce the catalytic activity of its nuclease domains, resulting in the cleavage of target transcripts. In certain embodiments, the gRNA or the crRNA can be designed to target a DNA sequence in the genome. In certain embodiments, the Cas protein comprises an inactive nuclease domain. In certain embodiments, the Cas protein does not comprise an active nuclease domain.

**[0070]** In certain embodiments, the guide RNA ("gRNA") comprises two short, non-coding RNA species referred to as crRNA and trans-acting RNA ("tracrRNA"). In an exemplary system, the gRNA forms a complex with a Cas protein of the present disclosure. The gRNA:Cas protein complex binds a target polynucleotide sequence in the genome. The gRNA can be supplied separately that works with the Cas protein in guiding the Cas protein (or the fusion) to a target location on a genomic sequence.

[0071] In one aspect of the present disclosure, a fusion protein provided herein that comprises a first fragment that is a truncated R2Cm or R2Ag protein and a second fragment that can be a DNA-binding domain or a CRISPR Cas protein. In certain embodiments, the second fragment is N-terminal to the first fragment. In certain embodiments, the second fragment can be alternatively C-terminal to the first fragment. The first and second fragments can be fused directly or via a linker sequence.

### ***C. Heterologous Endonuclease or Nickase***

[0072] In some embodiments, the native endonuclease of the R2Cm or R2Ag ORF can be replaced with a biological variant (*e.g.*, a sequence having at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98% or 99% sequence identity to the native endonuclease) or a heterologous endonuclease or a nickase.

[0073] An endonuclease is an enzyme that cleaves internal phosphodiester bonds of polynucleotides. These enzymes are either specific or non-specific to the sequences being cleaved. The endonucleases that are specific to a particular sequence are termed restriction endonucleases, which cleave large DNA molecules at specific sequences of four to six nucleotides.

[0074] In some embodiments the endonuclease is a eukaryotic endonuclease. Non-limiting examples include *Neurospora* endonuclease, S1 nuclease, P1-nuclease, Mung bean nuclease I, *Ustilago* nuclease, DNase I, AP endonuclease and Endo R. Other examples include FokI nuclease, type-II restriction 1-like endonucleases (RLE-type nuclease), and RLE-type endonucleases (REL).

[0075] Many known transposons and retrotransposons also include related endonucleases. In some embodiments, the heterologous endonuclease is one from R1, R2, R4, R5, R6, R7, R8, R9 or RT.

[0076] In some embodiments, the R2Cm or R2Ag endonuclease is substituted with a nickase. A nickase (or nicking endonuclease) is an enzyme that cuts one strand of a double-stranded DNA at a specific recognition nucleotide sequence, to produce DNA molecules that are nicked. Nickases can be prepared from Cas enzyme with site mutations (*e.g.*, D10A Cas9).

[0077] The heterologous DNA-binding domain/Cas and/or the heterologous endonuclease can be used to the truncated R2Cm or R2Ag ORF through a linker sequence. The term “linker sequence” refers to an innocuous length of nucleic acid or protein that joins two other sections of nucleic acid or protein.

[0078] Linkers within the scope of the present disclosure are characterized in terms of amino acid content, length, rigidity and secondary structure. Linkers within the scope of the present disclosure separate the first fragment and the second fragment and allow proper folding and functioning of each domain. In this manner, a linker can be tailored to the particular first and second fragments. According to one aspect, functional independence of the structural and fused (heterologous) domains is maximized by a suitable linker to limit steric interference between domains during the export and assembly processes of the bacterial cell. According to an additional aspect, cell stress is minimized by limiting the overall length of the fusion protein. Longer linker sequences and higher induction levels stress the biosynthetic machinery of the cells, inhibiting cell growth and leading to cell lysis in extreme cases.

[0079] Linkers within the scope of the present disclosure facilitate functioning of the first and second fragments. Linkers within the scope of the present disclosure allow efficient protein processing and export through the bacterial curli secretion machinery as well as provide the proper spatial and physicochemical separation of the polypeptides to retain their respective functions.

[0080] Linkers within the scope of the present disclosure include amino acid residues. The amino acid residues may be any of the naturally occurring amino acid residues. Amino acid residues may also be synthetic amino acids known to those of skill in the art. Representative amino acids which may be used in linkers include Glycine, Alanine, Valine, Leucine, Isoleucine, Serine, Cysteine, Selenocysteine, Threonine, Methionine, Proline, Phenylalanine, Tyrosine, Tryptophan, Histidine, Lysine, Arginine, Aspartate, Glutamate, Asparagine, and Glutamine.

[0081] According to one aspect, a linker sequence is a polypeptide sequence of at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 24, 48 or more amino acids. In some embodiments, the linker sequence comprises from about 3 amino acids to about 50 amino acids. In some embodiments, the linker sequence comprises from about 3 amino acids to about 40 amino

acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 30 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 25 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 24 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 23 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 22 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 21 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 20 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 19 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 18 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 17 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 16 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 15 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 14 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 13 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 12 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 11 amino acids. In some embodiments, the linker sequence comprises from about 5 amino acids to about 10 amino acids.

**[0082]** In some embodiments, the linker sequence comprises a flexible polypeptide, *e.g.* a polypeptide not having a rigid secondary and/or tertiary structure. In some embodiments, the linker sequence comprises glycine and serine residues. In some embodiments at least 50% of the amino acids comprised by the linker sequence are glycine or serine residues, *e.g.* at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95%, or more are glycine or serine residues. In some embodiments, the linker sequence consists of glycine and serine residues.

**[0083]** In some embodiments, the truncated R2Cm or R2Ag ORF and the one or more heterologous elements are not delivered and expressed together. Instead, they can be delivered and expressed separately, or are conjugated or bound to one another.

**[0084]** Therefore, in one aspect, the present disclosure provides a complex comprising a first polypeptide comprising the truncated R2Cm or R2Ag ORF and a second polypeptide

comprising the targeting domain (*e.g.*, a DNA-binding domain or a CRISPR Cas protein) and/or the endonuclease/nickase, wherein the first polypeptide is conjugated or bound to the second polypeptide. The term “conjugate” or “bind” as used herein refers to conjugation of a first conjugation moiety and a second conjugation moiety which form a complex and stabilized by chemical reactions with few of several available chemical groups (chemical conjugation) or via one or more covalent bonds (biological conjugation).

**[0085]** In certain embodiments, the first polypeptide includes a first conjugation moiety and the second polypeptide comprises a second conjugation moiety, wherein the first polypeptide and the second polypeptide are brought together by the conjugation moieties and form a stable complex.

**[0086]** In certain embodiments, the first polypeptide and the second polypeptide may respectfully but regardless of order, further include (a) a receptor and a corresponding ligand, (b) a peptide and a corresponding antibody, or (c) an enzyme and a corresponding substrate.

### **Recombinant R2 RNA and Constructs**

**[0087]** Recombinant R2 RNA and constructs useful for gene delivery are also provided. Each form of these RNA molecules is different from the natural counterparts.

**[0088]** In one embodiment, an open reading frame (ORF) or protein-coding sequence is inserted into an R2Cm or R2Ag RNA. One such recombinant RNA is a *cis* configuration as illustrated in **FIG. 4**. Accordingly, one embodiment provides an RNA that includes (a) a first fragment encoding the sequence of SEQ ID NO:5 or an amino acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to SEQ ID NO:5; (b) a 3'UTR comprising nucleotides 302-371 of SEQ ID NO:7 or a nucleic acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to 302-371 of SEQ ID NO:7; and (c) a second fragment disposed between (a) and (b), which encodes a heterologous protein.

**[0089]** In some embodiments, the RNA can include one or more additional elements from the wild-type R2Ag. An example is other, N-terminal, portions of the 3'UTR. In some embodiments, the RNA includes a complete 3'UTR of the wild-type R2Ag. In some embodiments, the RNA does not include the complete 3'UTR of the wild-type R2Ag, and includes, upstream to the 70-nt core, at least 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt,

90 nt, 100 nt, 110 nt, 120 nt, 130 nt, 140 nt, 150 nt, 160 nt, 170 nt, 180 nt, 190 nt, 200 nt, 210 nt, 220 nt, 230 nt, 240 nt, 250 nt, 260 nt, 270 nt, 280 nt, 290 nt, 300 nt or 301 nt deletion of the wild-type R2Ag.

**[0090]** In some embodiments, the RNA further includes a 5'UTR or a fragment thereof. In some embodiments, the 5'UTR is the wild-type 5'UTR as shown in SEQ ID NO:8. On the other hand, in some embodiments, the RNA includes a portion of the wild-type R2Ag, *i.e.*, having a deletion in the 5'UTR. For instance, the RNA may include a portion of SEQ ID NO:8 upstream of the ORF, but includes an at least 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt, 90 nt, 100 nt, 110 nt, 120 nt, 130 nt, 140 nt, 150 nt, 160 nt, 170 nt, 180 nt, 190 nt, 200 nt, 210 nt, 220 nt, 230 nt, 240 nt, 250 nt, 260 nt, 270 nt, 280 nt, 290 nt, or 297 nt deletion. When there is truncation or deletion in the 5'UTR sequence, in some embodiments, a 5' portion of the 5'UTR is retained. The retained 5' portion, in some embodiments, includes the 5' 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt, 90 nt, or 100 nt of the wild-type sequence. In some embodiments, the entire wild-type 5'UTR can be replaced by a 5'UTR from another RNA, such as another R2.

