



**Declarations under Rule 4.17:**

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

**Published:**

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

## COMPOSITIONS AND METHODS FOR HIGH-EFFICIENCY RECOMBINATION OF RNA MOLECULES

### CROSS-REFERENCE TO RELATED APPLICATIONS

5           The present application is a continuation-in-part of PCT/US2020/025430, filed March 27, 2020, which claims priority to US Provisional Application No. 62/933,714 filed November 11, 2019, all herein incorporated by reference in their entireties.

### FIELD

10           The present disclosure provides systems, kits, compositions, and methods that allow for recombination of two or more RNA molecules, allowing expression of a full-length protein.

### BACKGROUND

15           Gene therapy is a promising method for treating genetic diseases caused by loss-of-function mutations. Replacement genes are typically reintroduced into target cells using vectors such as AAV because the virus is generally safe and efficient at entering cells. However, in the case of AAV it is difficult to encapsulate more than about 5000 nucleotides using conventional capsids. Since the length of genes that encode large proteins often exceed the packaging constraints of AAV, many genetic diseases remain untreatable. Strategies to overcome this limitation have been explored in the past, but  
20           proved inefficient, led to expression of high levels of potentially toxic truncated protein, or both. Safe, high efficiency strategies for delivery of large proteins to treat disease are needed.

### SUMMARY

25           Provided herein are compositions for expressing a target protein. In one example, the composition includes (a) a first RNA molecule, the RNA molecule comprising from 5' to 3': (i) a coding sequence for an N-terminal portion of the target protein; (ii) a splice donor; and (iii) a first dimerization domain; and (b) a second RNA molecule, the RNA molecule comprising from 5' to 3': (i) a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice acceptor; and (v) a coding  
30           sequence for a C-terminal portion of the target protein.

          In some examples, the first and second dimerization domains bind by direct binding, indirect binding, or both.

          In some examples, the dimerization domains are kissing loop domains or hypodiverse domains.

          In some examples, the first and/or second RNA molecule comprise at least one splice enhancer.

Also provided are compositions for expressing a target protein comprising: (a) a first synthetic DNA molecule that encodes the first RNA molecule of any one of claims 1 to 16, wherein the first synthetic DNA molecule comprises (i) a first promoter operably linked to a sequence encoding the first RNA molecule; and (b) a second synthetic DNA molecule that encodes the second RNA molecule of any one of claims 1 to 16, wherein the second synthetic DNA molecule comprises (i) a second promoter operably linked to a sequence encoding the second RNA molecule.

Also provided are systems for expressing a target protein comprising the described compositions.

Also provided are methods of using the disclosed systems or the RNAs encoded by the systems to express a protein in a cell. Such a method can include introducing the system into a cell, and expressing the synthetic first and second RNA molecules in the same cell. In some examples, the cell is in a subject, and the method treats a disease in the subject such as a genetic disease caused by a mutation in a gene encoding the target protein. In some examples the genetic disease is Duchenne Muscular Dystrophy, Hemophilia A, Stargardt's Disease, or Usher Syndrome.

The foregoing and other objects and features of the disclosure will become more apparent from the following detailed description, which proceeds with reference to the accompanying figures.

### BRIEF DESCRIPTION OF THE DRAWINGS

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

FIG. 1A depicts a schematic of vector designs (left) and RNA interactions and splicing (right). Left: 5' trans-splice (trsp) DNA vector: Open arrows are two opposing promoters. RFP coding domain and 3'UTR with poly adenylation elements are expressed opposite from the N-terminal portion of YFP (n-yfp), followed by a splice donor sequence (SD), a downstream intronic splicing enhancer (DISE), and two intronic splicing enhancers (2xISE), a binding domain (BD, also referred to as dimerization domain), and a stable stem loop BoxB element (boxB), a self-cleaving hammerhead ribozyme (HHRz), ending with a 3' UTR containing poly adenylation elements. The n-yfp segment has a small intron inserted (white segment within n-yfp). 3' trsp DNA vector: Open arrows are two opposing promoters. BFP coding domain and 3'UTR with poly adenylation elements are expressed opposite from complementary binding domain (anti-BD, also referred to as dimerization domain), followed by three intronic splicing enhancer sequences (3xISE), a branch point (BP), a polypyrimidine tract (PPT), a splice acceptor sequence (SA), the c-terminal portion of the YFP coding sequence, ending with a 3' UTR containing poly adenylation elements. Right: pre-mRNA interactions (5' trsp-RNA + 3' trsp-RNA) and trans-splicing to generate an mRNA encoding YFP protein are shown.



FIG. 1B depicts transfection of only the N-terminal expression plasmid does not lead to YFP fluorescence.

FIG. 1C depicts transfection of only the C-terminal expression plasmid does not lead to YFP fluorescence.

FIG. 1D depicts expression of N-terminal and C-terminal fragments without binding domains shows low levels of YFP induction.

FIG. 1E depicts rationally designed dimerization/binding domain in a looped configuration (hypodiverse sequence consisting of either all pyrimidines or all purines that are interrupted by complementary sequences that form double stranded stem structures).

FIG. 1F depicts 3D rendering of the “looped” dimerization domain configuration.

FIG. 1G depicts negative control with no binding domain on the C-terminal half.

FIG. 1H depicts negative control with no binding domain on the N-terminal half.

FIG. 1I depicts matching binding domains in a looped configuration on both N- and C-terminal half shows strong YFP induction in 90% of the cells.

FIGS. 1J-1N depict data equivalent to that in FIGS. 1E-1I for a configuration of a binding domain with a 150 nucleotide hypodiverse sequence comprised exclusively of pyrimidine (or alternatively exclusively purine) containing sequence resulting in a fully open configuration.

FIG 1J depicts a 150 nucleotide hypodiverse pyrimidine sequence resulting in a fully open configuration for complimentary base pairing.

FIG 1K depicts a 3D rendering of the 150 nucleotide hypodiverse pyrimidine sequence from (1J).

FIG 1L depicts a control HEK293T cell transfection with the C-terminal-YFP encoding construct lacking a complimentary hypodiverse binding domain. Few transfected cells express YFP.

FIG 1M depicts a control HEK293T cell transfection with the N-terminal-YFP encoding construct lacking a complimentary hypodiverse binding domain. Few transfected cells express YFP.

FIG 1N depicts a HEK293T cell transfection with N-terminal-YFP and C-terminal-YFP constructs that both have complimentary hypodiverse dimerization binding domains. Many cells express YFP at high levels.

FIG. 1O depicts representative fluorescence images for cells shown in FIG. 1G. The positive markers for transfection (RFP+BFP) are expressed, but YFP protein is not reconstituted efficiently.

FIG. 1P depicts representative fluorescence images for cells shown in FIG. 1L. The positive markers for transfection (RFP+BFP) are expressed, and YFP protein is reconstituted at high levels in cells that are both RFP and BFP double positive.

FIG. 1Q depicts a comparison of conditions shown in FIG. 1D, FIGS. 1G-II, and FIGS. 1L-1N. N: no binding domain, Loop: looped hypodiverse binding domain configuration, Lin: linear hypodiverse configuration.

FIG. 2A depicts schematic of vector designs. The protein coding sequence of a yellow fluorescent protein (YFP) is split into an N-terminal, a middle fragment (m-yfp) and a C-terminal fragment. The junction of RNAs encoding the n and m fragments is joined by a looped design binding domain (BD1) and the junction between m and c fragments is joined by a looped binding domain (BD2). The pyrimidine (Y) and purine (R) sequences are arranged in such a way as to avoid self-circularization of the m-fragment and avoid direct recombination of the N- and C-fragment. The N-terminal fragment is co-expressed with red fluorescent protein as a transfection control, the C-terminal fragment is coexpressed with blue fluorescent protein as a transfection control. Promoter sequences are indicated with open arrows. Splice donor (SD) and splice acceptor (SA) sites are indicated. Intronic splicing elements including splice enhancers, polypyrimidine tracts and branch points are included, analogous to the elements used upstream (5') of the SA and downstream (3') of the SD in FIG. 1A.

FIG. 2B depicts human cell line transfection of plasmids I+II+III (see FIG. 2A) efficiently reconstituting high level YFP expression in 80% of the transfected cells.

FIG. 2C depicts representative fluorescent image of expression of the n and m fragment (plasmid I+II, see FIG. 2A) shows no yfp fluorescence (negative control).

FIG. 2D depicts representative fluorescent image of expression of the m and c fragment (plasmid II+III, see FIG. 2A) shows no yfp fluorescence (negative control).

FIG. 2E depicts representative fluorescent image showing that strong YFP fluorescence is induced by co-transfection of all three fragments (plasmid I+II+III, see FIG. 2A).

FIGS. 3A-3D depict efficient reconstitution of yellow fluorescent protein (YFP) from two fragments (SEQ ID NOS: 1 and 2) expressed from two AAV2/8s after systemic administration in the newborn (P3) mouse pup. (A) depicts AAV 1 encoding the n-terminal half fragment of YFP, and AAV 2 encoding the c-terminal half fragment. AAV 1+AAV 2 were mixed at equal titer and injected intravenously into mice. Tissue sample were collected 3 weeks following injection. (B) depicts YFP fluorescence in the liver of the juvenile mouse at the time of sacrifice (green). Uninjected liver is shown for comparison (control: no YFP detected). DRAQ5 nuclear stain is shown in magenta for context. (C) depicts strong YFP fluorescence in the heart muscle at the time of sacrifice (green). Top panels show macroscopic view and red autofluorescence for context (in magenta). Bottom panel shows cross-section with DRAQ5 nuclear stain for context (in magenta). Uninjected mouse heart lacking YFP is shown for control. (D) depicts strong YFP fluorescence in the skeletal muscles of the leg at the time of sacrifice. Uninjected mouse legs are shown for comparison (negative control, no YFP detected). Top panels show macroscopic view with red autofluorescence in magenta. Bottom panel shows microscopic

image of a cross-section through the leg. Bottom panel shows DRAQ5 nuclear stain in magenta for context.

FIGS. 4A-4B depict efficient reconstitution of yellow fluorescent protein (YFP) from three fragments (SEQ ID NOS: 145, 146 and 2, respectively) in the mouse tibialis anterior muscle after intramuscular injection of three AAV2/8 in the newborn (P3) mouse pup. (A) depicts a schematic of three AAV particles with separate N-, M-, and C-terminal fragments of YFP (analogous to Fig 2A). (B) Shows strong YFP fluorescence in a longitudinal section of the tibialis anterior muscle of a mouse injected with all three viral particles. DRAQ5 nuclear stain is shown in magenta for context.

FIGS. 5A-5F depict efficient reconstitution of yellow fluorescent protein (YFP) from two and from three fragments in adult mouse tibialis anterior muscle. (A) depicts N-terminal and C-terminal halves of YFP coding sequence are equipped with synthetic RNA-dimerization and recombination domains. (B) depicts two AAV transfer plasmids expressing these two fragments were electroporated transcutaneously into adult mouse tibialis anterior (TA) muscle and strong fluorescence was detected at 5 days post electroporation. (C) depicts no fluorescence was detectable in contralateral non-injected TA. (D) depicts n-terminal, middle, and c-terminal YFP coding sequence are equipped with synthetic RNA-dimerization and recombination domains linking each fragment to its adjacent fragment(s). (E) depicts transcutaneous electroporation of three AAV transfer plasmids expressing these three fragments. Strong YFP fluorescence is detected indicating efficient reconstitution of YFP from three fragments. (F) depicts fluorescence in contralateral non-injected TA. Fluorescent channel is overlaid onto grey scale photographs for context.

FIG. 6A is a schematic drawing providing an exemplary system for the disclosed RNA recombination methods, using two nucleic acid molecules 110, 150, wherein the target protein is divided into two portions and each portion is encoded by a different nucleic acid molecule. In some examples, the nucleic acid molecules 110, 150, of the system are DNA, and include promoters 112, 152. In some examples, the nucleic acid molecules 110, 150, of the system are RNA, and thus lack the promoters 112, 152. Drawing not to scale.

FIG. 6B is a schematic drawing providing an exemplary dimerization domain (*e.g.*, 122, 154 of FIG. 6A) that includes hypodiverse sequences interspersed with sequences that can form a stem, which results in local RNA loops that are open and available for basepairing in the absence of pseudoknot formation. Drawing not to scale.

FIG. 6C is a schematic drawing showing the interaction and hybridization (base pairing) between a pre-mRNA dimerization domain 122 of molecule 110 (FIG. 6A) and a pre-mRNA dimerization domain 154 of molecule 150 (FIG. 6A) allows the spliceosome components to recombine N-terminal coding sequence 114 and C-terminal coding sequence 164. This results in the 3' end of the

N-terminal protein coding sequence 114 fusing to the 5' end of the C terminal protein sequence 164, and a seamless junction between the N- and C-terminal portions. Drawing not to scale.

FIG. 6D is a schematic drawing providing an exemplary system for the disclosed RNA recombination methods, using three nucleic acid molecules 110, 200, 150, wherein the target protein is divided into three portions (N-terminal, middle, C-terminal) and each portion is encoded by a different nucleic acid molecule. Prior to transcription, nucleic acid molecules 110, 150, 200 of the system are DNA, and include promoters 112, 152, 202. Following transcription, nucleic acid molecules 110, 150, 200 of the system are RNA, and thus lack the promoters 112, 152, 202. Drawing not to scale.

FIG. 6E is a schematic drawing showing the interaction and hybridization (base pairing) between dimerization domain 122 of molecule 110 (FIG. 6D) and dimerization domain 204 of molecule 200 (FIG. 6D), and between dimerization domain 226 of molecule 200 (FIG. 6D) and dimerization domain 154 of molecule 150 (FIG. 6D), allows the spliceosome components to recombine N-terminal coding sequence 114, middle coding sequence 216, and C-terminal coding sequence 164. This results in the 3' end of the N terminal coding sequence 114 fusing to the 5' end of the middle protein sequence 216, and the 3' end of the middle coding sequence 216 fusing to the 5' end of the C-terminal sequence 164, and a seamless junction between the N-, middle, and C-terminal portions. In some examples, for example following transcription, the elements shown are RNA. Drawing not to scale.

FIG. 6F is a schematic drawing providing an exemplary system for the disclosed RNA recombination methods, using two nucleic acid molecules 110, 150, wherein the target protein is divided into two portions and each portion is encoded by a different nucleic acid molecule. In this example, the DNA has been transcribed into RNA, such that nucleic acid molecules 110, 150, of the system are RNA, and thus lack the promoters 112, 152 present in the DNA (see FIG. 6A). Drawing not to scale.

FIG. 7A is a schematic drawing providing an exemplary system for the disclosed RNA recombination methods, that like FIG. 6A uses two nucleic acid molecules 500, 600, but the dimerization domains are aptamers 512, 602, that recognize the same target molecule 700. In some examples, for example following transcription, the elements shown are RNA. Drawing not to scale.

FIG. 7B is a schematic drawing providing an exemplary system for the disclosed RNA recombination methods, that, related to FIG. 7A, uses dimerization domains that recognize the same target molecule. Here, the target recognized by the dimerization domain is a specific RNA molecule (instead of molecule 700 in FIG. 7A, e.g., protein or small molecule). Each domain recognizes a different portion of an mRNA molecule only expressed in target cells (*i.e.*, cells where target protein expression is desired), such as a cancer-specific transcript. In some examples, for example following transcription, the elements shown are RNA. Drawing not to scale.

FIG. 7C is a schematic drawing providing an exemplary system for the disclosed RNA recombination methods, that like FIG. 6A and 7A, uses two nucleic acid molecules 800, 900, and shows the dimerization domains 812, 902 hybridizing to an oligonucleotide 1000 that prevents the dimerization domains from interacting with one another, and therefore prevents or reduces recombination of the N-terminal coding sequence 802 and C-terminal coding sequence 914. In some examples, for example following transcription, the elements shown are RNA. Drawing not to scale.

FIG. 8 is a bar graph comparing reconstitution of YFP protein expression in the presence (w/) or absence (w/o) of a WPRE3 sequence in the 3' untranslated region. N=3 replicates per sample are shown.

FIG. 9A is a schematic drawing providing an example for the use of dimerization domain (*e.g.*, 122, 154 of FIG. 6A) that includes kissing loop interaction for high affinity dimerization. Using the teachings provided herein, one will appreciate that any of the disclosed coding portions (*e.g.*, YFP) can be replaced with other target protein coding sequences. Drawing not to scale.

FIG. 9B depicts RFP, BFP, and YFP signal in HEK293T cells transfected with both halves of the split YFP. Equipped with either a linear dimerization domain adhering to the hypodiverse design principle or a structured dimerization domain designed for kissing loop-loop interactions. Strong yellow fluorescent signal indicates efficient reconstitution.

FIGS. 10A-10Z are exemplary synthetic nucleic acid molecules that can be used with the systems and methods. In some examples, a synthetic nucleic acid molecule as at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, at last 99% or 100% sequence identity to the sequence of any one of SEQ ID NOS: 1 (FIGS. 10A-10B), 2 (FIGS. 10C-10E), 7 (FIG. 10E), 8 (FIG. 10F), 9 (FIG. 10G), 10 (FIG. 10H), 11 (FIG. 10I), 12 (FIG. 10J), 13 (FIG. 10K), 14 (FIG. 10L), 15 (FIG. 10M), 16 (FIG. 10N), 17 (FIG. 10O), 18 (FIG. 10P), 19 (FIG. 10Q), 20 (FIGS. 10R-10U), and 21 (FIGS. 10V-10Z), but with a different target protein coding sequence. Thus an intronic region using with any of the systems or methods provided herein can have at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, at last 99% or 100% sequence identity to any intronic sequence of SEQ ID NOS: 1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or 21. For example, FIGS. 10A-D show exemplary (A,B) first (SEQ ID NO: 1) and (C,D) second (SEQ ID NO: 2) synthetic molecules that can be used to express full-length YFP, while SEQ ID NO: 3 and 4 provide the corresponding synthetic intron portion without the YFP coding portion. In some examples, a synthetic intron sequence has at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, at last 99% or 100% sequence identity to SEQ ID NO: 3 or 4. Thus, the coding sequence portion of any synthetic molecule provided herein (*e.g.*, nt 544 to 1032 of SEQ ID NO: 1 and nt 905 to 1141 of SEQ ID NO: 2), can be replaced with another coding sequence portion.

FIG. 11 is a bar graph showing the reconstitution efficiency of different length random complimentary base-pairing binding domains (50 bp, 100 bp, 150 bp, 200 bp, 300 bp, 400 bp, and 500 bp). YFP median fluorescence intensity is compared between cells with matching RFP and BFP transfection levels. n=3 samples per condition. n=3 samples per condition.

FIGS. 12A-12B show that inclusion of a splice enhancer into the synthetic intron increases the reconstitution efficiency. FIG. 12A is a schematic drawing of the 5'-N and 3'-C-terminal constructs used (SEQ ID NO: 1 and 2). (see FIG. 1A for abbreviations). FIG. 12B is a bar graph showing the resulting YFP fluorescence following transfection of SEQ ID NO: 1 and 2 into cells, or various truncations thereof indicated by Δ. n=3 samples per condition.

FIGS. 13A-13D shows midline-crossing cortical neuron tracing by reconstitution of full-length flp recombinase (Flpo) from two fragments (SEQ ID NOS: 147 and 148). (A) Schematic representation of the 5'- and 3'-sequences used to reconstitute flpo (analogous to constructs in Fig 12A) (B) Schematic representation of a flp-reporter mouse line injected with the N- and C-flpo encoding AAV virus injected into left and right regions of the cortex, respectively. (C and D) show neuronal cell body and axon labeling of cortical neurons that project to the contralateral hemisphere of the brain and therefore were infected by both the N-flpo and C-flpo viruses. Hoechst staining (nuclei) is shown for context.

FIGS. 14A-14D show expression of oversized cargo (i.e. proteins encoded by long RNAs) in cell culture and *in vivo* in the mouse primary motor cortex. (A) Schematic representation of the 5'- and 3'-sequences used to reconstitute YFP, which include long stuffer sequences (uninterrupted open reading frames; SEQ ID NOS: 22 and 23, respectively). (B) Quantitative real-time PCR analysis of reconstitution efficiency of the oversize YFP constructs in HEK 293t cells. N=3 per condition. (C) Reconstituted YFP protein expression from full-length oversized YFP expression and split-REJ expression assessed by flow cytometry of transiently transfected HEK 293t cells. Median yellow fluorescence intensity is compared between cell populations with equal transfection control (blue and red) fluorescence for the different conditions. Y-axis shows median yellow fluorescence intensity [a.u.]. N=3 per condition. (D) Schematic of injections into mouse primary motor cortex, and images of brain tissue 10 days following injection, showing successful reconstitution of a long (2401 aa) YFP protein *in vivo*.

FIGS. 15A-15C show efficient reconstitution of full-length human coagulation factor VIII (FVIII) with N-terminal HA tag (substituting the N-terminal signal peptide) (2317 aa). (A) Schematic representation of the 5'- and 3'-sequences used to reconstitute FVIII (SEQ ID NOS: 24 and 25, respectively). (B) PCR amplification of the junction. (C) Western blot showing expression of FVIII. Lanes 1-3: expression of full-length FVIII (290kDa band shows full length, unprocessed FVIII). Lanes 4-6: expression of reconstituted FVIII (band at 290kDa shows successfully reconstituted FVIII). Lanes

7 and 8: expression of the N-terminus only shows absence of full-length FVIII band at 290 kDa. For all lanes: Expected proteolytic processing products are observed ranging from ~75kDa to ~210kDa. FVIII is probed for using a mouse anti-HA primary antibody. All lanes were loaded with 5micrograms of cleared cell protein extract. GAPDH (rabbit anti-GAPDH) is probed for as loading control.

5        FIGS. 16A-16F show efficient reconstitution of full-length human Abca4 with C-terminal FLAG-tag (2300 aa). (A) Schematic representation of the 5' - and 3' -sequences used to reconstitute Abca4 (SEQ ID NOS: 20 and 21, respectively), and a Sanger sequencing trace across the junction. (B) PCR amplification of the junction. (C) Schematic representation of the probes used to assay recombination of the 5' - and 3' -fragments. (D) PCR quantification of reconstitution efficiency after  
10 two days of expression in HEK 293t cells. N=2 per condition. (E) Western blot showing expression of Abca4. Lanes 1-3: expression of full-length Abca4 (~260kDa band shows full length Abca4). Lanes 4-6: expression of reconstituted Abca4 (band at 260kDa shows successfully reconstituted Abca4). Lanes 7 and 8: no transfection control (*i.e.*, HEK 293t lysate only) shows absence of any signal. Abca4 is probed for using a mouse anti-FLAG primary antibody. All lanes were loaded with 5micrograms of  
15 cleared cell protein extract. GAPDH (rabbit anti-GAPDH) is probed for as loading control. (F) Quantification of the western blot in (E) normalized for differential BFP concentration. Data is shown as normalized to the average of full-length expression control.

FIGS. 17A and 17B provide (A) HIV-1 based kissing loop dimerization domain (N-fragment, SEQ ID NO: 139, C-fragment SEQ ID NO: 140); and (B) HIV-2 based kissing loop dimerization  
20 domain (N-fragment, SEQ ID NO: 141, C-fragment SEQ ID NO: 142).

FIGS. 18A-18C show efficient reconstitution of full-length murine Otof with C-terminal FLAG-tag (2019 aa). The DNA sequences of the 5' and 3' molecules used are shown in SEQ ID NOS: 155 and 156. (A) Western blot showing expression of Otof. Lanes 1-3: expression of full-length Otof (~250kDa band shows full length Otof). Lanes 4-6: expression of reconstituted Otof (band at 250k Da shows successfully reconstituted Otof). Lane 7: no transfection control (*i.e.*, HEK 293t lysate only) shows absence of any signal. Otof is probed for using a mouse anti-FLAG primary antibody. All lanes were loaded with 5micrograms of cleared cell protein extract. GAPDH (rabbit anti-GAPDH) is probed for as loading control. (B) Raw quantification of the western blot and (C) normalized for differential BFP concentration. Data is shown as normalized to the average of full-length expression control.

30        FIGS. 19A-19C show efficient reconstitution of full-length human Myo7a with C-terminal FLAG-tag (2243 aa). The DNA sequences of the 5' and 3' molecules used are shown in SEQ ID NOS: 157 and 158. (A) Western blot showing expression of Myo7a. Lanes 1-3: expression of full-length Myo7a (~270kDa band shows full length Myo7a). Lanes 4-6: expression of reconstituted Myo7a (band at 270k Da shows successfully reconstituted Myo7a). Lane 7: no transfection control (*i.e.*, HEK 293t  
35 lysate only) shows absence of any signal. Myo7a is probed for using a mouse anti-FLAG primary

antibody. All lanes were loaded with 5micrograms of cleared cell protein extract. GAPDH (rabbit anti-GAPDH) is probed for as loading control. (B) Raw quantification of the western blot and (C) normalized for differential BFP concentration. Data is shown as normalized to the average of full-length expression control.

5           FIGS. 20A-20D show efficient reconstitution of full-length DCas9-VPR (1951 aa). The DNA sequences of the 5' and 3' molecules used are shown in SEQ ID NOS: 159 and 160. (A) Western blot showing expression of DCas9-VPR. Lanes 1-3: expression of full-length DCas9-VPR (~250kDa band shows full length DCas9-VPR). Lanes 4-6: expression of reconstituted DCas9-VPR (band at 250kDa shows successfully reconstituted DCas9-VPR). Lane 7: no transfection control (*i.e.*, HEK 293t lysate only) shows absence of any signal. DCas9-VPR is probed for using a mouse anti-Cas9 primary antibody. All lanes were loaded with 5micrograms of cleared cell protein extract. GAPDH (rabbit anti-GAPDH) is probed for as loading control. (B) Raw quantification of the western blot and (C) normalized for differential BFP concentration. Data is shown as normalized to the average of full-length expression control. (D) Example of transcriptional activation of a YFP expressing plasmid in HEK 293t cells. Full-length (upper panels) or two-way split REJ-dual dCas9-VPR (lower panels) is transiently transfected together with non-targeting guide RNA (left panels) or UAS-targeting guide RNA (right panels) expressing plasmids. All cells are also transfected with a UAS-YFP plasmid that is transcriptionally inactive until dCas9-VPR is targeted to the upstream region of a minimal promoter which results in expression of yellow fluorescent protein. Red fluorescent protein (RFP) is expressed with the N-terminal fragment of dCas9-VPR, Blue fluorescent protein (BFP) is expressed with the full-length dCas9-VPR or the C-terminal fragment of dCas9-VPR, respectively. RFP and BFP serve as transfection control. Upon expression of both full-length as well as two-way split dCas9-VPR paired with a UAS-targeting guide RNA, yellow fluorescent protein expression is observed, confirming functionality of the reconstituted full-length protein.

25           FIGS. 21A-21D show efficient reconstitution of full-length humanized Prime Editor (2118 aa). The DNA sequences of the 5' and 3' molecules used are shown in SEQ ID NOS: 161 and 162. (A) Western blot showing expression of Prime Editor. Lanes 1-3: expression of full-length Prime Editor (~260kDa band shows full length Prime Editor). Lanes 4-6: expression of reconstituted Prime Editor (band at 260k Da shows successfully reconstituted Prime Editor). Lane 7: no transfection control (*i.e.*, HEK 293t lysate only) shows absence of any signal. Prime Editor is probed for using a mouse anti-Cas9 primary antibody. All lanes were loaded with 5micrograms of cleared cell protein extract. GAPDH (rabbit anti-GAPDH) is probed for as loading control. (B) Raw quantification of the western blot and (C) normalized for differential BFP concentration. Data is shown as normalized to the average of full-length expression control. (D) Shows Prime Editor induced G to T transversion mutations induced in the FANCF and the VEGFA3 loci of HEK293t cells. The top panel shows the sequence



context for the FANCF and VEGFA3 loci respectively. The grey arrow indicates the sequence targeted by the prime editor guide RNA (pegRNA). The protospacer adjacent motif (PAM) is indicated with a grey box. The G that is targeted for transversion to T is highlighted in the sequence. Genomic loci are sequenced using Sanger sequence in three conditions. The top panel shows a representative sanger trace for unedited wild type condition. The second from the top panel shows a representative sanger trace that represents the full-length expressed prime editor construct. The area highlighted with the black box shows the appearance of a T band in the sanger sequence, indicative of successful incorporation of the edit in a portion of the cells. The lowest panels show representative sanger traces for cells edited with a two-way split reconstituted prime editor. The appearance of a T trace (black box) demonstrates functionality of the prime editor when reconstituted from two fragments.

FIGS. 22A-22C show efficient reconstitution of full-length humanized Cytosine Base Editor (AncBE4) (1854 aa). The DNA sequences of the 5' and 3' molecules used are shown in SEQ ID NOS: 163 and 164. (A) Western blot showing expression of AncBE4. Lanes 1-3: expression of full-length AncBE4 (~230kDa band shows full length AncBE4). Lanes 4-6: expression of reconstituted AncBE4 (band at 230k Da shows successfully reconstituted AncBE4). Lane 7: no transfection control (*i.e.*, HEK 293t lysate only) shows absence of any signal. AncBE4 is probed for using a mouse anti-Cas9 primary antibody. All lanes were loaded with 5micrograms of cleared cell protein extract. GAPDH (rabbit anti-GAPDH) is probed for as loading control. (B) Raw quantification of the western blot. Data is shown as normalized to the average of full-length expression control. (C) Shows AncBE4 induced C to T transition mutations induced in the EMX1 and the HEK site 3 loci of HEK293t cells. The top panel shows the sequence context for the EMX1 and HEK site 3 loci respectively. The grey arrow indicates the sequence targeted by the AncBE4 guide RNA (sgRNA). The protospacer adjacent motif (PAM) is indicated with a grey box. The Cs that are targeted for transition to T are highlighted in the sequence. Genomic loci are sequenced using Sanger sequence in three conditions. The top panel shows a representative sanger trace for unedited wild type condition. The second from the top panel shows a representative sanger trace that represents the full-length expressed AncBE4 construct. The area highlighted with the black box shows the appearance of a T band in the sanger sequence, indicative of successful incorporation of the edit in a portion of the cells. The lowest panels show representative sanger traces for cells edited with a two-way split reconstituted AncBE4. The appearance of a T trace (black box) demonstrates functionality of the AncBE4 when reconstituted from two fragments.

FIGS. 23A-23C show efficient reconstitution of full-length humanized Adenine Base Editor (Abe8e) (1606 aa). The DNA sequences of the 5' and 3' molecules used are shown in SEQ ID NOS: 165 and 166. (A) Western blot showing expression of Abe8e. Lanes 1-3: expression of full-length Abe8e (~230kDa band shows full length Abe8e). Lanes 4-6: expression of reconstituted Abe8e (band at 230k Da shows successfully reconstituted Abe8e). Lane 7: no transfection control (*i.e.*, HEK 293t

lysate only) shows absence of any signal. Abe8e is probed for using a mouse anti-Cas9 primary antibody. All lanes were loaded with 5micrograms of cleared cell protein extract. GAPDH (rabbit anti-GAPDH) is probed for as loading control. (B) Raw quantification of the western blot. Data is shown as normalized to the average of full-length expression control. (C) Shows Abe8e induced A to G transition mutations induced in the BCL11A and the HGB1/2 loci of HEK293t cells. The top panel shows the sequence context for the BCL11A and HGB1/2 loci respectively. The grey arrow indicates the sequence targeted by the Abe8e guide RNA (sgRNA). The protospacer adjacent motif (PAM) is indicated with a grey box. The As that are targeted for transition to G are highlighted in the sequence. Genomic loci are sequenced using Sanger sequence in three conditions. The top panel shows a representative sanger trace for unedited wild type condition. The second from the top panel shows a representative sanger trace that represents the full-length expressed Abe8e construct. The area highlighted with the black box shows the appearance of a G band in the sanger sequence, indicative of successful incorporation of the edit in a portion of the cells. The lowest panels show representative sanger traces for cells edited with a two-way split reconstituted Abe8e. The appearance of a G trace (black box) demonstrates functionality of the Abe8e when reconstituted from two fragments.

FIGS. 24A-24C Influence of downstream intronic splicing enhancers (DISE) and intronic splicing enhancers (ISE) and acceptor sequences on the efficiency of RNA end joining. (A) Schematic depiction of screen setup. The 5' fragment is an RNA molecule which is transcribed from a DNA construct using the human CMV promoter and enhancer. The RNA molecule produced contains a long stuffer open reading frame to simulate large cargo size. This stuffer sequence ends in a 2A self-cleaving peptide sequence and is followed by the coding region for the 5' fragment of a Yellow Fluorescent Protein (n-yfp). The 5' fragment of yfp ends in a splice donor site (SD). This splice donor site is followed by the 5' intronic portion of the RNA end joining module. For the purpose of determining the impact of DISE and ISE sequences on RNA end joining reaction efficiency, the 5' intronic portion is subdivided into three fragments: from 5' to 3': ds: downstream segment; m: mid intronic segment; dd: donor distal segment. The 5' intronic portion is followed by a trimodal kissing loop RNA dimerization domain. The message is terminated with a short poly adenylation signal. The overall length of this 5' RNA molecule is ~4kb to simulate a large cargo reconstitution scenario. The 3' fragment is an RNA molecule which is transcribed from a DNA construct using the human CMV promoter and enhancer. The 3' fragment starts with a trimodal kissing loop RNA dimerization domain that is complementary to the one on the 5' fragment encoding RNA molecule. The dimerization domain is followed by the 3' intronic portion of the RNA end joining module. This 3' intronic portion is subdivided into three segments: ad: acceptor distal segment; m: mid-intronic segment; ap: acceptor proximal segment. The acceptor proximal segment contains variations of the branch point and polypyrimidine tracts which are both essential for the spliceosome mediated RNA joining reaction. The

splice acceptor (SA) site is followed by the 3' yfp coding sequence which is followed by a self-cleaving 2A sequence that is followed by a long stuffer open reading frame. The message is terminated by an SV40 poly adenylation signal. The overall length of the 3' RNA molecule is ~4kb to simulate a large cargo reconstitution scenario. The association of the two RNA molecules (the 5' fragment and the 3' fragment) is mediated by the trimodal kissing loop RNA dimerization domain, the recruitment of the spliceosome and the RNA end joining reaction are mediated by the intronic segments. Successful RNA end joining results in reconstitution of the yfp open reading frame and subsequent translation of YFP. (B) Median YFP fluorescence intensity as determined by flow cytometry is shown for a number of intron configurations. In the first grouping (bars 1 to 9) a selection of potential downstream intronic splicing enhancer sequences were paired with a consensus splice donor site (GTAAGTATT in the DNA construct and GUAAGUAUU in the RNA sequence), shown in bars 1-8. These are compared to a consensus splice donor that is followed by scrambled sequence composed of equal parts of all four bases (ds9). In the second grouping, m1-m16 a selection of potential intronic splicing enhancers was compared to a scrambled sequence (m16). In the last grouping a selection of potential strong branch point, polypyrimidine tract, and splice acceptors were compared. The reference constructs were composed of scrambled sequence in all non-variable positions with a consensus donor followed by scrambled sequence in the ds position and a consensus splice acceptor sequence (where the whole polypyrimidine tract is composed of Ts in the DNA construct and Us in the RNA fragment respectively). (C) Listing of different DISE, ISE, and splice acceptor elements used.

## SEQUENCE LISTING

The nucleic and amino acid sequences listed in the accompanying sequence listing are shown using standard letter abbreviations for nucleotide bases, and three letter code for amino acids, as defined in 37 C.F.R. 1.822. Only one strand of each nucleic acid sequence is shown, but the complementary strand is understood as included by any reference to the displayed strand. The Sequence Listing is submitted as an ASCII text file, created on September 30, 2020, 157 KB, which is incorporated by reference herein. In the accompanying sequence listing:

**SEQ ID NOS: 1 and 2** are N- and C-terminal sequences, respectively, used to express full-length YFP. **SEQ ID NO: 1**, CMV promoter nt 1 to 543, YFP coding sequence nt 544 to 1032, synthetic intron nt 1033 to 1436, and untranslated poly A region nt 1437 to 1491. **SEQ ID NO: 2**, CMV promoter nt 1 to 522, synthetic intron nt 523 to 904, YFP coding sequence nt 905 to 1141, and nt 1142 to 1302 is the untranslated poly A region.

**SEQ ID NOS: 3 and 4** are 5'- and 3'-intronic sequences, respectively, that can be used to express a desired full-length protein, wherein a N-terminal portion of the full-length protein can be

added at nt 1 of SEQ ID NO: 3, and C-terminal portion of the full-length protein can be added at nt 382 of SEQ ID NO: 4.

**SEQ ID NOS: 5 and 6** are N- and C-terminal coding sequences, respectively, used to express full-length YFP.

5 **SEQ ID NO: 7** is an exemplary synthetic intron dimerization domain (FIG. 10E).

**SEQ ID NO: 8** is an exemplary synthetic intron without intronic splicing enhancers (FIG. 10F).

**SEQ ID NO: 9** is an exemplary synthetic intron without intronic splicing enhancers (FIG. 10G).

**SEQ ID NO: 10** is an exemplary synthetic intron without intronic splicing enhancers (FIG. 10H).

10 **SEQ ID NO: 11** is an exemplary synthetic intron without binding domain (FIG. 10I).