**[0091]** Another embodiment provides an RNA that includes (a) a first fragment encoding the sequence of SEQ ID NO:1 or an amino acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to SEQ ID NO:1; (b) a 3'UTR comprising nucleotides 194-264 of SEQ ID NO:3 or a nucleic acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to 194-264 of SEQ ID NO:3; and (c) a second fragment disposed between (a) and (b), which encodes a heterologous protein.

**[0092]** In some embodiments, the RNA can include one or more additional elements from the wild-type R2Cm. An example is other, N-terminal, portions of the 3'UTR. In some embodiments, the RNA includes a complete 3'UTR of the wild-type R2Cm. In some embodiments, the RNA does not include the complete 3'UTR of the wild-type R2Cm, and includes, upstream to the 70-nt core, at least 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt, 90 nt, 100 nt, 110 nt, 120 nt, 130 nt, 140 nt, 150 nt, 160 nt, 170 nt, 180 nt, 190 nt, or 193 nt deletion of the wild-type R2Cm.

**[0093]** In some embodiments, the RNA further includes a 5'UTR or a fragment thereof. In some embodiments, the 5'UTR is the wild-type 5'UTR as shown in SEQ ID NO:4. On the other hand, in some embodiments, the RNA includes a portion of the wild-type R2Cm, *i.e.*,

having a deletion in the 5'UTR. For instance, the RNA may include a portion of SEQ ID NO:4 upstream of the ORF, but includes an at least 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt, 90 nt, 100 nt, 110 nt, 120 nt, 130 nt, 140 nt, 150 nt, 160 nt, 170 nt, 180 nt, 190 nt, 200 nt, 210 nt, 220 nt, 230 nt, 240 nt, 250 nt, 260 nt, or 267 nt deletion. When there is truncation or deletion in the 5'UTR sequence, in some embodiments, a 5' portion of the 5'UTR is retained. The retained 5' portion, in some embodiments, includes the 5' 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt, 90 nt, or 100 nt of the wild-type sequence. In some embodiments, the entire wild-type 5'UTR can be replaced by a 5'UTR from another RNA, such as another R2.

**[0094]** In some embodiments, the RNA further includes a PolyA sequence. In some embodiments, the RNA further includes one or two flanking sequences complementary to a target site on the target genome for integration. In some embodiments, the target site is on a 28S rDNA.

**[0095]** In some embodiments, the second fragment of (c) which encodes the heterologous protein has the same orientation as the first fragment of (a) that encodes the ORF (SEQ ID NO:1 or 5). In some embodiments, the second fragment of (c) which encodes the heterologous protein has the opposite orientation as the first fragment of (a) that encodes the ORF (SEQ ID NO:1 or 5). In some embodiments, the second fragment further includes a promoter.

**[0096]** Engineered R2Ag and R2Cm, with or without a heterologous protein-coding sequence, are also provided. The engineering, in some embodiments, entails deletion of all or a portion of the 3'UTR, deletion of all or a portion of the 5'UTR, or deletion of either or both of the flanking sequences.

**[0097]** Accordingly, in one embodiment, provided an RNA that includes (a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:5 or an amino acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to SEQ ID NO:5; and (b) a 3'UTR that includes nucleotides 302-371 of SEQ ID NO:7 or a nucleic acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to 302-371 of SEQ ID NO:7, wherein the RNA does not include at least 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt, 90 nt, 100 nt, 110 nt, 120 nt, 130 nt, 140 nt, 150 nt, 160 nt, 170 nt, 180 nt, 190 nt, 200 nt, 210 nt, 220 nt, 230 nt, 240 nt, 250 nt, 260 nt, 270 nt, 280 nt, 290 nt,

300 nt or 301 nt nucleotides among nucleotides 1-301 of SEQ ID NO:7 upstream of nucleotides 302-371 of SEQ ID NO:7. In some embodiments, the RNA does not include any of SEQ ID NO:7.

**[0098]** Also provided, in one embodiment, is an RNA that includes (a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:5 or an amino acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to SEQ ID NO:5; and (b) a 3'UTR that includes nucleotides 302-371 of SEQ ID NO:7 or a nucleic acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to 302-371 of SEQ ID NO:7, wherein the RNA does not include at least 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt, 90 nt, 100 nt, 110 nt, 120 nt, 130 nt, 140 nt, 150 nt, 160 nt, 170 nt, 180 nt, 190 nt, 200 nt, 210 nt, 220 nt, 230 nt, 240 nt, 250 nt, 260 nt, 270 nt, 280 nt, 290 nt, or 297 nt nucleotides of SEQ ID NO:8 upstream of the ORF.

**[0099]** Further provided, in one embodiment, is an RNA that includes (a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:5 or an amino acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to SEQ ID NO:5; and (b) a 3'UTR that includes nucleotides 302-371 of SEQ ID NO:7 or a nucleic acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to 302-371 of SEQ ID NO:7, wherein the RNA does not include either or both sequences upstream of the ORF and downstream of the 3'UTR complementary to a 28S rDNA gene.

**[0100]** Yet, in one embodiment, provided an RNA that includes (a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:1 or an amino acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to SEQ ID NO:1; and (b) a 3'UTR that includes nucleotides 194-264 of SEQ ID NO:3 or a nucleic acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to 194-264 of SEQ ID NO:3, wherein the RNA does not include at least 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt, 90 nt, 100 nt, 110 nt, 120 nt, 130 nt, 140 nt, 150 nt, 160 nt, 170 nt, 180 nt, 190 nt, or 193 nt nucleotides among nucleotides 1-193 of SEQ ID NO:3 upstream of nucleotides 194-264 of SEQ ID NO:3. In some embodiments, the RNA does not include any of SEQ ID NO:3.

**[0101]** Also provided, in one embodiment, is an RNA that includes (a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:1 or an amino acid sequence having at

least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to SEQ ID NO:1; and (b) a 3'UTR that includes nucleotides 194-264 of SEQ ID NO:3 or a nucleic acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to 194-264 of SEQ ID NO:3, wherein the RNA does not include at least 10 nt, 20 nt, 30 nt, 40 nt, 50 nt, 60 nt, 70 nt, 80 nt, 90 nt, 100 nt, 110 nt, 120 nt, 130 nt, 140 nt, 150 nt, 160 nt, 170 nt, 180 nt, 190 nt, 200 nt, 210 nt, 220 nt, 230 nt, 240 nt, 250 nt, 260 nt, or 267 nt nucleotides of SEQ ID NO:4 upstream of the ORF.

**[0102]** Further provided, in one embodiment, is an RNA that includes (a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:1 or an amino acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to SEQ ID NO:1; and (b) a 3'UTR that includes nucleotides 194-264 of SEQ ID NO:3 or a nucleic acid sequence having at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99% sequence identity to 194-264 of SEQ ID NO:3, wherein the RNA does not include either or both sequences upstream of the ORF and downstream of the 3'UTR complementary to a 28S rDNA gene.

**[0103]** In some embodiments, the RNA further includes a PolyA sequence. In some embodiments, the RNA further includes one or two flanking sequences complementary to a target site on the target genome for integration. In some embodiments, the target site is on a 28S rDNA.

**[0104]** In some embodiments, DNA molecules (as well as vectors) that can be transcribed to prepare an RNA molecule of the disclosure are also provided.

### **Protein-Coding Sequence for Delivery**

**[0105]** In one embodiment, the R2Cm or R2Ag ORF, or an engineered version thereof, can be provided as a mRNA. The mRNA can further include the corresponding 5'UTR and/or 3'UTR of the native R2Cm or R2Ag.

**[0106]** In some embodiments, to insert a different sequence (*e.g.*, a second protein-coding sequence) to a target genome, the mRNA can further include that second protein-coding sequence. In some embodiments, the second protein-coding sequence is 3' to the R2Cm or R2Ag ORF. In some embodiments, the second protein-coding sequence is included in the mRNA in a reverse-complement orientation such that it cannot be directly translated by the mRNA. Rather it is translated after integration to the target genome. In some embodiments,

the second protein-coding sequence is inserted at 3' to the R2Cm or R2Ag ORF via a self-cleaving peptide sequence, such as 2A self-cleaving peptide.

**[0107]** Alternatively, in some embodiments, the second protein-coding sequence is provided on a separate mRNA that flanks the second protein-coding sequence with the 5'UTR and/or 3'UTR of the native R2Cm or R2Ag.

**[0108]** In some embodiments, the second protein is a therapeutic protein that can be helpful in treating a disease or condition in a patient. Examples include:

- Insulin - used to treat diabetes
- Phenylalanine hydroxylase - used to treat PKU (Phenylketonuria) disease
- Erythropoietin - used to treat anemia
- Filgrastim (G-CSF) - used to treat neutropenia
- Interferons - used to treat cancer, multiple sclerosis, and viral infections
- Tumor necrosis factor (TNF) inhibitors - used to treat autoimmune diseases, such as rheumatoid arthritis and Crohn's disease
- Monoclonal antibodies - used to treat cancer, autoimmune diseases, and some viral infections
- Follitropin alfa - used to treat infertility
- Coagulation factor - used to treat bleeding disorders, such as Hemophilia
- Enzyme replacement therapy enzymes - used to treat lysosomal storage disorders, such as Gaucher's disease
- Growth hormone - used to treat growth hormone deficiency.

**[0109]** In some embodiments, the polynucleotide (*e.g.*, mRNA) further includes a downstream flanking sequence that is 3' to the protein-coding sequence. Addition of a downstream flanking sequence can improve the retrotransposition efficiency. In certain embodiments, the protein-coding sequence may comprise, at the 3' end of the gene encoding sequence, one or more consecutive nucleotides (rDNA sequences) of the “downstream sequence” starting from the EN cleavage site. The protein-coding sequence may comprise, at the 3' end of the gene encoding sequence, different lengths of the downstream sequence depending on the type of polynucleotide or vector used for delivery. In certain embodiments, a length of the downstream sequence of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 180,

200, 250 nt can be included to the 3' mRNA of the fusion protein-encoding sequence. In some embodiments, the length can be 1-250, 1-200, 4-180, 10-150, 15-150, 40-90, 50-90, 70-90 or 71-90 nt, without limitation.

[0110] In some embodiments, the polynucleotide (*e.g.*, mRNA) further includes an upstream flanking sequence that is 5' to the protein-coding sequence. Addition of an upstream flanking sequence may also improve the retrotransposition efficiency. In certain embodiments, the protein-coding sequence may comprise, at the 5' end of the gene encoding sequence, one or more consecutive nucleotides (rDNA sequences) of the "upstream sequence" starting from the EN cleavage site. The protein-coding sequence may comprise, at the 5' end of the gene encoding sequence, different lengths of the downstream sequence depending on the type of polynucleotide or vector used for delivery. In certain embodiments, a length of the downstream sequence of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 180, 200, 250 nt can be included to the 5' mRNA of the fusion protein-encoding sequence. In some embodiments, the length can be 1-250, 1-200, 4-180, 10-150, 15-150, 40-90, 50-90, 70-90 or 71-90 nt, without limitation.

[0111] In some embodiments, the R2Cm 3'UTR has the nucleic acid sequence of SEQ ID NO:3. In some embodiments, the 3'UTR has at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99% or 100% sequence identity to SEQ ID NO:3. In some embodiments, the 3'UTR is directly connected to the downstream flanking sequence.

[0112] In some embodiments, the polynucleotide (*e.g.*, mRNA) further includes a 5' untranslated region (UTR) 5' to the protein-coding sequence. When an upstream flanking sequence is also included, preferably the upstream flanking sequence is 5' to the 5'UTR.

[0113] In some embodiments, the R2Cm 5'UTR has the nucleic acid sequence of SEQ ID NO:4. In some embodiments, the 5'UTR has at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99% or 100% sequence identity to SEQ ID NO:4.

[0114] In some embodiments, the R2Ag 3'UTR has the nucleic acid sequence of SEQ ID NO:7. In some embodiments, the 3'UTR has at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99% or 100% sequence identity to SEQ ID NO:7.

[0115] In some embodiments, the R2Ag 5'UTR has the nucleic acid sequence of SEQ ID NO:8. In some embodiments, the 5'UTR has at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99% or 100% sequence identity to SEQ ID NO:8.

### Gene Delivery and Treatments

[0116] The R2 retrotransposition systems disclosed here are suitable for delivering exogenous genes to a target cell, such as a mammalian cell.

[0117] Accordingly, in some embodiments, the present disclosure provides a method for integrating an exogenous polynucleotide to the genome of a target cell. In one embodiment, the method entails introducing to the cell the polynucleotide comprising the R2Cm or R2Ag retrotransposon of the instant disclosure. In some embodiments, the retrotransposon is delivered to the target cell as an mRNA molecule directly.

[0118] In some embodiments, when the polynucleotide comprises a CRISPR Cas protein encoding sequence, the method further comprises introducing to the cell a guide RNA (gRNA) that binds to the CRISPR Cas protein and recognizes a target DNA sequence in the genome.

[0119] mRNAs may be synthesized according to any of a variety of known methods. For example, the mRNAs may be synthesized via *in vitro* transcription (IVT). Briefly, IVT is typically performed with a linear or circular DNA template containing a promoter, a pool of ribonucleotide triphosphates, a buffer system that may include DTT and magnesium ions, and an appropriate RNA polymerase (*e.g.*, T3, T7 or SP6 RNA polymerase), DNase I, pyrophosphatase, and/or RNase inhibitor. The exact conditions will vary according to the specific application.

[0120] The mRNA may be synthesized as unmodified or modified mRNA. Typically, mRNAs are modified to enhance stability. Modifications of mRNA can include, for example, modifications of the nucleotides of the RNA. A modified mRNA can thus include, for example, backbone modifications, sugar modifications or base modifications. In some embodiments, antibody encoding mRNAs (*e.g.*, heavy chain and light chain encoding mRNAs) may be synthesized from naturally occurring nucleotides and/or nucleotide analogues (modified nucleotides) including, but not limited to, purines (adenine (A), guanine (G)) or pyrimidines (thymine (T), cytosine (C), uracil (U)), and as modified nucleotides

analogues or derivatives of purines and pyrimidines, such as *e.g.* 1-methyl-adenine, 2-methyl-adenine, 2-methylthio-N-6-isopentenyl-adenine, N6-methyl-adenine, N6-isopentenyl-adenine, 2-thio-cytosine, 3-methyl-cytosine, 4-acetyl-cytosine, 5-methyl-cytosine, 2,6-diaminopurine, 1-methyl-guanine, 2-methyl-guanine, 2,2-dimethyl-guanine, 7-methyl-guanine, inosine, 1-methyl-inosine, pseudouracil (5-uracil), dihydro-uracil, 2-thio-uracil, 4-thio-uracil, 5-carboxymethylaminomethyl-2-thio-uracil, 5-(carboxyhydroxymethyl)-uracil, 5-fluoro-uracil, 5-bromo-uracil, 5-carboxymethylaminomethyl-uracil, 5-methyl-2-thio-uracil, 5-methyl-uracil, N-uracil-5-oxyacetic acid methyl ester, 5-methylaminomethyl-uracil, 5-methoxyaminomethyl-2-thio-uracil, 5'-methoxycarbonylmethyl-uracil, 5-methoxy-uracil, uracil-5-oxyacetic acid methyl ester, uracil-5-oxyacetic acid (v), 1-methyl-pseudouracil, queosine, 13-D-mannosyl-queosine, wybutoxosine, and phosphoramidates, phosphorothioates, peptide nucleotides, methylphosphonates, 7-deazaguanosine, 5-methylcytosine and inosine. The preparation of such analogues is known to a person skilled in the art *e.g.* from the U.S. Pat. Nos. 4,373,071, 4,401,796, 4,415,732, 4,458,066, 4,500,707, 4,668,777, 4,973,679, 5,047,524, 5,132,418, 5,153,319, 5,262,530 and 5,700,642, the disclosure of which is included here in its full scope by reference.

**[0121]** In some embodiments, the mRNAs may contain RNA backbone modifications. Typically, a backbone modification is a modification in which the phosphates of the backbone of the nucleotides contained in the RNA are modified chemically. Exemplary backbone modifications typically include, but are not limited to, modifications from the group consisting of methylphosphonates, methylphosphoramidates, phosphoramidates, phosphorothioates (*e.g.* cytidine 5'-O-(1-thiophosphate)), boranophosphates, positively charged guanidinium groups etc., which means by replacing the phosphodiester linkage by other anionic, cationic or neutral groups.

**[0122]** In some embodiments, the mRNAs may contain sugar modifications. A typical sugar modification is a chemical modification of the sugar of the nucleotides it contains including, but not limited to, sugar modifications chosen from the group consisting of 2'-deoxy-2'-fluoro-oligoribonucleotide (2'-fluoro-2'-deoxycytidine 5'-triphosphate, 2'-fluoro-2'-deoxyuridine 5'-triphosphate), 2'-deoxy-2'-deamine-oligoribonucleotide (2'-amino-2'-deoxycytidine 5'-triphosphate, 2'-amino-2'-deoxyuridine 5'-triphosphate), 2'-O-alkyloligoribonucleotide, 2'-deoxy-2'-C-alkyloligoribonucleotide (2'-O-methylcytidine 5'-triphosphate, 2'-methyluridine 5'-triphosphate), 2'-C-alkyloligoribonucleotide, and isomers

thereof (2'-aracytidine 5'-triphosphate, 2'-arauridine 5'-triphosphate), or azidotriphosphates (2'-azido-2'-deoxycytidine 5'-triphosphate, 2'-azido-2'-deoxyuridine 5'-triphosphate).