**SEQ ID NO: 12** is an exemplary synthetic intron with dimerization domain (FIG. 10J).

**SEQ ID NO: 13** is an exemplary synthetic intron with dimerization domain (FIG. 10K).

**SEQ ID NO: 14** is an exemplary synthetic intron without intronic splicing enhancers (FIG. 10L).

15 **SEQ ID NO: 15** is an exemplary synthetic intron with DISE only (FIG. 10M).

**SEQ ID NO: 16** is an exemplary synthetic intron without HHrz (FIG. 10N).

**SEQ ID NO: 17** is an exemplary synthetic intron without intronic splicing enhancers (FIG. 10O).

**SEQ ID NO: 18** is an exemplary U12 dependent intron with binding domain (FIG. 10P).

20 **SEQ ID NO: 19** is an exemplary U12 dependent intron with binding domain (FIG. 10Q).

**SEQ ID NOS: 20 and 21** are the N- and C-terminal DNA sequences, respectively, used to express RNAs (pre-mRNAs) resulting in full-length Abca4. In SEQ ID NO: 20, the sequence corresponding to the N-terminal Abca4 coding region is at nt 22 to 3702, and nt 3703 to 3912 is the synthetic intron, and 3921 to 3969 is the untranslated poly A region. SEQ ID NO: 20 also comprises a splice donor at nt 3703-3711, a Rat FGFR2 DISE at nt 3714-3737, a cTNT intronic splicing enhancer at nt 3747-3770, an M2 intronic splicing enhancer at nt 3782-3794, and a kissing loop dimerization domain at nt 3801-3975. In SEQ ID NO: 21, nt 1 to 228 is the synthetic intron, nt 229 to 3366 is the C-terminal Abca4 coding region, 3367 to 3447 is the FLAG epitope tag, and nt 3476 to 3607 is the untranslated poly A region (signal). SEQ ID NO: 21 also comprises a kissing loop dimerization domain at nt 3-114, an M2 intronic splicing enhancer at nt 121-133, a cTNT intronic splicing enhancer at nt 140-163, an M2 intronic splicing enhancer at nt 175-187, a Branch Point Motif at nt 194-201, a poly pyrimidine tract at nt 207-226, and a splice acceptor at nt 228.

**SEQ ID NOS: 22 and 23** are the N- and C-terminal DNA sequences, respectively, used to express RNAs (pre-mRNAs) resulting in a long full-length YFP, wherein each includes splice enhancers. In SEQ ID NO: 22, the N-terminal YFP coding region is nt 22 to 3702, nt 3703 to 3912 is

the synthetic intron, and 3921 to 3969 is the untranslated poly A region. SEQ ID NO: 22 also comprises a splice donor at nt 3703-3711, a Rat FGFR2 DISE at nt 3714-3737, a cTNT intronic splicing enhancer at nt 3747-3770, an M2 intronic splicing enhancer at 3782-3794, and a kissing loop dimerization domain at 3801-3975. In SEQ ID NO: 23, nt 1 to 225 is the synthetic intron, nt 226 to 3747 C-terminal YFP coding region, nt 3748 to 3912 is the untranslated poly A region. SEQ ID NO: 23 comprises a kissing loop dimerization domain at nt 3-114, an M2 intronic splicing enhancer at nt 118-130, a cTNT intronic splicing enhancer at nt 137-160, a M2 intronic splicing enhancer at nt 172-184, a Branch Point Motif at nt 191-198, a poly pyrimidine tract at nt 204-223, and a splice acceptor at nt 225.

**SEQ ID NOS: 24 and 25** are the N- and C-terminal sequences, respectively, used to express RNAs (pre-mRNAs) resulting in full-length human Factor VIII. In SEQ ID NO: 24, N-terminal FVIII coding region with N-terminal HA epitope tag nt are at nt 22 to 3561, nt 3562 to 3771 is the synthetic intron, and nt 3780 to 3828 is the untranslated poly A region. SEQ ID NO: 24 also comprises a splice donor at nt 3562-3570, a Rat FGFR2 DISE at nt 3573-3596, a cTNT intronic splicing enhancer at nt 3606-3629, an M2 intronic splicing enhancer at nt 3641-3653, and a kissing loop dimerization domain at nt 3660-3834. In SEQ ID NO: 25, nt 1 to 225 is the synthetic intron, nt 226 to 3636 is the C-terminal FVIII coding region, and nt 3665 to 3797 is the untranslated poly A region. SEQ ID NO: 25 also comprises a splice donor at nt 3703-3711, a Rat FGFR2 DISE at nt 3714-3737, a cTNT intronic splicing enhancer at nt 3747-3770, an M2 intronic splicing enhancer at 3782-3794, and a kissing loop dimerization domain at nt 3801-3975.

**SEQ ID NOS: 26-136** are exemplary splicing enhancers that can be used with the systems provided herein (*e.g.*, 118, 120, 156 of FIG. 6A).

**SEQ ID NOS: 137 and 138** are exemplary splice donor sequences.

**SEQ ID NOS: 139 and 140** are the N- and C-fragment respectively, of an HIV-1 based kissing loop dimerization domain.

**SEQ ID NOS: 141 and 142** are the N- and C-fragment, respectively, of an HIV-2 based kissing loop dimerization domain.

**SEQ ID NO: 143** is an exemplary cryptic splice acceptor sequence.

**SEQ ID NO: 144** is an exemplary branch point consensus sequence.

**SEQ ID NOS: 145 and 146** are the N- and middle sequences, respectively, used to express a full-length YFP, along with SEQ ID NO: 2 (C-terminal fragment). In SEQ ID NO: 145, nt 1 to 543 is the CMV promoter sequence, nt 544 to 849 N-terminal YFP coding region, and nt 850 to 1305 is the synthetic intron. In SEQ ID NO: 146, nt 1 to 522 is the CMV promoter sequence, nt 523 to 901 is the synthetic intron, nt 902 to 1084 is the middle YFP coding region, and nt 1085 to 1543 is the untranslated poly A region.

**SEQ ID NOS: 147 and 148** are the 5' and 3'- synthetic sequences, respectively, used to express a full-length Flpo. In SEQ ID NO: 147, nt 1 to 540 is the CMV promoter sequence, nt 541 to 1112 N-terminal Flpo coding region, and nt 1113 to 1571 is the synthetic intron. In SEQ ID NO: 148, nt 1 to 522 is the CMV promoter sequence, nt 523 to 904 is the synthetic intron, nt 905 to 1604 is the C-terminal Flpo coding region, nt 1605 to 1765 is the untranslated poly A region.

**SEQ ID NOS: 149 and 150** are exemplary hypodiverse sequences.

**SEQ ID NOS: 151 and 152** are exemplary splice donor consensus sequences.

**SEQ ID NO: 153** is an exemplary kissing loop based on the HIV-2 kissing loop dimerization domain (SEQ ID NOS: 141 and 142, FIG. 17B).

**SEQ ID NO: 154** is an exemplary Kozak enhanced start codon.

**SEQ ID NOS: 155 and 156** are exemplary constructs that can be used to express a murine Otof coding sequence in vivo. SEQ ID NO: 155 is used to produce the N-terminal Otof RNA. It comprises a human CMV enhancer and promoter at nt 1-522, a putative transcription start site at nt 523, and a poly adenylation signal at nt 4263-4311. It encodes the N-terminal Otof RNA elements as follows: 5' untranslated region including Kozak sequence nt 523-546; 5' Otoferlin coding sequence nt 547-4044; 5' synthetic intron sequence nt 4045-4142; 5' trimodal kissing loop dimerization domain nt 4143-4254; and linker at nt 4255-4262. SEQ ID NO: 156 is used to produce the C-terminal Otof RNA. It comprises a human CMV enhancer and promoter at nt 1-522, a putative transcription start site at nt 523, and a poly adenylation signal at nt 3335-3467. It encodes the C-terminal Otof RNA elements as follows: 3' trimodal kissing loop dimerization domain nt 525-636; 3' synthetic intron sequence nt 637-747; 3' Otoferlin coding sequence nt 748-3225; C-terminal 3xFlag tag nt 3226-3306; and linker at nt 3307-3334.

**SEQ ID NOS: 157 and 158** are exemplary constructs that can be used to express a human MYOSIN VIIA (Myo7a) coding sequence in vivo. SEQ ID NO: 157 is used to produce the N-terminal Myo7a RNA. It comprises a human CMV enhancer and promoter at nt 1-522, a putative transcription start site at nt 523, and poly adenylation signal at nt 4344-4392. It encodes the N-terminal Myo7A RNA elements as follows: 5' untranslated region including Kozak sequence nt 523-543; 5' Myo7a coding sequence nt 544-4125; 5' synthetic intron sequence nt 4126-4223; 5' trimodal kissing loop dimerization domain nt 4224-4335; and linker at nt 4336-4343. SEQ ID NO: 158 is used to produce the C-terminal Myo7a RNA. It comprises a human CMV enhancer and promoter at nt 1-522, a putative transcription start site at nt 523, and a poly adenylation signal at nt 3923-4055. It encodes the C-terminal Myo7a RNA elements as follows: 3' trimodal kissing loop dimerization domain nt 525-636; 3' synthetic intron sequence nt 637-747; 3' Myo7a coding sequence nt 748-3813; C-terminal 3xFlag tag nt 3814-3894; and linker at nt 3895-3922.

**SEQ ID NOS: 159 and 160** are exemplary constructs that can be used to express a full-length enzymatically dead Cas9 fused to a VPR transcriptional activator domain (dCas9-VPR) coding sequence in vivo. SEQ ID NO: 159 is used to produce the N-terminal dCas9-VPR RNA. It comprises a human CMV enhancer and promoter at nt 1-522, a putative transcription start site at nt 523, and poly

5 adenylation signal at nt 4112-4161. It encodes the N-terminal dCas9-VPR RNA elements as follows: 5' untranslated region including Kozak sequence nt 523-543; 5' dCas9-VPR coding sequence nt 544-3894; 5' synthetic intron sequence nt 3895-3992; 5' trimodal kissing loop dimerization domain nt 3993-4104; and linker nt 4105-4112. SEQ ID NO: 160 is used to produce the C-terminal dCas9-VPR RNA. It comprises a human CMV enhancer and promoter at nt 1-522, a putative transcription start site at nt

10 523, and poly adenylation signal at nt 3278-3410. It encodes the C-terminal dCas9-VPR RNA elements as follows: 3' trimodal kissing loop dimerization domain nt 525-636; 3' synthetic intron sequence nt 637-747; 3' dCas9-VPR coding sequence nt 748-3249; and linker at nt 3250-3277.

**SEQ ID NOS: 161 and 162** are exemplary constructs that can be used to express a full-length humanized Cas9 Prime Editor (Prime Editor) coding sequence in vivo. SEQ ID NO: 161 encodes the

15 N-terminal Prime Editor sequence as follows: Human CMV enhancer and promoter nt 1-522; putative transcription start site nt 523; 5' untranslated region including Kozak sequence nt 523-543; 5' Prime Editor coding sequence nt 544-3894; 5' synthetic intron sequence nt 3895-3992; 5' trimodal kissing loop dimerization domain nt 3993-4104; linker nt 4105-4112; poly adenylation signal nt 4112-4161. SEQ ID NO: 162 encodes the C-terminal Prime Editor sequence as follows: Human CMV enhancer

20 and promoter nt 1-522; putative transcription start site nt 523; 3' trimodal kissing loop dimerization domain nt 525-636; 3' synthetic intron sequence nt 637-747; 3' Prime Editor coding sequence nt 748-3750; linker nt 3751-3778; poly adenylation signal nt 3779-3911.

**SEQ ID NOS: 163 and 164** are exemplary constructs that can be used to express a full-length humanized Cytosine Base Editor (AncBE4) coding sequence in vivo. SEQ ID NO: 163 encodes the N-

25 terminal AncBE4 sequence as follows: Human CMV enhancer and promoter nt 1-522; putative transcription start site nt 523; 5' untranslated region including Kozak sequence nt 523-540; 5' AncBE4 coding sequence nt 541-2892; 5' synthetic intron sequence nt 2893-2990; 5' trimodal kissing loop dimerization domain nt 2991-3102; linker nt 3103-3110; poly adenylation signal nt 3111-3159. SEQ ID NO: 164 encodes the C-terminal AncBE4 sequence as follows: Human CMV enhancer and promoter

30 nt 1-522; putative transcription start site nt 523; 3' trimodal kissing loop dimerization domain nt 525-636; 3' synthetic intron sequence nt 637-747; 3' AncBE4 coding sequence nt 748-3957; linker nt 3958-3982; poly adenylation signal nt 3983-4115.

**SEQ ID NOS: 165 and 166** are exemplary constructs that can be used to express a full-length humanized Adenine Base Editor (Abe8e) coding sequence in vivo. SEQ ID NO: 165 encodes the N-

35 terminal Abe8e sequence as follows: Human CMV enhancer and promoter nt 1-522; putative

transcription start site nt 523; 5' untranslated region including Kozak sequence nt 523-540; 5' Abe8e coding sequence nt 541-2706; 5' synthetic intron sequence nt 2707-2804; 5' trimodal kissing loop dimerization domain nt 2805-2916; linker nt 2917-2924; poly adenylation signal nt 2925-2973. SEQ ID NO: 166 encodes the C-terminal Abe8e sequence as follows: Human CMV enhancer and promoter nt 1-522; putative transcription start site nt 523; 3' trimodal kissing loop dimerization domain nt 525-636; 3' synthetic intron sequence nt 637-747; 3' Abe8e coding sequence nt 748-3399; linker nt 3400-3427; poly adenylation signal nt 3428-3560.

**SEQ ID NO: 167** is an exemplary kissing loop domain

(GATTTTTGACCTGCTCGATTGTCCACTGCGAGCAGGTCTTTTGGAGTCGGGCGAGGCGGA  
AGCCCGACTCCTTTTGGCATGCACGCTAGCCGCGTCGTGCATGCCTTTTATC).

**SEQ ID NO: 168** is an exemplary ISE, M2 (GGGTTATGGGACC).

**SEQ ID NO: 169** is an exemplary ISE, cTNT (GGCTGAGGGAAGGACTGTCCTGGG).

**SEQ ID NO: 170** is an exemplary DISE, Rat FGFR2 (CTCTTTCTTTCCATGGGTTGGCCT).

**SEQ ID NOS: 171 and 172** are exemplary constructs that can be used to express a full-length

YFP coding sequence. **SEQ ID NO: 171** encodes the N-terminal YFP sequence as follows: Human CMV enhancer and promoter nt 1-522; putative transcription start site nt 523; 5' untranslated region including Kozak sequence nt 523-543; 5' Stuffer open reading frame nt 544-3654; self cleaving 2A sequence nt 3655-3729; 5' yellow fluorescent protein segment nt 3730-4224; 5' synthetic intron sequence (variable) nt 4225-4294; 5' trimodal kissing loop dimerization domain (uppercase): 4295-4406; linker nt 4407-4414; poly adenylation signal nt 4415-4463. **SEQ ID NO: 172** encodes the C-terminal YFP sequence as follows: Name: 3' intron screening split YFP; Human CMV enhancer and promoter nt 1-522; Putative transcription start site nt 523; 3' trimodal kissing loop dimerization domain nt 525-636; 3' synthetic intron sequence (variable) nt 637-706; 3' yfp coding sequence nt 707-940; self-cleaving 2A sequence nt 941-1006; 3' stuffer open reading frame nt 1007-4228; linker nt 4229-4265; poly adenylation signal nt 4257-4388.

**SEQ ID NOS: 173-180** are exemplary intronic splicing enhancer sequences.

**SEQ ID NO: 181** is a scrambled sequence.

**SEQ ID NOS: 182-196** are exemplary intronic splicing enhancer sequences.

**SEQ ID NO: 197-198** are scrambled sequences.

**SEQ ID NOS: 199-203** are exemplary intronic splicing enhancer sequences.

**SEQ ID NO: 204** is a scrambled sequence.

**SEQ ID NO: 205** is an exemplary branch point sequence (TACTAACA).

**SEQ ID NO: 206** is an exemplary polyadenylation signal

AATAAAATATCTTTATTTTCATTACATCTGTGTGTTGGTTTTTGTGTG.



## DETAILED DESCRIPTION

Unless otherwise noted, technical terms are used according to conventional usage. Definitions of common terms in molecular biology can be found in Benjamin Lewin, *Genes VII*, published by Oxford University Press, 1999; Kendrew *et al.* (eds.), *The Encyclopedia of Molecular Biology*, published by Blackwell Science Ltd., 1994; and Robert A. Meyers (ed.), *Molecular Biology and Biotechnology: a Comprehensive Desk Reference*, published by VCH Publishers, Inc., 1995; and other similar references.

As used herein, the singular forms “a,” “an,” and “the,” refer to both the singular as well as plural, unless the context clearly indicates otherwise. As used herein, the term “comprises” means “includes.” Thus, “comprising a nucleic acid molecule” means “including a nucleic acid molecule” without excluding other elements. It is further to be understood that any and all base sizes given for nucleic acids are approximate, and are provided for descriptive purposes, unless otherwise indicated. Although many methods and materials similar or equivalent to those described herein can be used, particular suitable methods and materials are described below. In case of conflict, the present specification, including explanations of terms, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting. All references, including patent applications and patents, and GenBank Accession Nos., are herein incorporated by reference in their entirety.

In order to facilitate review of the various embodiments of the disclosure, the following explanations of specific terms are provided:

**Administration:** To provide or give a subject an agent, such as a therapeutic nucleic acid molecule provided herein, or other therapeutic agent, by any effective route. Exemplary routes of administration include, but are not limited to, injection (such as subcutaneous, intramuscular, intradermal, intraperitoneal, intrathecal, intratumoral, intraosseous, and intravenous), transdermal, intranasal, and inhalation routes. Administration can be systemic or local.

**Aptamer:** Nucleic acid molecules (such as DNA or RNA) that bind a specific target agent or molecule with high affinity and specificity. Aptamers can be used in the disclosed nucleic acid molecules as a dimerization domain. In one example, two aptamers can bind to each other, e.g., by standard basepairing, non-canonical base pair interactions, non-base pairing interactions, or a combination thereof, to mediate dimerization. In one example, aptamers allow RNA dimerization (and subsequent recombination) only in the presence of one or more targets recognized by the aptamer. Aptamers have been obtained through a combinatorial selection process called systematic evolution of ligands by exponential enrichment (SELEX) (see for example Ellington *et al.*, *Nature* 1990, 346, 818-822; Tuerk and Gold *Science* 1990, 249, 505-510; Liu *et al.*, *Chem. Rev.* 2009, 109, 1948-1998;

Shamah *et al.*, *Acc. Chem. Res.* 2008, 41, 130-138; Famulok, *et al.*, *Chem. Rev.* 2007, 107, 3715-3743; Manimala *et al.*, *Recent Dev. Nucleic Acids Res.* 2004, 1, 207-231; Famulok *et al.*, *Acc. Chem. Res.* 2000, 33, 591-599; Hesselberth, *et al.*, *Rev. Mol. Biotech.* 2000, 74, 15-25; Wilson *et al.*, *Annu. Rev. Biochem.* 1999, 68, 611-647; Morris *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 1998, 95, 2902-2907). In such a process, DNA or RNA molecules that are capable of binding a target molecule of interest are selected from a nucleic acid library consisting of  $10^{14}$ - $10^{15}$  different sequences through iterative steps of selection, amplification and mutation. The affinity of the aptamers towards their targets can rival that of antibodies, with dissociation constants in as low as the picomolar range (Morris *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 1998, 95, 2902-2907; Green *et al.*, *Biochemistry* 1996, 35, 14413-14424).

Aptamers that are specific to a wide range of targets from small organic molecules such as adenosine, to proteins such as thrombin, and even viruses and cells have been identified (Liu *et al.*, *Chem. Rev.* 2009, 109, 1948-1998; Lee *et al.*, *Nucleic Acids Res.* 2004, 32, D95-D100; Navani and Li, *Curr. Opin. Chem. Biol.* 2006, 10, 272-281; Song *et al.*, *TrAC, Trends Anal. Chem.* 2008, 27, 108-117). For example, aptamers are available that recognize metal ions such as Zn(II) (Ciesiolka *et al.*, *RNA* 1: 538-550, 1995) and Ni(II) (Hofmann *et al.*, *RNA*, 3:1289-1300, 1997); nucleotides such as adenosine triphosphate (ATP) (Huizenga and Szostak, *Biochemistry*, 34:656-665, 1995); and guanine (Kiga *et al.*, *Nucleic Acids Res.*, 26:1755-60, 1998); co-factors such as NAD (Kiga *et al.*, *Nucleic Acids Res.*, 26:1755-60, 1998) and flavin (Lauhon and Szostak, *J. Am. Chem. Soc.*, 117:1246-57, 1995); antibiotics such as viomycin (Wallis *et al.*, *Chem. Biol.* 4: 357-366, 1997) and streptomycin (Wallace and Schroeder, *RNA* 4:112-123, 1998); proteins such as HIV reverse transcriptase (Chaloin *et al.*, *Nucleic Acids Res.*, 30:4001-8, 2002) and hepatitis C virus RNA-dependent RNA polymerase (Biroccio *et al.*, *J. Virol.* 76:3688-96, 2002); toxins such as cholera whole toxin and staphylococcal enterotoxin B (Bruno and Kiel, *BioTechniques*, 32: pp. 178-180 and 182-183, 2002); and bacterial spores such as the anthrax (Bruno and Kiel, *Biosensors & Bioelectronics*, 14:457-464, 1999).

**Binding:** An association between two substances or molecules, such as the hybridization of one nucleic acid molecule to another (or itself), such as between two dimerization domains, or the binding of an aptamer to its target. An oligonucleotide molecule binds or stably binds to another nucleic acid molecule if there are a sufficient number of complementary base pairs between the oligonucleotide molecule and the target nucleic acid to permit detection of that binding. In some examples, binding between nucleic acid molecules may occur directly. In some examples, binding between nucleic acid molecules may occur indirectly, e.g., through an intermediate molecule. Either direct binding or indirect binding may occur by standard base pairing, by non-canonical base pair interactions, by non-base pair interactions, or a combination thereof. Non-canonical base pair interactions may occur by any means of stabilization known to those of skill in the art, including but not limited to Hoogsteen base pairs and wobble base pairs. Non-base pair interactions can include binding

through an intermediate molecule. In some examples, direct binding is between kissing loop dimerization domains. In some examples, direct binding is between hypodiverse dimerization domains. In some examples, direct binding is between aptamer regions. In some examples, direct binding between aptamer regions involves non-canonical base pair interactions. In some examples, direct binding between aptamer regions involves standard base pairing and non-canonical base pair interactions. In some examples, indirect binding occurs through a nucleic acid bridge. In some examples the nucleic acid bridge is an mRNA. A nonlimiting example of a nucleic acid bridge is depicted in Fig. 7B. In some examples, indirect binding occurs through an aptamer molecule. A nonlimiting example of indirect binding through an aptamer molecule is depicted in Fig. 7A. In some embodiments, indirect binding through an aptamer molecule involves non-base pair interactions between the aptamer molecule and the binding regions. In some embodiments, indirect binding through an aptamer molecule involves non-base pair interactions between the aptamer molecule and the binding regions, and base pairing interactions between the binding regions.

**C-terminal portion:** A region of a protein sequence that includes a contiguous stretch of amino acids that begins at or near the C-terminal residue of the protein. A C-terminal portion of the protein can be defined by a contiguous stretch of amino acids (*e.g.*, a number of amino acid residues).

**Cancer:** A malignant tumor characterized by abnormal or uncontrolled cell growth. Other features often associated with cancer include metastasis, interference with the normal functioning of neighboring cells, release of cytokines or other secretory products at abnormal levels and suppression or aggravation of inflammatory or immunological response, invasion of surrounding or distant tissues or organs, such as lymph nodes, etc. "Metastatic disease" refers to cancer cells that have left the original tumor site and migrate to other parts of the body for example *via* the bloodstream or lymph system.

**Complementarity:** The ability of a nucleic acid to form hydrogen bond(s) with another nucleic acid sequence by either traditional Watson-Crick base pairing or other non-traditional types. A percent complementarity indicates the percentage of residues in a nucleic acid molecule which can form hydrogen bonds (*e.g.*, Watson-Crick base pairing) with a second nucleic acid sequence (*e.g.*, 5, 6, 7, 8, 9, 10 out of 10 being 50%, 60%, 70%, 80%, 90%, and 100% complementary). "Perfectly complementary" means that all the contiguous residues of a nucleic acid sequence will hydrogen bond with the same number of contiguous residues in a second nucleic acid sequence. "Substantially complementary" as used herein refers to a degree of complementarity that is at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99%, or 100% over a region of 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, or more nucleotides, or refers to two nucleic acids that hybridize under stringent conditions. Thus, in some examples, a first dimerization domain and a second dimerization domain have perfect complementary to one another (*e.g.*, 100%). In other

examples, a first dimerization domain and a second dimerization domain are substantially complementary to one another (*e.g.*, at least 80%).

**Contact:** Placement in direct physical association, including a solid or a liquid form.

Contacting can occur *in vitro* or *ex vivo*, for example, by adding a reagent to a sample (such as one containing cells), or *in vivo* by administering to a subject.

**Downregulated or knocked down:** When used in reference to the expression of a molecule, such as a target nucleic acid or protein, refers to any process which results in a decrease in production of the target RNA or protein, but in some examples not complete elimination of the target RNA product or target RNA function. In one example, downregulation or knock down does not result in complete elimination of detectable target nucleic acid/protein expression or activity. In some examples, downregulation or knock down of a target nucleic acid includes processes that decrease translation of the target RNA and thus can decrease the presence of corresponding proteins. The disclosed system can be used to downregulate any target nucleic acid/protein of interest.

Downregulation or knock down includes any detectable decrease in the target nucleic acid/protein. In certain examples, detectable target nucleic acid/protein in a cell or cell free system decreases by at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% (such as a decrease of 40% to 90%, 40% to 80% or 50% to 95%) as compared to a control (such an amount of target nucleic acid/protein detected in a corresponding untreated cell or sample). In one example, a control is a relative amount of expression in a normal cell (*e.g.*, a non-recombinant cell that does not include a nucleic acid molecule for RNA recombination provided herein).

**Effective amount:** The amount of an agent (such as a system providing multiple vectors, each encoding a different portion of a therapeutic protein, such as dystrophin) that is sufficient to effect beneficial or desired results. An effective amount also can refer to an amount of correctly joined RNA or therapeutic protein produced that is sufficient to effect beneficial or desired results.

An effective amount (also referred to as a therapeutically effective amount) may vary depending upon one or more of: the subject and disease condition being treated, the weight and age of the subject, the severity of the disease condition, the manner of administration and the like, which can be determined by one of ordinary skill in the art. The beneficial therapeutic effect can include enablement of diagnostic determinations; amelioration of a disease, symptom, disorder, or pathological condition; reducing or preventing the onset of a disease, symptom, disorder or condition; and generally counteracting a disease, symptom, disorder or pathological condition.

In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein, sufficient to treat a disease, such as a genetic disease or cancer. In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is amount

sufficient to increase the survival time of a treated patient, for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 99%, at least 100%, at least 200%, at least 300%, at least 400%, at least 500%, or at least 600% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to increase the survival time of a treated patient, for example by at least 6 months, at least 9 months, at least 1 year, at least 1.5 years, at least 2 years, at least 2.5 years, at least 3 years, at least 4 years, at least 5 years, at least 10 years, at least 12 years, at least 15 years, or at least 20 years (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to increase mobility of a treated patient (such as a DMD patient), for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 99%, at least 100%, at least 200%, at least 300%, at least 400%, at least 500%, or at least 600% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to increase mobility of a treated patient (such as a DMD patient), for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 99%, at least 100%, at least 200%, at least 300%, at least 400%, at least 500%, or at least 600% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to increase cognitive ability of a treated patient (such as a DMD patient), for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 99%, at least 100%, at least 200%, at least 300%, at least 400%, at least 500%, or at least 600% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to increase respiratory function of a treated patient (such as a DMD patient), for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 99%, at least 100%, at least 200%, at least 300%, at least 400%, at least 500%, or at least 600% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to increase blood clotting of a treated patient (such as a hemophilia patient), for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 99%, at least 100%, at least 200%, at least 300%, at least

400%, at least 500%, or at least 600% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to increase vision of a treated patient (such as a Usher or Stargardt patient), for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 99%, at least 100%, at least 200%, at least 300%, at least 400%, at least 500%, or at least 600% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to increase hearing of a treated patient (such as a Usher patient), for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 99%, at least 100%, at least 200%, at least 300%, at least 400%, at least 500%, or at least 600% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein).

In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to reduce calf muscle size of a treated DMD patient, for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, or at least 95% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In one embodiment, an “effective amount” of two or more synthetic nucleic acid molecules provided herein is an amount sufficient to reduce cardiomyopathy muscle size of a treated DMD patient, for example by at least 10%, at least 20%, at least 25%, at least 50%, at least 70%, at least 75%, at least 80%, at least 90%, or at least 95% (as compared to no administration of the two or more synthetic nucleic acid molecules provided herein). In some examples, combinations of these effects are achieved.

**Increase or Decrease:** A statistically significant positive or negative change, respectively, in quantity from a control value (such as a value representing no therapeutic agent, such as no administration of the two or more synthetic nucleic acid molecules provided herein). An increase is a positive change, such as an increase at least 50%, at least 100%, at least 200%, at least 300%, at least 400% or at least 500% as compared to the control value. A decrease is a negative change, such as a decrease of at least 20%, at least 25%, at least 50%, at least 75%, at least 80%, at least 90%, at least 95%, at least 98%, at least 99%, or at least 100% decrease as compared to a control value. In some examples the decrease is less than 100%, such as a decrease of no more than 90%, no more than 95%, or no more than 99%.

**Hybridization:** Hybridization of a nucleic acid occurs when two nucleic acid molecules undergo an amount of hydrogen bonding to each other. The stringency of hybridization can vary according to the environmental conditions surrounding the nucleic acids, the nature of the hybridization

method, and the composition and length of the nucleic acids used. Calculations regarding hybridization conditions required for attaining particular degrees of stringency are discussed in Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001); and Tijssen, *Laboratory Techniques in Biochemistry and Molecular Biology*—

5 *Hybridization with Nucleic Acid Probes* Part I, Chapter 2 (Elsevier, New York, 1993). The  $T_m$  is the temperature at which 50% of a given strand of nucleic acid is hybridized to its complementary strand.

**Isolated:** An “isolated” biological component (such as a nucleic acid molecule or a protein) has been substantially separated, produced apart from, or purified away from other biological components in the cell or tissue of an organism in which the component occurs, such as other cells (*e.g.*, RBCs),  
10 chromosomal and extrachromosomal DNA and RNA, and proteins. Nucleic acids and proteins that have been “isolated” include nucleic acids and proteins purified by standard purification methods. The term also embraces nucleic acids and proteins prepared by recombinant expression in a host cell as well as chemically synthesized nucleic acids and proteins.

**Kissing loop/kissing stem loop:** An RNA structure that forms when bases between two hairpin  
15 loops form pair interactions. These intermolecular “kissing interactions” occur when the unpaired nucleotides in one hairpin loop, base pair with the unpaired nucleotides in another hairpin loop to form a stable interaction complex. See FIG. 9A for an example.

**N-terminal portion:** A region of a protein sequence that includes a contiguous stretch of amino acids that begins at the N-terminal residue of the protein. An N-terminal portion of the protein  
20 can be defined by a contiguous stretch of amino acids (*e.g.*, a number of amino acid residues).

**Non-naturally occurring, synthetic, or engineered:** Terms used herein as interchangeably and indicate the involvement of the hand of man. The terms, when referring to nucleic acid molecules or polypeptides indicate that the nucleic acid molecule or the polypeptide is at least substantially free from at least one other component with which they are naturally associated in nature and as found in nature.  
25 In addition, the terms can indicate that the nucleic acid molecules or polypeptides have a sequence not found in nature.

**Nucleic acid molecule:** A deoxyribonucleotide (DNA) or ribonucleotide (RNA) polymer, which can include natural nucleotides/ribonucleotides and/or analogues of natural nucleotides/ribonucleotides that hybridize to nucleic acid molecules in a manner similar to naturally  
30 occurring nucleotides. A nucleic acid molecule can be a single stranded (ss) DNA or RNA molecule or a double stranded (ds) nucleic acid molecule. RNA or mRNA as used herein may refer to a pre-mRNA molecule, or a mature RNA transcript. A pre-mRNA molecule comprises sequences to be removed by processing, *e.g.*, intron sequences removed by splicing following binding of the dimerization domains described herein. Nucleic acid molecules described herein can be DNA molecules from which an RNA  
35 is transcribed from a promoter on the DNA, *e.g.*, in the context of a DNA expression vector.

**Operably linked:** A first nucleic acid sequence is operably linked with a second nucleic acid sequence when the first nucleic acid sequence is placed in a functional relationship with the second nucleic acid sequence. For instance, a promoter sequence is operably linked to a nucleic acid sequence if the promoter affects the expression of the nucleic acid sequence, for example, the promoter effects transcription of a pre-mRNA, which when spliced may result in expression of a protein (such as a portion of a DMD, factor 8, factor 9, or ABCA4 coding sequence).

**Pharmaceutically acceptable carriers:** The pharmaceutically acceptable carriers useful in this invention are conventional. *Remington's Pharmaceutical Sciences*, by E. W. Martin, Mack Publishing Co., Easton, PA, 15th Edition (1975), describes compositions and formulations suitable for pharmaceutical delivery of a therapeutic agent, such as a nucleic acid molecule disclosed herein.

In general, the nature of the carrier will depend on the particular mode of administration being employed. For instance, parenteral formulations usually comprise injectable fluids that include pharmaceutically and physiologically acceptable fluids such as water, physiological saline, balanced salt solutions, aqueous dextrose, glycerol or the like as a vehicle. In addition to biologically-neutral carriers, pharmaceutical compositions to be administered can contain minor amounts of non-toxic auxiliary substances, such as wetting or emulsifying agents, preservatives, and pH buffering agents and the like, for example sodium acetate or sorbitan monolaurate.

**Polypeptide, peptide and protein:** Refer to polymers of amino acids of any length. The polymer may be linear or branched, it may include modified amino acids, and it may be interrupted by non-amino acids. The terms also encompass an amino acid polymer that has been modified; for example, disulfide bond formation, glycosylation, lipidation, acetylation, phosphorylation, or any other manipulation, such as conjugation with a labeling component. As used herein the term "amino acid" includes natural and/or unnatural or synthetic amino acids, including glycine and both the D or L optical isomers, and amino acid analogs and peptidomimetics. In one example, a protein is one associated with disease, such as a genetic disease (*e.g.*, see Table 1). In one example, a protein is a therapeutic protein, such as one used in the treatment of a disease, such as cancer. In one example a protein is at least 50 aa in length, at least 100 aa in length, at least 500 aa in length, at least 1000 aa in length, at least 1500 aa in length, such as at least 2000 aa, at least 2500 aa, at least 3000 aa, or at least 5000 aa.

**Polypyrimidine tract:** A region of pre-messenger RNA (mRNA) that promotes the assembly of the spliceosome, the protein complex specialized for carrying out RNA splicing during the process of post-transcriptional modification. This tract can be primarily pyrimidine nucleotides, such as uracil, and in some examples is 15–20 base pairs long, located about 5–40 base pairs before the 3' end of the intron to be spliced.



**Promoter/Enhancer:** An array of nucleic acid control sequences which direct transcription of a nucleic acid sequence. A promoter includes necessary nucleic acid sequences near the start site of transcription, such as, in the case of a polymerase II type promoter, a TATA element. A promoter also optionally includes distal enhancer or repressor elements which can be located as much as several thousand base pairs from the start site of transcription. In some examples a promoter sequence + its corresponding coding sequence is larger than the capacity for an AAV. In some examples a promoter sequence of a target protein is at least 3500 nt, at least 4000 nt, at least 5000 nt, or even at least 6000 nt.

A “constitutive promoter” is a promoter that is continuously active and is not subject to regulation by external signals or molecules. In contrast, the activity of an “inducible promoter” is regulated by an external signal or molecule (for example, a transcription factor). Both constitutive and inducible promoters can be used in the methods and systems provided herein (see *e.g.*, Bitter *et al.*, *Methods in Enzymology* 153:516-544, 1987). A tissue-specific promoter can be used in the methods and systems provided herein, for example to direct expression primarily in a desired tissue or cell of interest, such as muscle, neuron, bone, skin, blood, specific organs (*e.g.*, liver, pancreas), or particular cell types (*e.g.*, lymphocytes). In some examples, a promoter used herein is endogenous to the target protein expressed. In some examples, a promoter used herein is exogenous to the target protein expressed.

Also included are promoter elements which are sufficient to render promoter-dependent gene expression controllable for cell-type specific, tissue-specific, or inducible by external signals or agents; such elements may be located in the 5' or 3' regions of the gene. Promoters produced by recombinant DNA or synthetic techniques can also be used to provide for transcription of the nucleic acid sequences.

Exemplary promoters that can be used with the methods and systems provided herein include, but are not limited to an SV40 promoter, cytomegalovirus (CMV) promoter (optionally with the CMV enhancer), a pol III promoter (*e.g.*, U6 and H1 promoters), a pol II promoter (*e.g.*, the retroviral Rous sarcoma virus (RSV) LTR promoter (optionally with the RSV enhancer), the dihydrofolate reductase promoter, the  $\beta$ -actin promoter, the phosphoglycerol kinase (PGK) promoter, and the EF1 $\alpha$  promoter).