**[0123]** In some embodiments, the mRNAs may contain modifications of the bases of the nucleotides (base modifications). A modified nucleotide which contains a base modification is also called a base-modified nucleotide. Examples of such base-modified nucleotides include, but are not limited to, 2-amino-6-chloropurine riboside 5'-triphosphate, 2-aminoadenosine 5'-triphosphate, 2-thiocytidine 5'-triphosphate, 2-thiouridine 5'-triphosphate, 4-thiouridine 5'-triphosphate, 5-aminoallylcytidine 5'-triphosphate, 5-aminoallyluridine 5'-triphosphate, 5-bromocytidine 5'-triphosphate, 5-bromouridine 5'-triphosphate, 5-iodocytidine 5'-triphosphate, 5-iodouridine 5'-triphosphate, 5-methylcytidine 5'-triphosphate, 5-methyluridine 5'-triphosphate, 6-azacytidine 5'-triphosphate, 6-azauridine 5'-triphosphate, 6-chloropurine riboside 5'-triphosphate, 7-deazaadenosine 5'-triphosphate, 7-deazaguanosine 5'-triphosphate, 8-azaadenosine 5'-triphosphate, 8-azidoadenosine 5'-triphosphate, benzimidazole riboside 5'-triphosphate, N1-methyladenosine 5'-triphosphate, N1-methylguanosine 5'-triphosphate, N6-methyladenosine 5'-triphosphate, O6-methylguanosine 5'-triphosphate, pseudouridine 5'-triphosphate, puromycin 5'-triphosphate or xanthosine 5'-triphosphate.

**[0124]** In some embodiments, the mRNAs include a 5' cap structure. A 5' cap is typically added as follows: first, an RNA terminal phosphatase removes one of the terminal phosphate groups from the 5' nucleotide, leaving two terminal phosphates; guanosine triphosphate (GTP) is then added to the terminal phosphates via a guanylyl transferase, producing a 5'5'5 triphosphate linkage; and the 7-nitrogen of guanine is then methylated by a methyltransferase. Examples of cap structures include, but are not limited to, m7G(5')ppp(5'(A,G(5')ppp(5)A and G(5')ppp(5')G.

**[0125]** In some embodiments, the retrotransposon of the present disclosure is delivered by a vector such as a plasmid vector and a viral vector (*e.g.*, adenoviral vector).

**[0126]** In some embodiments, the target mammalian cell is a human cell. In some embodiments, the human cell is an induced pluripotent stem cell (iPSC). When an mRNA is directly introduced to the target stem cell, in a preferred embodiment, the downstream flanking sequence of the retrotransposon has a length of at least 40 nt, such as 40 to 90 nt, 45 to 85 nt, or 50 to 80 nt.

[0127] In some embodiments, the human cell is not a stem cell. When an mRNA is directly introduced to the target non-stem cell, in a preferred embodiment, the downstream flanking sequence of the retrotransposon has a length of at least 70 nt, such as 70 to 95 nt, 70 to 90 nt, or 71 to 85 nt.

[0128] In some embodiments, the R2Cm or R2Ag ORF and the exogenous gene are included in the same retrotransposon polynucleotide. In some embodiments, only the exogenous gene is included in the retrotransposon polynucleotide, while the R2Cm or R2Ag ORF is delivered separately on a different polynucleotide.

[0129] In some embodiments, the delivery is *in vitro*. In some embodiments, the delivery is *in vivo* or *ex vivo*. In some embodiments, the retrotransposon polynucleotide is delivered by parenteral means, or locally.

[0130] The term “parenteral” as used herein refers to modes of administration which include intravenous, intramuscular, intraperitoneal, intrasternal, subcutaneous and intra-articular injection and infusion.

[0131] Delivery can be systemic or local. In addition, it may be desirable to introduce the retrotransposon polynucleotide of the disclosure into the central nervous system by any suitable route, including intraventricular and intrathecal injection; intraventricular injection may be facilitated by an intraventricular catheter, for example, attached to a reservoir, such as an Ommaya reservoir. Pulmonary administration can also be employed, *e.g.*, by use of an inhaler or nebulizer, and formulation with an aerosolizing agent.

## EXAMPLES

[0132] The following examples are included to demonstrate specific embodiments of the disclosure. It should be appreciated by those of skill in the art that the techniques disclosed in the examples which follow represent techniques to function well in the practice of the disclosure, and thus can be considered to constitute specific modes for its practice. However, those of skill in the art should, in light of the present disclosure, appreciate that many changes can be made in the specific embodiments which are disclosed and still obtain a like or similar result without departing from the spirit and scope of the disclosure.

**Example 1: Discovery and Testing of New R2 Retrotransposons**

[0133] This example describes the discovery and testing of two new R2 retrotransposons from bird species.

[0134] The amino acid sequence of the open reading frame (ORF) of R2OI, in particular those of the main functional domains, was used to search for analogs in the GenBank database. Thirty-three total hits were returned from the bioinformatic approach, belonging to 11 different organisms. Notably, 29 of the hits are from eight different bird species (*Aves*), and none were from mammalian species.

[0135] Based on manual sequence analysis and preliminary testing, two of the hits, which were not known as retrotransposases, were subjected to additional evaluation. They were from *Corvus moneduloides* (New Caledonian crow) and *Accipiter gentilis* (Eurasian goshawk), respectively, and are herein referred to as R2Cm and R2Ag. Their amino acid sequences are provided in **Table 1**. Through bioinformatic analysis, the domain structures of these newly discovered R2 ORF, as well as the associated 5' untranslated region (5'UTR) and 3'UTR sequences, were obtained and also shown in **Table 1**.

**Table 1. Sequences of R2Cm and R2Ag Retrotransposons**

| R2Cm ORF – SEQ ID NO:1                                                                         |        |        |
|------------------------------------------------------------------------------------------------|--------|--------|
| MGISKPFLGPPIRLFQVSVGTQTSNSDWVDLVS RDFKPGTSSKSCGVDLVSLTSKHQVDPVAQNFPSPGSP                       |        | 70     |
| NSAENESDLVNLEFYRSDLCLDNVGGHFSVNWEGLSRLPEVFEHPTPQPSVEGGDHLPKDSEVIVLS DGE                        |        | 140    |
| ESGEHECGSALETSPPEPESTSETSPPEPATSTVPATSPPPPEDHEKQAPGDI PVVTIPDKNPP <b>CPCCGT</b>                |        | 210    |
| <u>RVGGVGGLIEHLREAHGRRGVCFRCAKCGKESTNCHSIVCHVPKCRAKTETAPAGEWICEECGRDFQTKI</u>                  |        | 280    |
| N-ZnF1                                                                                         | N-ZnF2 | N-ZnF3 |
| <u>GLGQHKRLAH</u> HPMIRNQERIAASHPKETSNRGAHKRC <b><u>WTKEEEELLIKLEAQFEGNKNINKLIAEHIPTKT</u></b> |        | 350    |
| SANT DNA-Binding Domain                                                                        |        |        |
| <b><u>AKQISDK</u></b> RRLLPRKPTEDVCEEPGVSHRTRRAAAGLKIEPGPSNQPGKAEDKGPEISQRLRKPAAGRGT           |        | 420    |
| EPEIRHHPDKRTKQLGNIDQRREGQLQAHYQEKLGEWLSAGKIKTFPRAFKQLIKQEQPVVNTTAQDC                           |        | 490    |
| FGCLEAISHTKPA SRDKDKRESQGGKPTRPYQKWMKKR <b><u>QVVRGKYLRFORL FHLDRGKLAKIILDDVEC</u></b>         |        | 560    |
| Binds 3'UTR                                                                                    |        |        |
| LSCDISPSKIYSDFKARWETPGKFDSLGFQINGKADNSA <b><u>FRDLITAKEIEKNVREMSRNSAPGPDGITL</u></b>           |        | 630    |

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |      |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| Template-Switching                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |      |
| <u>RD</u> <u>IK</u> <u>K</u> <u>M</u> <u>D</u> <u>P</u> <u>E</u> <u>F</u> <u>S</u> <u>Q</u> <u>T</u> <u>M</u> <u>E</u> <u>I</u> <u>F</u> <u>N</u> <u>L</u> <u>W</u> <u>L</u> <u>T</u> <u>S</u> <u>G</u> <u>K</u> <u>I</u> <u>P</u> <u>D</u> <u>M</u> <u>V</u> <u>R</u> <u>GCRTVLIPKSTRPERLKDINNWRPITIGSILLRLFSRIL</u>                                                                                                                                                                                                                                                                                                                                 | 700  |
| <u>TAR</u> <u>L</u> <u>T</u> <u>K</u> <u>A</u> <u>C</u> <u>P</u> <u>L</u> <u>N</u> <u>P</u> <u>R</u> <u>Q</u> <u>R</u> <u>G</u> <u>F</u> <u>I</u> <u>R</u> <u>A</u> <u>A</u> <u>G</u> <u>C</u> <u>S</u> <u>E</u> <u>N</u> <u>L</u> <u>K</u> <u>L</u> <u>L</u> <u>Q</u> <u>T</u> <u>I</u> <u>I</u> <u>R</u> <u>S</u> <u>A</u> <u>K</u> <u>R</u> <u>D</u> <u>H</u> <u>K</u> <u>P</u> <u>L</u> <u>G</u> <u>V</u> <u>V</u> <u>F</u> <u>V</u> <u>D</u> <u>I</u> <u>A</u> <u>K</u> <u>A</u> <u>F</u> <u>D</u> <u>T</u> <u>V</u> <u>S</u> <u>H</u> <u>R</u> <u>H</u> <u>I</u> <u>L</u> <u>Q</u> <u>G</u> <u>L</u> <u>Q</u> <u>E</u> <u>R</u>                 | 770  |
| Reverse-transcriptase (RT)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |      |
| <u>C</u> <u>V</u> <u>D</u> <u>P</u> <u>H</u> <u>I</u> <u>V</u> <u>S</u> <u>L</u> <u>V</u> <u>S</u> <u>N</u> <u>M</u> <u>Y</u> <u>E</u> <u>N</u> <u>I</u> <u>R</u> <u>T</u> <u>Y</u> <u>I</u> <u>T</u> <u>T</u> <u>G</u> <u>K</u> <u>T</u> <u>Q</u> <u>T</u> <u>D</u> <u>H</u> <u>I</u> <u>Q</u> <u>I</u> <u>Q</u> <u>V</u> <u>G</u> <u>V</u> <u>K</u> <u>Q</u> <u>G</u> <u>D</u> <u>P</u> <u>M</u> <u>S</u> <u>P</u> <u>F</u> <u>L</u> <u>F</u> <u>N</u> <u>L</u> <u>A</u> <u>M</u> <u>D</u> <u>P</u> <u>L</u> <u>L</u> <u>C</u> <u>E</u> <u>L</u> <u>E</u> <u>T</u> <u>S</u> <u>G</u> <u>K</u> <u>G</u> <u>Y</u> <u>H</u> <u>R</u> <u>G</u> <u>P</u> | 840  |
| <u>S</u> <u>S</u> <u>V</u> <u>T</u> <u>A</u> <u>M</u> <u>A</u> <u>F</u> <u>A</u> <u>D</u> <u>D</u> <u>L</u> <u>V</u> <u>L</u> <u>L</u> <u>S</u> <u>D</u> <u>S</u> <u>W</u> <u>E</u> <u>N</u> <u>M</u> <u>K</u> <u>R</u> <u>N</u> <u>I</u> <u>R</u> <u>I</u> <u>V</u> <u>E</u> <u>T</u> <u>F</u> <u>C</u> <u>D</u> <u>L</u> <u>T</u> <u>G</u> <u>L</u> <u>K</u> <u>T</u> <u>Q</u> <u>G</u> <u>E</u> <u>K</u> <u>C</u> <u>H</u> <u>G</u> <u>F</u> <u>F</u> <u>I</u> <u>K</u> <u>P</u> <u>T</u> <u>K</u> <u>D</u> <u>S</u> <u>Y</u> <u>T</u> <u>V</u> <u>N</u> <u>D</u> <u>C</u> <u>A</u> <u>A</u> <u>W</u> <u>T</u> <u>I</u> <u>N</u> <u>G</u> <u>T</u> | 910  |
| <u>P</u> <u>L</u> <u>N</u> <u>M</u> <u>I</u> <u>H</u> <u>P</u> <u>G</u> <u>E</u> <u>S</u> <u>E</u> <u>K</u> <u>Y</u> <u>L</u> <u>G</u> <u>L</u> <u>Q</u> <u>I</u> D P W T G L A K S D L S T K L E S W L E K I G K S P L K P L Q K D I L K T Y T I P R L T Y L A D H S D                                                                                                                                                                                                                                                                                                                                                                               | 980  |
| LKAGFLEALDQTVRTKVKEWLHLPACTCDAI FYSSTRDGGGLGLTKLAGLIPSVQARRLHRLAQSSDETMK                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | 1050 |
| LFLEKDQMEQLHKKLWVQAGGDREKIPSIWEAKPANVPSTSYSADITSEWEAPSLKSKFPKPCNWRKKEF                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 1120 |
| KNWTKLVSQGRGIENFEGDKISNHWIQYRGIPHRKLLTALQLRANVYPTREFLARGRKDKYTAK <u>CRHCE</u>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 1190 |
| ADNETCAHIIIGNCPITQ <u>D</u> <u>A</u> <u>R</u> <u>I</u> <u>K</u> <u>R</u> <u>H</u> <u>N</u> <u>Y</u> <u>I</u> <u>C</u> <u>K</u> <u>V</u> <u>L</u> <u>S</u> <u>E</u> <u>E</u> <u>A</u> <u>K</u> <u>K</u> <u>D</u> <u>W</u> <u>V</u> <u>V</u> <u>F</u> <u>Q</u> <u>E</u> <u>P</u> <u>H</u> <u>I</u> <u>R</u> <u>D</u> <u>T</u> <u>N</u> <u>K</u> <u>E</u> <u>L</u> <u>Y</u> <u>K</u> <u>P</u> <u>D</u> <u>L</u> <u>I</u> <u>F</u> <u>V</u> <u>K</u> <u>D</u> <u>A</u> <u>R</u> <u>A</u> <u>F</u> <u>V</u>                                                                                                                                                | 1260 |
| C-ZnF                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |      |
| <u>V</u> <u>D</u> <u>V</u> <u>T</u> <u>V</u> <u>R</u> <u>Y</u> <u>E</u> <u>A</u> <u>S</u> <u>K</u> <u>T</u> <u>S</u> <u>L</u> <u>E</u> <u>E</u> <u>A</u> <u>A</u> <u>A</u> <u>E</u> <u>K</u> <u>V</u> <u>N</u> <u>K</u> <u>Y</u> <u>N</u> <u>H</u> <u>L</u> <u>A</u> <u>E</u> <u>E</u> <u>I</u> <u>R</u> <u>E</u> <u>L</u> <u>T</u> <u>N</u> <u>A</u> <u>K</u> <u>D</u> <u>V</u> <u>V</u> <u>F</u> <u>M</u> <u>G</u> <u>F</u> <u>L</u> <u>G</u> <u>A</u> <u>R</u> <u>G</u> <u>K</u> <u>W</u> <u>Y</u> <u>H</u> <u>G</u> <u>N</u> <u>F</u> <u>E</u> <u>L</u> <u>L</u> <u>G</u> <u>A</u> <u>L</u> <u>G</u> <u>L</u> <u>S</u> <u>K</u> <u>T</u>          | 1330 |
| Endonuclease (EN)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |      |
| <u>R</u> <u>Q</u> <u>E</u> <u>K</u> <u>A</u> <u>A</u> <u>R</u> <u>T</u> <u>L</u> <u>S</u> <u>S</u> <u>I</u> <u>A</u> <u>L</u> <u>T</u> <u>S</u> <u>S</u> <u>V</u> <u>D</u> <u>I</u> <u>V</u> <u>H</u> <u>M</u> <u>F</u> <u>A</u> <u>S</u> <u>R</u> <u>S</u> <u>R</u> <u>S</u> <u>C</u> <u>K</u> <u>V</u> <u>A</u>                                                                                                                                                                                                                                                                                                                                     | 1364 |
| <b>R2Cm Reverse-Transcriptase (R2Cm RT) – SEQ ID NO:2</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |      |
| GCRTVLIPKSTRPERLKDINNWRPITIGSILLRLFSRILTARLTACPLNPRQGRGFIRAAGCSENKLLQ<br>TIIRSAKRDKHPLGVVFDIAKAFDTVSHRHILQGLQERCVDPHIVSLVSNMYENIRTYITTGKTQTDHI<br>QIQVGKQGDPMSPFLFNLAMDPLLCELETSGKGYHRGPSSVTAMAFADDLVLLSDSWENMKRNIRIVET<br>FCDLTGLKTQGEKCHGFFIKPTKDSYTVNDCAAWTINGTPLNMIHPGESEKYLGLQI                                                                                                                                                                                                                                                                                                                                                                  |      |
| <b>R2Cm 3'UTR – SEQ ID NO:3 (in DNA sequence); underlined is R2Cm 3' Core (194-264)</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |      |
| TAGACTGTTACTTTATTTTACTTTGTAAATGTTGTCCTGTAACATTTTATAACATATCTTCAATAAAGTT<br>TCGGCGGGGGCTAGGGACGCAGGGCAGCCACAAGCCAGAACATCGCTCACGGTAAGTAGGGACAGGACGG<br>AACTTTTCATAACGCGTCAATTACCACCCGATTTGGACCAATTTACCTGATTTGTCCAAGGTGGACGGGC<br>CACCTCTACTTAACCCGGAAGGAACATATAGTACTTATATGTGTTCGAAAAA                                                                                                                                                                                                                                                                                                                                                                    |      |
| <b>R2Cm 5'UTR – SEQ ID NO:4 (in DNA sequence)</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |      |
| CGGCTTAGGAGAGTATAGTTACGAAGTGGGCATCGCTGCAGAGATCGCACCTCCTCGTGGTCCCCTGGA<br>TACCCTGCGTAGGGTAACAAAGTCGATCTCTGCCCCAGTACGGAGTCGTTGGGAATCACCAGTTCAACGA<br>TCTCCTGCTAAATTTGGTGAAACAAATTCGAGGTAAAGCCGCACGGCTATTGCCCTAAACCTGGCCCCA<br>AGGTTTCAAGAAGAGGGCAAAAGATCGGTAAGTGGTGAAGGGGGCCACTACCGTTCTA                                                                                                                                                                                                                                                                                                                                                                |      |
| <b>R2Ag ORF – SEQ ID NO:5</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |      |
| MGRPMRISFVSVATQTSPLDLRDNVSSSIQDHKTGGLLEARFEIRGLEFVNSDLEGLTSQQSPTLTDLL                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 70   |