**Recombinant:** A recombinant nucleic acid molecule or protein sequence is one that has a sequence that is not naturally occurring or has a sequence that is made by an artificial combination of two otherwise separated segments of sequence (*e.g.*, a viral vector that includes a portion of a dystrophin coding sequence, such as about a third, half, or two-thirds of a coding sequence). This artificial combination can be accomplished by, for example, chemical synthesis or the artificial manipulation of isolated segments of nucleic acids, such as by genetic engineering techniques. Similarly, a recombinant or transgenic cell is one that contains a recombinant nucleic acid molecule.

**Sequence identity:** The similarity between amino acid (or nucleotide) sequences is expressed in terms of the similarity between the sequences, otherwise referred to as sequence identity. Sequence

identity is frequently measured in terms of percentage identity (or similarity or homology); the higher the percentage, the more similar the two sequences are.

Methods of alignment of sequences for comparison are known. Various programs and alignment algorithms are described in: Smith and Waterman, *Adv. Appl. Math.* 2:482, 1981; Needleman and Wunsch, *J. Mol. Biol.* 48:443, 1970; Pearson and Lipman, *Proc. Natl. Acad. Sci. U.S.A.* 85:2444, 1988; Higgins and Sharp, *Gene* 73:237, 1988; Higgins and Sharp, *CABIOS* 5:151, 1989; Corpet *et al.*, *Nucleic Acids Research* 16:10881, 1988; and Pearson and Lipman, *Proc. Natl. Acad. Sci. U.S.A.* 85:2444, 1988. Altschul *et al.*, *Nature Genet.* 6:119, 1994, presents a detailed consideration of sequence alignment methods and homology calculations.

The NCBI Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, *J. Mol. Biol.* 215:403, 1990) is available from several sources, including the National Center for Biotechnology Information (NCBI, Bethesda, MD) and on the internet, for use in connection with the sequence analysis programs blastp, blastn, blastx, tblastn and tblastx. A description of how to determine sequence identity using this program is available on the NCBI website on the internet.

Variants of a native protein or coding sequence (such as a DMD, factor 8, factor 9, or ABCA4 sequence) are typically characterized by possession of at least about 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98% or at least 99% sequence identity counted over the full length alignment with the amino acid sequence using the NCBI Blast 2.0, gapped blastp set to default parameters. For comparisons of amino acid sequences of greater than about 30 amino acids, the Blast 2 sequences function is employed using the default BLOSUM62 matrix set to default parameters, (gap existence cost of 11, and a per residue gap cost of 1). When aligning short peptides (fewer than around 30 amino acids), the alignment should be performed using the Blast 2 sequences function, employing the PAM30 matrix set to default parameters (open gap 9, extension gap 1 penalties). Proteins with even greater similarity to the reference sequences will show increasing percentage identities when assessed by this method, such as at least 95%, at least 98%, or at least 99% sequence identity. When less than the entire sequence is being compared for sequence identity, homologs and variants will typically possess at least 80% sequence identity over short windows of 10-20 amino acids, and may possess sequence identities of at least 85% or at least 90% or at least 95% depending on their similarity to the reference sequence. Methods for determining sequence identity over such short windows are available at the NCBI website on the internet. These sequence identity ranges are provided for guidance only; it is possible that strongly significant homologs could be obtained that fall outside of the ranges provided.

Variants of the disclosed nucleic acid sequences (such as synthetic intron sequences and coding sequences) are typically characterized by possession of at least about 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98% or at least 99% sequence identity counted over the full length alignment with the nucleic acid sequence using the NCBI Blast 2.0, gapped blastn set to default

parameters. One of skill in the art will appreciate that these sequence identity ranges are provided for guidance only; it is possible that functional sequences could be obtained that fall outside of the ranges provided.

**Subject:** A mammal, for example a human. Mammals include, but are not limited to, murines, simians, humans, farm animals, sport animals, and pets. In one embodiment, the subject is a non-human mammalian subject, such as a monkey or other non-human primate, mouse, rat, rabbit, pig, goat, sheep, dolphin, dog, cat, horse, or cow. In some examples, the subject is a laboratory animal/organism, such as a mouse, rabbit, or rat. In some examples, the subject treated using the methods disclosed herein is a human.

In some examples, the subject has genetic disease, such as one listed in Table 1, that can be treated using the methods disclosed herein. In some examples, the subject treated using the methods disclosed herein is a human subject having a genetic disease. In some examples, the subject treated using the methods disclosed herein is a human subject having cancer

**Therapeutic agent:** Refers to one or more molecules or compounds that confer some beneficial effect upon administration to a subject. The disclosed synthetic nucleic acid molecules and systems provided herein are therapeutic agents. The beneficial therapeutic effect can include enablement of diagnostic determinations; amelioration of a disease, symptom, disorder, or pathological condition; reducing or preventing the onset of a disease, symptom, disorder or condition; and generally counteracting a disease, symptom, disorder or pathological condition.

**Transduced, Transformed and Transfected:** A virus or vector “transduces” a cell when it transfers nucleic acid molecules into a cell. A cell is “transformed” or “transfected” by a nucleic acid transduced into the cell when the nucleic acid becomes stably replicated by the cell, either by incorporation of the nucleic acid into the cellular genome, or by episomal replication.

These terms encompasses all techniques by which a nucleic acid molecule can be introduced into such a cell, including transfection with viral vectors, transformation with plasmid vectors, and introduction of naked DNA by electroporation, lipofection, particle gun acceleration and other methods in the art. In some example the method is a chemical method (*e.g.*, calcium-phosphate transfection), physical method (*e.g.*, electroporation, microinjection, particle bombardment), fusion (*e.g.*, liposomes), receptor-mediated endocytosis (*e.g.*, DNA-protein complexes, viral envelope/capsid-DNA complexes) and biological infection by viruses such as recombinant viruses (Wolff, J. A., ed, *Gene Therapeutics*, Birkhauser, Boston, USA, 1994). Methods for the introduction of nucleic acid molecules into cells are known (*e.g.*, see U.S. Patent No. 6,110,743). These methods can be used to transduce a cell with the disclosed nucleic acid molecules.

**Transgene:** An exogenous gene, for example supplied by a vector, such as AAV. In one example, a transgene encodes a portion of a target protein, such as about a third, half, or two-thirds of a

target protein, for example operably linked to a promoter sequence. In one example, a transgene includes a portion of a dystrophin coding sequence, such as about a third, half, or two-thirds of a dystrophin coding sequence (or other therapeutic coding sequence, such as one encoding a protein listed in Table 1), for example operably linked to a promoter sequence.

5           **Treating, Treatment, and Therapy:** Any success or indicia of success in the attenuation or amelioration of an injury, pathology or condition, including any objective or subjective parameter such as abatement, remission, diminishing of symptoms or making the condition more tolerable to the patient, slowing in the rate of degeneration or decline, making the final point of degeneration less debilitating, improving a subject's physical or mental well-being, or prolonging the length of survival.  
10       The treatment may be assessed by objective or subjective parameters; including the results of a physical examination, blood and other clinical tests, and the like. In some examples, treatment with the disclosed methods results in a decrease in the number or severity of symptoms associated with a genetic disease, such as increasing the survival time of a treated patient with the genetic disease.

          In some examples, treatment with the disclosed methods results in a decrease in the number or  
15       severity of symptoms associated with DMD or other genetic disease, such as increasing survival, increasing the mobility (*e.g.*, walking, climbing), improving cognitive ability, reducing calf muscle size, reduce cardiomyopathy, improving vision, improving hearing, improving blood clotting, or improve respiratory function. In some examples, combinations of these effects are achieved.

**Tumor, neoplasia, malignancy or cancer:** A neoplasm is an abnormal growth of tissue or  
20       cells which results from excessive cell division. Neoplastic growth can produce a tumor. The amount of a tumor in an individual is the “**tumor burden**” which can be measured as the number, volume, or weight of the tumor. A tumor that does not metastasize is referred to as “**benign**.” A tumor that invades the surrounding tissue and/or can metastasize is referred to as “**malignant**.” A “**non-cancerous tissue**” is a tissue from the same organ wherein the malignant neoplasm formed, but does not have the  
25       characteristic pathology of the neoplasm. Generally, noncancerous tissue appears histologically normal. A “**normal tissue**” is tissue from an organ, wherein the organ is not affected by cancer or another disease or disorder of that organ. A “**cancer-free**” subject has not been diagnosed with a cancer of that organ and does not have detectable cancer.

          Exemplary tumors, such as cancers, that can be treated with the disclosed methods and systems  
30       include solid tumors, such as breast carcinomas (*e.g.* lobular and duct carcinomas), sarcomas, carcinomas of the lung (*e.g.*, non-small cell carcinoma, large cell carcinoma, squamous carcinoma, and adenocarcinoma), mesothelioma of the lung, colorectal adenocarcinoma, stomach carcinoma, prostatic adenocarcinoma, ovarian carcinoma (such as serous cystadenocarcinoma and mucinous  
35       cystadenocarcinoma), ovarian germ cell tumors, testicular carcinomas and germ cell tumors, pancreatic adenocarcinoma, biliary adenocarcinoma, hepatocellular carcinoma, bladder carcinoma (including, for

instance, transitional cell carcinoma, adenocarcinoma, and squamous carcinoma), renal cell adenocarcinoma, endometrial carcinomas (including, *e.g.*, adenocarcinomas and mixed Mullerian tumors (carcinosarcomas)), carcinomas of the endocervix, ectocervix, and vagina (such as adenocarcinoma and squamous carcinoma of each of same), tumors of the skin (*e.g.*, squamous cell carcinoma, basal cell carcinoma, malignant melanoma, skin appendage tumors, Kaposi sarcoma, cutaneous lymphoma, skin adnexal tumors and various types of sarcomas and Merkel cell carcinoma), esophageal carcinoma, carcinomas of the nasopharynx and oropharynx (including squamous carcinoma and adenocarcinomas of same), salivary gland carcinomas, brain and central nervous system tumors (including, for example, tumors of glial, neuronal, and meningeal origin), tumors of peripheral nerve, soft tissue sarcomas and sarcomas of bone and cartilage, and lymphatic tumors (including B-cell and T-cell malignant lymphoma). In one example, the tumor is an adenocarcinoma.

The methods and systems can also be used to treat liquid tumors, such as a lymphatic, white blood cell, or other type of leukemia. In a specific example, the tumor treated is a tumor of the blood, such as a leukemia (for example acute lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), acute myelogenous leukemia (AML), chronic myelogenous leukemia (CML), hairy cell leukemia (HCL), T-cell prolymphocytic leukemia (T-PLL), large granular lymphocytic leukemia, and adult T-cell leukemia), lymphomas (such as Hodgkin's lymphoma and non-Hodgkin's lymphoma), and myelomas).

**Upregulated:** When used in reference to the expression of a molecule, such as a target nucleic acid/protein, refers to any process which results in an increase in production of the target nucleic acid/protein. In some examples, upregulation or activation of a target RNA includes processes that increase translation of the target RNA and thus can increase the presence of corresponding proteins.

Upregulation includes any detectable increase in target nucleic acid/protein. In certain examples, detectable target nucleic acid/protein expression in a cell or cell free system increases by at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, at least 90%, at least 95%, at least 100%, at least 200%, at least 400%, or at least 500% as compared to a control (such an amount of target nucleic acid/protein detected in a corresponding sample not treated with a nucleic acid molecule provided herein). In one example, a control is a relative amount of expression in a normal cell (*e.g.*, a non-recombinant cell that does not include a system provided herein).

**Under conditions sufficient for:** A phrase that is used to describe any environment that permits a desired activity. In one example the desired activity is increased expression or activity of a protein needed to treat a disease. In one example the desired activity is treatment of or slowing the progression of a genetic disease such as DMD (or other genetic disease listed in Table 1) *in vivo*, for example using the disclosed methods and systems.

**Vector:** A nucleic acid molecule into which a foreign nucleic acid molecule can be introduced without disrupting the ability of the vector to replicate and/or integrate in a host cell. Vectors include, but are not limited to, nucleic acid molecules that are single-stranded, double-stranded, or partially double-stranded; nucleic acid molecules that comprise one or more free ends, no free ends (*e.g.*, circular); nucleic acid molecules that comprise DNA, RNA, or both; and other varieties of polynucleotides.

A vector can include nucleic acid sequences that permit it to replicate in a host cell, such as an origin of replication. A vector can also include one or more selectable marker genes and other genetic elements. An integrating vector is capable of integrating itself into a host nucleic acid. An expression vector is a vector that contains the necessary regulatory sequences to allow transcription and translation of inserted gene or genes.

One type of vector is a "plasmid," which refers to a circular double stranded DNA loop into which additional DNA segments can be inserted, such as by standard molecular cloning techniques. Another type of vector is a viral vector, wherein virally-derived DNA or RNA sequences are present in the vector for packaging into a virus (*e.g.*, retroviruses, replication defective retroviruses, adenoviruses, replication defective adenoviruses, and adeno-associated viruses). Viral vectors also include polynucleotides carried by a virus for transfection into a host cell. In some embodiments, the vector is a lentivirus (such as an integration-deficient lentiviral vector) or adeno-associated viral (AAV) vector.

In some embodiments, the vector is an AAV, such as AAV serotypes AAV9 or AAVrh.10. In some embodiments, the vector is one that can penetrate the blood-brain barrier, for example following intravenous administration. The adeno-associated virus serotype rh.10 (AAV.rh10) vector partially penetrates the blood-brain barrier, providing high levels and spread of transgene expression.

## II. Overview of Several Embodiments

One approach to curing patients who suffer from genetic diseases is gene replacement therapy (generally referred to as gene therapy). In such an approach, the defective gene is replaced by an intact version of it, delivered through *e.g.*, a viral vector, which achieves sustained expression from months to years. Although adeno associated viruses (AAVs) have been used for clinical gene replacement therapy, they have a limited packaging capacity (*e.g.*, about less than 5 kb). Thus, strategies to overcome this packaging limitation are needed to achieve gene replacement of genes that exceed the about 5 kb size limit. For example some promoters alone, coding sequences alone, or the combined promoter + coding sequence, exceed the about 5 kb size limit of an AAV. Thus, such proteins encoded by such promoters and coding sequences can be expressed using the disclosed systems.

Prior methods to overcome the cargo limitations of AAV do not appear to achieve the efficiency required to produce adequate levels of target protein in sufficient numbers of cells to treat disease. For

example as dystrophin is about 11kb, it needs to be delivered in a minimum of three fragments to be compatible with AAV packaging limitations.

Splicing mediated recombination of two RNA molecules using naturally occurring intron sequences for one or both of the RNA fragments is inefficient. First, these natural intron sequences are sequences from naturally occurring introns and are comprised of a mix of all four RNA nucleotides. Such sequences tend to fold up into structures that can obstruct trans-interaction by forming strong intramolecular base pairs rather than being available for intermolecular interactions. Second, these naturally occurring intron sequences have not evolved to strongly attract the spliceosome components, since exon rather than introns drive the exon definition in higher eukaryotes. These two limitations of previous strategies are addressed herein by designing synthetic intronic sequences that are not found in nature. These synthetic sequences contain elements that strongly attract and stimulate spliceosome recruitment on the one hand while minimizing the secondary structure (and in some examples other structure, such as tertiary structure) that obstructs bringing the two RNA fragments together.

The inventors developed a novel nucleic acid based element that can be used to efficiently reconstitute the coding sequence of large genes from multiple serial fragments. The disclosed methods and systems differ from prior methods. The disclosed highly efficient synthetic introns utilize an optimal arrangement of RNA elements (or DNA encoding these elements) that efficiently drive the RNA splicing reaction between non-covalently linked RNAs (pre-mRNAs). The method/system is a significant advancement over previous attempts to harness trans-splicing because it generates high levels of functional protein that more closely approximate the therapeutic levels of a protein to treat genetic diseases. The innovation is based on selecting non-natural RNA domains that inherently are incapable of forming strong cis-binding interactions that interfere with trans-interactions with a second RNA having a complementary strand (also having inherently low cis-binding capacity). These optimized dimerization domains and/or synthetic introns can include non-natural sequences (*e.g.*, sequences not found in human cells and/or not found in another biological system) used in combination with optimized motifs that facilitate RNA splicing (including splice donor, splice acceptor, splice enhancer, and splice branch point sequences). A synthetic nucleic acid can be a non-natural nucleic acid sequence, *e.g.*, a sequence not found in human cells and/or not found in another biological system). By optimizing the trans-dimerization of the RNA strands in the context of the appropriate RNA motifs that mediate efficient splicing, it is demonstrated herein for the first time that two or three different RNAs can be precisely and efficiently covalently linked in the same cell producing high levels of functional proteins in vivo and in vitro. Unlike the “hybrid” approach that provides an inefficient combination at the DNA level via DNA recombination that is ultimately followed by RNA splicing in cis to excise the DNA recombination site from the mature transcript, the disclosed method/system promotes a more efficient reaction in which two protein coding RNA fragments are joined together on

the pre-mRNA level with less risk of producing recombination products that encode non-functional and/or deleterious products.

The data demonstrate that by using efficient synthetic RNA-dimerization and recombination domains (sRdR domains, also referred to as RNA end-joining (REJ) domains), a gene of interest can efficiently reconstitute from two or three separate gene fragments expressed in the same cell. These results show the ability of the disclosed methods and systems to reconstitute large genes like dystrophin or the blood clotting Factor VIII, or the ATP binding cassette subfamily A member 4 (Abca4) using AAVs, in order to treat Duchenne Muscular Dystrophy and Hemophilia A, or Stargardt's Disease respectively. Based on these observations, other genetic diseases can be similarly treated, such as ones benefiting from expression of a large protein (*e.g.*, see disorders listed in Table 1). Other applications include research and biotechnology applications.

To address some of the limitations with existing strategies for reconstitution of fragmented genes from multiple AAVs, provided herein is a system that serially aligns and recombines two or more individual synthetic RNA molecules in the target cell. Each individual synthetic RNA molecule includes a synthetic intron sequence, containing a dimerization domain and elements needed for RNA splicing, which upon binding of dimerization domains to one another in the correct order, mediates efficient RNA recombination of individual fragments. In one example, reconstitution of a coding sequence from two fragments is achieved by appending a first synthetic intron (A) to the 3' end of the N-terminal coding fragment and a complimentary second synthetic domain (A') to the 5' end of the C-terminal coding fragment. The two RNAs are recombined by a cell's intrinsic RNA splicing machinery (*i.e.*, the spliceosome machinery). The synthetic intron domains contain two functional elements: (1) a dimerization domain to mediate base pairing between the two halves that are to be recombined and (2) a domain optimized to efficiently recruit the splicing machinery to mediate efficient reconstitution of the two RNA molecules. In some examples, a synthetic intron includes a sequence having at least 50% at least 60%, at least 70%, at least 75%, 80%, at least 85%, at least 90%, at least 95%, at least 98%, at least 99%, or 100% sequence identity to any synthetic intron provided in SEQ ID NOS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 145, 146, 147, 148, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, and 166 (*e.g.*, see FIGS. 10A-10Z). In some examples, a synthetic intron is an RNA molecule encoded by a sequence having at least 50% at least 60%, at least 70%, at least 75%, 80%, at least 85%, at least 90%, at least 95%, at least 98%, at least 99%, or 100% sequence identity to any synthetic intron provided in SEQ ID NOS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 145, 146, 147, 148, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, and 166, but without the provided promoter sequence). One skilled in the art will appreciate that any of the molecules provided in SEQ ID NOS: 1, 2, 20, 21, 22, 23, 24, 25, 145, 146, 147, 148, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, and 166 can be modified to replace



the protein coding portions (e.g., 114 and 164 of FIG. 6A) with another protein coding sequence of interest (e.g., YFP coding sequence of SEQ ID NO: 1, 2, 22 or 23 can be replaced with a therapeutic protein coding sequence). Thus, also provided herein are synthetic intron molecules having at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, at least 99%, or 100% sequence identity to any synthetic intron portion provided in SEQ ID NO: 1, 2, 20, 21, 22, 23, 24, 25, 145, 146, 147, 148, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, and 166 (e.g., nt 3703-3975 of SEQ ID NO: 22 and nt 1-225 of SEQ ID NO: 23). Also provided are synthetic intron RNA molecules encoded by a sequence having at least 50% at least 60%, at least 70%, at least 75%, 80%, at least 85%, at least 90%, at least 95%, at least 98%, at least 99%, or 100% sequence identity to any synthetic intron provided in SEQ ID NOS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 145, 146, 147, 148, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, and 166, but without the provided promoter sequence).

Exemplary dimerization domains were bioinformatically selected to minimize/optimize their internal secondary/tertiary structure. The dimerization domains tested contained long stretches of low diversity nucleotide sequences to avoid intramolecular annealing. By avoiding intramolecular annealing, these dimerization domains are present in an open configuration and therefore are available for pairing with the corresponding complementary dimerization domain sequence. The synthetic intron domains contain intronic splice enhancing elements which lead to efficient recruitment of the splicing machinery.

The disclosed RNA molecules are designed to have at least an open and available single-stranded region that is available to bind to the complementary dimerization domain to allow efficient splicing and recombination of the RNAs. In some examples, this is achieved by utilizing only purines or only pyrimidines for the binding domains. Due to the inability of purines to pair with themselves (and pyrimidines likewise) these stretches of RNA have an open predicted structure.

RNA molecules are present as a single strand in the cells. Being single stranded they are inherently prone to hybridize to themselves and thereby form strong secondary and tertiary structures. The most stable base pairs will be G with C, A with U, and the G with U wobble pair. Thermodynamically, the pairing of two bases is favored over an open configuration. To design efficient synthetic nucleic acid molecules, two dimerization domains having complementarity to one another are present in an open configuration such that the dimerization domains are available for inter-molecular base pairing. To avoid intra-molecular base pairing in between other parts of the synthetic nucleic acid molecules, a long stretch of non-diverse sequences containing incompatible bases can be included. For example, a long stretch of pyrimidines (*i.e.*, C and T) or purines (*i.e.*, A and G) can be present in the synthetic nucleic acid molecules. Pyrimidines cannot form canonical base pairs with other pyrimidines, purines cannot form canonical base pairs with other purines. Such a stretch of purines or pyrimidines

can range from a couple bases to a couple hundreds of bases. Since these stretches cannot intra-molecularly bind, they are available for inter-molecular base pairing with a complementary fragment. For example, the synthetic nucleic acid molecules A and A' may be configured with A containing a pyrimidine stretch (e.g., 5'-CCUU(...)CCUU-3') and A' containing the complementary purine sequence (e.g., 5'-AAGG(...)AAGG-3').

The disclosed synthetic nucleic acid molecules (e.g., RNA or DNA encoding the RNA) are designed to minimize any off-target binding to incorrect sites in the genome. Off target binding can be reduced by altering the sequence of the nucleic acid molecule.

The same design principle, that is the use of hypodiverse stretches of RNA bases to achieve open synthetic nucleic acid configurations, can be extended to using stretches of single bases e.g. using a series of Gs that would base pair with a series of Cs and a series of As that would base pair with a series of Us, in the dimerization domains.

To increase recombination of two or more synthetic nucleic acid molecules, the following methods can be used. RNA splicing depends on the recruitment of spliceosome components to the 5' end of the intron (the splice donor site) and the 3' end of the intron (the splice acceptor site, with its associated branch point sequence and the polypyrimidine tract). Different ribonucleoproteins are recruited to the intron through base pairing of protein associated small nuclear RNA (snRNA) with intronic sequences. By placing perfect match consensus sequences into the RNA dimerization and recombination domains, the recruitment of spliceosome components can be facilitated which in turn enhances the efficiency of spliceosome mediated recombination. Previously characterized intronic splice enhancer sequences can recruit additional splicing promoting factors that are referred to as intronic splice enhancers.

In some examples, instead of using naturally occurring RNA sequences for the RNA splicing sequences, consensus sequences are used. For example, consensus sequences can be used for any of the sequences that are involved in splicing, including splice donor, splice acceptor, splice enhancer and splice branch point sequences. With these synthetic nucleic acid molecules, two (or more) RNA molecules can be serially joined together in a cell ex vivo, in vitro, or in vivo. Outside of the encoded synthetic intronic domains, synthetic nucleic acid molecules can include any promoter and coding sequence. For example, two synthetic nucleic acid molecules could carry two halves of a single gene. This was tested in vitro and in vivo by reconstituting two halves of a yellow fluorescent protein (YFP), and was shown to be efficient (see FIGS. 3A-3D).

The modular nature of the synthetic nucleic acid molecules allowed for testing the efficiency of achieving serial recombination (i.e., >2) of multiple RNA fragments using a combinatorial set of optimized complimentary dimerization domains (FIGS. 4A-4B). A three-way split yellow fluorescent protein was efficiently reconstituted and expressed at high levels in >80% of transfected cells.

These results demonstrate that a single RNA molecule can be reconstituted from at least three different synthetic nucleic acid molecules, such as when expression of a disease causing gene (or therapeutic protein) that has a promoter and/or a coding sequence that is too long to fit into a single gene therapy vector such as AAV.

5 In some examples, the synthetic nucleic acid molecules, e.g., synthetic DNA molecules, of the inventive compositions, systems, kits, and methods, are produced by transcription of an RNA virus genome by reverse transcriptase.

The disclosed system allows for the efficient RNA recombination between individual fragments. In some examples, reconstitution (*i.e.*, splicing or recombination) efficiency achieved using  
10 the compositions, systems or methods of the disclosure is determined using any suitable method known to one of skill in the art. In some examples, reconstitution efficiency is represented by a measure of correctly joined RNA relative to a control RNA, or a measure of full-length protein or protein activity relative to that of a control protein. In some examples the control RNA is the unjoined RNA, wherein reconstitution efficiency is represented by a measure of joined RNA relative to unjoined RNA. This  
15 measurement can be made by detecting and comparing junction RNA and the unjoined 3' RNA species 3' (*e.g.*, junction RNA: 3' RNA). In some examples wherein more than two RNAs are joined, joining at either or all junctions are evaluated. In some examples, reconstitution efficiency is represented by a measure of full-length or active protein relative to a protein fragment or inactive protein.

In some examples, the reconstitution, recombination or splicing efficiency (a measure of the  
20 correct joining of the two or more different coding sequences present on different RNA molecules, and/or the production of the desired full-length protein) is about 10% to about 100%. In some examples, the reconstitution efficiency is about 10% to about 15%, about 10% to about 20%, about 10% to about 25%, about 10% to about 30%, about 10% to about 40%, about 10% to about 50%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 10% to about 100%, about 15% to about 20%, about 15% to about 25%, about 15% to about 30%, about  
25 15% to about 40%, about 15% to about 50%, about 15% to about 60%, about 15% to about 70%, about 15% to about 80%, about 15% to about 90%, about 15% to about 100%, about 20% to about 25%, about 20% to about 30%, about 20% to about 40%, about 20% to about 50%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 20% to about  
30 100%, about 25% to about 30%, about 25% to about 40%, about 25% to about 50%, about 25% to about 60%, about 25% to about 70%, about 25% to about 80%, about 25% to about 90%, about 25% to about 100%, about 30% to about 40%, about 30% to about 50%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 30% to about 100%, about 40% to about 50%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about  
35 40% to about 90%, about 40% to about 100%, about 50% to about 60%, about 50% to about 70%,

about 50% to about 80%, about 50% to about 90%, about 50% to about 100%, about 60% to about 70%, about 60% to about 80%, about 60% to about 90%, about 60% to about 100%, about 70% to about 80%, about 70% to about 90%, about 70% to about 100%, about 80% to about 90%, about 80% to about 100%, or about 90% to about 100%. In some examples, the reconstitution efficiency is about 10%, about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 100%. In some examples, the reconstitution efficiency is at least about 10%, about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, or about 90%. In some examples, the reconstitution efficiency is at most about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 100%.

In some examples, the reconstitution, recombination or splicing efficiency (in this example a measure of the correct joining of two different coding sequences present on different RNA molecules, and/or the production of the desired full-length protein, wherein the two different coding sequences encode a transcript of about 3200 nt to 9000 nt, such as about 4000 to 9000 nt, about 4400 to 9000 nt, about 3200 to 4000 nt, about 3200 to 3600 nt, for example about 4500 nt, about 4000 nt, about 3800 nt, about 3600 nt, or about 3200 nt), is about 10% to about 100%. In some examples, the reconstitution efficiency using a two-part system is about 10% to about 15%, about 10% to about 20%, about 10% to about 25%, about 10% to about 30%, about 10% to about 40%, about 10% to about 50%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 10% to about 100%, about 15% to about 20%, about 15% to about 25%, about 15% to about 30%, about 15% to about 40%, about 15% to about 50%, about 15% to about 60%, about 15% to about 70%, about 15% to about 80%, about 15% to about 90%, about 15% to about 100%, about 20% to about 25%, about 20% to about 30%, about 20% to about 40%, about 20% to about 50%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 20% to about 100%, about 25% to about 30%, about 25% to about 40%, about 25% to about 50%, about 25% to about 60%, about 25% to about 70%, about 25% to about 80%, about 25% to about 90%, about 25% to about 100%, about 30% to about 40%, about 30% to about 50%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 30% to about 100%, about 40% to about 50%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about 40% to about 90%, about 40% to about 100%, about 50% to about 60%, about 50% to about 70%, about 50% to about 80%, about 50% to about 90%, about 50% to about 100%, about 60% to about 70%, about 60% to about 80%, about 60% to about 90%, about 60% to about 100%, about 70% to about 80%, about 70% to about 90%, about 70% to about 100%, about 80% to about 90%, about 80% to about 100%, or about 90% to about 100%. In some examples, the reconstitution efficiency is about 10%, about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%,

about 80%, about 90%, or about 100%. In some examples, the reconstitution efficiency is at least about 10%, about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, or about 90%. In some examples, the reconstitution efficiency is at most about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 100%.

In some examples, the reconstitution, recombination or splicing efficiency (in this example a measure of the correct joining of the two different coding sequences present on different RNA molecules, and/or the production of the desired full-length protein, wherein the two different coding sequences encode a transcript of about 4000 nt), is about 40% to about 60%, such as about 40% to about 50%, about 42% to about 47%, for example about 45%.

In some examples, the reconstitution, recombination or splicing efficiency (in this example a measure of the correct joining of the two different coding sequences present on different RNA molecules, and/or the production of the desired full-length protein, wherein the two different coding sequences encode a transcript of about 3800 nt), is about 40% to about 60%, such as about 40% to about 50%, about 42% to about 47%, for example about 45%.

In some examples, the reconstitution, recombination or splicing efficiency (in this example a measure of the correct joining of the two different coding sequences present on different RNA molecules, and/or the production of the desired full-length protein, wherein the two different coding sequences encode a transcript of about 3600 nt), is about 25% to about 50%, such as about 30% to about 40%, for example about 35%.

In some examples, the reconstitution, recombination or splicing efficiency (in this example a measure of the correct joining of the two different coding sequences present on different RNA molecules, and/or the production of the desired full-length protein, wherein the two different coding sequences encode a transcript of about 3200 nt), is about 25% to about 50%, such as about 30% to about 40%, for example about 35%.

In some examples, the reconstitution, recombination or splicing efficiency (in this example a measure of the correct joining of three different coding sequences present on different RNA molecules, and/or the production of the desired full-length protein, wherein the three different coding sequences encode a transcript of about 3200 nt to about 13,500 nt, such as about 4000 nt to about 5,000 nt, about 4000 nt to about 13,500 nt, about 6000 nt to about 12,000 nt, about 6000 nt to about 10,000 nt, or about 8000 nt to about 12,000 nt, for example up to about 13,500 nt), is about 10% to about 100%. In some examples, the reconstitution efficiency using a three-part system is about 10% to about 15%, about 10% to about 20%, about 10% to about 25%, about 10% to about 30%, about 10% to about 40%, about 10% to about 50%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 10% to about 100%, about 15% to about 20%, about 15% to about 25%, about

15% to about 30%, about 15% to about 40%, about 15% to about 50%, about 15% to about 60%, about 15% to about 70%, about 15% to about 80%, about 15% to about 90%, about 15% to about 100%, about 20% to about 25%, about 20% to about 30%, about 20% to about 40%, about 20% to about 50%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 20% to about 100%, about 25% to about 30%, about 25% to about 40%, about 25% to about 50%, about 25% to about 60%, about 25% to about 70%, about 25% to about 80%, about 25% to about 90%, about 25% to about 100%, about 30% to about 40%, about 30% to about 50%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 30% to about 100%, about 40% to about 50%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about 40% to about 90%, about 40% to about 100%, about 50% to about 60%, about 50% to about 70%, about 50% to about 80%, about 50% to about 90%, about 50% to about 100%, about 60% to about 70%, about 60% to about 80%, about 60% to about 90%, about 60% to about 100%, about 70% to about 80%, about 70% to about 90%, about 70% to about 100%, about 80% to about 90%, about 80% to about 100%, or about 90% to about 100%. In some examples, the reconstitution efficiency is about 10%, about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 100%. In some examples, the reconstitution efficiency is at least about 10%, about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, or about 90%. In some examples, the reconstitution efficiency is at most about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 100%.

In some examples, the reconstitution, recombination or splicing efficiency (in this example a measure of the correct joining of four different coding sequences present on different RNA molecules, and/or the production of the desired full-length protein, wherein the four different coding sequences encode a transcript of about 3200 nt to about 18,000 nt, such as about 4000 nt to about 18,000 nt, about 4000 nt to about 5,000 nt, about 10,000 nt to about 18,000 nt, about 15,000 nt to about 18,000nt, or about 12,000 nt to about 15,000 nt, for example up to about 18,000 nt), is about 10% to about 100%. In some examples, the reconstitution efficiency using a four-part system is about 10% to about 15%, about 10% to about 20%, about 10% to about 25%, about 10% to about 30%, about 10% to about 40%, about 10% to about 50%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 10% to about 100%, about 15% to about 20%, about 15% to about 25%, about 15% to about 30%, about 15% to about 40%, about 15% to about 50%, about 15% to about 60%, about 15% to about 70%, about 15% to about 80%, about 15% to about 90%, about 15% to about 100%, about 20% to about 25%, about 20% to about 30%, about 20% to about 40%, about 20% to about 50%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 20% to about 100%, about 25% to about 30%, about 25% to about 40%, about 25%

to about 50%, about 25% to about 60%, about 25% to about 70%, about 25% to about 80%, about 25% to about 90%, about 25% to about 100%, about 30% to about 40%, about 30% to about 50%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 30% to about 100%, about 40% to about 50%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about 40% to about 90%, about 40% to about 100%, about 50% to about 60%, about 50% to about 70%, about 50% to about 80%, about 50% to about 90%, about 50% to about 100%, about 60% to about 70%, about 60% to about 80%, about 60% to about 90%, about 60% to about 100%, about 70% to about 80%, about 70% to about 90%, about 70% to about 100%, about 80% to about 90%, about 80% to about 100%, or about 90% to about 100%. In some examples, the reconstitution efficiency is about 10%, about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 100%. In some examples, the reconstitution efficiency is at least about 10%, about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, or about 90%. In some examples, the reconstitution efficiency is at most about 15%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, or about 100%. In some examples, the compositions, systems or methods of the disclosure are evaluated by determining an RNA or protein production level using any suitable method known to one of skill in the art. In some examples, the RNA production level is represented by a measure of correctly joined RNA relative to a control RNA, or a measure of full-length protein relative to a control. In some examples the control RNA is a corresponding mutant RNA or an endogenous RNA. For example, the ratio of the amount of joined RNA to the amount of mutant or endogenous RNA produced in the transfected cell is compared with same ratio in nontransfected cells, to determine the production level of the correctly joined RNA. In some examples, the ratio of the amount of the correctly joined RNA, full-length protein, or the protein activity, to the amount of the control RNA, or the amount or activity of the control protein, are compared.

In some examples, the RNA production level achieved is 5% to 100%. In some examples, the RNA production level achieved is about 5% to about 100%. In some examples, the RNA production level achieved is about 5% to about 10%, about 5% to about 20%, about 5% to about 25%, about 5% to about 30%, about 5% to about 40%, about 5% to about 50%, about 5% to about 60%, about 5% to about 70%, about 5% to about 80%, about 5% to about 90%, about 5% to about 100%, about 10% to about 20%, about 10% to about 25%, about 10% to about 30%, about 10% to about 40%, about 10% to about 50%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 10% to about 100%, about 20% to about 25%, about 20% to about 30%, about 20% to about 40%, about 20% to about 50%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 20% to about 100%, about 25% to about 30%, about

25% to about 40%, about 25% to about 50%, about 25% to about 60%, about 25% to about 70%, about 25% to about 80%, about 25% to about 90%, about 25% to about 100%, about 30% to about 40%, about 30% to about 50%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 30% to about 100%, about 40% to about 50%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about 40% to about 90%, about 40% to about 100%, about 50% to about 60%, about 50% to about 70%, about 50% to about 80%, about 50% to about 90%, about 50% to about 100%, about 60% to about 70%, about 60% to about 80%, about 60% to about 90%, about 60% to about 100%, about 70% to about 80%, about 70% to about 90%, about 70% to about 100%, about 80% to about 90%, about 80% to about 100%, or about 90% to about 100%.

In some examples, the RNA production level achieved is about 5%, about 10%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 100%. In some examples, the RNA production level achieved is at least about 5%, about 10%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, or about 90%. In some examples, the RNA production level achieved is at most about 10%, about 20%, about 25%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, or about 100%.

In some examples, the protein production level is represented by a measure of the amount of full-length protein or protein activity relative to that of a control protein. In some examples the control protein is a corresponding mutant protein or an endogenous protein. For example, the ratio of the amount of full-length protein or protein activity to the amount of mutant or endogenous protein produced in the transfected cell is compared with same ratio in nontransfected cells. In some examples, the control protein is the full-length protein produced in, *e.g.*, a cell that is engineered to express a control full-length protein (wherein the cell is not transfected with the inventive constructs) or a non-transfected cell from a normal subject that expresses a control full-length protein, and the protein production level is determined by measuring the amount or activity of the protein in the transfected cell and comparing it to that of the control protein. In some examples, the control protein is a mutant form of the protein, produced in a cell that is transfected or nontransfected with the construct, and the amount of full-length protein or protein activity is compared with that of the control protein to determine the protein production level. In some examples, the amount of full-length protein or protein activity is compared with that of an endogenous, or housekeeping, protein to determine the protein production level.

In some examples, the protein production level achieved is about 1% to about 100%. In some examples, the protein production level achieved is about 10% to about 100%. In some examples, the protein production level achieved is about 10% to about 20%, about 10% to about 30%, about 10% to about 40%, about 10% to about 50%, about 10% to about 60%, about 10% to about 70%, about 10% to about 75%, about 10% to about 80%, about 10% to about 85%, about 10% to about 90%, about 10% to



about 100%, about 20% to about 30%, about 20% to about 40%, about 20% to about 50%, about 20% to about 60%, about 20% to about 70%, about 20% to about 75%, about 20% to about 80%, about 20% to about 85%, about 20% to about 90%, about 20% to about 100%, about 30% to about 40%, about 30% to about 50%, about 30% to about 60%, about 30% to about 70%, about 30% to about 75%, about 30% to about 80%, about 30% to about 85%, about 30% to about 90%, about 30% to about 100%, about 40% to about 50%, about 40% to about 60%, about 40% to about 70%, about 40% to about 75%, about 40% to about 80%, about 40% to about 85%, about 40% to about 90%, about 40% to about 100%, about 50% to about 60%, about 50% to about 70%, about 50% to about 75%, about 50% to about 80%, about 50% to about 85%, about 50% to about 90%, about 50% to about 100%, about 60% to about 70%, about 60% to about 75%, about 60% to about 80%, about 60% to about 85%, about 60% to about 90%, about 60% to about 100%, about 70% to about 75%, about 70% to about 80%, about 70% to about 85%, about 70% to about 90%, about 70% to about 100%, about 75% to about 80%, about 75% to about 85%, about 75% to about 90%, about 75% to about 100%, about 80% to about 85%, about 80% to about 90%, about 80% to about 100%, about 85% to about 90%, about 85% to about 100%, or about 90% to about 100%. In some examples, the protein production level achieved is about 10%, about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 75%, about 80%, about 85%, about 90%, or about 100%. In some examples, the protein production level achieved is at least about 10%, about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 75%, about 80%, about 85%, or about 90%. In some examples, the protein production level achieved is at most about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 75%, about 80%, about 85%, about 90%, or about 100%.

In some examples, the protein activity level achieved is about 50% to about 100%. In some examples, the protein activity level achieved is about 50% to about 100%. In some examples, the protein activity level achieved is about 50% to about 55%, about 50% to about 60%, about 50% to about 65%, about 50% to about 70%, about 50% to about 75%, about 50% to about 80%, about 50% to about 85%, about 50% to about 90%, about 50% to about 95%, about 50% to about 100%, about 55% to about 60%, about 55% to about 65%, about 55% to about 70%, about 55% to about 75%, about 55% to about 80%, about 55% to about 85%, about 55% to about 90%, about 55% to about 95%, about 55% to about 100%, about 60% to about 65%, about 60% to about 70%, about 60% to about 75%, about 60% to about 80%, about 60% to about 85%, about 60% to about 90%, about 60% to about 95%, about 60% to about 100%, about 65% to about 70%, about 65% to about 75%, about 65% to about 80%, about 65% to about 85%, about 65% to about 90%, about 65% to about 95%, about 65% to about 100%, about 70% to about 75%, about 70% to about 80%, about 70% to about 85%, about 70% to about 90%, about 70% to about 95%, about 70% to about 100%, about 75% to about 80%, about 75% to about 85%, about 75% to about 90%, about 75% to about 95%, about 75% to about 100%, about

80% to about 85%, about 80% to about 90%, about 80% to about 95%, about 80% to about 100%, about 85% to about 90%, about 85% to about 95%, about 85% to about 100%, about 90% to about 95%, about 90% to about 100%, or about 95% to about 100%. In some examples, the protein activity level achieved is about 50%, about 55%, about 60%, about 65%, about 70%, about 75%, about 80%, about 85%, about 90%, about 95%, or about 100%. In some examples, the protein activity level achieved is at least about 50%, about 55%, about 60%, about 65%, about 70%, about 75%, about 80%, about 85%, about 90%, or about 95%. In some examples, the protein activity level achieved is at most about 55%, about 60%, about 65%, about 70%, about 75%, about 80%, about 85%, about 90%, about 95%, or about 100%.

In some examples, the amount of correctly joined RNA or full-length protein produced in a cell is sufficient to ameliorate or cure a condition or disease in a subject, as understood by one of skill in the art for the particular condition or disease. In some examples, the amount of correctly joined RNA or full-length protein produced in a cell is an effective amount. In some examples, this amount is equivalent to about 50% to 100% the amount of the RNA or protein produced in a normal cell. In some examples, this amount is equivalent to about 40% to about 100% the amount of the RNA or protein produced in a normal cell. In some examples, this amount is equivalent to about 40% to about 45%, about 40% to about 50%, about 40% to about 55%, about 40% to about 60%, about 40% to about 65%, about 40% to about 70%, about 40% to about 75%, about 40% to about 80%, about 40% to about 85%, about 40% to about 90%, about 40% to about 100%, about 45% to about 50%, about 45% to about 55%, about 45% to about 60%, about 45% to about 65%, about 45% to about 70%, about 45% to about 75%, about 45% to about 80%, about 45% to about 85%, about 45% to about 90%, about 45% to about 100%, about 50% to about 55%, about 50% to about 60%, about 50% to about 65%, about 50% to about 70%, about 50% to about 75%, about 50% to about 80%, about 50% to about 85%, about 50% to about 90%, about 50% to about 100%, about 55% to about 60%, about 55% to about 65%, about 55% to about 70%, about 55% to about 75%, about 55% to about 80%, about 55% to about 85%, about 55% to about 90%, about 55% to about 100%, about 60% to about 65%, about 60% to about 70%, about 60% to about 75%, about 60% to about 80%, about 60% to about 85%, about 60% to about 90%, about 60% to about 100%, about 65% to about 70%, about 65% to about 75%, about 65% to about 80%, about 65% to about 85%, about 65% to about 90%, about 65% to about 100%, about 70% to about 75%, about 70% to about 80%, about 70% to about 85%, about 70% to about 90%, about 70% to about 100%, about 75% to about 80%, about 75% to about 85%, about 75% to about 90%, about 75% to about 100%, about 80% to about 85%, about 80% to about 90%, about 80% to about 100%, about 85% to about 90%, about 85% to about 100%, or about 90% to about 100% the amount of the RNA or protein produced in a normal cell. In some examples, this amount is equivalent to about 40%, about 45%, about 50%, about 55%, about 60%, about 65%, about 70%, about 75%, about 80%, about 85%,

about 90%, or about 100% the amount of the RNA or protein produced in a normal cell. In some examples, this amount is equivalent to about at least about 40%, about 45%, about 50%, about 55%, about 60%, about 65%, about 70%, about 75%, about 80%, about 85%, or about 90% the amount of the RNA or protein produced in a normal cell. In some examples, this amount is equivalent to about at most about 45%, about 50%, about 55%, about 60%, about 65%, about 70%, about 75%, about 80%, about 85%, about 90%, or about 100% the amount of the RNA or protein produced in a normal cell.

The measurements of RNA or protein used to determine recombination efficiency or production level can be made by any suitable method known to those of skill in the art. In some examples, recombination efficiency or production level is determined by measuring an amount of functional protein expressed, for example by Western blotting. In some examples, recombination efficiency or production level is determined by measuring the RNA transcript, for example using two probe based quantitative real-time PCR. For example, the first assay spans a sequence fully contained in the 3' exonic coding sequence (labelled 3' probe). The second assay spans the junction between the 5' and the 3' exonic coding sequence (labelled junction probe). Reconstitution efficiency can be calculated as the ratio of (junction probe count)/(3' probe count). "Reconstitution efficiency," "recombination efficiency," and "splicing efficiency" are used interchangeably herein.

In some examples, a dimerization domain is about 20 to about 1000 nt, or about 50 to about 160 nt, or about 50 to about 500 nt, or about 50 to 1000 nt, wherein reconstitution efficiency results in production of an effective amount of correctly joined RNA or full-length protein. In some examples, a dimerization domain is about 50 to about 160 nt, wherein reconstitution efficiency results in production of an effective amount of correctly joined RNA or full-length protein.

Achieving efficient recombination between multiple RNA molecules allows for packaging and delivery of transgenes into AAVs, which exceed the packaging limit of a single AAV. AAV packaging limits represent a major hurdle for gene therapy approaches for diseases caused by the absence/defect of large genes. One application of this system is expression of large disease-causing genes using viral vectors with restricted packaging capacity. Disease and genes include but are not limited to (Disease (gene, OMIM gene identifier)): 1) Duchenne muscular dystrophy and Becker muscular dystrophy (dystrophin, OMIM:300377); 2) Dysferlinopathies (Dysferlin, OMIM:603009); 3) Cystic fibrosis (CFTR, OMIM:602421); 4) Usher's Syndrome 1B (Myosin VIIA, OMIM:276903); 5) Stargardt disease 1 (ABCA4, OMIM:601691); 6) Hemophilia A (Coagulation Factor VIII, OMIM:300841); 7) Von Willebrand disease (von Willebrand Factor, OMIM:613160); 8) Marfan Syndrome (Fibrillin 1, OMIM:134797); 9) Von Recklinghausen disease (neurofibromatosis-1, OMIM:162200), and hearing loss (OTOF, OMIM: 603681). Others are provided in Table 1. In addition, Cas9 proteins (such as those exemplified in Examples 20-23, can be expressed using the disclosed systems provided herein, for

example to treat genomic point mutations or activate or overexpress genes. Delivery of a transgene can be achieved by splitting it into multiple fragments using the approach provided herein.

Additional applications of the disclosed methods and systems include intersectional gene delivery for targeted gene expression. One can make use of differential infection/expression patterns of two viruses encoding a fragmented gene. The reconstituted protein will get expressed in an overlapping population of cells that represents the intersection of what either virus would express in on its own. Examples for such an application may include: (1) delivery of two halves (or three thirds, or other portions) of a protein using retrogradely transported viral vectors from two (or more) projection targets to label bifurcating dual projection neurons, (2) delivery of one fragment under the control of a promoter that is active in population A and the second fragment from a promoter active in population B to specifically tag/manipulate the AUB population, (3) delivery of the first half of a protein with a viral vector that has a tropism for population A and the second half with a viral vector that has a tropism for population B to specifically tag/manipulate the AUB population. Or, combinations of these approaches.

In one example the dimerization domains are aptamer sequences, for example to facilitate dimerization in the presence of a (a) small molecular trigger recognized by the aptamers, or a (b) protein that is present in the cell binding to the two halves and therefore stimulating dimerization.

In some embodiments the RNA-RNA interactions necessary for end-joining can be controlled positively or negatively by other nucleotides such as (a) an antisense oligonucleotide sequence with homology to the two halves (ssDNA triggered dimerization). In such an example, an antisense oligonucleotide having a complementary sequence to both halves bridges the two molecules together, thus facilitating spliceosome mediated recombination of the two molecules, (b) an antisense oligonucleotide sequence with homology to one of the two joining-RNAs could occlude RNA-dimerization of the two molecules and serve as an off-switch for gene expression, or (c) an endogenous cellular RNA with homology to the two halves (RNA triggered dimerization). In such an example, a cellular RNA (e.g., mRNA or retroelement) having a complementary sequence to both halves bridges the two molecules together, thus facilitating spliceosome mediated recombination of the two molecules.

These molecule, protein, or RNA mediated interactions allow for controllable/fine tuned gene expression levels: Through titrating in molecules that interact with the binding domains (*e.g.*, antisense oligonucleotides, small molecules, endogenous cellular RNAs), dimerization efficiency between the two halves can be modulated to regulate expression levels independent of promoter activity. Such an installment can be used if a narrow range of protein expression levels are needed.

### III. Systems

Provided herein is a system that can be used to recombine two or more RNA molecules, such as at least two, at least three, at least four, or at least five different RNA molecules (such as 2, 3, 4, 5, 6, 7, 8, 9 or 10 different RNA molecules) using synthetic introns containing dimerization sequences. Unlike fragmentation and reconstitution of two fragments at the protein level, the disclosed approach does not require extensive protein engineering to find a suitable split point. Reconstitution on an RNA level allows for seamless joining of two fragments of a protein. The disclosed methods and systems allow for large genes (and corresponding proteins), such as those greater than about 4.5 kb, at least 5 kb, at least 5.5 kb, at least 6 kb, at least kb, at least 8 kb, at least 8 kb, at least 10 kb, at least 13.5 kb, or at least 18 kb, to be divided into two or more fragments or portions, which can each be introduced into a cell or subject via separate vectors, such as multiple AAV. In one example, the system includes two portions for recombining two RNA molecules, for example wherein the target protein is encoded by at least about 4500 nt to about 9000 nt, such as 4000 nt to 5000 nt. In one example, the system includes three portions for recombining three RNA molecules, for example wherein the target protein is encoded by up to about 13,500 nt, such as about 4500 nt to about 13,500 nt or 4000 nt to 5000 nt. In one example, the system includes four portions for recombining four RNA molecules, for example wherein the target protein is encoded by up to about 18,000 nt, such as about 4500 nt to about 18,000 nt or 4000 nt to 5000 nt.. This helps to overcome the limited space available in vectors. In some examples, an endogenous promoter length limits the capability of its corresponding gene to be expressed in an AAV. In some examples, a coding sequence length limits its capability to be expressed in an AAV. In some examples, an endogenous promoter length and its coding sequence length limits their capability to be expressed together in an AAV. The disclosed systems can be used to express such long sequences that have been previously difficult to express in AAV.

In some examples, the target protein to be reconstituted is a protein associated with disease, such as a monogenic disease, recessive genetic disease, a disease caused by a mutation in a large gene (*e.g.*, greater than about 4500 nt, such as those of at least 5 kb, at least 5.5 kb, at least 6 kb, at least kb, at least 8 kb, at least 8 kb, at least 10 kb, at least 13.5 kb, or at least 18 kb), and/or disease caused by a gene (such as a promoter + coding sequence) that exceed AAV's capacity (*e.g.*, greater than 5000 nt). Examples of such diseases include, but are not limited to, hemophilia A (caused by mutations in the F8 gene, 7kb coding region), hemophilia B (caused by mutations in the F9 gene), Duchenne muscular dystrophy (caused by mutations in the dystrophin gene, 11 kb coding region), sickle cell anemia (caused by mutation in beta globin domain of hemoglobin, which has a promoter of about 3.5 kb), Stargardt disease (caused by mutations in the ABCA4 gene, 6.9 kb coding region), Usher syndrome (caused by a mutation in MYO7A, 7 kb coding region, resulting in hearing loss and visual impairment).

In one example, the target protein to be reconstituted is one that can treat a disease, such as a cancer, such as a cancer of the breast, lung, prostate, liver, kidney, brain, bone, ovary, uterus, skin, or

colon. In one example, the therapeutic target protein to be reconstituted is a toxin, such as an AB toxin, such as diphtheria toxin A or pseudomonas exotoxin A, or a form that lacks receptor binding activity (*e.g.*, diphtheria toxin DAB389, DAB486, DT388, DT390, or pseudomonas exotoxin A PE38 or PE40).

In some examples, an RNA sequence encoding the target protein and used in the disclosed methods and systems are codon optimized for expression in a target organism or cell, such as codon optimized for expression in a human, canine, pig, feline, mouse, or rat cell. Thus, in some examples, the RNA coding sequence includes preferred codons (*e.g.*, does not include rare codons with low utilization). Codon optimization can be performed by identifying abundant tRNA levels in the target organism or cells. In some examples, an RNA sequence encoding the protein is de-enriched for cryptic splice donor and acceptor sites to maximize an RNA recombination reaction.

In some examples, a protein is divided into two portions, such as about two equal halves (or other proportions, such as portion A expressing about 1/3 and portion B expressing about 2/3, or portion A expressing about 1/4 and portion B expressing about 3/4, etc.). However, it is not required that each portion be the same number of nucleotides (or encode the same number of amino acids). In such an example, the method can use two synthetic nucleic acid molecules (*e.g.*, RNA or DNA encoding such RNA), one which includes a coding sequence for an N-terminal portion of the protein, and another which includes a coding sequence for a C-terminal portion of the protein. Based on this foundation, one skilled in the art will appreciate that in addition to dividing a protein into two fragments or portions, proteins of interest can be divided or split into more than two fragments, such as three fragments. The design principle of the intronic sequences of three RNA molecules is similar to that of the two, but instead a different pair of dimerization domains for one of the two junctions is utilized. Thus, for example, an N-terminal protein coding sequence is followed by an intronic sequence with a specific binding domain (*e.g.*, first dimerization sequence), the middle coding sequence includes an intronic sequence with a complementary sequence to the first dimerization sequence (second dimerization sequence). The middle coding fragment is followed by another intronic fragment with another dimerization sequence (third dimerization sequence, different from the second dimerization sequence). The third fragment includes the C-terminal coding sequence of the protein, and includes an intronic region with a dimerization sequence (fourth dimerization sequence) complementary to the third dimerization sequence. In the use of more than one middle portion, the two middle portions may be referred to as a middle portion and a first middle portion, or as a first middle portion and a second middle portion, or as a first middle portion, a second middle portion and a third middle portion, etc., in a way understood to distinguish the respective portions.

In one example, a desired protein is divided into an N-terminal portion and a C-terminal portion (*e.g.*, divided in roughly half, or unequal apportionment, such as 1/3 and 2/3 or 1/4 and 3/4), which can be reconstituted using the disclosed systems and methods. Referring to FIG. 6A, in such an example,

the system includes at least two synthetic nucleic acid molecules 110, 150. Each nucleic acid molecule 110, 150 can be composed of DNA or RNA (if RNA, promoter 112, 152 are absent). In some examples, each of 110, 150 is about at least 100 nucleotides/ribonucleotides (nt) in length, such as at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, at least 8000 nt, at least 10,000 nt, such as 200 to 10,000 nt, 200 to 8000 nt, 500 to 5000 nt, or 200 to 1000 nt. The molecules 110, 150 can include natural and/or non-natural nucleotides or ribonucleotides.

Molecule 110 is the 5'-located molecule of the system, as it includes a splice donor 116. In embodiments where molecule 110 is DNA, it includes a promoter 112 operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a coding sequence for an N-terminal portion of the target protein 114, wherein the coding sequence for an N-terminal portion of the target protein 114 comprises a splice junction at a 3'-end of the target protein coding sequence, SD 116, optional DISE 118, optional ISE 120, dimerization domain 122, and optional polyadenylation sequence 124. Any promoter 112 (or enhancer) can be used, such as one that utilizes RNA polymerase II, such as a constitutive or inducible promoter. In some examples, promoter 112 is a tissue-specific promoter, such as one constitutively active in muscle tissue (such as skeletal or cardiac), optical tissue (such as retinal tissue), inner ear tissue, liver tissue, pancreatic tissue, lung tissue, skin tissue, bone, or kidney tissue. In some examples, promoter 112 is a cell-specific promoter, such as one constitutively active in a cancer cell, or a normal cell. In some examples, promoter 112 is an endogenous promoter of the target protein expressed, and in some example is long (*e.g.*, at least 2500 nt, at least 3000 nt, at least 4000 nt, at least 5000 nt, or at least 7500 nt). In some examples, promoter 112 is at least about 50 nucleotides (nt) in length, such as at least 100, at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, at least 8000 nt, at least 9000 nt, or at least 10,000 nt, such as 50 to 10,000 nt, 100 to 5000 nt, 500 to 5000 nt, or 50 to 1000 nt in length. In some examples, molecule 110 is DNA, and is at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, or at least 8000 nt, such as 200 to 10,000 nt, 200 to 8000 nt, 500 to 5000 nt, or 200 to 1000 nt in length. As shown in FIG. 6F, in embodiments where molecule 110 is RNA, for example after transcription of the DNA into RNA, molecule 110 does not include promoter 112, and 114 is the RNA encoded by the coding sequence for an N-terminal portion of the target protein. In some examples, molecule 110 is RNA, does not include promoter 112, and is at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, or at least 8000 nt, such as 200 to 10,000 nt, 200 to 8000 nt, 500 to 5000 nt, or 200 to 1000 nt in length. The molecule 110 (with or without promoter 112) can include natural and/or non-natural nucleotides or ribonucleotides.

The splice junction around the 3' end of the N-terminal coding sequence (or RNA sequence encoded thereby) 114 can match the consensus sequence found in the target cell or organism into which molecules 110, 150 are introduced. In humans the splice junction sequence is AG (adenine-guanine) or UG (uracil-guanine) at position -1 and -2 of the 5' splice site for U2-dependent introns or AG, UG, CU (cytosine-uracil), or UU for U12-dependent introns. Thus, in some examples, the splice junction is 2 nt in length, and the 3' end of the N-terminal coding portion 114 is AG, UG, CU or UU. In some examples a DNA molecule encoding a portion of a target protein comprises sequences that encode parts of multiple splice junctions, e.g., at the 3' end of the DNA molecule encoding the N-terminal portion of the target protein, and at the 5' end of the DNA molecule encoding the C-terminal portion of the target protein.

The remaining 3'-terminal portion of molecule 110 is intronic, 130. In some examples, intronic sequence 130 is about at least 10 nt, such as at least 20 nt, at least 50 nt, at least 100 nt, at least 250 nt, at least 250 nt, at least 300 nt, at least 400 nt, or at least 500 nt in length, such as 20 to 500, 20 to 250, 20 to 100, 50 to 100, or 50 to 200 nt in length. Immediately following N-terminal coding sequence (or RNA encoded thereby) 114 is a splice donor (SD) 116 (such as a SD consensus sequence, such as a SD human consensus sequence). Thus SD 116 of intronic sequence 130 is 3' to N-terminal coding sequence 114. SD 116 forms a recognition sequence for the spliceosome components to bind to the RNA molecule. The sequence of SD 116 can be a SD consensus sequence found in the target cell or organism into which molecules 110, 150 are introduced. In some examples, SD 116 is at least 2 nt, such as at least 5 nt, or at least 10 nt in length, such as 2 to 10, 2 to 8, 2 to 5 or 5 to 10 nt. The SD 116 can be used to recruit U2 or U12 dependent splicing machinery. In one example, U2 dependent splicing is used in human cells, and the SD 116 sequence includes or is GUAAGUAUU. In one example, U12 dependent splicing is used in human cells, and the SD 116 sequence includes or is AUAUCCUUUUUA (SEQ ID NO: 137) or GUAUCCUUUUUA (SEQ ID NO: 138). Throughout, it is understood that RNA sequences can be described using nucleotides A,G,T and C, and that DNA sequences can be described using nucleotides A,G,U and C.

Intronic sequence 130 optionally includes one or both of a set of splicing enhancer sequences referred to as downstream intronic splice enhancer (DISE) 118 and intronic splice enhancer (ISE) 120, which stimulate action (e.g., increase activity) of the spliceosome. In some examples, intronic sequence 130 includes at least two splicing enhancer sequences, such as at least 3, at least 4, or at least 5 splicing enhancer sequences. Exemplary splicing enhancer sequences include DISE 118 and ISE 120. In some examples, inclusion of one or more splicing enhancer sequences 118, 120 in intronic sequence 130 increases splicing efficiency by at least 20%, at least 30%, at least 40%, at least 50%, at least 75%, at least 80%, at least 90% or at least 95%. Exemplary splicing enhancer sequences that can be used are provided in SEQ ID NOS: 26-136, 151, and 152, as well as GGGTTT, GGTGGT, TTTGGG,



GAGGGG, GGTATT, GTAACG, GGGGGTAGG, GGAGGGTTT, GGGTGGTGT TTCAT, CCATTT, TTTTAAA, TGCAT, TGCATG, TGTGTT, CTAAC, TCTCT, TCTGT, TCTTT, TGCATG, CTAAC, CTGCT, TAACC, AGCTT, TTCATTA, GTTAG, TTTTGC, ACTAAT, ATGTTT, CTCTG, GGG, GGG(N)2-4GGG, TGGG, YCAY, UGCAUG, or  $3x(G_{3-6}N_{1-7})$ . In some examples, if DISE 118 is present, can be at least 3 nt, at least 4 nt, at least 5 nt, at least 10 nt, at least 25 nt, at least 50 nt, at least 75 nt, or at least 100 nt in length, such as 3 to 10, 3 to 11, 4 to 11, 5 to 11, 10 to 50, 5 to 100, 10 to 25, 10 to 20, or 20 to 75 nt, the sequence of DISE 118 is or comprises CUCUUUCUUUTCCAUGGGUUGGCU (SEQ ID NO: 134), TGCATG, CTAAC, CTGCT, TAACC, AGCTT, TTCATTA, GTTAG, TTTTGC, ACTAAT, ATGTTT or CTCTG. In some examples, if ISE 120 is present, it can be about at least 3 nt, at least 4 nt, at least 5 nt, at least 10 nt, such as at least 20 nt, at least 25 nt, at least 30 nt, at least 40 nt, or at least 50 nt in length, such as 3 to 10, 3 to 11, 4 to 11, 5 to 11, 10 to 50, 20 to 25, 10 to 25, 10 to 20, or 20 to 40 nt in length. In one example, the sequence of ISE 120 is or comprises GGCUGAGGGAAGGACUGUCCUGGG (SEQ ID NO: 135), GGGUUAUGGGACC (SEQ ID NO: 136), TTCAT, CCATTT, TTTTAAA, TGCAT, TGCATG, TGTGTT, CTAAC, TCTCT, TCTGT, or TCTTT. In some examples, intronic sequence 130 includes at least two, at least 3, or at least 4 ISEs 120. In some examples, ISE 120 is or comprises at least one sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ NO: 173, 174, 175, 176, 177, 178, 179, 180, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 199, 200, 201, 202, or 203, such as at least 2, at least 3 of such sequences, such as 1, 2, 3, 4 or 5 of such sequences. In some examples, DISE 118 is or comprises at least one sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ NO: 173, 174, 175, 176, 177, 178, 179, 180, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 199, 200, 201, 202, or 203, such as at least 2, at least 3 of such sequences, such as 1, 2, 3, 4 or 5 of such sequences.

The SD 116 (and if present also enhancer sequences 118, 120) is followed 3' by a dimerization domain 122 used to bring the N-terminal coding sequence (or RNA encoded thereby) 114, and C-terminal coding sequence 154 to be combined, together. Intronic sequence 130 portion of molecule 110 can optionally include at the 3'-end a polyadenylation site 124, which terminates transcription of that fragment. In some examples, polyadenylation sequence 124 is a polyA sequence of at least 15 As, such as 15 to 30 or 15 to 20 As.

In some examples, first dimerization domain 122 (and second dimerization domain 154 of molecule 150) includes a plurality of unpaired nucleotides (that is, unpaired within the structure of the molecule 110 itself). Having unpaired nucleotides in the dimerization domain allows the 5' (or first) dimerization domain 122 and the 3' (or second) dimerization domain 154 to interact through base

pairing. Through this interaction, molecules 110 and 150 are kept in proximity which prompts the spliceosome to recombine the two molecules by joining the N-terminal coding region (or RNA encoded thereby) 114 and the C terminal coding region (or RNA encoded thereby) 164.

In one example, dimerization domain 122 (and 154) includes “hypodiverse sequences,” which contain a limited diversity of nucleotides and are thus unlikely to form stem loops with themselves in the secondary structure of each molecule 110, 150. Such a hypodiverse dimerization domain 122 (and 154) can be a relatively open configuration, independent of the sequences of the DNA encoding the N- and C-terminus of the protein (or RNA encoded thereby) 114, 164. This allows the nucleotides of the first dimerization domain 122 to be available to form base pairs with the corresponding second dimerization domain 154 of molecule 150, allowing subsequent joining of the N-terminal coding sequence (or RNA encoded thereby) 114 and C-terminal coding sequence (or RNA encoded thereby) 164. In some examples, first and second dimerization domain 122, 154 includes hypodiverse sequences interspersed with sequences that can form a stem, which results in local RNA loops that are open and available for basepairing in the absence of pseudoknot formation (FIG. 6B). Exemplary hypodiverse sequences include a repeated series of Us (such as 30 to 500 Us), a repeated series of As (such as 30 to 500 As), a repeated series of Gs (such as 30 to 500 Gs), a repeated series of Cs (such as 30 to 500 Cs), a mixture containing only As and Gs (such as 30 to 500 As and Gs, *e.g.*, AAAGAAGGAA(...)) (SEQ ID NO: 149) which can be repeated), a mixture containing only Cs and Us (such as 30 to 500 Cs and Us, *e.g.*, CUUUCUUUUCUU(...)) (SEQ ID NO: 150) which can be repeated). Other exemplary hypodiverse sequences include complementary sequences that form helices flanked by hypodiverse sequences.

In some examples, first and second dimerization domain 122, 154 only include purines or only include pyrimidines. In one example, the first dimerization domain 122 only includes purines, while the second dimerization domain 154 only includes pyrimidines. In another example, the first dimerization domain 122 only includes pyrimidines, while the second dimerization domain 154 only includes purines. Due to the inability of purines to pair with themselves (and pyrimidines likewise) these stretches of RNA have an open predicted structure.

In some examples, first and second dimerization domain 122, 154 do not include cryptic splice acceptors that could compete with RNA recombination, such as sequences similar to the splice donor consensus sequence NNNAGGUNNNN (SEQ ID NO: 151) or NNNUGGUNNNN (SEQ ID NO: 152) (wherein N refers to any nucleotide). In some examples, first dimerization domain 122 is no more than 1000 nt, such as no more than 750 nt, or more than 500 nt, such as 6 to 1000 nt, 10 to 1000 nt, 20 to 1000 nt, 30 to 1000 nt, 30 to 750 nt, 30 to 500 nt, 50 to 500 nt, 50 to 100 nt, or 100 to 250 nt. In some examples, first dimerization domain 122 is greater than 50 nt, such as at least 51 nt, at least 100 nt, at least 150 nt, at least 161 nt, or at least 170 nt, such as 51 to 159 nt, 51 to 150 nt, 51 to 120 nt, 51 to 100

nt, or 51 to 70 nt. In some examples, first dimerization domain 122 is greater than 160 nt, such as at least 161 nt, at least 170 nt, at least 180 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, at least 600 nt, at least 700 nt, at least 800 nt, at least 900 nt, or at least 1000 nt, such as 161 to 100 nt, 161 to 500 nt, 161 to 300 nt, 161 to 200 nt, or 161 to 170 nt. In some examples, first dimerization domain 122 is less than 50 nt, such as 6 to 49 nt, 6 to 45 nt, 6 to 40 nt, 6 to 30 nt, 6 to 20 nt, or 6 to 10 nt.

In some examples, a dimerization domain is 20 to 160 nt, 50-500 nt, or 500-1000 nt. In some examples, a dimerization domain is about 20 nt to about 160 nt. In some examples, a dimerization domain is about 20 nt to about 40 nt, about 20 nt to about 50 nt, about 20 nt to about 70 nt, about 20 nt to about 90 nt, about 20 nt to about 100 nt, about 20 nt to about 110 nt, about 20 nt to about 120 nt, about 20 nt to about 130 nt, about 20 nt to about 140 nt, about 20 nt to about 150 nt, about 20 nt to about 160 nt, about 40 nt to about 50 nt, about 40 nt to about 70 nt, about 40 nt to about 90 nt, about 40 nt to about 100 nt, about 40 nt to about 110 nt, about 40 nt to about 120 nt, about 40 nt to about 130 nt, about 40 nt to about 140 nt, about 40 nt to about 150 nt, about 40 nt to about 160 nt, about 50 nt to about 70 nt, about 50 nt to about 90 nt, about 50 nt to about 100 nt, about 50 nt to about 110 nt, about 50 nt to about 120 nt, about 50 nt to about 130 nt, about 50 nt to about 140 nt, about 50 nt to about 150 nt, about 50 nt to about 160 nt, about 70 nt to about 90 nt, about 70 nt to about 100 nt, about 70 nt to about 110 nt, about 70 nt to about 120 nt, about 70 nt to about 130 nt, about 70 nt to about 140 nt, about 70 nt to about 150 nt, about 70 nt to about 160 nt, about 90 nt to about 100 nt, about 90 nt to about 110 nt, about 90 nt to about 120 nt, about 90 nt to about 130 nt, about 90 nt to about 140 nt, about 90 nt to about 150 nt, about 90 nt to about 160 nt, about 100 nt to about 110 nt, about 100 nt to about 120 nt, about 100 nt to about 130 nt, about 100 nt to about 140 nt, about 100 nt to about 150 nt, about 100 nt to about 160 nt, about 110 nt to about 120 nt, about 110 nt to about 130 nt, about 110 nt to about 140 nt, about 110 nt to about 150 nt, about 110 nt to about 160 nt, about 120 nt to about 130 nt, about 120 nt to about 140 nt, about 120 nt to about 150 nt, about 120 nt to about 160 nt, about 130 nt to about 140 nt, about 130 nt to about 150 nt, about 130 nt to about 160 nt, about 140 nt to about 150 nt, about 140 nt to about 160 nt, or about 150 nt to about 160 nt. In some examples, a dimerization domain is about 20 nt, about 40 nt, about 50 nt, about 70 nt, about 90 nt, about 100 nt, about 110 nt, about 120 nt, about 130 nt, about 140 nt, about 150 nt, or about 160 nt. In some examples, a dimerization domain is at least about 20 nt, about 40 nt, about 50 nt, about 70 nt, about 90 nt, about 100 nt, about 110 nt, about 120 nt, about 130 nt, about 140 nt, or about 150 nt. In some examples, a dimerization domain is at most about 40 nt, about 50 nt, about 70 nt, about 90 nt, about 100 nt, about 110 nt, about 120 nt, about 130 nt, about 140 nt, about 150 nt, or about 160 nt.

In some examples, a dimerization domain is about 50 nt to about 500 nt. In some examples, a dimerization domain is about 50 nt to about 100 nt, about 50 nt to about 150 nt, about 50 nt to about 200 nt, about 50 nt to about 250 nt, about 50 nt to about 300 nt, about 50 nt to about 350 nt, about 50 nt

to about 400 nt, about 50 nt to about 500 nt, about 100 nt to about 150 nt, about 100 nt to about 200 nt, about 100 nt to about 250 nt, about 100 nt to about 300 nt, about 100 nt to about 350 nt, about 100 nt to about 400 nt, about 100 nt to about 500 nt, about 150 nt to about 200 nt, about 150 nt to about 250 nt, about 150 nt to about 300 nt, about 150 nt to about 350 nt, about 150 nt to about 400 nt, about 150 nt to about 500 nt, about 200 nt to about 250 nt, about 200 nt to about 300 nt, about 200 nt to about 350 nt, about 200 nt to about 400 nt, about 200 nt to about 500 nt, about 250 nt to about 300 nt, about 250 nt to about 350 nt, about 250 nt to about 400 nt, about 250 nt to about 500 nt, about 300 nt to about 350 nt, about 300 nt to about 400 nt, about 300 nt to about 500 nt, about 350 nt to about 400 nt, about 350 nt to about 500 nt, or about 400 nt to about 500 nt. In some examples, a dimerization domain is about 50 nt, about 100 nt, about 150 nt, about 200 nt, about 250 nt, about 300 nt, about 350 nt, about 400 nt, or about 500 nt. In some examples, a dimerization domain is at least about 50 nt, about 100 nt, about 150 nt, about 200 nt, about 250 nt, about 300 nt, about 350 nt, or about 400 nt. In some examples, a dimerization domain is at most about 100 nt, about 150 nt, about 200 nt, about 250 nt, about 300 nt, about 350 nt, about 400 nt, or about 500 nt.

In some examples, the sequence of first and second dimerization domains 122 and 154 are determined by in silico structure prediction screening (*e.g.*, RNA folding structure prediction is used to screen a library of possible dimerization domain sequences; sequences with a large proportion of unpaired nucleotides in both the dimerization domain and the corresponding anti-dimerization domain are selected), hypodiverse nucleotide design (*e.g.*, dimerization domain designed to include a stretch of hypodiverse sequence, such as a repeat sequence of only U, only A, only C, only G, only R (G and A), or only Y (U and C), the sequence cannot fold onto itself), or empirical screening (*e.g.*, a library of dimerization domains and corresponding anti-dimerization domains are synthesized and screened for maximal recombination efficiency).