|                                                                                                               |      |
|---------------------------------------------------------------------------------------------------------------|------|
| AVLPEVYEVPNLSLRDKNTCPTPVEDENEEESSPLGAAAGTSPLPSTPVSRSCEDPPAESTPMVKIPD                                          | 140  |
| KNPS <u>C</u> PCCGNVTGKLSGLIE <u>H</u> LKRAHGGKKKILFQ <u>C</u> ARCGRTDVKHHSIACHFPKCKGKGPVEASPERSWK <u>C</u> E | 210  |
| N-ZnF1 N-ZnF2                                                                                                 |      |
| VCEKTFESKIGLGQ <u>H</u> KRLV <u>H</u> FPVTRNLERITATRPKTGTMRGAHRC <u>WTEDEIELLRKLDKQYEGQKNINK</u>              | 280  |
| N-ZnF3 SANT DNA-Binding                                                                                       |      |
| <u>LIAEHLPTKTTKQISDK</u> RRDLHKSPVKKGKDQGQMKHLELEGEHLVEGEHLVCGSEMEGGLRSRYRETV                                 | 350  |
| domain                                                                                                        |      |
| SEWLTSGKLSILHGMFEQVIEGEKDPAIVAEKAMKDCLACLASKSLELKAGQDRPRQVARTARPSEPCKG                                        | 420  |
| WMKKRAVKGQFLRFQRLFYLDRGKLAGIILDNVESLCKGIPQDEIHDVFKSRWESPGTFKGLGAFRSKG                                         | 490  |
| Binds 3'UTR                                                                                                   |      |
| LADNSAFEAMITAKEISKNIQEMNKNSAPGPDRLSLRDIIKIDPHCVQLMEVFNLWLVTGSPDVVK <u>ECR</u>                                 | 560  |
| Template-Switching                                                                                            |      |
| <u>TVLISKSSLPERQMDINNWRPITIGSIILRLFSRILTARLAKACPINPRQKGFIRSAGCAENKLLQLLI</u>                                  | 630  |
| <u>KNAKKEHRPLGVVFDLAKAFDTVSHHHLIAVLKQKGVDDHHIITLITNMYQNTRTRITMKNEQSDPIGIG</u>                                 | 700  |
| Reverse-transcriptase (RT)                                                                                    |      |
| <u>IGVKQGDPMSPLLFNLSVDPLLCCKLEEGSGYQHDSKGVTSMAFADDLVLLSGSWEQMOSNIKILEAFCE</u>                                 | 770  |
| <u>LTGLRTQGEKCHGFYIQPTKDSYTTINDCPPWQINGTPLTMIEPGSSEKYLGLRI</u> DPWTGISKPELLMKIR                               | 840  |
| DWLQRIGEAPLKPFQKVDILKGYTIPRLIYLADQAEVKATSLEELDITRTTVKEWLHLPSTCDAILYS                                          | 910  |
| STRDGGLGITRLSGLIPSIQARRLHRIAQSSDDVRSIVKREGIEKDFEKLWLTAGGDRNKIPSIWDPSA                                         | 980  |
| IRVTSEDPGEDELASEWELPTPKTIFPKPCDWRKQEFINWTKLQSQGRGIVNFEKDKISNNWLERYRGIP                                        | 1050 |
| HRKLLTALQLRANVYPTREFLARGRQESHIS <u>CRHCQ</u> AENESSAHIIGYCPAVQ <u>DARIKRHNQLCDLLIA</u>                        | 1120 |
| C-ZnF                                                                                                         |      |
| <u>EVKKKDWIVFQEPIPKDKQNEIYKPDLVFVKGDQAMVVDVTIRYESKPSSLADAAAEKVKKYCHLKSQIQ</u>                                 | 1190 |
| Endonuclease (EN)                                                                                             |      |
| <u>ELTNAKNIKFKGFVPGARGKWYSGNYEVLKALGISTSQRQERARCMSRKALFTSVDDIIHIFVSHSRSLVR</u>                                | 1260 |
| <u>T</u>                                                                                                      | 1261 |
| <b>R2Ag Reverse-Transcriptase (R2Ag RT) – SEQ ID NO:6</b>                                                     |      |
| ECRTVLISKSSLPERQMDINNWRPITIGSIILRLFSRILTARLAKACPINPRQKGFIRSAGCAENKLLQ                                         |      |
| LLIKNAKKEHRPLGVVFDLAKAFDTVSHHHLIAVLKQKGVDDHHIITLITNMYQNTRTRITMKNEQSDPI                                        |      |

|                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| GIQIGVKQGDPMSPLLFNLSVDPLLCKLEEEGSGYQHDSKGVTSMAFADDLVLLSGSWEGMQSNIKILEA<br>FCELTGLRTQGEKCHGFYIQPTKDSYTIINDCPPWQINGTPLTMIEPGSSEKYLGLRI                                                                                                                                                                                                                                                                         |
| <b>R2Ag 3'UTR – SEQ ID NO:7 (in DNA sequence); underlined is R2Ag 3' Core (302-371)</b>                                                                                                                                                                                                                                                                                                                      |
| TTATTGTTATGTATTCTGGTTGTATTGTTTTTGCTGTATTACTGTTGATTGTTCTTATTGTATCTATT<br>GCTCTGTCTCCTATTGTTCTTGTGTTTTATTGTTATTTGTTGCCTTGCTGTTTGAATATTGATTTTCAATA<br>AAACTACCCCTGGTACTCCTGCCCTGTTACCCTGGCCCTGGAAAGGTTGGGAGAGTAAGCTGGAACCTCGT<br>ATTTCATGACGTGTACATCAAGCCAAATTTTAAAGACAAAATATGAATTGGAATTGACAAATTTACCAA<br>TTTGGACCACTTACTGGACTAGACACAGGTGGACGGCCACCTTTACTTAACTCGGAAAAGAAACACATA<br><u>TACTTTTATGTGTTTGTAAAA</u> |
| <b>R2Ag 5'UTR – SEQ ID NO:8 (in DNA sequence)</b>                                                                                                                                                                                                                                                                                                                                                            |
| ATGGGATCATAGTTACTAACTGGGCTGAGCTGCAGAGGTTACACCTCCTCGTGGTCCCCTGGATGCCCT<br>TCGGGGTAACCAAGTTAACCTCTGCCCCAGTGGGGGAAGTATAACTACTTCCCTCCCTATCGTCTGGCCTC<br>GCCAACCATTTCGATAGATTGAACCGGTGATATAAACTCCTGGCAACGACTTTCAGTCGCTGCCCAAAGC<br>CTCGCTACGGGCACAGTTGGGCTCAGTCACACAGCACTTGAGGGTGTGTTGTTTGACTTATACTACTCC<br>TCTATCTGGACTGGCAAA                                                                                    |

[0136] As shown in the table, each of the R2Cm and R2Ag ORFs includes, from the N- to the C-terminus, three N-terminal zinc finger domains (N-ZnF1, N-ZnF2 and N-ZnF3), a SANT DNA-binding domain, a 3' UTR binding domain, a template switching domain, a reverse-transcriptase (RT), a C-terminal zinc finger domain (C-ZnF), and an endonuclease (EN).

[0137] The activities of these newly discovered R2 retrotransposons were tested with cultured human cell line, 293T. Cells were grown in Dulbecco's minimal essential medium (DMEM) supplemented with 10% fetal bovine serum (Gibco) and 1% penicillin/streptomycin (Invitrogen) in 5% CO<sub>2</sub> at 37 °C. Cells were grown as adherent cell cultures and passaged every 72 h at a 1:5 dilution (volume of cells: final volume of medium) until confluency was reached, in order to maintain log-phase growth.

[0138] Cells were seeded at 60% confluence in 24-well cluster plates one day before transfection. For transfection, OptiMEM medium (Invitrogen) was pre-warmed to room temperature. Transfection of mRNA into cells was performed with Lipofectamine® MessengerMAX™ (Invitrogen). mRNA *Lipofectamine*® MessengerMAX™ Reagent (1.5 µL of each) were mixed with 25 µL OptiMEM medium and incubated for 10 min at temperature. 0.5 µg of mRNA was mixed with OptiMEM medium in total volume 25 µL. Then, mixed 25 µL of lipofectamine reagent mixture with 25 µL mRNA mixture and incubated for 5 min at

temperature. Mixtures were applied to each well. Helper virus infected the plasmid transfected cells at the same time and continue to incubate for 72 h.

[0139] Integration of the donor sequence into the human genome was measured with quantitative PCR (qPCR) in isolated genomic DNA. The results are shown in **Table 2**. For both R2, the target insertion site is within the 28S rDNA.

**Table 2. Integration Efficiency of R2Cm and R2Ag**

| R2   | Copies of Inserted Sequence |                |
|------|-----------------------------|----------------|
|      | Wild-Type R2                | Mutant Control |
| R2Cm | +++                         | -              |
| R2Ag | +++                         | -              |

[0140] As shown, each of the newly discovered R2 retrotransposons had excellent retrotransposition activities.

**Example 2. Transfection Efficiency**

[0141] This example tested the transfection efficiency of R2Ag.

[0142] In this example, the HEK293T cells were used as the target cells, and the R2Ag retrotransposon system carried a coding sequence for GFP. Two published retrotransposons, ZoAI (Zhang *et al.*, “Harnessing eukaryotic retroelement proteins for transgene insertion into human safe-harbor loci,” *Nature Biotechnology* (2024) doi.org/10.1038/s41587-024-02137-y) and R2-Tg (Chen *et al.*, “All-RNA-mediated targeted gene integration in mammalian cells with rationally engineered R2 retrotransposons,” *Cell* (2024) 187:4674-89), were included as benchmarks. HEK cells were transfected with two mRNA in ‘trans’ with CMV-GFP-PolyA cassette or a negative control. Cells were imaged and subjected to flow cytometry analysis at Day 3 post transfection, with no observed reduction in viability relative to transfection of GFP mRNA.

[0143] A sample GFP image is shown in **FIG. 1A**. The GFP expression levels in the negative control (Reverse Transcriptase-mutant; **FIG. 1B**, top) and in the functional R2Ag (**FIG. 1B**, bottom) are shown. For comparison, **FIG. 1C** shows the 293T transfection efficiency of R2-Tg (adapted from Zhang *et al.*, 2024, Extended Data Fig. 3(i)). As shown in the figures, R2Ag achieved a ~24.5% transfection efficiency, considerably higher than the 10% transfection efficiency of the benchmark ZoAI.

[0144] In a similar experiment, the transfection efficiency of R2Ag in comparison to R2-Tg was tested in primary T cells. The results are shown in **FIG. 2A** (R2Ag) and **FIG. 2B** (adapted from Chen et al., 2024, Fig. 5k). Again, R2Ag mediated superior gene delivery efficiency over R2-Tg.

[0145] The transfection efficiency of R2Ag was further tested in primary human hepatocytes (PHH). As shown in **FIG. 3**, the GFP cargo was delivered by R2Age into primary human cells resulting in 1.1% and 0.56% GFP+ cells in T cells and hepatocytes, respectively.

### Example 3. R2Ag *trans* and *cis* Configurations

[0146] This example demonstrates that the R2Ag retrotransposition system works in both *trans* and *cis* configurations.

[0147] A *trans* configuration is illustrated in **FIG. 4**, upper panel, which includes two separate RNA molecules. One of the molecules includes the R2 ORF and the other includes a transgene flanked by the 5'- and 3'-UTR, along with other useful elements such as the target homology region. In a *cis* configuration (lower panel), the transgene is included, in either forward or reverse orientation, downstream of the R2 ORF.

[0148] The transfection efficiencies of the *cis* configuration, in which the transgene was disposed in both forward and reverse orientations, were tested in HEK293T cells. Capped RNA was transcribed with the mMESSAGE mMACHINE T7 Transcription Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. DNA templates were PCR-amplified using KOD One (MERCK), and 28S flanking and 22 adenines (22As) were added by primers. HEK293T was reverse-transfected with 500 ng RNA using Lipofectamin MessengerMAX. Cells were collected 2 d after transfection and analyzed by CytoFLEX Flow Cytometer (Beckman).

[0149] The results are shown in **FIG. 5**. Both the forward and reverse orientations achieved high delivery and expression efficiencies.

### Example 4. Minimum 3'UTR Sequence

[0150] This example sought to identify the minimum 3'UTR sequence required for transposition efficiency.

[0151] Five different trans constructs were prepared, as illustrated in **FIG. 6**. Relative to the first construct on the top, constructs #2 (150-Rev-630) and #3 (300-Rev-630) had deletions in the 5' region, and constructs #4 (560-Rev-160) and #5 (560-Rev-320) had deletions in the 3' UTR. Interestingly, deletion of the partial 3'UTR sequences while keeping a 70-nt core (nucleotides 302-371 of SEQ ID NO:7) (#4) resulted in higher GFP signal, indicating improved efficiency. Sequence alignment shows that the 3'UTR includes a highly homologous 71-nt core (nucleotides 194-264 of SEQ ID NO:3).

[0152] On the other hand, truncation or mutation of the 5'UTR sequence did not alter the transposition efficiency or specificity.

#### **Example 5. Confirmation of Full-Length Integration**

[0153] This example demonstrates that R2Ag can integrate full-length cargos as confirmed by long-read sequencing.

[0154] Upon R2Ag-mediated transfection as described above, a limited-dilution cell sample was expanded and maintained over 27 days in culture that possessed 50% GFP+ cells at the time of genomic DNA extraction. High molecular weight gDNA was extracted and sequenced using Oxford Nanopore long-read sequencing, which resulted in a ~30x whole genome coverage with an N50 of 10Kb. The reads were initially mapped to the 7.15Kb reference with both the insert and flanking 28s rDNA sequences (shown in **FIG. 7**), and then remapped to the 2.1Kb GFP insert reference. Among the total 58 reads mapped, 14 short ones (around or shorter than 200bp) were removed. For the remaining 44 reads, 23 full-length sequences were identified, at a 52% rate of full-length integration.

#### **Example 6. Determination of R2Ag Target Specificity and Efficiency**

[0155] This example screened for integration of R2Ag in the HEK genome and calculated the percentage of on-target vs. off-target integrations.

[0156] HEK cells were cultured in 6-well plate format and retrotransposition was performed using transient lipofectamine transfection of a 2-RNA system with 500 ng R2Ag helper RNA and a 1.8 µg of a donor mRNA. At Day 3 after transfection, genomic DNA was isolated. Integration site location and relative abundance were determined by S-EPTS/LM-PCR, next generation sequencing and bioinformatical analysis at Protogene (Burlington,

MA). Briefly, identification of integration sites involved tabulation of the chromosome number, sequence orientation (plus or minus strand) and locus of the IS (based on human reference genome hg38). Data is depicted to illustrate the on-target site of integration (at the human 28S genomic locus) or identified off-target integration sites.

[0157] The results are shown in **FIG. 8**. R2Ag yielded in 99.5% integration efficiency with all of off-target site integration less than 0.2%.

[0158] The integration efficiency of R2Ag as compared to reference R2 was also evaluated in this example. Transfection was conducted in 6-well plate format, with ~500 ng (~0.65 pmol) helper mRNA in addition to donor RNA. One or three days following RNA transfection, quantitative PCR was used to determine the amount of the 3' junction upon integration (the wild-type 28S was used as control).

[0159] As shown in **FIG. 9**, R2Ag exhibited an integration efficiency greatly higher than R2Ol (R2 isolated from the HdrR strain of Japanese medaka fish, *Oryzias latipes*).

\*       \*       \*

[0160] Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs.

[0161] The inventions illustratively described herein may suitably be practiced in the absence of any element or elements, limitation or limitations, not specifically disclosed herein. Thus, for example, the terms “comprising”, “including,” “containing”, etc. shall be read expansively and without limitation. Additionally, the terms and expressions employed herein have been used as terms of description and not of limitation, and there is no intention in the use of such terms and expressions of excluding any equivalents of the features shown and described or portions thereof, but it is recognized that various modifications are possible within the scope of the invention claimed.

[0162] Thus, it should be understood that although the present invention has been specifically disclosed by preferred embodiments and optional features, modification, improvement and variation of the inventions embodied therein herein disclosed may be resorted to by those skilled in the art, and that such modifications, improvements and variations are considered to be within the scope of this invention. The materials, methods,

and examples provided here are representative of preferred embodiments, are exemplary, and are not intended as limitations on the scope of the invention.

**[0163]** The invention has been described broadly and generically herein. Each of the narrower species and subgeneric groupings falling within the generic disclosure also form part of the invention. This includes the generic description of the invention with a proviso or negative limitation removing any subject matter from the genus, regardless of whether or not the excised material is specifically recited herein.

**[0164]** In addition, where features or aspects of the invention are described in terms of Markush groups, those skilled in the art will recognize that the invention is also thereby described in terms of any individual member or subgroup of members of the Markush group.

**[0165]** All publications, patent applications, patents, and other references mentioned herein are expressly incorporated by reference in their entirety, to the same extent as if each were incorporated by reference individually. In case of conflict, the present specification, including definitions, will control.

**[0166]** It is to be understood that while the disclosure has been described in conjunction with the above embodiments, that the foregoing description and examples are intended to illustrate and not limit the scope of the disclosure. Other aspects, advantages and modifications within the scope of the disclosure will be apparent to those skilled in the art to which the disclosure pertains.