In some examples, the sequence of first and second dimerization domains 122, 154 are designed to contain complementary RNA hairpin structures (also called stem loops) that can form strong kissing loop interactions with their counter parts. In some examples, kissing loops are used when three or more dimerization domains are used to join three or more portions of a coding sequence, such as four or more or five or more dimerization domains, such as 3, 4, 5, 6, 7, 8, 9 or 10 dimerization domains (*e.g.*, FIG. 6E). Each hairpin loop (or stem loop) of a kissing loop is composed of at least two complementary sequences (*e.g.*, form a stem) separated by a region of non-complementary sequence (*e.g.*, form a loop). In some examples, a dimerization domain can be composed of 1 or more (such as at least 2, at least 3, at least 4, or at least 5, such as 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20) loops. In some examples with multiple loops, all or some of the loops can be repeated. In some examples with multiple loops, all or some loops can be different. In some examples, each complementary sequence is about 4 to 100 nt, which are separated by a loop of about 3 to 20 nt. Base-pairing between the two

complementary sequences results in a helix (or stem), for example of at least 4 bp, at least 5 bp, at least 10 bp, at least 20 bp, at least 30 bp, at least 40 bp, at least 50 bp, at least 75 bp, at least 90 bp, or at least 100 bp, such as 4 to 100 bp, 5 to 75 bp, or 10 to 50 bp. In some examples, the loop portion is at least 3 nt, at least 5 nt, at least 10 nt, at least 15 nt, or at least 20 nt, such as 3 to 20 nt, 5 to 15 nt or 5 to 10 nt, wherein the loop is not base paired. Complementary sequences between two hairpin loops result in base pairing, and generation of a kissing loop/kissing stem loop interaction. In some examples, the complementary sequences between the two hairpin loops occurs between at least 3 nucleotides of one loop with at least 3 nucleotides of a second loop, such as at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 19, or at least 20 nt (such as 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20) of the first loop, with at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 19, or at least 20 nt (such as 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20) of the second loop. In some examples, the complementary sequences between the two hairpin loops occurs between at least 15%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 95%, or 100% of the total loop sequence.

In some instances, the stems of the kissing loops are chosen to base pair in trans between the two RNA molecules. In such an example, after forming a kissing loop interaction of one hairpin loop on one molecule with another hairpin loop on a second molecule, the respective stem (or helix) regions of the initial hairpin loops can base pair in trans between the two RNA molecules through strand replacement/invasion and extended duplex formation. In some examples, within the initial loop sequence, up to about 85% of nucleotides can remain unpaired after extended duplex formation (*e.g.*, about 15% of the nt are paired between the two loops). In some examples, the kissing loop is based on the HIV-1 DIS loop (SEQ ID NOS: 139 and 140, FIG. 17A), and includes two A nucleotides on the 5' side of 6 nucleotides of complementary sequence, followed by one A nucleotide on the 3' side (*e.g.*, AANNNNNA where N can be any of A, U, G, or C). In some examples, the kissing loop is based on the HIV-2 kissing loop dimerization domain (SEQ ID NOS: 141 and 142, FIG. 17B), and includes a G and an A nucleotide on the 5' side of six nucleotides of complementary sequence followed by three A nucleotides on the 3' side (*e.g.*, GANNNNNAAA (SEQ ID NO: 153) where N can be A, U, G, or C).

In one configuration, extended duplex formation is favored by inclusion of mismatches in the initial stems that result in higher percentage of matching in the extended duplex. Thus, in some examples, the helix or stem region of a hairpin loop can contain up to 30% of base pairs that are not paired initially (*e.g.*, no more than 30%, no more than 20%, no more than 15%, no more than 10%, no more than 5%, or no more than 1%, such as 1 to 30%, 5 to 30%, 10 to 30%, or 25 to 30% of base pairs are not paired initially). These regions of non-pairing can form bulges, mismatches, or internal loops.

In addition to an interaction of two hairpin loops (kissing loop interaction), other forms of loop interactions can be utilized for the first and second dimerization domains 122, 154. In one example the loops are bulges, where one strand of a base paired helix contains one or more nucleotides that bulge out from the stem structure. Exemplary bulges are at least 1 nt, at least 2 nt, at least 3 nt, at least 4 nt, at least 5 nt, at least 10 nt or at least 20 nt, such as 1 to 20 nt, 1 to 15 nt, 1 to 10 nt, or 5 to 10 nt, or 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 nt. In one example the loops are internal loops, for example, where 1 or more nucleotides in a helix are mismatched, resulting in a helix interrupted by an internal loop at the positions of mismatch. In some examples the helix is at least 4 nt on each of the strands (*e.g.*, at least 5 nt, at least 10 nt, at least 20 nt, at least 30 nt, at least 40 nt, at least 50 nt, at least 75 nt, at least 90 nt, or at least 100 nt, such as 4 to 100 nt, 5 to 75 nt, or 10 to 50 nt. such as 4 to 100 nt), on either side of the internal loop that is at least 1 nt (*e.g.*, at least 2 nt, at least 3 nt, at least 4 nt, at least 5 nt, at least 10 nt or at least 20 nt, such as 1 to 20 nt, 1 to 15 nt, 1 to 10 nt, or 5 to 10 nt, or 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 nt on each of the strands). In one example the loops are multi-branched loops, wherein three helices or stems form a triangle with one or more unpaired nucleotides connecting the three helices. In some examples, each of the helices is at least 4 bp (*e.g.*, at least 5 bp, at least 10 bp, at least 20 bp, at least 30 bp, at least 40 bp, at least 50 bp, at least 75 bp, at least 90bp, or at least 100 bp, such as 4 to 100 bp, 5 to 75 bp, or 10 to 50 bp), and the unpaired nucleotides that form the triangle are at least 3 nt (*e.g.*, at least 4 nt, at least 5 nt, at least 10 nt, at least 20, at least 15, at least 30, at least 40, at least 50, or at least 60 nt, such as 3 to 60 nt, 3 to 30 nt, 3 to 25 nt, or 5 to 20 nt, such as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2, 25, 30, 35, 40, 45, 50, 55 or 60 nucleotides). A kissing interaction can occur between any two of these types of loops (*e.g.*, between two or more binding domains that each include one or more helices). In some examples, helices within one dimerization domain (*e.g.*, first dimerization domain 122) have a direct counterpart in the other binding domain (*e.g.*, second dimerization domain 154) to allow for extended duplex formation after initial loop kissing interaction. In some examples, dimerization domains containing helices to generate loops, form a single kissing stem loop upon interaction between the two or more dimerization domains (*e.g.*, 122, 154 of FIG. 6A). In some examples, dimerization domains containing helices form multiple loops for kissing loop interactions upon interaction between the two or more dimerization domains (*e.g.*, 122, 154 of FIG. 6A). In some examples, one or more dimerization domains (*e.g.*, 122 of FIG. 6A) contain helices destabilized by the inclusion of bulges, single base bulges, mismatches or internal loops, or G-U wobble pairs, but match to the other binding domain (*e.g.*, 154 of FIG. 6A), to favor extended duplex formation after initial kissing/pairing. In some examples, one or more dimerization domains (*e.g.*, 122 of FIG. 6A) contain destabilized helices, which when stabilized (*e.g.*, theophylline switch kissing loop) expose a loop that can interact with a second dimerization domain (*e.g.*, 122 of FIG. 6A) via loop-loop interactions (*e.g.*, kissing/pairing).

In some examples these stem loops contain at least 10 nt, such as at least 20 nt, at least 25 nt, at least 50 nt, at least 75 nt, or at least 100 nt in length, such as 10 to 50, 20 to 25, 10 to 100, 10 to 20, or 20 to 40 nt in length. Each dimerization domain can contain at least 1 individual stem loop, such as at least 2, at least 5, at least 10, at least 15, or at least 20, such as 1 to 20, 2 to 5 or 1 to 10 individual stem loops.

In some examples, 3 to 10 portions of a coding sequence are joined by 2 to 9 kissing loops, e.g., 3 portions are joined by 2 kissing loops, 4 portions are joined by 3 kissing loops, etc., wherein each of the 2 to 9 kissing loops are different. In some examples, a kissing loop comprises multiple stem loops, e.g., 2 to 20 stem loops. In some examples, each of the multiple stem loops in the kissing loop are the same. In some examples, each of the multiple stem loops in the kissing loop are different. In some examples, a dimerization domain comprises 1 to 20 stem loops. In some examples, a dimerization domain comprises 1 stem loop to 20 stem loops. In some examples, a dimerization domain comprises 1 stem loop to 2 stem loops, 1 stem loop to 3 stem loops, 1 stem loop to 4 stem loops, 1 stem loop to 5 stem loops, 1 stem loop to 6 stem loops, 1 stem loop to 7 stem loops, 1 stem loop to 8 stem loops, 1 stem loop to 9 stem loops, 1 stem loop to 10 stem loops, 1 stem loop to 15 stem loops, 1 stem loop to 20 stem loops, 2 stem loops to 3 stem loops, 2 stem loops to 4 stem loops, 2 stem loops to 5 stem loops, 2 stem loops to 6 stem loops, 2 stem loops to 7 stem loops, 2 stem loops to 8 stem loops, 2 stem loops to 9 stem loops, 2 stem loops to 10 stem loops, 2 stem loops to 15 stem loops, 2 stem loops to 20 stem loops, 3 stem loops to 4 stem loops, 3 stem loops to 5 stem loops, 3 stem loops to 6 stem loops, 3 stem loops to 7 stem loops, 3 stem loops to 8 stem loops, 3 stem loops to 9 stem loops, 3 stem loops to 10 stem loops, 3 stem loops to 15 stem loops, 3 stem loops to 20 stem loops, 4 stem loops to 5 stem loops, 4 stem loops to 6 stem loops, 4 stem loops to 7 stem loops, 4 stem loops to 8 stem loops, 4 stem loops to 9 stem loops, 4 stem loops to 10 stem loops, 4 stem loops to 15 stem loops, 4 stem loops to 20 stem loops, 5 stem loops to 6 stem loops, 5 stem loops to 7 stem loops, 5 stem loops to 8 stem loops, 5 stem loops to 9 stem loops, 5 stem loops to 10 stem loops, 5 stem loops to 15 stem loops, 5 stem loops to 20 stem loops, 6 stem loops to 7 stem loops, 6 stem loops to 8 stem loops, 6 stem loops to 9 stem loops, 6 stem loops to 10 stem loops, 6 stem loops to 15 stem loops, 6 stem loops to 20 stem loops, 7 stem loops to 8 stem loops, 7 stem loops to 9 stem loops, 7 stem loops to 10 stem loops, 7 stem loops to 15 stem loops, 7 stem loops to 20 stem loops, 8 stem loops to 9 stem loops, 8 stem loops to 10 stem loops, 8 stem loops to 15 stem loops, 8 stem loops to 20 stem loops, 9 stem loops to 10 stem loops, 9 stem loops to 15 stem loops, 9 stem loops to 20 stem loops, 10 stem loops to 15 stem loops, 10 stem loops to 20 stem loops, or 15 stem loops to 20 stem loops. In some examples, a dimerization domain comprises 1 stem loop, 2 stem loops, 3 stem loops, 4 stem loops, 5 stem loops, 6 stem loops, 7 stem loops, 8

stem loops, 9 stem loops, 10 stem loops, 15 stem loops, or 20 stem loops. In some examples, a dimerization domain comprises at least 1 stem loop, 2 stem loops, 3 stem loops, 4 stem loops, 5 stem loops, 6 stem loops, 7 stem loops, 8 stem loops, 9 stem loops, 10 stem loops, or 15 stem loops. In some examples, a dimerization domain comprises at most 2 stem loops, 3 stem loops, 4 stem loops, 5 stem loops, 6 stem loops, 7 stem loops, 8 stem loops, 9 stem loops, 10 stem loops, 15 stem loops, or 20 stem loops.

Other mechanisms can be used to allow the two or more dimerization domains (*e.g.*, 122, 154 of FIG. 6A) to bind or interact with one another sufficient for recombination of the coding sequences to occur. In some examples, the two or more dimerization domains (*e.g.*, 122, 154 of FIG. 6A) are nucleic acid aptamers (such as RNA aptamers) that can interact with one another, for example through a non-base pairing interaction, or can bind to a common molecule (*e.g.*, protein, ATP, metal ion, co-factor, or synthetic ligand). In some examples, two or more dimerization domains (*e.g.* 122, 154 of FIG. 6A) do not hybridize to one another, but can both (or all) hybridize to the same bridge nucleic acid molecule. In some examples, such a bridge nucleic acid molecule can be exogenously provided to the cells, tissues, or organism. In some examples, such a bridge nucleic acid molecule can be a DNA or RNA sequence inside the cell, such as a transcript or genomic locus. In some examples, the two or more dimerization domains (*e.g.*, 122, 154 of FIG. 6A) are sequences that can interact with one another, for example through a non-base pairing interaction.

Molecule 150 is the 3'-located molecule, and includes a splice acceptor (SA) 162 and a second dimerization domain 154. In embodiments where molecule 150 is DNA, it includes a second promoter 152 followed by intronic sequence 170. Promoter 152 can be operably linked to intronic sequence 170. Any promoter 152 can be used, such as a constitutive or inducible promoter. In some examples, promoter 152 is a tissue-specific promoter, such as one constitutively active in muscle tissue (such as skeletal or cardiac), optical tissue (such as retinal tissue), inner ear tissue, liver tissue, pancreatic tissue, lung tissue, skin tissue, bone, or kidney tissue. In some examples, promoter 112 is a cell-specific promoter, such as one constitutively active in a cancer cell, or a normal cell. In some examples, promoter 112 is an endogenous promoter of the target protein expressed, and in some example is long (*e.g.*, at least 2500 nt, at least 3000 nt, at least 4000 nt, at least 5000 nt, or at least 7500 nt). In some examples, promoter 112 is at least about 50 nucleotides (nt) in length, such as at least 100, at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, at least 8000 nt, at least 9000 nt, or at least 10,000 nt, such as 50 to 10,000 nt, 100 to 5000 nt, 500 to 5000 nt, or 50 to 1000 nt in length. In some examples promoter 112 and promoter 152 are the same promoter. In other examples, promoter 112 and promoter 152 are the different promoters. In some examples, molecule 150 is DNA, and is at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, or at



least 8000 nt, such as 200 to 10,000 nt, 200 to 8000 nt, 500 to 5000 nt, or 200 to 1000 nt in length. As shown in FIG. 6F, in embodiments where molecule 150 is RNA, for example after expression of the DNA into RNA, molecule 150 no longer includes promoter 152, and 164 is the RNA encoded by the coding sequence for a C-terminal portion of the target protein. In some examples, molecule 150 is RNA, does not include promoter 152, and is at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, or at least 8000 nt, such as 200 to 10,000 nt, 200 to 8000 nt, 500 to 5000 nt, or 200 to 1000 nt in length. Molecule 150 (with or without promoter 152) can include natural and/or non-natural nucleotides or ribonucleotides.

The intronic sequence 170 includes a second dimerization domain 154, optional ISE 156, branching point 158, polypyrimidine tract 160, followed by a splice acceptor sequence 162. In some examples, intronic sequence 130 is about at least 10 nt, such as at least 20 nt, at least 30 nt, at least 50 nt, at least 100 nt, at least 250 nt, at least 250 nt, at least 300 nt, at least 400 nt, or at least 500 nt in length, such as 20 to 500, 20 to 250, 20 to 100, 50 to 100, 30 to 500, or 50 to 200 nt in length.

Second dimerization domain 154 has a sequence that is the reverse complement of first dimerization domain 122 sequence of molecule 110. Thus, same design features and considerations of first dimerization domain 122 discussed above also apply to second dimerization domain 154. For example, in some examples the second dimerization domain 154 contains a stem loop that can form a kissing loop interaction the first dimerization domain 122. In some examples, second dimerization domain 154 does not include cryptic splice acceptors (*e.g.*, NNNAGGUNNN; SEQ ID NO: 143) that could compete with RNA recombination. In some example, second dimerization domain 154 has a hypodiverse sequence. In some examples, second dimerization domain 154 is no more than 1000 nt, such as no more than 750 nt, or more than 500 nt, such as 30 to 1000 nt, 30 to 750 nt, 30 to 500 nt, 50 to 500 nt, 50 to 100 nt, or 100 to 250 nt. In some examples, second dimerization domain 154 is greater than 50 nt, such as at least 51 nt, at least 100 nt, at least 150 nt, at least 161 nt, or at least 170 nt, such as 51 to 159 nt, 51 to 150 nt, 51 to 120 nt, 51 to 100 nt, or 51 to 70 nt. In some examples, second dimerization domain 154 is greater than 160 nt, such as at least 161 nt, at least 170 nt, at least 180 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, at least 600 nt, at least 700 nt, at least 800 nt, at least 900 nt, or at least 1000 nt, such as 161 to 100 nt, 161 to 500 nt, 161 to 300 nt, 161 to 200 nt, or 161 to 170 nt. In some examples, second dimerization domain 154 is less than 50 nt, such 6 to 49 nt, 6 to 45 nt, 6 to 40 nt, 6 to 30 nt, 6 to 20 nt, or 6 to 10 nt.

3' - to second dimerization domain 154 is an optional ISE 156, branch point sequence 158 (such as a branch point consensus sequence), polypyrimidine tract 160, followed by a splice acceptor sequence 162. ISE 156, like ISE 120 and DISE 118 of molecule 110, stimulates the spliceosome to catalyze the recombination reaction. In some examples, intronic sequence 150 includes at least two ISE 156, such as at least 3, at least 4, or at least 5 ISEs 156. Exemplary splicing enhancer sequences include

ISE 156. In some examples, inclusion of one or more splicing enhancer sequences 156 in intronic sequence 150 increases recombination or splicing efficiency by at least 10%, at least 20%, at least 30%, at least 40%, or at least 50%. Exemplary splicing enhancer sequences that can be used are provided in SEQ ID NOS: 26-136, 151, and 152, as well as GGGTTT, GGTGGT, TTTGGG, GAGGGG, GGTATT, GTAACG, GGGGGTAGG, GGAGGGTTT, GGGTGGTGT TTCAT, CCATTT, TTTTAAA, TGCAT, TGCATG, TGTGTT, CTAAC, TCTCT, TCTGT, TCTTT, TGCATG, CTAAC, CTGCT, TAACC, AGCTT, TTCATTA, GTTAG, TTTTGC, ACTAAT, ATGTTT, CTCTG, GGG, GGG(N)<sub>2-4</sub>GGG, TGGG, YCAY, UGCAUG, or 3x(G<sub>3-6</sub>N<sub>1-7</sub>). In some examples, if ISE 156 is present, it can be about least 3 nt, at least 4 nt, at least 5 nt, at least 10 nt, such as at least 20 nt, at least 25 nt, at least 30 nt, at least 40 nt, or at least 50 nt in length, such as 3 to 10, 3 to 11, 4 to 11, 5 to 11, 10 to 50, 20 to 25, 10 to 25, 10 to 20, or 20 to 40 nt in length. In one example, the sequence of ISE 156 is or comprises GGCUGAGGGAAGGACUGUCCUGGG (SEQ ID NO: 135), GGGUUAUGGGACC (SEQ ID NO: 136), TTCAT, CCATTT, TTTTAAA, TGCAT, TGCATG, TGTGTT, CTAAC, TCTCT, TCTGT, or TCTTT. In some examples ISE 120 and ISE 156 are the same sequence. In other examples, ISE 120 and ISE 156 are the different sequences.

3'- to second dimerization domain 154 (and ISE 156 if present) is branch point sequence 158 (such as a branch point consensus sequence), a polypyrimidine tract 160, followed by a splice acceptor sequence 162 (such as a splice acceptor consensus sequence). The sequence of branch point 158 is based on the consensus sequence of the species of the target cell or organism. For example, for human splicing, the consensus sequence can include or be YUNAY. Thus, a sequence that it uses can be CUAAC for U2-dependent introns, or for U12-dependent introns UUUUCCUUAACU (SEQ ID NO: 144).

Polypyrimidine tract 160 includes C, U, or both C and U nucleotides, such as C<sub>n</sub>U<sub>y</sub>, wherein n+y is greater than or equal to 10 nucleotides, and can include nucleotides -3 to -22 relative to the 3'- splice junction. In some examples, polypyrimidine tract 160 includes at least 80% Y nucleotides (*i.e.*, U, C, or both U and C). In some examples, polypyrimidine tract 160 is a polyC or polyU sequence. In some examples, polypyrimidine tract 160 is a polyU sequence of at least 15 Us, such as 15 to 30 or 15 to 20 Us. Branch point 158 and polypyrimidine tract 160 are essential splicing components. The sequence of SA 162 can be based on the consensus sequence of the species of the target cell or organism. For example, in humans, the SA sequence can be AG in positions -1 and -2 relative to the 3'- splice site for U2-dependnet introns and AC or AG for U12-dependnet introns. Thus, in some examples, SA 162 can be 2 nt in length, such as AG or AC.

Immediately following SA 162 is an exonic sequence which includes a DNA sequence encoding a C-terminal portion of a target protein 164 having a splice junction at its 5' end. The splice junction at the 5' end of DNA sequence encoding a C-terminal portion of a target protein 164, that can

match the consensus sequence found in the target cell or organism into which molecules 110, 150 are introduced. In some example splice junction can be GA or GU at position +1 and +2 of the 3' splice site for U2-dependent introns or GU or AU for U12-dependent introns. Thus, in some examples, the splice junction is 2 nt in length, and the 5' end of the C-terminal coding portion 164 is GA, GU, or AU.

5 The exonic sequence following intronic portion 170 of molecule 150 includes a second coding portion (*e.g.*, half) of the target protein, *e.g.*, the C terminal fragment 164, and optional polyadenylation sequence 166. Thus, molecule 150 includes sequence 164 encoding a C-terminal portion of a target protein. The 3'-end of molecule 150 optionally includes a polyadenylation sequence 166, which promotes the assembly of the spliceosome. In some examples, polyadenylation sequence 166 is a  
 10 polyA sequence of at least 15 As, such as 15 to 30 or 15 to 20 As. In some examples polyadenylation sequence 166 and polyadenylation sequence 124 are the same sequence. In other examples, polyadenylation sequence 166 and polyadenylation sequence 124 are the different sequences.

In some examples, the N-terminal coding region 114 and/or the C terminal coding region 164 is a native coding sequence. For example, the coding sequence is one that is found in the cell or organism  
 15 into which the disclosed system is introduced. (*e.g.*, a human coding sequence when introduced into a human cell or subject). In some examples, the N-terminal coding region 114 and/or the C terminal coding region 164 is codon optimized relative to a native coding sequence, for example to maximize tRNA availability, or to de-enrich for cryptic splice sites (*e.g.*, to reduce or avoid incorrect splicing and promote the correct junction formation). In some examples, a portion of the N-terminal coding region  
 20 114 and/or the C terminal coding region 164 is codon optimized relative to a native coding sequence, for example the about 200 nt adjacent to each junction (*e.g.*, the 3'-end of 114, and the 5' end of 164) can be codon optimized or altered to contain exonic splice enhancer sites (ESE) (which would bind SR proteins). For example, the coding sequence can be one not found in the cell or organism into which the disclosed system is introduced. (*e.g.*, a human coding sequence when introduced into a mouse cell  
 25 or subject).

In some examples, the N-terminal coding region 114 and/or the C terminal coding region 164 include an intron that is either natural or synthetic in nature and contains both a splice donor and acceptor site. For example, an intron embedded inside the to the coding sequence to be expressed can be included upstream (*e.g.*, about 200 nt upstream) of sequence 116, inside the N-terminal coding  
 30 region 114, an intron embedded inside the coding sequence to be expressed can be included downstream (*e.g.*, about 200 nt downstream) of the sequence 162 and inside the C-terminal coding region 164, or both. Inclusion of such introns can be used to stimulate splicing machinery attachment to the trans-splicing intron donor and acceptor. In some examples, such (stimulatory-) introns could be derived from the host in which 110 and 150 are expressed. In some examples, such (stimulatory-)  
 35 introns could be derived from other organisms, or viral in origin, or synthetic in origin.

In some examples, inclusion of a sequence to stabilize the molecule 150 (*e.g.*, placed between 164 and 166 in the 3' untranslated region of 150 in FIG. 6A) can increase expression efficiency of the recombined product by at least 25%, at least 30%, at least 40%, at least 50%, at least 60%, or at least 75%, such as 25 to 95%, 25 to 75%, 25 to 60%, 25 to 50%, 40 to 95%, 40 to 60%, or 50 to 60%. In some examples, woodchuck post-transcriptional regulatory element (WPRE) or truncations thereof (*e.g.* WPRE3) are included in the 3'-UTR as a stabilizing element to enhance recombined product expression efficiency. In some example a WPRE sequence has at least 80%, at least 85%, at least 90%, at least 95%, or 100% sequence identity to nt 1093 to 1684 of GenBank accession no. J04514 or to the 247 bp sequence of WPRE3.

As shown in FIG. 6C, interaction and hybridization (base pairing) between first dimerization domain 122 of molecule 110 and second dimerization domain 154 of molecule 150, allows the spliceosome components to recombine N-terminal coding sequence 114 and C-terminal coding sequence 164. Specifically the 3' end of the N terminal protein coding sequence 114 is fused to the 5' end of the C terminal protein sequence 164 as a seamless junction between the two portions.

FIG. 6D shows a schematic of a system wherein a target protein is divided into three portions, an N-terminal, middle, and C-terminal portion (wherein each portion can be similar or different in size). One skilled in the art will appreciate that a protein can thus be divided into any number of desired segments or portions, and an appropriate number of molecules designed using the information provided herein. In such an example, the system includes at least three synthetic nucleic acid molecules 110, 200, and 150, wherein molecule 110 includes molecule 114 which encodes the N-terminal portion of the protein, molecule 200 includes molecule 216 which encodes the middle portion of the protein, and molecule 150 includes molecule 164 which encodes the C-terminal portion of the protein. Each nucleic acid molecule 110, 200, 150 can be composed of DNA, and following translation, can be RNA with promoters 112, 202, 152 absent. In some examples, each of 110, 200, 150 (with or without promoters 112, 202, 152) is at least about 100 nucleotides/ribonucleotides (nt) in length, such as at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, or at least 8000 nt, such as 200 to 10,000 nt, 200 to 8000 nt, 500 to 5000 nt, or 200 to 1000 nt. The molecules 110, 150, 200 (with or without promoters 112, 202, 152) can include natural and/or non-natural nucleotides or ribonucleotides. In addition to using two (or more) orthogonal dimerization domains, one of the two introns can be a U2-type intron and the second intron can be a U12-type intron. Splice donor and acceptors of U2 and U12 dependent introns show minimal cross reactivity since the consensus recognition sequences between the two types of introns are different. Both strategies (*i.e.*, the orthogonal dimerization domains, and the U2 vs U12 type introns) promote recombination of the three fragments in the correct order (*e.g.*, to avoid the first fragment to directly join up to the last fragment and to avoid the middle fragment circularizing onto itself).

Molecule 110 of FIG. 6D includes the same features disclosed above for FIG. 1A, namely a promoter 112 operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a coding sequence for an N-terminal portion of the target protein 114, wherein the coding sequence for an N-terminal portion of the target protein 114 comprises a splice junction at a 3'-end of the target protein coding sequence, , SD 116, optional DISE 118, optional ISE 120, dimerization domain 122, and optional polyadenylation sequence 124, but wherein first dimerization domain 122 has reverse complementary to third dimerization domain 204 of molecule 200. As shown in FIG. 6F, in embodiments where molecule 110 is RNA, for example after expression of the DNA into RNA, molecule 110 does not include promoter 112, and 114 is the RNA encoded by the coding sequence for an N-terminal portion of the target protein. Molecule 110 (with or without promoter 112) can include natural and/or non-natural nucleotides or ribonucleotides.

Molecule 150 of FIG. 6D includes the same features disclosed above for FIG. 1A, namely promoter 152 operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a second dimerization domain 154, optional ISE 156, a branch point sequence 158, a polypyrimidine tract 160, a splice acceptor (SA) 162; and a coding sequence for a C-terminal portion of the target protein 164, wherein the coding sequence for the C-terminal portion of the target protein comprises a splice junction at a 5'-end of the target protein coding sequence, and optionally polyadenylation sequence 166. The second dimerization domain 154 has reverse complementary to fourth dimerization domain 226 of molecule 200. Molecule 150 (with or without promoter 152) can include natural and/or non-natural nucleotides or ribonucleotides.

Molecule 200 allows for the joining of the N- and C-terminal coding regions 114, 164, by providing dimerization domains having reverse complementarity to dimerization domains 122, 154 of molecule 110 and molecule 150, respectively. Molecule 200 includes features from both molecule 110 and molecule 150, including two intronic sequences 230, 240. Specifically, in embodiments where molecule 200 is DNA, molecule 200 includes promoter 210 (which can be the same or different than promoter 112 and/or 152) operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': third dimerization domain 204 (which is the reverse complement to first dimerization domain 122 of molecule 110 in FIG. 6D), optional ISE 206, branch point 208, polypyrimidine tract 210, SA 212, a coding sequence for a middle portion of the target protein 216, wherein the coding sequence for the middle portion of the target protein 216 comprises a splice junction at a 5'-end of the target protein coding sequence and a splice junction at a 3'-end of the target protein coding sequence, SD 220, optional DISE 222, optional ISE 224, fourth dimerization domain 226 (which is the reverse complement to fourth dimerization domain 154 of molecule 150 in FIG. 6D), and optional polyadenylation sequence 228. In some examples, molecule 200 is DNA, and is at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at

least 7000, or at least 8000 nt, such as 200 to 10,000 nt, 200 to 8000 nt, 500 to 5000 nt, or 200 to 1000 nt in length. In embodiments where molecule 200 is RNA, for example after expression of the DNA into RNA, molecule 200 no longer includes promoter 202, and 216 is the RNA encoded by the coding sequence for a middle portion of the target protein. In some examples, molecule 200 is RNA, does not include promoter 202, and is at least 200, at least 300, at least 500, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, or at least 8000 nt, such as 200 to 10,000 nt, 200 to 8000 nt, 500 to 5000 nt, or 200 to 1000 nt in length. The molecule 200 (with or without promoter 202) can include natural and/or non-natural nucleotides or ribonucleotides.

As shown in FIG. 6E, interaction and hybridization (base pairing) between first dimerization domain 122 of molecule 110 and third dimerization domain 204 of molecule 200, and interaction and hybridization (base pairing) between fourth dimerization domain 226 of molecule 200 and second dimerization domain 154 of molecule 150, allows the spliceosome components to recombine N-terminal coding sequence 114, middle coding sequence 216, and C-terminal coding sequence 164. Specifically the 3' end of the N terminal protein coding sequence 114 is fused to the 5' end of the middle protein sequence 216, and the 3' end of middle protein sequence 216, is fused to the 5' end of the C-terminal protein sequence 164 as a seamless junction between the three portions.

Alternative dimerization domains are shown in FIGS. 7A-7B and 9A. That is, as an alternative to using dimerization domains that hybridize to one another (*e.g.*, 112 to 204, 226 to 154, FIGS. 6D, 6E), in one example aptamer sequences are used. As shown in FIG. 7A, in both synthetic nucleic acid molecules 500, 600, aptamer sequences 512, 602 are used instead of the dimerization domains, and the aptamers come together via their interaction with a target (such as adenosine, dopamine, or caffeine). In such an example, the aptamer sequence 512, 602 of each molecule 500, 600 can be the same, or even be different sequences. Molecule 500 of FIG. 7A includes the same features disclosed above for molecule 110 of FIG. 6A, which when DNA includes a promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a coding sequence for an N-terminal portion of the target protein 502, wherein the coding sequence for an N-terminal portion of the target protein 502 comprises a splice junction at a 3'-end of the target protein coding sequence, SD 506, optional DISE 508, optional ISE 510, a first aptamer 512 instead of a first dimerization domain, and optional polyadenylation sequence. In embodiments when molecule 500 is RNA, for example when transcribed from the DNA molecule, molecule 500 does not include a promoter (*e.g.*, as shown in FIG. 7A). Similarly, molecule 600 of FIG. 7A includes the same features disclosed above for molecule 150 of FIG. 6A, which when DNA includes a promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': r aptamer 602 instead of second dimerization domain 154, optional ISE 604, branch point 606, polypyrimidine tract 608, SA 610, DNA encoding a C-terminal portion of a target protein 614 with a splice junction at its 5'-end, and optional

polyadenylation sequence 616. In embodiments when molecule 600 is RNA, for example when transcribed from the DNA molecule, molecule 500 does not include a promoter (e.g., as shown in FIG. 7A). Interaction of the two aptamers 512, 602, with each other or molecule 700 allows the spliceosome components to recombine N-terminal coding sequence 502 and C-terminal coding sequence 614.

5 Specifically the 3' end of the N terminal protein coding sequence 502 is fused to the 5' end of the C terminal protein sequence 614 as a seamless junction between the two portions. Molecules 500 and 600 can include natural and/or non-natural nucleotides or ribonucleotides.

In some examples, aptamer sequences 512, 602 recognize (e.g., specifically bind) the same target 700 (FIG. 7A), or can even recognize different targets (wherein a synthetic molecule is also administered with the system provided herein, which includes each molecule specifically recognized by each aptamer, or the part of the molecule recognized by the aptamer, such as a caffeine/dopamine hybrid molecule). Exemplary targets recognized by aptamers include cellular proteins, small molecules, exogenous proteins, or an RNA molecule.

FIG. 7B shows an example similar to FIG. 7A. The dimerization domains (512, 602 FIG. 7A) recognize an RNA molecule. In the example shown in FIG. 7B, each domain recognizes a different portion of an mRNA molecule only expressed in target cells (cells where target protein expression is desired), such as a cancer-specific transcript. In such an example, the coding sequences comprised by the RNAs (502, 614 of FIG. 7A) only recombine in the presence of the specific RNA molecule recognized by the dimerization domains. Here, the target protein would only be expressed in cancer cells, not normal cells. Such a system allows for control of the target protein expression (e.g., a therapeutic protein for cancer, such as a toxin or a cytotoxic enzyme such as thymidine kinase with ganciclovir; thus in some examples the target protein is a toxin or thymidine kinase) in cancer cells, reducing undesirable side effects of expression the target protein in normal, non-cancer cells.

FIG. 7C provides an exemplary “off-switch” example. Here, the hybridization/binding of dimerization domains 812, 902 (which are reverse complements of one another) of synthetic nucleic acid molecules 800, 900 can be reduced by providing an anti-binding domain oligonucleotide (e.g., RNA or DNA) 1000 (which can be two different anti-binding domain oligonucleotides 1000, one that is the reverse complement of 812, and one that is the reverse complement of 912) that competes for the binding/hybridization. Anti-binding domain oligonucleotide 1000 can thus act as an “off-switch” for reconstitution of the protein encoded by N- and C-terminal coding portions 802 and 914, respectively. Molecule 800 of FIG. 7C includes the same features disclosed above for a molecule 110 of FIG. 6A that is an RNA molecule (and thus lacks a promoter), which RNA molecule comprises from 5' to 3': a coding sequence for an N-terminal portion of the target protein 802, wherein the coding sequence for an N-terminal portion of the target protein 802 comprises a splice junction at a 3'-end of the target protein coding sequence, SD 806, optional DISE 808, optional ISE 810, dimerization domain 812, and optional

polyadenylation sequence 814. Similarly, molecule 900 of FIG. 7C includes the same features disclosed above for a molecule 150 of FIG. 6A that is an RNA molecule (and thus lacks a promoter), which RNA molecule comprises from 5' to 3': anti-dimerization domain 902, optional ISE 904, branch point 906, polypyrimidine tract 908, SA 910, RNA encoding a C-terminal portion of a target protein 914, and optional polyadenylation sequence 916. The two dimerization domains 812, 902 cannot interact/hybridize to each other in the presence of the anti-binding domain oligonucleotides 1000, and therefore prevents or reduces recombination of the N-terminal coding sequence 802 and C-terminal coding sequence 914. Such an application can be used to reduce or eliminate expression of the protein encoded by the system. Molecules 800 and 900 can include natural and/or non-natural nucleotides or ribonucleotides.

FIG. 9A provides an exemplary dimerization domain that uses kissing loop interactions instead of reverse complementary sequence hybridization for dimerization. Kissing loop interactions are formed when the bases in the loops of two RNA hairpins form interacting pairs between two RNA molecules. The molecule on the left hand side labelled with n-yfp represents an RNA molecule that encodes the n-terminal fragment of yfp, linked to a synthetic intron that contains a splice donor site, a downstream intronic splicing enhancer element, and two intronic splicing enhancer elements. The dimerization domain this molecule contains three RNA hairpin loops that each are composed of a stem (where the RNA hybridizes onto itself) and a loop (in which the RNA is not hybridized to itself). In this example, the dimerization domain contains three stem and loop elements (also referred to as hairpin loops) and is referred to as a trimodal kissing loop dimerization domain. The molecule on the right hand side labelled with c-yfp represents an RNA molecule that encodes the c terminal portion of yfp. From 5' to 3' this molecule is composed of a trimodal kissing loop dimerization domain that contains a set of three hairpin loops. The loop portions can form kissing loop interactions with the corresponding loops on the complementary n-yfp molecule. The trimodal kissing loop dimerization domain is followed by a synthetic intron sequence that contains three intronic splicing enhancer sequences, a branch point sequence, a polypyrimidine tract, and a splice acceptor site. The synthetic intron sequence is followed by the c-terminal yfp coding sequence, which is followed by a 3' untranslated region that contains a poly adenylation signal. At the top of the figure, a representative 3-dimensional rendering of a kissing loop interaction is shown. This rendering illustrates how the kinked form of the hairpin loop exposes the loop residues towards the outside which renders them available for the kissing loop interaction.

Upon association of the two molecules, the spliceosome mediates a trans-splicing reaction which results in the joining of the n-terminal and the c-terminal yfp coding sequence which then allows for expression of the full-length fluorescent protein.



Although FIGS. 6A-7C and 9A show embodiments where a system uses two synthetic nucleic acid molecules are used (*i.e.*, the target protein coding sequence is split between two synthetic nucleic acid molecules), one skilled in the art will appreciate that such embodiments can be used similarly with more than two synthetic nucleic acid molecules, such as three, four, five, six, seven, eight, nine, or 10 synthetic nucleic acid molecules using the teachings herein.

In some examples, the system includes a nucleic acid molecule that suppresses expression of un-assembled/un-recombined fragments. In such an example, if the two or more portions of a full-length coding sequence (*e.g.*, 114 of 110, 164 of 150 of FIG. 6A, respectively), did not recombine, the nucleic acid molecule would suppress expression of each portion of a full-length coding sequence that was not recombined into a full-length protein. For example, such a suppressive nucleic acid molecule can destabilize the RNA once outside the nucleus, prevent translation, stimulate translation from a shifted start codon, contain microRNA target sites, or contain protein degron or destabilization domains that when translated suppress the protein activity or flag it for degradation.

In one example, destabilization of the un-recombined RNA molecule is achieved by including a self-cleaving RNA sequence (*e.g.*, Hammerhead ribozyme or HDV ribozyme) into the synthetic intron, for example at any position within intronic sequence 130 of FIG. 6A or 6F. In one example, cleaving the RNA molecule leads to a loss of the RNA stabilizing poly A tail, which can suppress expression of an un-recombined protein from open reading frame 114 of FIG. 6A or 6F. In one example, a self-cleaving RNA sequence is included at any position within s intronic sequence 170 of FIG. 6A or 6F to cleave off the 5' terminal CAP which in one example can lead to reduced expression of an open reading frame that includes parts or the whole of coding sequence 164 of FIG. 6A or 6F. In one example self-cleaving RNA sequences are substituted with an RNA cleaving enzyme target site, such as a Csy4 target site.

In some examples, a suppressive nucleic acid molecule includes a start codon (ATG) or a Kozak enhanced start codon (GCCGCCACCATG (SEQ ID NO: 154) or GCCACCATG or ACCATG) at any position within intronic sequence 170 of FIG. 6A or 6F that directs translation of an open reading frame that is shifted -1, -2, +1, or +2 nucleotides relative to the open reading frame sequence 164 of FIG. 6A or 6F. In one example, un-assembled fragment expression is reduced or suppressed by using this decoy start codon strategy to direct translation away from the to be suppressed open reading frame of sequence 164 of FIG. 6A or 6F.

In some examples, a suppressive nucleic acid molecule includes one or more micro RNA target sites at any position within intronic sequence 130 of FIG. 6A or 6F, and/or at any position within intronic sequence 170 of FIG. 6A or 6F. If a particular molecule (*e.g.*, 110 or 150 in FIG. 6A or 6F) is exported from the nucleus, it becomes subject to micro RNA / small hairpin RNA dependent

degradation which can suppress unintended un-joined fragment expression by degrading/suppressing un-joined RNA that was exported from the nucleus. In one example, such a micro RNA target sequence can be complementary to a micro RNA known to be expressed in the cell, or tissue, or animal into which the molecules 110 and 150 of FIG. 6A or 6F are introduced. In one example, this micro RNA target sequence is complementary to a sequence that is introduced into the cell, or tissue, or animal. In one example, such a microRNA can be expressed from an RNA-polymerase III dependent promoter in the form of a small hairpin RNA. In one example, such a microRNA can be expressed from an RNA polymerase II dependent promoter and embedded in a micro RNA processing loop (*e.g.*, mir30 scaffold).

In some examples, destabilization of the un-recombined protein product from an open reading frame (*e.g.*, 114 in FIG. 6) can be achieved by depleting stop codon occurrence in intronic sequence 130 of FIG. 6A or 6F and an additional inclusion of an RNA sequence coding for an in frame protein signal that can flag a protein for degradation (*e.g.*, a degron sequence) that is placed at any position within intronic sequence 130 of FIG. 6A or 6F and which is in frame with the open reading frame that is extended out from sequence 114 of FIG. 6A or 6F. In one example a degron sequence can be that of a PEST sequence, or that of the CL1 degron sequence. Degron sequences used can employ proteasome-dependent, proteasome-independent, ubiquitin-dependent, or ubiquitin-independent pathways. In one example, un-recombined protein destabilization is enhanced by inclusion of several of the same or different degron sequences.

In some examples, destabilization of the un-recombined protein product from open reading frame sequence 164 in FIG. 6A is achieved by introduction of a start codon (ATG) followed by a degron sequence at any position within intronic sequence 170 in FIG. 6A which is in frame with an open reading frame within sequence 164 in FIG. 6. In this example, the degron sequence will be N-terminally joined to the un-recombined protein fragment that will be suppressed by being flagged for degradation.

#### IV. Compositions and Kits

Compositions and kits are provided that include two or more of the synthetic nucleic acid molecules provided herein, wherein the synthetic nucleic acid molecule encode a full-length protein when recombined. In some examples, the two or more of the synthetic nucleic acid molecules provided herein are DNA. In some examples, the two or more of the synthetic nucleic acid molecules provided herein are RNA, and do not include promoter sequences. In one example, the composition or kit includes two of the synthetic nucleic acid molecules provided herein, wherein each of the two synthetic nucleic acid molecules encodes a different portion of a target protein (*i.e.*, N- terminal and C-terminal, wherein the whole coding sequence is generated when recombination between the two molecules

occurs), such as one listed in Table 1 (or a therapeutic protein, such as a toxin or thymidine kinase). In one example, the composition or kit includes three of the synthetic nucleic acid molecules provided herein, wherein each of the three synthetic nucleic acid molecules encodes a different portion of a target protein (*i.e.*, N- terminal, middle, and C-terminal, wherein the whole coding sequence is generated when recombination between the three molecules occurs), such as one listed in Table 1 (or a therapeutic protein, such as a toxin or thymidine kinase). In one example, the composition or kit includes four or more of the synthetic nucleic acid molecules provided herein, wherein each of the four or more synthetic nucleic acid molecules encodes a different portion of a target protein (*i.e.*, N- terminal, first middle, second middle (and optionally additional middle), and C-terminal, wherein the whole coding sequence is generated when recombination between the four or more synthetic nucleic acid molecules occurs), such as one listed in Table 1 (or a therapeutic protein, such as a toxin or thymidine kinase). In one example, the composition or kit includes two or more sets of two or more of the synthetic nucleic acid molecules provided herein, wherein each set of synthetic nucleic acid molecules encodes a different target protein, such as two or more listed in Table 1 (and/or a therapeutic protein, such as a toxin or thymidine kinase).

In one example, each synthetic nucleic acid molecule in the composition or kit is part of a vector, such as AAV or other gene therapy vector. In one example, the composition or kit includes a cell, such as a bacterial cell or eukaryotic cell, that includes two or more disclosed synthetic nucleic acid molecules, wherein the synthetic nucleic acid molecules encode a full-length target protein when recombined.

Such compositions can include a pharmaceutically acceptable carrier (*e.g.*, saline, water, glycerol, DMSO, or PBS). In some examples, the composition is a liquid, lyophilized powder, or cryopreserved.

In some examples, the kit includes a delivery system (*e.g.*, liposome, a particle, an exosome, or a microvesicle) to direct cell type specific uptake/enhance endosomal escape/enable blood-brain barrier crossing etc. In some examples, the kits further include cell culture or growth media, such as media appropriate for growing bacterial, plant, insect, or mammalian cells. In some examples, such parts of a kit are in separate containers. Exemplary containers include plastic or glass vials or tubes.

In some examples, each of two or more the synthetic nucleic acid molecules provided herein are in separate containers. In some examples, each of two or more sets of two or more of the synthetic nucleic acid molecules provided herein are in separate containers.

## V. Methods of Treatment

The disclosed methods and systems can be used to express any protein of interest, for example when a protein is too large to be expressed by a therapeutic virus (*e.g.*, AAV) or when a complete gene

sequence (*e.g.*, endogenous promoter + coding sequence) is too large to be expressed by a therapeutic virus (*e.g.*, AAV). In such cases, the coding sequence of the target protein may be divided into two or more portions using the disclosed systems, and recombined in the correct order, allowing for the protein to be expressed when and where desired.

5 The subject to be treated can be any mammal, such as one with a monogenetic disorder, such as one listed in Table 1. In one example, the subject has cancer. Thus, humans, cats, pigs, rats, mice, cows, goats, and dogs, can be treated with the disclosed methods. In some examples, the subject is a human infant less than 6 months of age. In some examples, the subject is a human infant less than 1 year of age. In some examples, the subject is a human juvenile. In some examples, the subject is a  
10 human adult at least 18 years of age. In some examples, the subject is female. In some examples, the subject is male.

The two or more synthetic nucleic acid molecules provided herein used to treat a subject can be matched to the subject treated. Thus, for example, if the subject to be treated is a dog, a dog coding sequence for the target protein can be used and the intronic sequence can be optimized for expression in  
15 dog cells, and if the subject to be treated is a human, a human coding sequence for the target protein can be used and the intronic sequence can be optimized for expression in human cells.

The two or more synthetic nucleic acid molecules provided herein can be administered as part of a vector, such as an adeno-associated vector (AAV), for example AAV serotype rh.10. In some examples, vectors (*e.g.*, AAV) including one of the two or more synthetic nucleic acid molecules  
20 provided herein are administered systemically, such as intravenously. Thus, if a coding sequence is divided between two synthetic nucleic acid molecules provided herein, two AAV's are administered, each AAV including one of the two synthetic nucleic acid molecules provided herein.

A therapeutically effective amount of two or more synthetic nucleic acid molecules provided herein is administered, for example in AAVs. In some examples, the two or more synthetic nucleic acid  
25 molecules provided herein when part of a viral vector (*e.g.*, AAV) is administered at a dose of at least  $1 \times 10^{11}$  genome copies (gc), at least  $1 \times 10^{12}$  gc, at least  $2 \times 10^{12}$  gc, at least  $1 \times 10^{13}$  gc, at least  $2 \times 10^{13}$  gc per subject, or at least  $1 \times 10^{14}$  gc per subject, such as  $2 \times 10^{11}$  gc per subject,  $2 \times 10^{12}$  gc per subject,  $2 \times 10^{13}$  gc per subject, or  $2 \times 10^{14}$  gc per subject. In some examples, the two or more synthetic nucleic acid molecules provided herein when part of a viral vector (*e.g.*, AAV) is administered at a dose of at  
30 least  $1 \times 10^{11}$  gc/kg, at least  $5 \times 10^{11}$  gc/kg, at least  $1 \times 10^{12}$  gc/kg, at least  $5 \times 10^{12}$  gc/kg, at least  $1 \times 10^{13}$  gc/kg, or at least  $4 \times 10^{13}$  gc/kg, such as  $4 \times 10^{11}$  gc/kg,  $4 \times 10^{12}$  gc/kg, or  $4 \times 10^{13}$  gc/kg.

If adverse symptoms develop, such as AAV-capsid specific T cells in the blood, corticosteroids can be administered (*e.g.*, see Nathwani *et al.*, *N Engl J Med.* 365(25):2357-65, 2011).

Diseases that can be treated with the disclosed methods include any genetic disease of the blood  
35 (*e.g.* sickle cell disease, primary immunodeficiency diseases), HIV (such as HIV-1), and hematologic

malignancies or cancers. Examples of primary immunodeficiency diseases and their corresponding mutations include those listed in Al-Herz *et al.* (*Frontiers in Immunology*, volume 5, article 162, April 22, 2014, herein incorporated by reference in its entirety). Hematologic malignancies or cancers are those tumors that affect blood, bone marrow, and lymph nodes. Examples include leukemia (*e.g.*, acute lymphoblastic leukemia, acute myelogenous leukemia, chronic lymphocytic leukemia, chronic myelogenous leukemia, acute monocytic leukemia), lymphoma (*e.g.*, Hodgkin's lymphoma and non-Hodgkin's lymphoma), and myeloma. In some examples, the disease is a monogenetic disease. Table 1 provides a list of exemplary disorders and genes that can be targeted by the disclosed systems and methods. Additional examples are provided here: [rarediseases.info.nih.gov/diseases/diseases-by-category/5/congenital-and-genetic-diseases](https://rarediseases.info.nih.gov/diseases/diseases-by-category/5/congenital-and-genetic-diseases) (list herein incorporated by reference). Any genetic disease caused by a lack of protein (*e.g.*, recessive mutation) or an insufficiency of protein can benefit from the disclosed systems and methods. In cases where the coding region of the gene is relatively small, the disclosed systems and methods are useful to add regulatory sequences, such as tissue specific promoters or specific non-coding RNA segments, to direct gene expression to the appropriate cell types at the appropriate levels.

**Table 1: Exemplary disorders and corresponding mutations**

| Disease                    | Gene                                   | Mutation                                                                        |
|----------------------------|----------------------------------------|---------------------------------------------------------------------------------|
| <b>Blood cell disorder</b> |                                        |                                                                                 |
| sickle cell anemia         | $\beta$ -globin chain of hemoglobin    | SNP (A to T) that gives rise to point mutation (Glu->Val at 6 <sup>th</sup> aa) |
| hemophilia                 | any of clotting factors I through XIII |                                                                                 |
| hemophilia A               | clotting factor VIII                   | large deletions, insertions, inversions, and point mutations                    |
| hemophilia B               | clotting factor IX                     |                                                                                 |
| Alpha-Thalassemia          | HBA1 or HBA2                           | Mutation or a deletion in chromosome 16 p                                       |
| Beta-Thalassemia           | HBB                                    | Mutations in chromosome 11                                                      |
| Delta-Thalassemia          | HBD                                    | mutation                                                                        |
| von Willebrand Disease     | von Willebrand factor                  | mutations or deletion                                                           |
| pernicious anemia          | MTHFR                                  |                                                                                 |
| Fanconi anemia             | FANCA, FANCC, FANCD2, FANCG, FANCI     | FANCA: c.3788_3790del (p.Phe1263del); c.1115_1118delTTGG                        |

|                                          |                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                           |
|------------------------------------------|------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                          |                                                                                    | <p>(p.Val372fs); Exon 12–17del; Exon 12–31del; c.295C&gt;T (p.Gln99X)</p> <p>FANCC: c.711+4A&gt;T (originally reported as IVS4+4A&gt;T); c.67delG (originally reported as 322delG)</p> <p>FANCD2: c.1948-16T&gt;G</p> <p>FANCG; c.313G&gt;T (p.Glu105X); c.1077-2A&gt;G; c.1480+1G&gt;C; c.307+1G&gt;C; c.1794_1803del (p.Trp599fs); c.637_643del (p.Tyr213fs)</p> <p>FANCI: c.2392C&gt;T (p.Arg798X)</p> |
| Thrombocytopenic purpura                 | ADAMTS13                                                                           | Missense and nonsense mutations                                                                                                                                                                                                                                                                                                                                                                           |
| thrombophilia                            | Factor V Leiden<br>Prothrombin                                                     | <p>Mutation in the F5 gene at position 1691</p> <p>Prothrombin G20210A</p>                                                                                                                                                                                                                                                                                                                                |
| <b>Primary Immunodeficiency Diseases</b> |                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                           |
| T-B+ SCID                                | IL-2RG, JAK3, defect in gamma chain of receptors for IL-2, -4, -7, -9, -15 and -21 |                                                                                                                                                                                                                                                                                                                                                                                                           |
| T-B- SCID                                | RAG1, RAG2                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                           |

|                                                        |                                                                                                      |                                                                                                                              |
|--------------------------------------------------------|------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| WHIM syndrome                                          | CXCR4                                                                                                | heterozygous mutations ( <i>e.g.</i> , in the carboxy-terminus); carboxy-terminus truncation ( <i>e.g.</i> , 10-19 residues) |
| <b>Other Primary immune deficiency (PID) syndromes</b> |                                                                                                      |                                                                                                                              |
| IL-7 receptor severe combined immune deficiency (SCID) | IL7 receptor                                                                                         |                                                                                                                              |
| Adenosine deaminase deficiency (ADA) SCID              | ADA                                                                                                  |                                                                                                                              |
| Purine nucleoside phosphorylase (PNP) deficiency       | PNP                                                                                                  |                                                                                                                              |
| Wiskott-Aldrich syndrome (WAS)                         | WAS                                                                                                  | More than 300 mutations identified                                                                                           |
| Chronic granulomatous disease (CGD)                    | CYBA, CYBB, NCF1, NCF2, or NCF4                                                                      |                                                                                                                              |
| Leukocyte adhesion deficiency (LAD)                    | Beta-2 integrin                                                                                      |                                                                                                                              |
| HIV                                                    | C-C chemokine receptor type 5 (CCR5),<br>MSRB1<br>HIV long terminal repeats<br>CSCR4<br>P17<br>PSIP1 | Deletion of 32 bp in CCR5                                                                                                    |
| Duchenne muscular dystrophy                            | CCR5<br>DMD                                                                                          |                                                                                                                              |
| Glycogen storage disease type IA                       | G6Pase                                                                                               |                                                                                                                              |
| Retinal Dystrophy                                      | CEP290<br>ABCA4                                                                                      | C2991+1655A>G<br>5196+1216C>A;<br>5196+1056A>G;<br>5196+1159G>A;<br>5196+1137G>A;<br>938-619A>G; 4539+2064C>T                |

|                                                                                                                                                                                                                                               |                                                                                                  |                          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|--------------------------|
| X-linked immunodeficiency with magnesium defect, Epstein-Barr virus infection, and neoplasia (XMEN)                                                                                                                                           | MAGT1                                                                                            |                          |
| <b>MonoGenetic Disorders</b>                                                                                                                                                                                                                  |                                                                                                  |                          |
| Metachromatic leukodystrophy (MLD)                                                                                                                                                                                                            | arylsulfatase A (ARSA)                                                                           |                          |
| Adrenoleukodystrophy (ALD)                                                                                                                                                                                                                    | ABCD1                                                                                            |                          |
| Mucopolysaccharidoses (MPS) disorders<br>Hunter syndrome<br>Hurler syndrome<br>Scheie syndrome<br>Sanfilippo syndrome A, B, C, and D<br>Morquio syndrome A Morquio syndrome B<br>Maroteaux-Lamy syndrome<br>Sly syndrome<br>Natowicz syndrome | IDS<br>IDUA<br>IDUA<br>SGSH, NAGLU,<br>HGSNAT, GNS<br>GALNS<br>GLB1<br>ARSB<br><br>GUSB<br>HYAL1 |                          |
| Alpha mannosidosis                                                                                                                                                                                                                            | MAN2B1                                                                                           |                          |
| Nieman Pick disease types A, B, and C                                                                                                                                                                                                         | SMPD1, NPC1, NPC2                                                                                |                          |
| Cystic fibrosis                                                                                                                                                                                                                               | cystic fibrosis transmembrane conductance regulator (CFTR)                                       | $\Delta F508$            |
| Polycystic kidney disease                                                                                                                                                                                                                     | PKD-1, PDK-2, PDK-3                                                                              |                          |
| Tay Sachs Disease                                                                                                                                                                                                                             | HEXA                                                                                             | 1278insTATC              |
| Gaucher disease                                                                                                                                                                                                                               | GBA                                                                                              |                          |
| Huntington's disease                                                                                                                                                                                                                          | HTT                                                                                              | CAG repeat               |
| Neurofibromatosis types 1 and 2                                                                                                                                                                                                               | NF-1 and NF2                                                                                     | CGA->UGA->Arg1306Term in |



|                                        |                                            |               |
|----------------------------------------|--------------------------------------------|---------------|
|                                        |                                            | NF1           |
| Familial hypercholesterolemia          | APOB, LDLR,<br>LDLRAP1, and PCSK9          |               |
| <b>Cancers</b>                         |                                            |               |
| Chronic myeloid leukemia<br>(CML)      | BCR-ABL<br>ASXL1                           | fusion        |
| Acute myeloid leukemia (AML)           | Chromosome 11q23 or<br>t(9;11)             | translocation |
| Osteosarcoma                           | RUNX2                                      |               |
| Colorectal cancer                      | EPHA1                                      |               |
| Gastric cancer, melanoma               | PD-1                                       |               |
| Prostate cancer                        | Androgen receptor                          |               |
| Cervical cancer                        | E6, E7                                     |               |
| Glioblastoma                           | CD                                         |               |
| <b>Neurological disorders</b>          |                                            |               |
| Alzheimer's disease                    | NGF                                        |               |
| Metachromatic leukodystrophy           | ARSA                                       |               |
| Multiple sclerosis                     | MBP                                        |               |
| Wiskott-Aldrich syndrome               | WASP                                       |               |
| X-linked adrenoleukodystrophy          | ABCD1                                      |               |
| AACD deficiency                        | AADC                                       |               |
| Batten disease                         | CLN2                                       |               |
| Canavan disease                        | ASPA                                       |               |
| Giant axonal neuropathy                | GAN                                        |               |
| Leber's hereditary optic<br>neuropathy | MT-ND4                                     |               |
| MPS IIIA                               | SGSH, SUMF1                                |               |
| Parkinson's disease                    | GAD, NTRN, TH,<br>AADC, CH1, GDNF,<br>AADC |               |
| Pompe disease                          | GAA                                        |               |
| Spinal muscular atrophy type 1         | SMN                                        |               |

Using the disclosed methods and systems can be used to treat any of the disorders listed in Table 1, or other known genetic disorder. The disclosed methods can also be used to treat other

disorders, such as a cancer that can benefit from expression of a therapeutic protein in a cancer cell, such as a toxin or thymidine kinase. If the subject is administered two or more synthetic molecules provided herein that express a full-length thymidine kinase, the subject is also administered ganciclovir. Treatment does not require 100% removal of all characteristics of the disorder, but can be a reduction in such. Although specific examples are provided below, based on this teaching one will understand that symptoms of other disorders can be similarly affected. For example, the disclosed methods can be used to increase expression of a protein that is not expressed or has reduced expression by the subject, or decrease expression of a protein that is undesirably expressed or has reduced expression by the subject. For example, the disclosed methods can be used to treat or reduce the undesirable effects of a genetic disease.

For example, the disclosed methods and systems can treat or reduce the undesirable effects of sickle cell disease by expressing a full-length wild-type  $\beta$ -globin chain of hemoglobin. In one example the disclosed methods reduce the symptoms of sickle-cell disease in the recipient subject (such as one or more of, presence of sickle cells in the blood, pain, ischemia, necrosis, anemia, vaso-occlusive crisis, aplastic crisis, splenic sequestration crisis, and haemolytic crisis) for example a reduction of at least 10%, at least 20%, at least 50%, at least 70%, or at least 90% (as compared to no administration of the therapeutic nucleic acid molecule). In one example the disclosed methods decrease the number of sickle cells in the recipient subject, for example a decrease of at least 10%, at least 20%, at least 50%, at least 70%, at least 90%, or at least 95% (as compared to no administration of the therapeutic nucleic acid molecule).

For example, the disclosed methods and systems can treat or reduce the undesirable effects of thrombophilia by expressing a full-length wild-type factor V Leiden or prothrombin gene. In one example the disclosed methods reduce the symptoms of thrombophilia in the recipient subject (such as one or more of, thrombosis, such as deep vein thrombosis, pulmonary embolism, venous thromboembolism, swelling, chest pain, palpitations) for example a reduction of at least 10%, at least 20%, at least 50%, at least 70%, or at least 90% (as compared to no administration of the therapeutic nucleic acid molecule). In one example the disclosed methods decrease the activity of coagulation factors in the recipient subject, for example a decrease of at least 10%, at least 20%, at least 50%, at least 70%, at least 90%, or at least 95% (as compared to no administration of the therapeutic nucleic acid molecule).

For example, the disclosed methods and systems can treat or reduce the undesirable effects of CD40 ligand deficiency by expressing a full-length wild-type CD40 ligand gene. In one example the disclosed methods reduce the symptoms of CD40 ligand deficiency in the recipient subject (such as one or more of, elevate serum IgM, low serum levels of other immunoglobulins, opportunistic infections, autoimmunity and malignancies) for example a reduction of at least 10%, at least 20%, at least 50%, at

least 70%, or at least 90% (as compared to no administration of the therapeutic nucleic acid molecule s). In one example the disclosed methods increase the amount or activity of CD40 ligand deficiency in the recipient subject, for example an increase of at least 10%, at least 20%, at least 50%, at least 70%, at least 90%, at least 100%, at least 200% or at least 500% (as compared to no administration of the therapeutic nucleic acid molecule).

For example, the disclosed methods can be used to treat or reduce the undesirable effects of a primary immunodeficiency disease resulting from a genetic defect. For example, the disclosed methods and systems (which can use two or more synthetic nucleic acid molecules to express a functional protein missing or defective in the subject, for example using AAV) can treat or reduce the undesirable effects of a primary immunodeficiency disease. In one example the disclosed methods reduce the symptoms of a primary immunodeficiency disease in the recipient subject (such as one or more of, a bacterial infection, fungal infection, viral infection, parasitic infection, lymph gland swelling, spleen enlargement, wounds, and weight loss) for example a reduction of at least 10%, at least 20%, at least 50%, at least 70%, or at least 90% (as compared to no administration of the therapeutic nucleic acid molecule). In one example the disclosed methods increase the number of immune cells (such as T cells, such as CD8 cells) in the recipient subject with a primary immune deficiency disorder, for example an increase of at least 10%, at least 20%, at least 50%, at least 70%, at least 90%, at least 95%, at least 100%, at least 200%, at least 300%, at least 400%, or at least 500% (as compared to no administration of the therapeutic nucleic acid molecule). In one example the disclosed methods reduce the number of infections ((such as bacterial, viral, fungal, or combinations thereof) in the recipient subject over a set period of time (such as over 1 year) with a primary immune deficiency disorder, for example a decrease of at least 10%, at least 20%, at least 50%, at least 70%, at least 90%, or at least 95%, (as compared to no administration of the therapeutic nucleic acid molecule).

For example, the disclosed methods can be used to treat or reduce the undesirable effects of a monogenetic disorder. For example, the disclosed methods (which can use two or more synthetic nucleic acid molecules to express a functional protein missing or defective in the subject, for example using AAV) can treat or reduce the undesirable effects of a monogenetic disorder. In one example the disclosed methods reduce the symptoms of a monogenetic disorder in the recipient subject, for example a reduction of at least 10%, at least 20%, at least 50%, at least 70%, or at least 90% (as compared to no administration of the therapeutic nucleic acid molecule). In one example the disclosed methods increase the amount of normal protein not normally expressed by the recipient subject with a monogenetic disorder, for example an increase of at least 10%, at least 20%, at least 50%, at least 70%, at least 90%, at least 95%, at least 100%, at least 200%, at least 300%, at least 400%, or at least 500% (as compared to no administration of the therapeutic nucleic acid molecule).

For example, the disclosed methods can be used to treat or reduce the undesirable effects of a hematological malignancy in the recipient subject. In one example the disclosed methods reduce the number of abnormal white blood cells (such as B cells) in the recipient subject (such as a subject with leukemia), for example a reduction of at least 10%, at least 20%, at least 50%, at least 70%, or at least 90% (as compared to no administration of the disclosed therapies). In one example, administration of the disclosed therapies can be used to treat or reduce the undesirable effects of a lymphoma, such as reduce the size of the lymphoma, volume of the lymphoma, rate of growth of the lymphoma, metastasis of the lymphoma, for example a reduction of at least 10%, at least 20%, at least 50%, at least 70%, or at least 90% (as compared to no administration of the disclosed therapies). In one example, administration of disclosed therapies can be used to treat or reduce the undesirable effects of multiple myeloma, such as reduce the number of abnormal plasma cells in the recipient subject, for example a reduction of at least 10%, at least 20%, at least 50%, at least 70%, or at least 90% (as compared to no administration of the disclosed therapies).

For example, the disclosed methods can be used to treat or reduce the undesirable effects of a malignancy, such as one that results from a genetic defect in the recipient subject. In one example the disclosed methods reduce the number of cancer cells, the size of a tumor, the volume of a tumor, or the number of metastases, in the recipient subject (such as a subject with a cancer listed herein), for example a reduction of at least 10%, at least 20%, at least 50%, at least 70%, or at least 90% (as compared to no administration of the disclosed therapies). In one example, administration of the disclosed therapies can be used to treat or reduce the undesirable effects of a lymphoma, such as reduce the size of the tumor, volume of the tumor, rate of growth of the cancer, metastasis of the cancer, for example a reduction of at least 10%, at least 20%, at least 50%, at least 70%, or at least 90% (as compared to no administration of the disclosed therapies).

For example, the disclosed methods can be used to treat or reduce the undesirable effects of a neurological disease that results from a genetic defect in the recipient subject. In one example the disclosed methods increase neurological function in the recipient subject (such as a subject with a neurological disease listed above), for example an increase of at least 10%, at least 20%, at least 50%, at least 70%, at least 90%, at least 100%, at least 200%, at least 300%, at least 400%, or at least 500% (as compared to no administration of the disclosed therapies).

### **Treatment of Duchenne Muscular Dystrophy (DMD)**

Duchenne muscular dystrophy (DMD, MIM:310200) is a lethal hereditary disease characterized by progressive muscle weakness and degeneration. As the disease progresses, degenerating muscle fibres are replaced by fat and fibrotic tissue. DMD is rooted in deficiency of the gene dystrophin (MIM:300377). The dystrophin gene spans a region of 22kbp, and is prone to mutations. Thus, DMD

can in some cases sporadically manifest even in patients without a familial history of the disease-causing mutation. DMD is one of four conditions known as dystrophinopathies. The other three diseases that belong to this group are Becker Muscular dystrophy (BMD, a mild form of DMD); an intermediate clinical presentation between DMD and BMD; and DMD-associated dilated cardiomyopathy (heart-disease) with little or no clinical skeletal, or voluntary, muscle disease. Thus, in some examples a patient with DMD, BMD, an intermediate clinical presentation between DMD and BMD; or DMD-associated dilated cardiomyopathy (heart-disease) with little or no clinical skeletal, or voluntary, muscle disease, is treated with the disclosed systems and methods.

The disclosed methods and systems can be used to treat the monogenic cause of DMD by expressing dystrophin. Dystrophin has a long coding region, such as dystrophin. Current methods of expressing dystrophin from a single AAV utilize shortened/truncated versions of dystrophin (micro-dystrophin and mini- dystrophin). Several of these truncated dystrophin delivery therapies are being tested in Phase I/II clinical trials (NCT03362502, NCT00428935, NCT03368742, NCT03375164). Although these truncated versions of dystrophin may ameliorate the worst consequences of dystrophin deficiency in DMD, they are not expected to have full functionality when compared to full-length dystrophin as the truncated versions are missing key domains in the rod and hinge region of the full-length protein. The disclosed methods and systems alleviate the size restriction of the transgenic payload of AAV by using “multiplexed” AAV combinations, because multiple AAV viruses can efficiently infect the same cell when introduced at high multiplicity of infection (MOI, i.e., high titer).

Thus, in some examples, a composition that includes two or more AAVs, each containing one of a set of disclosed synthetic molecules, is administered (*e.g.*, i.v.) to a DMD subject in a therapeutically effective amount, such as a set that includes two, three, four or five different synthetic RNA molecules (each in a different AAV), which when recombined, result in a full-length dystrophin coding sequence.

## VI. Exemplary Embodiments

1. A system for expressing a target protein, comprising (a) a first synthetic nucleic acid molecule comprising a first promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a coding sequence for an N-terminal portion of the target protein; a splice donor; and a first dimerization domain; and (b) a second synthetic nucleic acid molecule comprising a second promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; a branch point sequence; a polypyrimidine tract; a splice acceptor; and a coding sequence for a C-terminal portion of the target protein.

2. A system for expressing a target protein, comprising: (a) a first synthetic nucleic acid molecule comprising a first promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a coding sequence for an N-terminal portion of the target protein; a splice donor; and a first dimerization domain; and (b) a second synthetic nucleic acid molecule comprising a second promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; a branch point sequence; a polypyrimidine tract; a splice acceptor; and a coding sequence for a middle portion of the target protein; a second splice donor; and a third dimerization domain; and (c) a third synthetic nucleic acid molecule comprising a third promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a fourth dimerization domain, wherein the fourth dimerization domain binds to the third dimerization domain; a branch point sequence; a polypyrimidine tract; a splice acceptor; and a coding sequence for a C-terminal portion of the target protein.

3. A system for expressing a target protein, comprising (a) a first synthetic nucleic acid molecule comprising a first promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a coding sequence for an N-terminal portion of the target protein, a splice donor, and a first dimerization domain; (b) a second synthetic nucleic acid molecule comprising a second promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; a branch point sequence; a polypyrimidine tract; a splice acceptor; and a coding sequence for a middle portion of the target protein; a second splice donor; and a third dimerization domain; and (c) a third synthetic nucleic acid molecule comprising a third promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a fourth dimerization domain, wherein the fourth dimerization domain binds to the third dimerization domain; a branch point sequence; a polypyrimidine tract; a splice acceptor; and a coding sequence for a first middle portion of the target protein; a second splice donor; and a fifth dimerization domain; and (c) a fourth synthetic nucleic acid molecule comprising a fourth promoter operably linked to a sequence encoding an RNA molecule, the RNA molecule comprising from 5' to 3': a sixth dimerization domain, wherein the sixth dimerization domain binds to the fifth dimerization domain; a branch point sequence; a polypyrimidine tract; a splice acceptor; and a coding sequence for a C-terminal portion of the target protein.

4. The system of any one of embodiments 1 to 3, wherein each promoter is independently selected.

5. The system of any one of embodiments 1 to 4, wherein:  
the first and second promoter are the same promoter;

the first and second promoter are different promoters;  
the first, second, and third promoters are the same promoter;  
the first, second, and third promoters are different promoters;  
the first, second, third, and fourth promoters are the same promoter; or  
the first, second, third and fourth promoters are different promoters.

6. The system of any one of embodiments 1 to 5, wherein each of the first, second, third, and fourth promoter is independently selected from: a constitutive promoter; a tissue-specific promoter; and a promoter endogenous to the target protein.

7. The system of any one of embodiments 1 to 6, wherein the first and second dimerization domains, third and fourth dimerization domains, and/or fifth and sixth dimerization domains, bind by direct binding, indirect binding, or a combination thereof.

8. The composition of claim 7, wherein direct binding or indirect binding comprises basepairing interactions, non-canonical base pairing interactions, non-base pairing interactions, or a combination thereof.

9. The composition of claim 7 or 8, wherein direct binding comprises base pairing interactions between kissing loops or hypodiverse regions.

10. The composition of claim 7 or 8, wherein direct binding comprises non-canonical basepairing interactions, non-canonical base pairing interactions, non-base pairing interactions, or a combination thereof, between aptamer regions.

11. The composition of claim 7 or 8, wherein indirect binding comprises basepairing interactions through a nucleic acid bridge.

12. The composition of claim 7 or 8, wherein indirect binding comprises non-base pairing interactions between an aptamer and an aptamer target, or between two aptamers.

13. The system of any one of embodiments 1 to 12, wherein the first, second, third, fourth, fifth and/or sixth dimerization domain does not comprise a cryptic splice acceptor.

14. The system of any one of embodiments 1 to 13, comprising at least one pair of directly binding or indirectly binding aptamer sequence dimerization domains.

15. The system of any one of embodiments 1 to 14, comprising at least one pair of kissing loop interaction dimerization domains.

16. The system of any one of embodiments 1 to 15, wherein the target protein is a protein associated with disease, or a therapeutic protein.

17. The system of embodiment 16, wherein the disease is a monogenic disease.

18. The system of embodiment 17, wherein the therapeutic protein is a toxin.

19. The system of any one of embodiments 16 to 18, wherein the disease and the target protein are one listed in Table 1.

20. The system of any one of embodiments 1 to 19, wherein the first, second, third, and/or fourth synthetic nucleic acid molecule further comprises a polyadenylation sequence at a 3'-end of the first, second, third, or fourth synthetic nucleic acid molecule.

21. The system of any one of embodiments 1 or 4 to 20, wherein the  
5 the first synthetic nucleic acid molecule further comprises one or both of a downstream intronic splice enhancer (DISE) 3' to the splice donor and 5' to the first dimerization domain, an intronic splice enhancer (ISE) 3' to the splice donor and 5' to the first dimerization domain; and/or

the second synthetic nucleic acid molecule further comprises one or both of an ISE 3' to the second dimerization domain and 5' to the branch point sequence, and a DISE 3' to the splice donor and  
10 5' to the dimerization domain;  
and any combination thereof.