**CLAIMS:**

1. A method for inserting an exogenous sequence to a genomic DNA of a cell, comprising introducing to the cell
  - (a) a polypeptide or a first polynucleotide encoding the polypeptide, wherein the polypeptide comprises
    - (i) a DNA binding domain or a CRISPR Cas protein,
    - (ii) a reverse-transcriptase comprising the amino acid sequence of SEQ ID NO:2 or 6, or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:2 or 6, and
    - (iii) an endonuclease or a nickase, and
  - (b) an RNA comprising (i) a template for the exogenous sequence and (ii) a 3' UTR, or a second polynucleotide encoding the RNA.
2. The method of claim 1, which comprises introducing the first polynucleotide and the second polynucleotide to the cell.
3. The method of claim 2, wherein the first polynucleotide and the second polynucleotide are on a same vector or on separate vectors.
4. The method of any one of claims 1 to 3, wherein the reverse-transcriptase comprises the amino acid sequence of SEQ ID NO:6 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:6.
5. The method of claim 4, wherein the 3'UTR comprise nucleotides 302-371 of SEQ ID NO:7, or a nucleic acid sequence having at least 85% sequence identity to nucleotides 302-371 of SEQ ID NO:7.
6. The method of claim 5, wherein the 3'UTR comprise SEQ ID NO:7, or a nucleic acid sequence having at least 85% sequence identity SEQ ID NO:7.
7. The method of claim 5, wherein the RNA of (b) does not include at least 50 nucleotides among nucleotides 1-301 of SEQ ID NO:7 upstream of nucleotides 302-371 of SEQ ID NO:7.

8. The method of any one of claims 1 to 3, wherein the reverse-transcriptase comprises the amino acid sequence of SEQ ID NO:2 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:2.
9. The method of claim 8, wherein the 3'UTR comprise nucleotides 194-264 of SEQ ID NO:3, or a nucleic acid sequence having at least 85% sequence identity to nucleotides 194-264 of SEQ ID NO:3.
10. The method of claim 9, wherein the 3'UTR comprise SEQ ID NO:3, or a nucleic acid sequence having at least 85% sequence identity SEQ ID NO:3.
11. The method of claim 9, wherein the RNA of (b) does not include at least 50 nucleotides among nucleotides 1-193 of SEQ ID NO:3 upstream of nucleotides 194-264 of SEQ ID NO:3.
12. The method of any preceding claim, wherein the polypeptide comprises the endonuclease.
13. The method of claim 12, wherein the endonuclease comprises amino acid residues 1208-1364 of SEQ ID NO:1 or amino acid residues 1175-1261 of SEQ ID NO:5 or an amino acid sequence having at least 85% sequence identity to amino acid residues 1208-1364 of SEQ ID NO:1 or amino acid residues 1175-1261 of SEQ ID NO:5.
14. The method of any one of claims 1-13, wherein the polypeptide comprises the nickase, which optionally is Cas nickase.
15. The method of any preceding claim, wherein the CRIRPR Cas protein is selected from the group consisting of Cas9, Cas12 and Cas13.
16. The method of any one of claims 1-14, wherein the DNA binding domain is selected from the group consisting of a transcription activator-like effector (TALE) DNA binding domain, a homeodomain protein, a zinc finger, a helix-turn-helix, a leucine zipper, and a DNA-binding domain of a homing endonuclease, a transposon or retro-transposon.

17. The method protein of claim 16, wherein the DNA binding domain is a TALE.
18. The method of claim 16, wherein the homing endonuclease is selected from the group consisting of I-SceI, I-CreI and I-DmoI.
19. The method of claim 16, wherein the retro-transposon is a long interspersed nuclear element (LINE), optionally selected from the group consisting of R1, R2, R4, R5, R6, R7, R8, and R9.
20. The method of any one of claims 1-11, wherein the polypeptide comprises the amino acid sequence of SEQ ID NO:1 or 5, or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:1 or 5.
21. The method of claim 20, wherein the polypeptide comprises the amino acid sequence of SEQ ID NO:1 or 5.
22. The method of any preceding claim, wherein the RNA further comprises a 5'UTR.
23. The method of claim 22, wherein the 5'UTR comprises the sequence of SEQ ID NO:4 or 8.
24. The method of any preceding claim, wherein the RNA further comprises one or two flanking sequences that flank a target site on a target genome at which the exogenous sequence is expected to be inserted.
25. The method of claim 24, wherein the target site is a site within a 28S rDNA gene.
26. The method of any preceding claim, wherein the cell is a mammalian cell, preferably a human cell.
27. The method of any preceding claim, wherein the exogenous sequence encodes a protein, preferably a therapeutic protein.

28. A chimeric polypeptide, comprising (i) a DNA binding domain or a CRISPR Cas protein, (ii) a reverse-transcriptase comprising the amino acid sequence of SEQ ID NO:2 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:2, and (iii) an endonuclease or a nickase, wherein the chimeric protein does not include any one of (a) amino acid residues 205-226 of SEQ ID NO:1, (b) amino acid residues 232-255 of SEQ ID NO:1, (c) amino acid residues 269-290 of SEQ ID NO:1, and (d) amino acid residues 317-357 of SEQ ID NO:1, wherein at least (i) or (ii) is heterologous to SEQ ID NO:1.
29. A chimeric polypeptide, comprising (i) a DNA binding domain or a CRISPR Cas protein, (ii) a reverse-transcriptase comprising the amino acid sequence of SEQ ID NO:6 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:6, and (iii) an endonuclease or a nickase, wherein the chimeric protein does not include any one of (a) amino acid residues 145-166 of SEQ ID NO:1, (b) amino acid residues 175-195 of SEQ ID NO:1, (c) amino acid residues 209-230 of SEQ ID NO:1, and (d) amino acid residues 257-297 of SEQ ID NO:5, wherein at least (i) or (ii) is heterologous to SEQ ID NO:5.
30. A method for inserting an exogenous sequence to a genomic DNA of a cell, comprising introducing to the cell (a) the polypeptide of claim 28 or 29 or a first polynucleotide encoding the polypeptide, and (b) an RNA comprising (i) a template for the exogenous sequence and (ii) a 5' UTR and/or a 3' UTR, or a second polynucleotide encoding the RNA.
31. An RNA or a DNA vector encoding the RNA, wherein the RNA comprises:
- (a) a first fragment encoding the sequence of SEQ ID NO:5 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:5;
  - (b) a 3'UTR comprising nucleotides 302-371 of SEQ ID NO:7 or a nucleic acid sequence having at least 85% sequence identity to 302-371 of SEQ ID NO:7; and
  - (c) a second fragment disposed between (a) and (b), which encodes a heterologous protein.
32. The RNA or DNA vector of claim 31, wherein the RNA does not include at least 50 nucleotides among nucleotides 1-301 of SEQ ID NO:7 upstream of nucleotides 302-371 of SEQ ID NO:7.

33. The RNA or DNA vector of claim 31 or 32, wherein the RNA does not include at least 50 nucleotides of SEQ ID NO:8 upstream of the first fragment.
34. An RNA or a DNA vector encoding the RNA, wherein the RNA comprises:
- (a) a first fragment encoding the sequence of SEQ ID NO:1 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:1;
  - (b) a 3'UTR comprising nucleotides 194-264 of SEQ ID NO:3 or a nucleic acid sequence having at least 85% sequence identity to 194-264 of SEQ ID NO:3; and
  - (c) a second fragment disposed between (a) and (b), which encodes a heterologous protein.
35. The RNA or DNA vector of claim 34, wherein the RNA does not include at least 50 nucleotides among nucleotides 1-193 of SEQ ID NO:3 upstream of nucleotides 194-264 of SEQ ID NO:3.
36. The RNA or DNA vector of claim 34 or 35, wherein the RNA does not include at least 50 nucleotides of SEQ ID NO:4 upstream of the first fragment.
37. The RNA or DNA vector of any one of claims 31-36, wherein the RNA further comprises one or two flanking sequences that flank a target site on a target genome at which the heterologous sequence is expected to be inserted.
38. The RNA or DNA vector of claim 37, wherein the target site is a site within a 28S rDNA gene.
39. An RNA or a DNA vector encoding the RNA, wherein the RNA comprises:
- (a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:5 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:5; and
  - (b) a 3'UTR comprising nucleotides 302-371 of SEQ ID NO:7 or a nucleic acid sequence having at least 85% sequence identity to 302-371 of SEQ ID NO:7, wherein the RNA does not include:
    - (1) at least 50 nucleotides among nucleotides 1-301 of SEQ ID NO:7 upstream of nucleotides 302-371 of SEQ ID NO:7;
    - (2) at least 50 nucleotides of SEQ ID NO:8 upstream of the ORF; or

(3) either or both sequences upstream of the ORF and downstream of the 3'UTR complementary to a 28S rDNA gene.

39. An RNA or a DNA vector encoding the RNA, wherein the RNA comprises:

(a) an open reading frame (ORF) encoding the sequence of SEQ ID NO:1 or an amino acid sequence having at least 85% sequence identity to SEQ ID NO:1; and

(b) a 3'UTR comprising nucleotides 194-264 of SEQ ID NO:3 or a nucleic acid sequence having at least 85% sequence identity to 194-264 of SEQ ID NO:3,

wherein the RNA does not include:

(1) at least 50 nucleotides among nucleotides 1-193 of SEQ ID NO:3 upstream of nucleotides 194-264 of SEQ ID NO:3;

(2) at least 50 nucleotides of SEQ ID NO:4 upstream of the ORF; or

(3) either or both sequences upstream of the ORF and downstream of the 3'UTR complementary to a 28S rDNA gene.