22. The system of any one of embodiments 2 or 4 to 20, wherein the  
the first synthetic nucleic acid molecule further comprises a DISE 3' to the first splice donor and 5' to the first dimerization domain, an ISE 3' to the first splice donor and 5' to the first dimerization  
15 domain, or both a DISE and ISE;

the second synthetic nucleic acid molecule further comprises an ISE 3' to the second dimerization domain and 5' to the first branch point sequence, a DISE 3' to the second splice donor and 5' to the second dimerization domain, an ISE 3' to the second splice donor and 5' to the third dimerization domain, or combinations thereof; and/or

20 the third synthetic nucleic acid molecule further comprises an ISE 3' to the fourth dimerization domain and 5' to the second branch point sequence;  
and any combination thereof.

23. The system of any one of embodiments 3 to 20, wherein  
the first synthetic nucleic acid molecule further comprises a DISE 3' to the first splice donor  
25 and 5' to the first dimerization domain, an ISE 3' to the first splice donor and 5' to the first dimerization domain, or both a DISE and ISE;

the second synthetic nucleic acid molecule further comprises an ISE 3' to the second dimerization domain and 5' to the first branch point sequence, a DISE 3' to the second splice donor and 5' to the second dimerization domain, an ISE 3' to the second splice donor and 5' to the third  
30 dimerization domain, or combinations thereof;

the third synthetic nucleic acid molecule further comprises ISE 3' to the fourth dimerization domain and 5' to the second branch point sequence; and/or

the fourth synthetic nucleic acid molecule further comprises a ISE 3' to the fifth dimerization domain and 5' to the third branch point sequence, a DISE 3' to the third splice donor and 5' to the fifth



dimerization domain, an ISE 3' to the third splice donor and 5' to the sixth dimerization domain, or combinations thereof;

and any combination thereof.

24. The system of any one of embodiments 1 to 23, wherein when the system is introduced into a cell the RNA molecules are produced and recombine in the proper order, resulting in a full-length coding sequence of the target protein.

25. The system of any one of embodiments 1 to 24, wherein each of the synthetic first, second, third and fourth nucleic acid molecules are part of a separate viral vector.

26. The system of embodiment 25, wherein the viral vector is AAV.

27. The system of any one of embodiments 1 to 26, wherein the first and/or third synthetic nucleic acid molecule further comprises a self-cleaving RNA sequence or an RNA-cleaving enzyme target sequence positioned anywhere 3' to the splice donor such that it cleaves off the 3' located polyadenylated tail to decrease or suppress protein fragment expression from a non-recombined RNA molecule;

the second and/or fourth synthetic nucleic acid molecule further comprises a self-cleaving RNA sequence or an RNA-cleaving enzyme target sequence positioned anywhere 5' to the branch point sequence such that it cleaves off the 5' located RNA cap to decrease or suppress protein fragment expression from a non-recombined RNA molecule;

the second and/or fourth synthetic nucleic acid molecule further comprises a start codon anywhere 5' to the branch point sequence that is shifted relative to the open reading frame 3' of the splice acceptor to decrease or suppress translation of a target protein fragment from a non-recombined RNA molecule;

the first and/or third synthetic nucleic acid molecule further comprises a micro RNA target site anywhere 3' to the splice donor such that an un-joined RNA fragment undergoes micro RNA dependent degradation once outside the nucleus;

the second and/or fourth synthetic nucleic acid molecule further comprises a micro RNA target site anywhere 3' to the coding sequence such that an un-joined RNA fragment undergoes micro RNA dependent degradation once outside the nucleus;

the first and/or third synthetic nucleic acid molecule further comprises a sequence encoding a degron protein degradation tag anywhere 3' to the splice donor such that it is in frame with the target protein open reading frame 5' of the splice donor site such that an un-joined protein fragment is tagged for degradation;

the second and/or fourth synthetic nucleic acid molecule further comprises a start codon and an in frame degron protein degradation tag anywhere 5' to the branch point sequence such that it is in

or combinations thereof.

29. The system of any one of embodiments 1 to 28, wherein the coding sequence for an N-terminal portion of the target protein, a middle portion of the target protein, or a C-terminal portion of the target protein encoded by a synthetic nucleic acid molecule of the system each has a size

30 independently selected from: about 1000 nt to about 4000 nt, about 1,000 nt to about 1,500 nt, about 1,000 nt to about 2,000 nt, about 1,000 nt to about 2,500 nt, about 1,000 nt to about 3,000 nt, about 1,000 nt to about 3,500 nt, about 1,000 nt to about 4,000 nt, about 1,500 nt to about 2,000 nt, about 1,500 nt to about 2,500 nt, about 1,500 nt to about 3,000 nt, about 1,500 nt to about 3,500 nt, about 1,500 nt to about 4,000 nt, about 2,000 nt to about 2,500 nt, about 2,000 nt to about 3,000 nt, about 2,000 nt to about 3,500 nt, about 2,000 nt to about 4,000 nt, about 2,500 nt to about 3,000 nt, about

35

2,500 nt to about 3,500 nt, about 2,500 nt to about 4,000 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 4,000 nt, about 3,500 nt to about 4,000 nt, about 1,000 nt, about 1,500 nt, about 2,000 nt, about 2,500 nt, about 3,000 nt, about 3,500 nt, and about 4,000 nt.

30. The system of any one of embodiments 1 to 29, wherein any one, two, three, or four RNAs  
5 encoded by any of the one, two, three, or four synthetic nucleic acid molecules of the system,  
respectively, each has a size independently selected from: about 2500 to 4500 nt, about 2,500 nt to  
about 2,750 nt, about 2,500 nt to about 3,000 nt, about 2,500 nt to about 3,250 nt, about 2,500 nt to  
about 3,500 nt, about 2,500 nt to about 3,750 nt, about 2,500 nt to about 4,000 nt, about 2,500 nt to  
about 4,250 nt, about 2,500 nt to about 4,500 nt, about 2,750 nt to about 3,000 nt, about 2,750 nt to  
10 about 3,250 nt, about 2,750 nt to about 3,500 nt, about 2,750 nt to about 3,750 nt, about 2,750 nt to  
about 4,000 nt, about 2,750 nt to about 4,250 nt, about 2,750 nt to about 4,500 nt, about 3,000 nt to  
about 3,250 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 3,750 nt, about 3,000 nt to  
about 4,000 nt, about 3,000 nt to about 4,250 nt, about 3,000 nt to about 4,500 nt, about 3,250 nt to  
about 3,500 nt, about 3,250 nt to about 3,750 nt, about 3,250 nt to about 4,000 nt, about 3,250 nt to  
15 about 4,250 nt, about 3,250 nt to about 4,500 nt, about 3,500 nt to about 3,750 nt, about 3,500 nt to  
about 4,000 nt, about 3,500 nt to about 4,250 nt, about 3,500 nt to about 4,500 nt, about 3,750 nt to  
about 4,000 nt, about 3,750 nt to about 4,250 nt, about 3,750 nt to about 4,500 nt, about 4,000 nt to  
about 4,250 nt, about 4,000 nt to about 4,500 nt, about 4,250 nt to about 4,500 nt, about 2,500 nt, about  
2,750 nt, about 3,000 nt, about 3,250 nt, about 3,500 nt, about 3,750 nt, about 4,000 nt, about 4,250 nt,  
20 and about 4,500 nt.

31. The system of any one of embodiments 1 and 4 to 30, wherein:  
the synthetic nucleic acid molecules have a total size selected from about 5000 nt to about 10,000 nt,  
about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt,  
about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt,  
25 about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,000 nt to about 9,500 nt,  
about 5,000 nt to about 10,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt,  
about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt,  
about 5,500 nt to about 8,500 nt, about 5,500 nt to about 9,000 nt, about 5,500 nt to about 9,500 nt,  
about 5,500 nt to about 10,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt,  
30 about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt,  
about 6,000 nt to about 9,000 nt, about 6,000 nt to about 9,500 nt, about 6,000 nt to about 10,000 nt,  
about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt,  
about 6,500 nt to about 8,500 nt, about 6,500 nt to about 9,000 nt, about 6,500 nt to about 9,500 nt,  
about 6,500 nt to about 10,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt,  
35 about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,000 nt to about 9,500 nt,

about 7,000 nt to about 10,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 7,500 nt to about 9,500 nt, about 7,500 nt to about 10,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,000 nt to about 9,500 nt, about 8,000 nt to about 10,000 nt, about 8,500 nt to about 9,000 nt, about 8,500 nt to about 9,500 nt, about 8,500 nt to about 10,000 nt, about 9,000 nt to about 9,500 nt, about 9,000 nt to about 10,000 nt, about 9,500 nt to about 10,000 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, about 9,000 nt, about 9,500 nt, and about 10,000 nt;

the total target protein coding sequence is selected from about 2000 nt to about 8000 nt, about 2,000 nt to about 3,000 nt, about 2,000 nt to about 3,500 nt, about 2,000 nt to about 4,000 nt, about 2,000 nt to about 4,500 nt, about 2,000 nt to about 5,000 nt, about 2,000 nt to about 5,500 nt, about 2,000 nt to about 6,000 nt, about 2,000 nt to about 6,500 nt, about 2,000 nt to about 7,000 nt, about 2,000 nt to about 7,500 nt, about 2,000 nt to about 8,000 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,500 nt, about 3,000 nt to about 5,000 nt, about 3,000 nt to about 5,500 nt, about 3,000 nt to about 6,000 nt, about 3,000 nt to about 6,500 nt, about 3,000 nt to about 7,000 nt, about 3,000 nt to about 7,500 nt, about 3,000 nt to about 8,000 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,500 nt, about 3,500 nt to about 5,000 nt, about 3,500 nt to about 5,500 nt, about 3,500 nt to about 6,000 nt, about 3,500 nt to about 6,500 nt, about 3,500 nt to about 7,000 nt, about 3,500 nt to about 7,500 nt, about 3,500 nt to about 8,000 nt, about 4,000 nt to about 4,500 nt, about 4,000 nt to about 5,000 nt, about 4,000 nt to about 5,500 nt, about 4,000 nt to about 6,000 nt, about 4,000 nt to about 6,500 nt, about 4,000 nt to about 7,000 nt, about 4,000 nt to about 7,500 nt, about 4,000 nt to about 8,000 nt, about 4,500 nt to about 5,000 nt, about 4,500 nt to about 5,500 nt, about 4,500 nt to about 6,000 nt, about 4,500 nt to about 6,500 nt, about 4,500 nt to about 7,000 nt, about 4,500 nt to about 7,500 nt, about 4,500 nt to about 8,000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, or about 7,500 nt to about 8,000 nt. the total target protein coding sequence is about 2,000 nt, about 3,000 nt, about 3,500 nt, about 4,000 nt, about 4,500 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, and about 8,000 nt; and/or

the RNA encoded by the two synthetic nucleic acid molecules has a total size selected from about 5,000 nt to about 9000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 5,500 nt to about 8,500 nt, about 5,500 nt to about 9,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 6,500 nt to about 8,500 nt, about 6,500 nt to about 9,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, or about 8,500 nt to about 9,000 nt. the RNA encoded by the two synthetic nucleic acid molecules has a total size of about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, and about 9,000 nt.

32. The system of any one of embodiments 2 and 4 to 30, wherein:

the synthetic nucleic acid molecules have a total size selected from about 7500 nt to about 15,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,500 nt, about 7,500 nt to about 10,000 nt, about 7,500 nt to about 10,500 nt, about 7,500 nt to about 11,000 nt, about 7,500 nt to about 11,500 nt, about 7,500 nt to about 12,000 nt, about 7,500 nt to about 12,500 nt, about 7,500 nt to about 13,000 nt, about 7,500 nt to about 14,000 nt, about 7,500 nt to about 15,000 nt, about 8,500 nt to about 9,500 nt, about 8,500 nt to about 10,000 nt, about 8,500 nt to about 10,500 nt, about 8,500 nt to about 11,000 nt, about 8,500 nt to about 11,500 nt, about 8,500 nt to about 12,000 nt, about 8,500 nt to about 12,500 nt, about 8,500 nt to about 13,000 nt, about 8,500 nt to about 14,000 nt, about 8,500 nt to about 15,000 nt, about 9,500 nt to about 10,000 nt, about 9,500 nt to about 10,500 nt, about 9,500 nt to about 11,000 nt, about 9,500 nt to about 11,500 nt, about 9,500 nt to about 12,000 nt, about 9,500 nt to about 12,500 nt, about 9,500 nt to about 13,000 nt, about 9,500 nt to about 14,000 nt, about 9,500 nt to about 15,000 nt, about 10,000 nt to about 10,500 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 11,500 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 12,500 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,500 nt to about 11,000 nt, about 10,500 nt to about 11,500 nt, about 10,500 nt to about 12,000 nt, about 10,500 nt to about 12,500 nt, about 10,500 nt to about 13,000 nt, about 10,500 nt to about 14,000 nt, about 10,500 nt to about 15,000 nt, about 11,000 nt to about 11,500 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 12,500 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about

14,000 nt, about 11,000 nt to about 15,000 nt, about 11,500 nt to about 12,000 nt, about 11,500 nt to about 12,500 nt, about 11,500 nt to about 13,000 nt, about 11,500 nt to about 14,000 nt, about 11,500 nt to about 15,000 nt, about 12,000 nt to about 12,500 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,500 nt to about 13,000 nt, about 12,500 nt to about 14,000 nt, about 12,500 nt to about 15,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, or about 14,000 nt to about 15,000 nt. the synthetic nucleic acid molecules have a total size of about 7,500 nt, about 8,500 nt, about 9,500 nt, about 10,000 nt, about 10,500 nt, about 11,000 nt, about 11,500 nt, about 12,000 nt, about 12,500 nt, about 13,000 nt, about 14,000 nt, and about 15,000 nt;

the total target protein coding sequence is selected from about 3000 nt to about 12,000 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 5,000 nt, about 3,000 nt to about 6,000 nt, about 3,000 nt to about 7,000 nt, about 3,000 nt to about 7,500 nt, about 3,000 nt to about 8,000 nt, about 3,000 nt to about 8,500 nt, about 3,000 nt to about 9,000 nt, about 3,000 nt to about 1,000 nt, about 3,000 nt to about 11,000 nt, about 3,000 nt to about 12,000 nt, about 4,000 nt to about 5,000 nt, about 4,000 nt to about 6,000 nt, about 4,000 nt to about 7,000 nt, about 4,000 nt to about 7,500 nt, about 4,000 nt to about 8,000 nt, about 4,000 nt to about 8,500 nt, about 4,000 nt to about 9,000 nt, about 4,000 nt to about 1,000 nt, about 4,000 nt to about 11,000 nt, about 4,000 nt to about 12,000 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,000 nt to about 1,000 nt, about 5,000 nt to about 11,000 nt, about 5,000 nt to about 12,000 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,000 nt to about 1,000 nt, about 6,000 nt to about 11,000 nt, about 6,000 nt to about 12,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,000 nt to about 1,000 nt, about 7,000 nt to about 11,000 nt, about 7,000 nt to about 12,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 7,500 nt to about 1,000 nt, about 7,500 nt to about 11,000 nt, about 7,500 nt to about 12,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,000 nt to about 1,000 nt, about 8,000 nt to about 11,000 nt, about 8,000 nt to about 12,000 nt, about 8,500 nt to about 9,000 nt, about 8,500 nt to about 1,000 nt, about 8,500 nt to about 11,000 nt, about 8,500 nt to about 12,000 nt, about 9,000 nt to about 1,000 nt, about 9,000 nt to about 11,000 nt, about 9,000 nt to about 12,000 nt, about 1,000 nt to about 11,000 nt, about 1,000 nt to about 12,000 nt, or about 11,000 nt to about 12,000 nt. the total target protein coding sequence is about 3,000 nt, about 4,000 nt, about 5,000 nt, about 6,000 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, about 9,000 nt, about 1,000 nt, about 11,000 nt, and about 12,000 nt; and/or

the RNA encoded by the three synthetic nucleic acid molecules has a total size selected from about 7500 nt to about 13,500 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 7,500 nt to about 9,500 nt, about 7,500 nt to about 10,000 nt, about 7,500 nt to about 10,500 nt, about 7,500 nt to about 11,000 nt, about 7,500 nt to about 11,500 nt, about 7,500 nt to about 12,000 nt, about 7,500 nt to about 12,500 nt, about 7,500 nt to about 13,000 nt, about 7,500 nt to about 13,500 nt, about 8,500 nt to about 9,000 nt, about 8,500 nt to about 9,500 nt, about 8,500 nt to about 10,000 nt, about 8,500 nt to about 10,500 nt, about 8,500 nt to about 11,000 nt, about 8,500 nt to about 11,500 nt, about 8,500 nt to about 12,000 nt, about 8,500 nt to about 12,500 nt, about 8,500 nt to about 13,000 nt, about 8,500 nt to about 13,500 nt, about 9,000 nt to about 9,500 nt, about 9,000 nt to about 10,000 nt, about 9,000 nt to about 10,500 nt, about 9,000 nt to about 11,000 nt, about 9,000 nt to about 11,500 nt, about 9,000 nt to about 12,000 nt, about 9,000 nt to about 12,500 nt, about 9,000 nt to about 13,000 nt, about 9,000 nt to about 13,500 nt, about 9,500 nt to about 10,000 nt, about 9,500 nt to about 10,500 nt, about 9,500 nt to about 11,000 nt, about 9,500 nt to about 11,500 nt, about 9,500 nt to about 12,000 nt, about 9,500 nt to about 12,500 nt, about 9,500 nt to about 13,000 nt, about 9,500 nt to about 13,500 nt, about 10,000 nt to about 10,500 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 11,500 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 12,500 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 13,500 nt, about 10,500 nt to about 11,000 nt, about 10,500 nt to about 11,500 nt, about 10,500 nt to about 12,000 nt, about 10,500 nt to about 12,500 nt, about 10,500 nt to about 13,000 nt, about 10,500 nt to about 13,500 nt, about 11,000 nt to about 11,500 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 12,500 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 13,500 nt, about 11,500 nt to about 12,000 nt, about 11,500 nt to about 12,500 nt, about 11,500 nt to about 13,000 nt, about 11,500 nt to about 13,500 nt, about 12,000 nt to about 12,500 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 13,500 nt, about 12,500 nt to about 13,000 nt, about 12,500 nt to about 13,500 nt, or about 13,000 nt to about 13,500 nt. the RNA encoded by the two synthetic nucleic acid molecules has a total size of about 7,500 nt, about 8,500 nt, about 9,000 nt, about 9,500 nt, about 10,000 nt, about 10,500 nt, about 11,000 nt, about 11,500 nt, about 12,000 nt, about 12,500 nt, about 13,000 nt, and about 13,500 nt.

33. The system of any one of embodiments 3 and 4 to 30, wherein:

the synthetic nucleic acid molecules have a total size selected from about 10,000 nt to about 20,000 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,000 nt to about 16,000 nt, about 10,000 nt to about 17,000 nt, about 10,000 nt to about 18,000 nt, about 10,000 nt to about 19,000 nt, about 10,000 nt to about 20,000 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 14,000 nt, about 11,000 nt to about 15,000 nt, about 11,000 nt to about 16,000 nt, about 11,000 nt to about 17,000 nt, about 11,000 nt to about 18,000

nt, about 11,000 nt to about 19,000 nt, about 11,000 nt to about 20,000 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,000 nt to about 16,000 nt, about 12,000 nt to about 17,000 nt, about 12,000 nt to about 18,000 nt, about 12,000 nt to about 19,000 nt, about 12,000 nt to about 20,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, about 13,000 nt to about 16,000 nt, about 13,000 nt to about 17,000 nt, about 13,000 nt to about 18,000 nt, about 13,000 nt to about 19,000 nt, about 13,000 nt to about 20,000 nt, about 14,000 nt to about 15,000 nt, about 14,000 nt to about 16,000 nt, about 14,000 nt to about 17,000 nt, about 14,000 nt to about 18,000 nt, about 14,000 nt to about 19,000 nt, about 14,000 nt to about 20,000 nt, about 15,000 nt to about 16,000 nt, about 15,000 nt to about 17,000 nt, about 15,000 nt to about 18,000 nt, about 15,000 nt to about 19,000 nt, about 15,000 nt to about 20,000 nt, about 16,000 nt to about 17,000 nt, about 16,000 nt to about 18,000 nt, about 16,000 nt to about 19,000 nt, about 16,000 nt to about 20,000 nt, about 17,000 nt to about 18,000 nt, about 17,000 nt to about 19,000 nt, about 17,000 nt to about 20,000 nt, about 18,000 nt to about 19,000 nt, about 18,000 nt to about 20,000 nt, or about 19,000 nt to about 20,000 nt. the synthetic nucleic acid molecules have a total size of about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, about 15,000 nt, about 16,000 nt, about 17,000 nt, about 18,000 nt, about 19,000 nt, and about 20,000 nt;

the total target protein coding sequence is selected from about 4000 nt to about 16,000 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 9,000 nt, about 5,000 nt to about 10,000 nt, about 5,000 nt to about 11,000 nt, about 5,000 nt to about 12,000 nt, about 5,000 nt to about 13,000 nt, about 5,000 nt to about 14,000 nt, about 5,000 nt to about 15,000 nt, about 5,000 nt to about 16,000 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 9,000 nt, about 6,000 nt to about 10,000 nt, about 6,000 nt to about 11,000 nt, about 6,000 nt to about 12,000 nt, about 6,000 nt to about 13,000 nt, about 6,000 nt to about 14,000 nt, about 6,000 nt to about 15,000 nt, about 6,000 nt to about 16,000 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 9,000 nt, about 7,000 nt to about 10,000 nt, about 7,000 nt to about 11,000 nt, about 7,000 nt to about 12,000 nt, about 7,000 nt to about 13,000 nt, about 7,000 nt to about 14,000 nt, about 7,000 nt to about 15,000 nt, about 7,000 nt to about 16,000 nt, about 8,000 nt to about 9,000 nt, about 8,000 nt to about 10,000 nt, about 8,000 nt to about 11,000 nt, about 8,000 nt to about 12,000 nt, about 8,000 nt to about 13,000 nt, about 8,000 nt to about 14,000 nt, about 8,000 nt to about 15,000 nt, about 8,000 nt to about 16,000 nt, about 9,000 nt to about 10,000 nt, about 9,000 nt to about 11,000 nt, about 9,000 nt to about 12,000 nt, about 9,000 nt to about 13,000 nt, about 9,000 nt to about 14,000 nt, about 9,000 nt to about 15,000 nt, about 9,000 nt to about 16,000 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,000 nt to about 16,000 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about



14,000 nt, about 11,000 nt to about 15,000 nt, about 11,000 nt to about 16,000 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,000 nt to about 16,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, about 13,000 nt to about 16,000 nt, about 14,000 nt to about 15,000 nt, about 14,000 nt to about 16,000 nt, or about 15,000 nt to about 16,000 nt. the total target protein coding sequence is about 5,000 nt, about 6,000 nt, about 7,000 nt, about 8,000 nt, about 9,000 nt, about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, about 15,000 nt, or about 16,000 nt. the total target protein coding sequence is at least about 5,000 nt, about 6,000 nt, about 7,000 nt, about 8,000 nt, about 9,000 nt, about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, and about 15,000 nt; and/or

the RNA encoded by the two synthetic nucleic acid molecules has a total size selected from about 10,000 nt to about 18,000 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,000 nt to about 16,000 nt, about 10,000 nt to about 17,000 nt, about 10,000 nt to about 18,000 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 14,000 nt, about 11,000 nt to about 15,000 nt, about 11,000 nt to about 16,000 nt, about 11,000 nt to about 17,000 nt, about 11,000 nt to about 18,000 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,000 nt to about 16,000 nt, about 12,000 nt to about 17,000 nt, about 12,000 nt to about 18,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, about 13,000 nt to about 16,000 nt, about 13,000 nt to about 17,000 nt, about 13,000 nt to about 18,000 nt, about 14,000 nt to about 15,000 nt, about 14,000 nt to about 16,000 nt, about 14,000 nt to about 17,000 nt, about 14,000 nt to about 18,000 nt, about 15,000 nt to about 16,000 nt, about 15,000 nt to about 17,000 nt, about 15,000 nt to about 18,000 nt, about 16,000 nt to about 17,000 nt, about 16,000 nt to about 18,000 nt, or about 17,000 nt to about 18,000 nt. the RNA encoded by the two synthetic nucleic acid molecules has a total size of about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, about 15,000 nt, about 16,000 nt, about 17,000 nt, and about 18,000 nt.

34. The system of any one of embodiments 1 to 33, wherein the RNA recombination efficiency is about 10% to about 95%, about 10% to about 20%, about 10% to about 30%, about 10% to about 35%, about 10% to about 40%, about 10% to about 45%, about 10% to about 50%, about 10% to about 55%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 20% to about 30%, about 20% to about 35%, about 20% to about 40%, about 20% to about 45%, about 20% to about 50%, about 20% to about 55%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 30% to about 35%, about 30% to about 40%, about 30% to about 45%, about 30% to about 50%, about 30% to about 55%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 35% to about

40%, about 35% to about 45%, about 35% to about 50%, about 35% to about 55%, about 35% to about 60%, about 35% to about 70%, about 35% to about 80%, about 35% to about 90%, about 40% to about 45%, about 40% to about 50%, about 40% to about 55%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about 40% to about 90%, about 45% to about 50%, about 45% to about 55%, about 45% to about 60%, about 45% to about 70%, about 45% to about 80%, about 45% to about 90%, about 50% to about 55%, about 50% to about 60%, about 50% to about 70%, about 50% to about 80%, about 50% to about 90%, about 55% to about 60%, about 55% to about 70%, about 55% to about 80%, about 55% to about 90%, about 60% to about 70%, about 60% to about 80%, about 60% to about 90%, about 70% to about 80%, about 70% to about 90%, about 80% to about 90%, about 10%, about 20%, about 30%, about 35%, about 40%, about 45%, about 50%, about 55%, about 60%, about 70%, about 80%, or about 90%, or about 95%.

35. The system of any one of embodiments 1 to 34, wherein the first dimerization domain and the second dimerization domain, the third dimerization domain and the fourth dimerization domain, and/or the fifth dimerization domain and the sixth dimerization domain, are each no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50 to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, or at least about 95%.

36. The system of any one of embodiments 1 to 35, wherein each dimerization domain is no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50 to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, or at least 90%.

37. A composition comprising the system of any one of embodiments 1 to 36.

38. A composition comprising an RNA molecule as described in any one of embodiments 1 to 37.

39. A composition comprising one, two, three, or four RNA molecules as described in any one of embodiments 1 to 37.

40. The composition of any one of embodiments 37 to 39, wherein the composition comprises first, second, third and optionally fourth synthetic nucleic acid or RNA molecules, each encoding at least a portion of dystrophin, factor 8, ABCA4, or MYO7A.

41. An RNA molecule as described in any one of embodiments 1 to 36.

42. A kit comprising the system of any one of embodiments 1 to 41, or composition of any one of embodiments 37 to 40, wherein any of the synthetic first, second, third and fourth nucleic acid molecules can be in separate containers, and optionally further comprising a buffer such as a pharmaceutically acceptable carrier.

5 43. A method of expressing a target protein in a cell, comprising:

introducing the system of any one of embodiments 1 to 36, or composition of any one of embodiments 35 to 37, into a cell, and expressing the synthetic first and second, first, second, and third, or first, second, third and fourth RNA molecules in the cell, wherein the target protein is produced in the cell.

10 44. The method of embodiment 43, wherein the cell is in a subject, and introducing comprises administering a therapeutically effective amount the system to the subject.

45. The method of embodiment 44, wherein the method treats a genetic disease caused by a mutation in a gene encoding the target protein in the subject, wherein the method results in expression of functional target protein in the subject.

15 46. The method of embodiment 45, wherein

the genetic disease is Duchenne muscular dystrophy and the target protein is dystrophin;

the genetic disease is Hemophilia A and the target protein is F8;

the genetic disease is Stargardt disease and the target protein is ABCA4; or

the genetic disease is Usher syndrome and the target protein is MYO7A.

20 47. A nucleic acid molecule comprising at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, at least 99%, or 100% sequence identity to a synthetic intron provided in any one of SEQ ID NOS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 145, 146, 147, 148, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, and 166.

25 48. The nucleic acid molecule of embodiment 47, wherein the synthetic intron is nt 3703 to 3975 of SEQ ID NO: 20, nt 1 to 228 of SEQ ID NO: 21, nt 3703 to 3975 of SEQ ID NO: 22, nt 1 to 225 of SEQ ID NO: 23, nt 3560 to 3828 of SEQ ID NO: 24, or nt 1-225 of SEQ ID NO: 25.

49. The synthetic nucleic acid molecule of embodiment 47 or 48, further comprising a portion of a protein coding sequence.

30 50. The synthetic nucleic acid molecule of embodiment 49, wherein the portion of the protein coding sequence comprises an N-terminal half, an N-terminal third, a middle portion, a C-terminal half, or a C-terminal third of the protein coding sequence.

51. A system of any one of embodiments 1 to 36, or composition of any one of embodiments 37 to 40, wherein at least one synthetic nucleic acid molecule comprises a synthetic intron comprising a nucleic acid molecule as set forth in any one of embodiments 47 to 50.

52. The composition, system, method, or kit of any preceding embodiment, wherein the synthetic nucleic acid is DNA that is produced by transcription of an RNA virus genome by reverse transcriptase.

## VII. Additional Exemplary Embodiments

5 1. A composition for expressing a target protein comprising (a) a first RNA molecule, the RNA molecule comprising from 5' to 3': (i) a coding sequence for an N-terminal portion of the target protein; (ii) a splice donor; and (iii) a first dimerization domain; and (b) a second RNA molecule, the RNA molecule comprising from 5' to 3': (i) a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; (ii) a branch point sequence; (iii) a polypyrimidine tract; 10 (iv) a splice acceptor; and (v) a coding sequence for a C-terminal portion of the target protein.

2. A composition for expressing a target protein comprising: (a) a first RNA molecule, the RNA molecule comprising from 5' to 3': (i) a coding sequence for an N-terminal portion of the target protein; (ii) a splice donor; and (iii) a first dimerization domain; and (b) a second RNA molecule, the RNA molecule comprising from 5' to 3': (i) a second dimerization domain, wherein the second 15 dimerization domain binds to the first dimerization domain; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice acceptor; (v) a coding sequence for a middle portion of the target protein; (vi) a second splice donor; and (vii) a third dimerization domain; and (c) a third RNA molecule, the RNA molecule comprising from 5' to 3': (i) a fourth dimerization domain, wherein the fourth dimerization domain binds to the third dimerization domain; (ii) a branch point sequence; (iii) a 20 polypyrimidine tract; (iv) a splice acceptor; and (v) a coding sequence for a C-terminal portion of the target protein.

3. A composition for expressing a target protein comprising (a) a first RNA molecule, the RNA molecule comprising from 5' to 3': (i) a coding sequence for an N-terminal portion of the target protein, (ii) a splice donor; and (iii) a first dimerization domain; (b) a second RNA molecule, the RNA molecule 25 comprising from 5' to 3': (i) a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice acceptor; (v) a coding sequence for a middle portion of the target protein; (vi) a second splice donor; and (vii) a third dimerization domain; and (c) a third RNA molecule, the RNA molecule comprising from 5' to 3': (i) a fourth dimerization domain, wherein the fourth dimerization domain 30 binds to the third dimerization domain; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice acceptor; (v) a coding sequence for a first middle portion of the target protein; (vi) a second splice donor; and (vii) a fifth dimerization domain; and (d) a fourth RNA molecule, the RNA molecule comprising from 5' to 3': (i) a sixth dimerization domain, wherein the sixth dimerization domain binds to the fifth dimerization domain; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice 35 acceptor; and (v) a coding sequence for a C-terminal portion of the target protein.

4. The composition of any one of embodiments 1 to 3, wherein the first and second dimerization domains, third and fourth dimerization domains, and/or fifth and sixth dimerization domains, bind by direct binding, indirect binding, or a combination thereof.

5. The composition of claim 4, wherein direct binding or indirect binding comprises basepairing interactions, non-canonical base pairing interactions, non-base pairing interactions, or a combination thereof.

6. The composition of claim 4 or 5, wherein direct binding comprises base pairing interactions between kissing loops or hypodiverse regions.

7. The composition of claim 4 or 5, wherein direct binding comprises non-canonical basepairing interactions, non-canonical base pairing interactions, non-base pairing interactions, or a combination thereof, between aptamer regions.

8. The composition of claim 4 or 5, wherein indirect binding comprises basepairing interactions through a nucleic acid bridge.

9. The composition of claim 4 or 5, wherein indirect binding comprises non-base pairing interactions between an aptamer and an aptamer target agent, or between two aptamers.

10. The composition of any one of embodiments 1 to 9, wherein the first, second, third, fourth, fifth and/or sixth dimerization domain does not comprise a cryptic splice acceptor.

11. The composition of any one of embodiments 1 to 10, comprising at least one pair of directly binding or indirectly binding aptamer sequence dimerization domains.

12. The composition of any one of embodiments 1 to 11, comprising at least one pair of kissing loop interaction dimerization domains.

13. The composition of any one of embodiments 1 to 12, wherein the target protein is a protein associated with disease, or a therapeutic protein.

14. The composition of embodiment 13, wherein the disease is a monogenic disease.

15. The composition of embodiment 14, wherein the therapeutic protein is a toxin.

16. The composition of any one of embodiments 13 to 15, wherein the disease and the target protein are one listed in Table 1.

17. The composition of any one of embodiments 1 to 16, wherein the first, second, third, and/or fourth RNA molecule further comprises a polyA tail at a 3'-end of the first, second, third, or fourth RNA molecule.

18. The composition of any one of embodiments 1 or 4 to 17, wherein the first RNA molecule further comprises one or both of a downstream intronic splice enhancer (DISE) 3' to the splice donor and 5' to the first dimerization domain, an intronic splice enhancer (ISE) 3' to the splice donor and 5' to the first dimerization domain; and/or

the second RNA molecule further comprises one or both of an ISE 3' to the second dimerization domain and 5' to the branch point sequence, and a DISE 3' to the splice donor and 5' to the dimerization domain;

or any combination thereof.

5 19. The composition of any one of embodiments 2 or 4 to 17, wherein the  
the first RNA molecule further comprises a DISE 3' to the first splice donor and 5' to the first dimerization domain, an ISE 3' to the first splice donor and 5' to the first dimerization domain, or both;  
the RNA molecule further comprises an ISE 3' to the second dimerization domain and 5' to the first branch point sequence, a DISE 3' to the second splice donor and 5' to the second dimerization  
10 domain, an ISE 3' to the second splice donor and 5' to the third dimerization domain, or any combination thereof; and/or

the third RNA molecule further comprises an ISE 3' to the fourth dimerization domain and 5' to the second branch point sequence;

or any combination thereof.

15 20. The composition of any one of embodiments 3 to 17, wherein  
the first RNA molecule further comprises a DISE 3' to the first splice donor and 5' to the first dimerization domain, an ISE 3' to the first splice donor and 5' to the first dimerization domain, or both;  
the second RNA molecule further comprises an ISE 3' to the second dimerization domain and 5' to the first branch point sequence, a DISE 3' to the second splice donor and 5' to the second  
20 dimerization domain, an ISE 3' to the second splice donor and 5' to the third dimerization domain, or any combination thereof;

the third RNA molecule further comprises an ISE 3' to the fourth dimerization domain and 5' to the second branch point sequence; and/or

the fourth RNA molecule further comprises an ISE 3' to the fifth dimerization domain and 5' to  
25 the third branch point sequence, a DISE 3' to the third splice donor and 5' to the fifth dimerization domain, an ISE 3' to the third splice donor and 5' to the sixth dimerization domain, or any combination thereof;

or any combination thereof.

30 24. The composition of any one of embodiments 1 to 23, wherein  
the first and/or third RNA molecule further comprises a self-cleaving RNA sequence or an RNA-cleaving enzyme target sequence positioned anywhere 3' to the splice donor such that it cleaves off the 3' located polyadenylated tail to decrease or suppress protein fragment expression from a non-recombined RNA molecule;

the second and/or fourth RNA molecule further comprises a self-cleaving RNA sequence or an  
35 RNA-cleaving enzyme target sequence positioned anywhere 5' to the branch point sequence such that it

cleaves off the 5' located RNA cap to decrease or suppress protein fragment expression from a non-recombined RNA molecule;

the second and/or fourth RNA molecule further comprises a start codon anywhere 5' to the branch point sequence that is shifted relative to the open reading frame 3' of the splice acceptor to decrease or suppress translation of a target protein fragment from a non-recombined RNA molecule;

the first and/or third RNA molecule further comprises a micro RNA target site anywhere 3' to the splice donor such that an un-joined RNA fragment undergoes micro RNA dependent degradation once outside the nucleus;

the second and/or fourth RNA molecule further comprises a micro RNA target site anywhere 3' to the coding sequence such that an un-joined RNA fragment undergoes micro RNA dependent degradation once outside the nucleus;

the first and/or third RNA molecule further comprises a sequence encoding a degron protein degradation tag anywhere 3' to the splice donor such that it is in frame with the target protein open reading frame 5' of the splice donor site such that an un-joined protein fragment is tagged for degradation;

the second and/or fourth RNA molecule further comprises a start codon and an in-frame degron protein degradation tag anywhere 5' to the branch point sequence such that it is in frame with the target protein open reading frame 3' of the splice acceptor site such that an un-joined protein fragment is tagged for degradation;

or any combination thereof.

25. A composition for expressing a target protein comprising: (a) a first synthetic DNA molecule that encodes the first RNA molecule of any one of embodiments 1 and 4 to 24, wherein the first synthetic DNA molecule comprises (i) a first promoter operably linked to a sequence encoding the first RNA molecule; and (b) a second synthetic DNA molecule that encodes the second RNA molecule of any one of embodiments 1 and 4 to 24, wherein the second synthetic DNA molecule comprises (i) a second promoter operably linked to a sequence encoding the second RNA molecule.

26. A composition for expressing a target protein comprising: (a) a first synthetic DNA molecule that encodes the first RNA molecule of any one of embodiments 2 and 4 to 24, wherein the first synthetic DNA molecule comprises (i) a first promoter operably linked to a sequence encoding the first RNA molecule; (b) a second synthetic DNA molecule that encodes the second RNA molecule of any one of embodiments 2 and 4 to 24, wherein the second synthetic DNA molecule comprises (i) a second promoter operably linked to a sequence encoding the second RNA molecule; and (c) a third synthetic DNA molecule that encodes the third RNA molecule of any one of embodiments 2 and 4 to 24, wherein the third synthetic DNA molecule comprises (i) a third promoter operably linked to a sequence encoding the third RNA molecule.

27. A composition for expressing a target protein comprising: (a) a first synthetic DNA molecule that encodes the first RNA molecule of any one of embodiments 3 and 4 to 24, wherein the first synthetic DNA molecule comprises (i) a first promoter operably linked to a sequence encoding the first RNA molecule; (b) a second synthetic DNA molecule that encodes the second RNA molecule of any one of embodiments 3 and 4 to 24, wherein the second synthetic DNA molecule comprises (i) a second promoter operably linked to a sequence encoding the second RNA molecule; (c) a third synthetic DNA molecule that encodes the third RNA molecule of any one of embodiments 3 and 4 to 24, wherein the third synthetic DNA molecule comprises (i) a third promoter operably linked to a sequence encoding the third RNA molecule; and (d) a fourth synthetic DNA molecule that encodes the fourth RNA molecule of any one of embodiments 3 and 4 to 24, wherein the fourth synthetic DNA molecule comprises (i) a fourth promoter operably linked to a sequence encoding the fourth RNA molecule.

28. The composition of any one of embodiments 25 to 27, wherein each promoter is independently selected.

29. The composition of any one of embodiments 25 to 28, wherein:

the first and second promoter are the same promoter;

the first and second promoter are different promoters;

the first, second, and third promoters are the same promoter;

the first, second, and third promoters are different promoters;

the first, second, third, and fourth promoters are the same promoter; or

the first, second, third and fourth promoters are different promoters.

30. The composition of any one of embodiments 25 to 29, wherein each of the first, second, third, and fourth promoter is independently selected from: a constitutive promoter; a tissue-specific promoter; and a promoter endogenous to the target protein.

31. A system for expressing a target protein comprising a composition of any one of embodiments 25 to 30.

32. The system of embodiment 31, wherein when the system is introduced into a cell the RNA molecules are produced and recombine in the proper order, resulting in a full-length coding sequence of the target protein.

33. The system of embodiment 31 or 32, wherein each of the first and second RNA molecules (in a two-part system), each of the first, second and third RNA molecules (in a three-part system), or each of the first, second, third and fourth, RNA molecules (in a four-part system) is transcribed from a separate viral vector.

34. The system of any one of embodiments 31 to 33, wherein the viral vector is AAV.



35. The system of any one of embodiments 31 to 34, wherein the first, second, third, or fourth synthetic DNA molecule of the system each has a size independently selected from: about 2500 nt to about 5000 nt, 2,500 nt to about 2,750 nt, about 2,500 nt to about 3,000 nt, about 2,500 nt to about 3,250 nt, about 2,500 nt to about 3,500 nt, about 2,500 nt to about 3,750 nt, about 2,500 nt to about 4,000 nt, about 2,500 nt to about 4,250 nt, about 2,500 nt to about 4,500 nt, about 2,500 nt to about 4,750 nt, about 2,500 nt to about 5,000 nt, about 2,750 nt to about 3,000 nt, about 2,750 nt to about 3,250 nt, about 2,750 nt to about 3,500 nt, about 2,750 nt to about 3,750 nt, about 2,750 nt to about 4,000 nt, about 2,750 nt to about 4,250 nt, about 2,750 nt to about 4,500 nt, about 2,750 nt to about 4,750 nt, about 2,750 nt to about 5,000 nt, about 3,000 nt to about 3,250 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 3,750 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,250 nt, about 3,000 nt to about 4,500 nt, about 3,000 nt to about 4,750 nt, about 3,000 nt to about 5,000 nt, about 3,250 nt to about 3,500 nt, about 3,250 nt to about 3,750 nt, about 3,250 nt to about 4,000 nt, about 3,250 nt to about 4,250 nt, about 3,250 nt to about 4,500 nt, about 3,250 nt to about 4,750 nt, about 3,250 nt to about 5,000 nt, about 3,500 nt to about 3,750 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,250 nt, about 3,500 nt to about 4,500 nt, about 3,500 nt to about 4,750 nt, about 3,500 nt to about 5,000 nt, about 3,750 nt to about 4,000 nt, about 3,750 nt to about 4,250 nt, about 3,750 nt to about 4,500 nt, about 3,750 nt to about 4,750 nt, about 3,750 nt to about 5,000 nt, about 4,000 nt to about 4,250 nt, about 4,000 nt to about 4,500 nt, about 4,000 nt to about 4,750 nt, about 4,000 nt to about 5,000 nt, about 4,250 nt to about 4,500 nt, about 4,250 nt to about 4,750 nt, about 4,250 nt to about 5,000 nt, about 4,500 nt to about 4,750 nt, about 4,500 nt to about 5,000 nt, about 4,750 nt to about 5,000 nt, about 2,500 nt, about 2,750 nt, about 3,000 nt, about 3,250 nt, about 3,500 nt, about 3,750 nt, about 4,000 nt, about 4,250 nt, about 4,500 nt, about 4,750 nt, and about 5,000 nt.

36. The system of any one of embodiments 31 to 35, wherein the coding sequence for an N-terminal portion of the target protein (in a two, three, or four-part system), a middle portion of the target protein (in a three-part system), a first middle portion of the target protein (in a four-part system), or a C-terminal portion of the target protein (in a two, three, or four-part system) encoded by a synthetic DNA molecule of the system each has a size independently selected from: each has a size independently selected from: about 1,000 nt to about 4,500 nt. each has a size independently selected from: about 1,000 nt to about 1,500 nt, about 1,000 nt to about 2,000 nt, about 1,000 nt to about 2,500 nt, about 1,000 nt to about 3,000 nt, about 1,000 nt to about 3,500 nt, about 1,000 nt to about 4,000 nt, about 1,000 nt to about 4,500 nt, about 1,500 nt to about 2,000 nt, about 1,500 nt to about 2,500 nt, about 1,500 nt to about 3,000 nt, about 1,500 nt to about 3,500 nt, about 1,500 nt to about 4,000 nt, about 1,500 nt to about 4,500 nt, about 2,000 nt to about 2,500 nt, about 2,000 nt to about 3,000 nt, about 2,000 nt to about 3,500 nt, about 2,000 nt to about 4,000 nt, about 2,000 nt to about 4,500 nt,

about 2,500 nt to about 3,000 nt, about 2,500 nt to about 3,500 nt, about 2,500 nt to about 4,000 nt, about 2,500 nt to about 4,500 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,500 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,500 nt, about 4,000 nt to about 4,500 nt, about 1,000 nt, about 1,500 nt, about 2,000 nt, about 2,500 nt, about 3,000 nt, about 3,500 nt, about 4,000 nt, or about 4,500 nt.

37. The system of any one of embodiments 31 to 36, wherein any one, two, three, or four RNA molecules encoded by any of the one, two, three, or four synthetic DNA molecules of the system, respectively, each has a size independently selected from: about 2500 to 4500 nt, about 2,500 nt to about 2,750 nt, about 2,500 nt to about 3,000 nt, about 2,500 nt to about 3,250 nt, about 2,500 nt to about 3,500 nt, about 2,500 nt to about 3,750 nt, about 2,500 nt to about 4,000 nt, about 2,500 nt to about 4,250 nt, about 2,500 nt to about 4,500 nt, about 2,750 nt to about 3,000 nt, about 2,750 nt to about 3,250 nt, about 2,750 nt to about 3,500 nt, about 2,750 nt to about 3,750 nt, about 2,750 nt to about 4,000 nt, about 2,750 nt to about 4,250 nt, about 2,750 nt to about 4,500 nt, about 3,000 nt to about 3,250 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 3,750 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,250 nt, about 3,000 nt to about 4,500 nt, about 3,250 nt to about 3,500 nt, about 3,250 nt to about 3,750 nt, about 3,250 nt to about 4,000 nt, about 3,250 nt to about 4,250 nt, about 3,250 nt to about 4,500 nt, about 3,500 nt to about 3,750 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,250 nt, about 3,500 nt to about 4,500 nt, about 3,750 nt to about 4,000 nt, about 3,750 nt to about 4,250 nt, about 3,750 nt to about 4,500 nt, about 4,000 nt to about 4,250 nt, about 4,000 nt to about 4,500 nt, about 4,250 nt to about 4,500 nt, about 2,500 nt, about 2,750 nt, about 3,000 nt, about 3,250 nt, about 3,500 nt, about 3,750 nt, about 4,000 nt, about 4,250 nt, and about 4,500 nt.

38. The system of any one of embodiments 31 to 37, the system comprising a composition of embodiment 25 and 28 to 30, wherein:

the first and second synthetic DNA molecules have a total size selected from about 5000 nt to about 10,000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,000 nt to about 9,500 nt, about 5,000 nt to about 10,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 5,500 nt to about 8,500 nt, about 5,500 nt to about 9,000 nt, about 5,500 nt to about 9,500 nt, about 5,500 nt to about 10,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,000 nt to about 9,500 nt, about 6,000 nt to about 10,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about

8,000 nt, about 6,500 nt to about 8,500 nt, about 6,500 nt to about 9,000 nt, about 6,500 nt to about 9,500 nt, about 6,500 nt to about 10,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,000 nt to about 9,500 nt, about 7,000 nt to about 10,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 7,500 nt to about 9,500 nt, about 7,500 nt to about 10,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,000 nt to about 9,500 nt, about 8,000 nt to about 10,000 nt, about 8,500 nt to about 9,000 nt, about 8,500 nt to about 9,500 nt, about 8,500 nt to about 10,000 nt, about 9,000 nt to about 9,500 nt, about 9,000 nt to about 10,000 nt, about 9,500 nt to about 10,000 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, about 9,000 nt, about 9,500 nt, and about 10,000 nt;

the total target protein coding sequence size is selected from about 2000 nt to about 8000 nt, about 2,000 nt to about 3,000 nt, about 2,000 nt to about 3,500 nt, about 2,000 nt to about 4,000 nt, about 2,000 nt to about 4,500 nt, about 2,000 nt to about 5,000 nt, about 2,000 nt to about 5,500 nt, about 2,000 nt to about 6,000 nt, about 2,000 nt to about 6,500 nt, about 2,000 nt to about 7,000 nt, about 2,000 nt to about 7,500 nt, about 2,000 nt to about 8,000 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,500 nt, about 3,000 nt to about 5,000 nt, about 3,000 nt to about 5,500 nt, about 3,000 nt to about 6,000 nt, about 3,000 nt to about 6,500 nt, about 3,000 nt to about 7,000 nt, about 3,000 nt to about 7,500 nt, about 3,000 nt to about 8,000 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,500 nt, about 3,500 nt to about 5,000 nt, about 3,500 nt to about 5,500 nt, about 3,500 nt to about 6,000 nt, about 3,500 nt to about 6,500 nt, about 3,500 nt to about 7,000 nt, about 3,500 nt to about 7,500 nt, about 3,500 nt to about 8,000 nt, about 4,000 nt to about 4,500 nt, about 4,000 nt to about 5,000 nt, about 4,000 nt to about 5,500 nt, about 4,000 nt to about 6,000 nt, about 4,000 nt to about 6,500 nt, about 4,000 nt to about 7,000 nt, about 4,000 nt to about 7,500 nt, about 4,000 nt to about 8,000 nt, about 4,500 nt to about 5,000 nt, about 4,500 nt to about 5,500 nt, about 4,500 nt to about 6,000 nt, about 4,500 nt to about 6,500 nt, about 4,500 nt to about 7,000 nt, about 4,500 nt to about 7,500 nt, about 4,500 nt to about 8,000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, or about 7,500 nt to about 8,000 nt.

the total target protein coding sequence is about 2,000 nt, about 3,000 nt, about 3,500 nt, about 4,000 nt,

about 4,500 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, and about 8,000 nt; and/or

the summed size of the RNA molecules encoded by the two synthetic DNA molecules is about 5,000 nt to about 9000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 5,500 nt to about 8,500 nt, about 5,500 nt to about 9,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 6,500 nt to about 8,500 nt, about 6,500 nt to about 9,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,500 nt to about 9,000 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, or about 9,000 nt.

39. The system of any one of embodiments 31 to 36, the system comprising a composition of any one of embodiments 26 and 28 to 30, wherein:

the first, second, and third synthetic DNA molecules have a total size of about 7500 nt to about 15,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,500 nt, about 7,500 nt to about 10,000 nt, about 7,500 nt to about 10,500 nt, about 7,500 nt to about 11,000 nt, about 7,500 nt to about 11,500 nt, about 7,500 nt to about 12,000 nt, about 7,500 nt to about 12,500 nt, about 7,500 nt to about 13,000 nt, about 7,500 nt to about 14,000 nt, about 7,500 nt to about 15,000 nt, about 8,500 nt to about 9,500 nt, about 8,500 nt to about 10,000 nt, about 8,500 nt to about 10,500 nt, about 8,500 nt to about 11,000 nt, about 8,500 nt to about 11,500 nt, about 8,500 nt to about 12,000 nt, about 8,500 nt to about 12,500 nt, about 8,500 nt to about 13,000 nt, about 8,500 nt to about 14,000 nt, about 8,500 nt to about 15,000 nt, about 9,500 nt to about 10,000 nt, about 9,500 nt to about 10,500 nt, about 9,500 nt to about 11,000 nt, about 9,500 nt to about 11,500 nt, about 9,500 nt to about 12,000 nt, about 9,500 nt to about 12,500 nt, about 9,500 nt to about 13,000 nt, about 9,500 nt to about 14,000 nt, about 9,500 nt to about 15,000 nt, about 10,000 nt to about 10,500 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 11,500 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 12,500 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,500 nt to about 11,000 nt, about 10,500 nt to about 11,500 nt, about 10,500 nt to about 12,000 nt, about 10,500 nt to about 12,500 nt, about 10,500 nt to about 13,000 nt, about 10,500 nt to about 14,000 nt,

about 10,500 nt to about 15,000 nt, about 11,000 nt to about 11,500 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 12,500 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 14,000 nt, about 11,000 nt to about 15,000 nt, about 11,500 nt to about 12,000 nt, about 11,500 nt to about 12,500 nt, about 11,500 nt to about 13,000 nt, about 11,500 nt to about 14,000 nt, about 11,500 nt to about 15,000 nt, about 12,000 nt to about 12,500 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,500 nt to about 13,000 nt, about 12,500 nt to about 14,000 nt, about 12,500 nt to about 15,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, or about 14,000 nt to about 15,000 nt, about 7,500 nt, about 8,500 nt, about 9,500 nt, about 10,000 nt, about 10,500 nt, about 11,000 nt, about 11,500 nt, about 12,000 nt, about 12,500 nt, about 13,000 nt, about 14,000 nt, or about 15,000 nt;

the total target protein coding sequence about 3000 nt to about 12,000 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 5,000 nt, about 3,000 nt to about 6,000 nt, about 3,000 nt to about 7,000 nt, about 3,000 nt to about 7,500 nt, about 3,000 nt to about 8,000 nt, about 3,000 nt to about 8,500 nt, about 3,000 nt to about 9,000 nt, about 3,000 nt to about 1,000 nt, about 3,000 nt to about 11,000 nt, about 3,000 nt to about 12,000 nt, about 4,000 nt to about 5,000 nt, about 4,000 nt to about 6,000 nt, about 4,000 nt to about 7,000 nt, about 4,000 nt to about 7,500 nt, about 4,000 nt to about 8,000 nt, about 4,000 nt to about 8,500 nt, about 4,000 nt to about 9,000 nt, about 4,000 nt to about 1,000 nt, about 4,000 nt to about 11,000 nt, about 4,000 nt to about 12,000 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,000 nt to about 1,000 nt, about 5,000 nt to about 11,000 nt, about 5,000 nt to about 12,000 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,000 nt to about 1,000 nt, about 6,000 nt to about 11,000 nt, about 6,000 nt to about 12,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,000 nt to about 1,000 nt, about 7,000 nt to about 11,000 nt, about 7,000 nt to about 12,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 7,500 nt to about 1,000 nt, about 7,500 nt to about 11,000 nt, about 7,500 nt to about 12,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,000 nt to about 1,000 nt, about 8,000 nt to about 11,000 nt, about 8,000 nt to about 12,000 nt, about 8,500 nt to about 9,000 nt, about 8,500 nt to about 1,000 nt, about 8,500 nt to about 11,000 nt, about 8,500 nt to about 12,000 nt, about 9,000 nt to about 1,000 nt, about 9,000 nt to about 11,000 nt, about 9,000 nt to about 12,000 nt, about 1,000 nt to about 11,000 nt, about 1,000 nt to about 12,000 nt, about 11,000 nt to about 12,000 nt, about 3,000 nt, about 4,000 nt, about 5,000 nt, about 6,000 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, about 9,000 nt, about 1,000 nt, about 11,000 nt, or about 12,000 nt; and/or

the summed size of the RNA molecules encoded by the three synthetic DNA molecules is about 7500 nt to about 13,500 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 7,500 nt to about 9,500 nt, about 7,500 nt to about 10,000 nt, about 7,500 nt to about 10,500 nt, about 7,500 nt to about 11,000 nt, about 7,500 nt to about 11,500 nt, about 7,500 nt to about 12,000 nt, about 7,500 nt to about 12,500 nt, about 7,500 nt to about 13,000 nt, about 7,500 nt to about 13,500 nt, about 8,500 nt to about 9,000 nt, about 8,500 nt to about 9,500 nt, about 8,500 nt to about 10,000 nt, about 8,500 nt to about 10,500 nt, about 8,500 nt to about 11,000 nt, about 8,500 nt to about 11,500 nt, about 8,500 nt to about 12,000 nt, about 8,500 nt to about 12,500 nt, about 8,500 nt to about 13,000 nt, about 8,500 nt to about 13,500 nt, about 9,000 nt to about 9,500 nt, about 9,000 nt to about 10,000 nt, about 9,000 nt to about 10,500 nt, about 9,000 nt to about 11,000 nt, about 9,000 nt to about 11,500 nt, about 9,000 nt to about 12,000 nt, about 9,000 nt to about 12,500 nt, about 9,000 nt to about 13,000 nt, about 9,000 nt to about 13,500 nt, about 9,500 nt to about 10,000 nt, about 9,500 nt to about 10,500 nt, about 9,500 nt to about 11,000 nt, about 9,500 nt to about 11,500 nt, about 9,500 nt to about 12,000 nt, about 9,500 nt to about 12,500 nt, about 9,500 nt to about 13,000 nt, about 9,500 nt to about 13,500 nt, about 10,000 nt to about 10,500 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 11,500 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 12,500 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 13,500 nt, about 10,500 nt to about 11,000 nt, about 10,500 nt to about 11,500 nt, about 10,500 nt to about 12,000 nt, about 10,500 nt to about 12,500 nt, about 10,500 nt to about 13,000 nt, about 10,500 nt to about 13,500 nt, about 11,000 nt to about 11,500 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 12,500 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 13,500 nt, about 11,500 nt to about 12,000 nt, about 11,500 nt to about 12,500 nt, about 11,500 nt to about 13,000 nt, about 11,500 nt to about 13,500 nt, about 12,000 nt to about 12,500 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 13,500 nt, about 12,500 nt to about 13,000 nt, about 12,500 nt to about 13,500 nt, about 13,000 nt to about 13,500 nt, about 7,500 nt, about 8,500 nt, about 9,000 nt, about 9,500 nt, about 10,000 nt, about 10,500 nt, about 11,000 nt, about 11,500 nt, about 12,000 nt, about 12,500 nt, about 13,000 nt, or about 13,500 nt.

40. The system of any one of embodiments 31 to 36, the system comprising a composition of any one of embodiments 27 and 28 to 30, wherein:

the first, second, third and fourth synthetic DNA molecules have a total size of about 10,000 nt to about 20,000 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,000 nt to about 16,000 nt, about 10,000 nt to about 17,000 nt, about 10,000 nt to about 18,000 nt, about 10,000 nt to about 19,000 nt, about 10,000 nt to about 20,000 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 14,000 nt, about 11,000 nt to about 15,000 nt, about 11,000 nt to about 16,000 nt, about 11,000 nt to about 17,000 nt, about 11,000 nt to

about 18,000 nt, about 11,000 nt to about 19,000 nt, about 11,000 nt to about 20,000 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,000 nt to about 16,000 nt, about 12,000 nt to about 17,000 nt, about 12,000 nt to about 18,000 nt, about 12,000 nt to about 19,000 nt, about 12,000 nt to about 20,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, about 13,000 nt to about 16,000 nt, about 13,000 nt to about 17,000 nt, about 13,000 nt to about 18,000 nt, about 13,000 nt to about 19,000 nt, about 13,000 nt to about 20,000 nt, about 14,000 nt to about 15,000 nt, about 14,000 nt to about 16,000 nt, about 14,000 nt to about 17,000 nt, about 14,000 nt to about 18,000 nt, about 14,000 nt to about 19,000 nt, about 14,000 nt to about 20,000 nt, about 15,000 nt to about 16,000 nt, about 15,000 nt to about 17,000 nt, about 15,000 nt to about 18,000 nt, about 15,000 nt to about 19,000 nt, about 15,000 nt to about 20,000 nt, about 16,000 nt to about 17,000 nt, about 16,000 nt to about 18,000 nt, about 16,000 nt to about 19,000 nt, about 16,000 nt to about 20,000 nt, about 17,000 nt to about 18,000 nt, about 17,000 nt to about 19,000 nt, about 17,000 nt to about 20,000 nt, about 18,000 nt to about 19,000 nt, about 18,000 nt to about 20,000 nt, about 19,000 nt to about 20,000 nt, about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, about 15,000 nt, about 16,000 nt, about 17,000 nt, about 18,000 nt, about 19,000 nt, or about 20,000 nt;

the total target protein coding sequence is about 4000 nt to about 16,000 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 9,000 nt, about 5,000 nt to about 10,000 nt, about 5,000 nt to about 11,000 nt, about 5,000 nt to about 12,000 nt, about 5,000 nt to about 13,000 nt, about 5,000 nt to about 14,000 nt, about 5,000 nt to about 15,000 nt, about 5,000 nt to about 16,000 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 9,000 nt, about 6,000 nt to about 10,000 nt, about 6,000 nt to about 11,000 nt, about 6,000 nt to about 12,000 nt, about 6,000 nt to about 13,000 nt, about 6,000 nt to about 14,000 nt, about 6,000 nt to about 15,000 nt, about 6,000 nt to about 16,000 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 9,000 nt, about 7,000 nt to about 10,000 nt, about 7,000 nt to about 11,000 nt, about 7,000 nt to about 12,000 nt, about 7,000 nt to about 13,000 nt, about 7,000 nt to about 14,000 nt, about 7,000 nt to about 15,000 nt, about 7,000 nt to about 16,000 nt, about 8,000 nt to about 9,000 nt, about 8,000 nt to about 10,000 nt, about 8,000 nt to about 11,000 nt, about 8,000 nt to about 12,000 nt, about 8,000 nt to about 13,000 nt, about 8,000 nt to about 14,000 nt, about 8,000 nt to about 15,000 nt, about 8,000 nt to about 16,000 nt, about 9,000 nt to about 10,000 nt, about 9,000 nt to about 11,000 nt, about 9,000 nt to about 12,000 nt, about 9,000 nt to about 13,000 nt, about 9,000 nt to about 14,000 nt, about 9,000 nt to about 15,000 nt, about 9,000 nt to about 16,000 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,000 nt to about 16,000 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 14,000 nt, about

11,000 nt to about 15,000 nt, about 11,000 nt to about 16,000 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,000 nt to about 16,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, about 13,000 nt to about 16,000 nt, about 14,000 nt to about 15,000 nt, about 14,000 nt to about 16,000 nt, or about 15,000 nt to about 16,000 nt. the total target protein coding sequence is about 5,000 nt, about 6,000 nt, about 7,000 nt, about 8,000 nt, about 9,000 nt, about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, about 15,000 nt, about 16,000 nt about 5,000 nt, about 6,000 nt, about 7,000 nt, about 8,000 nt, about 9,000 nt, about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, or about 15,000 nt; and/or

the summed size of the RNA molecules encoded by the four synthetic DNA molecules is about 10,000 nt to about 18,000 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,000 nt to about 16,000 nt, about 10,000 nt to about 17,000 nt, about 10,000 nt to about 18,000 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 14,000 nt, about 11,000 nt to about 15,000 nt, about 11,000 nt to about 16,000 nt, about 11,000 nt to about 17,000 nt, about 11,000 nt to about 18,000 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,000 nt to about 16,000 nt, about 12,000 nt to about 17,000 nt, about 12,000 nt to about 18,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, about 13,000 nt to about 16,000 nt, about 13,000 nt to about 17,000 nt, about 13,000 nt to about 18,000 nt, about 14,000 nt to about 15,000 nt, about 14,000 nt to about 16,000 nt, about 14,000 nt to about 17,000 nt, about 14,000 nt to about 18,000 nt, about 15,000 nt to about 16,000 nt, about 15,000 nt to about 17,000 nt, about 15,000 nt to about 18,000 nt, about 16,000 nt to about 17,000 nt, about 16,000 nt to about 18,000 nt, about 17,000 nt to about 18,000 nt, about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, about 15,000 nt, about 16,000 nt, about 17,000 nt, or about 18,000 nt.

41. The system of any one of embodiments 31 to 40, wherein the first dimerization domain and the second dimerization domain, the third dimerization domain and the fourth dimerization domain, and/or the fifth dimerization domain and the sixth dimerization domain, are each no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50 to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, or about 100%.



42. The system of any one of embodiments 31 to 41, wherein each dimerization domain is no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50 to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, at least 90%, or about 100%.

43. The system of any one of embodiments 31 to 42, wherein the RNA recombination efficiency is about 10% to about 100%, about 10% to about 20%, about 10% to about 30%, about 10% to about 35%, about 10% to about 40%, about 10% to about 45%, about 10% to about 50%, about 10% to about 55%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 20% to about 30%, about 20% to about 35%, about 20% to about 40%, about 20% to about 45%, about 20% to about 50%, about 20% to about 55%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 30% to about 35%, about 30% to about 40%, about 30% to about 45%, about 30% to about 50%, about 30% to about 55%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 35% to about 40%, about 35% to about 45%, about 35% to about 50%, about 35% to about 55%, about 35% to about 60%, about 35% to about 70%, about 35% to about 80%, about 35% to about 90%, about 40% to about 45%, about 40% to about 50%, about 40% to about 55%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about 40% to about 90%, about 45% to about 50%, about 45% to about 55%, about 45% to about 60%, about 45% to about 70%, about 45% to about 80%, about 45% to about 90%, about 50% to about 55%, about 50% to about 60%, about 50% to about 70%, about 50% to about 80%, about 50% to about 90%, about 55% to about 60%, about 55% to about 70%, about 55% to about 80%, about 55% to about 90%, about 60% to about 70%, about 60% to about 80%, about 60% to about 90%, about 70% to about 80%, about 70% to about 90%, about 80% to about 90%, about 10%, about 20%, about 30%, about 35%, about 40%, about 45%, about 50%, about 55%, about 60%, about 70%, about 80%, about 90%, about 95%, or about 100%.

44. A composition comprising a system of any one of embodiments 31 to 43.

45. The composition of embodiment 44, wherein the composition comprises first, second, third and optionally fourth RNA molecules, each encoding at least a portion of dystrophin, factor 8, ABCA4, or MYO7A.

46. A kit comprising the system of any one of embodiments 31 to 43, or composition of any one of embodiments 44 and 45, wherein any of the synthetic first, second, third and fourth nucleic acid molecules can be in separate containers, and optionally further comprising a buffer such as a pharmaceutically acceptable carrier.

47. A method of expressing a target protein in a cell, comprising:

introducing the system of any one of embodiments 31 to 43, or composition of any one of embodiments 44 and 45, into a cell, and expressing the first and second, first, second, and third, or first, second, third and fourth RNA molecules in the cell, wherein the target protein is produced in the cell.

48. The method of embodiment 47, wherein the cell is in a subject, and introducing comprises administering a therapeutically effective amount the system to the subject.

49. The method of embodiment 48, wherein the method treats a genetic disease caused by a mutation in a gene encoding the target protein in the subject, wherein the method results in expression of functional target protein in the subject.

50. The method of embodiment 49, wherein

the genetic disease is Duchenne muscular dystrophy and the target protein is dystrophin;

the genetic disease is Hemophilia A and the target protein is F8;

the genetic disease is Stargardt disease and the target protein is ABCA4; or

the genetic disease is Usher syndrome and the target protein is MYO7A.

51. A system of any one of embodiments 31 to 43, a composition of any one of embodiments 1 to 24, 44 and 45, a kit of embodiment 46, or a method of any one of embodiments 47 to 50, wherein one, two, three, or four RNA molecules comprise at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, at least 99%, or 100% sequence identity to a synthetic intron provided in any one of SEQ ID NOS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 145, 146, 147, 148, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, and 166.

52. A system of any one of embodiments 31 to 43, and 51, a composition of any one of embodiments 1-24, 44 and 45, a kit of embodiment 46, or a method of any one of embodiments 47 to 50, wherein one, two, three, or four RNA molecules comprise a synthetic intron selected from: nt 3703 to 3975 of SEQ ID NO: 20, nt 1 to 228 of SEQ ID NO: 21, nt 3703 to 3975 of SEQ ID NO: 22, nt 1 to 225 of SEQ ID NO: 23, nt 3560 to 3828 of SEQ ID NO: 24, and nt 1-225 of SEQ ID NO: 25.

53. A system of any of embodiments 31 to 43, 51, and 52, a composition of any one of embodiments 1 to 24, 44 and 45, or a method of any one of embodiments 47 to 50, wherein the one, two, three, or four RNA molecules further comprise a portion of a protein coding sequence.

54. A system of any of embodiments 31 to 43, and 51 to 53, a composition of any one of embodiments 1 to 24, 44 and 45, or a method of any one of embodiments 47 to 50, wherein the portion of the protein coding sequence comprises an N-terminal half, an N-terminal third, a middle portion, a first middle portion, a C-terminal half, or a C-terminal third of the protein coding sequence.

55. A system of any one of embodiments 31 to 43 and 51 to 54, a composition of any one of embodiments 1 to 24, 44 and 45, or a method of any one of embodiments 47 to 50, comprising: (a) a first RNA molecule, the RNA molecule comprising from 5' to 3': (i) a coding sequence for an N-terminal portion of the target protein; (ii) a splice donor; (ii-2) a DISE, an ISE, or both; and (iii) a first

dimerization domain; and (b) a second RNA molecule, the RNA molecule comprising from 5' to 3': (i) a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; (i-2) at least one ISE sequences; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice acceptor; and (v) a coding sequence for a C-terminal portion of the target protein.

5 56. A system of any one of embodiments 31 to 43 and 51 to 55, a composition of any one of embodiments 1 to 24, 44 and 45, or a method of any one of embodiments 47 to 50, comprising: (a) a first RNA molecule, the RNA molecule comprising from 5' to 3': (i) a coding sequence for an N-terminal portion of the target protein; (ii) a splice donor; (ii-2) a DISE, an ISE, and an ISE; and (iii) a first dimerization domain; and (b) a second RNA molecule, the RNA molecule comprising from 5' to  
10 3': (i) a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; (i-2) three ISE sequences; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice acceptor; and (v) a coding sequence for a C-terminal portion of the target protein.

57. The composition of any one of embodiments 1 and 4-24, wherein any one or two of the two RNA molecules, or the composition of any one of embodiments 2 and 4-24, wherein any one, two or  
15 three of the three RNA molecules, or the composition of any one of embodiments 3 and 4-24, wherein any one, two, three, or four of the four RNA molecules, each has a size independently selected from: about 2500 to 4500 nt, about 2,500 nt to about 2,750 nt, about 2,500 nt to about 3,000 nt, about 2,500 nt to about 3,250 nt, about 2,500 nt to about 3,500 nt, about 2,500 nt to about 3,750 nt, about 2,500 nt to about 4,000 nt, about 2,500 nt to about 4,250 nt, about 2,500 nt to about 4,500 nt, about 2,750 nt to  
20 about 3,000 nt, about 2,750 nt to about 3,250 nt, about 2,750 nt to about 3,500 nt, about 2,750 nt to about 3,750 nt, about 2,750 nt to about 4,000 nt, about 2,750 nt to about 4,250 nt, about 2,750 nt to about 4,500 nt, about 3,000 nt to about 3,250 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 3,750 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,250 nt, about 3,000 nt to about 4,500 nt, about 3,250 nt to about 3,500 nt, about 3,250 nt to about 3,750 nt, about 3,250 nt to  
25 about 4,000 nt, about 3,250 nt to about 4,250 nt, about 3,250 nt to about 4,500 nt, about 3,500 nt to about 3,750 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,250 nt, about 3,500 nt to about 4,500 nt, about 3,750 nt to about 4,000 nt, about 3,750 nt to about 4,250 nt, about 3,750 nt to about 4,500 nt, about 4,000 nt to about 4,250 nt, about 4,000 nt to about 4,500 nt, about 4,250 nt to about 4,500 nt, about 2,500 nt, about 2,750 nt, about 3,000 nt, about 3,250 nt, about 3,500 nt, about  
30 3,750 nt, about 4,000 nt, about 4,250 nt, and about 4,500 nt.

58. The composition of any one of embodiments 1 and 4-24, wherein:  
the total target protein coding sequence is about 2000 nt to about 8000 nt, about 2,000 nt to about 3,000 nt, about 2,000 nt to about 3,500 nt, about 2,000 nt to about 4,000 nt, about 2,000 nt to about 4,500 nt, about 2,000 nt to about 5,000 nt, about 2,000 nt to about 5,500 nt, about 2,000 nt to  
35 about 6,000 nt, about 2,000 nt to about 6,500 nt, about 2,000 nt to about 7,000 nt, about 2,000 nt to

about 7,500 nt, about 2,000 nt to about 8,000 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,500 nt, about 3,000 nt to about 5,000 nt, about 3,000 nt to about 5,500 nt, about 3,000 nt to about 6,000 nt, about 3,000 nt to about 6,500 nt, about 3,000 nt to about 7,000 nt, about 3,000 nt to about 7,500 nt, about 3,000 nt to about 8,000 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,500 nt, about 3,500 nt to about 5,000 nt, about 3,500 nt to about 5,500 nt, about 3,500 nt to about 6,000 nt, about 3,500 nt to about 6,500 nt, about 3,500 nt to about 7,000 nt, about 3,500 nt to about 7,500 nt, about 3,500 nt to about 8,000 nt, about 4,000 nt to about 4,500 nt, about 4,000 nt to about 5,000 nt, about 4,000 nt to about 5,500 nt, about 4,000 nt to about 6,000 nt, about 4,000 nt to about 6,500 nt, about 4,000 nt to about 7,000 nt, about 4,000 nt to about 7,500 nt, about 4,000 nt to about 8,000 nt, about 4,500 nt to about 5,000 nt, about 4,500 nt to about 5,500 nt, about 4,500 nt to about 6,000 nt, about 4,500 nt to about 6,500 nt, about 4,500 nt to about 7,000 nt, about 4,500 nt to about 7,500 nt, about 4,500 nt to about 8,000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, or about 7,500 nt to about 8,000 nt. the total target protein coding sequence is about 2,000 nt, about 3,000 nt, about 3,500 nt, about 4,000 nt, about 4,500 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, or about 8,000 nt; and/or

the summed size of the two RNA molecules is about 5,000 nt to about 9000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 5,500 nt to about 8,500 nt, about 5,500 nt to about 9,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 6,500 nt to about 8,500 nt, about 6,500 nt to about 9,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,500 nt to about 9,000 nt, about 5,000 nt, about

5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, or about 9,000 nt.

59. The composition of any one of embodiments 2 and 4-24, wherein:

the total target protein coding sequence size is about 3000 nt to about 12,000 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 5,000 nt, about 3,000 nt to about 6,000 nt, about 3,000 nt to about 7,000 nt, about 3,000 nt to about 7,500 nt, about 3,000 nt to about 8,000 nt, about 3,000 nt to about 8,500 nt, about 3,000 nt to about 9,000 nt, about 3,000 nt to about 1,000 nt, about 3,000 nt to about 11,000 nt, about 3,000 nt to about 12,000 nt, about 4,000 nt to about 5,000 nt, about 4,000 nt to about 6,000 nt, about 4,000 nt to about 7,000 nt, about 4,000 nt to about 7,500 nt, about 4,000 nt to about 8,000 nt, about 4,000 nt to about 8,500 nt, about 4,000 nt to about 9,000 nt, about 4,000 nt to about 1,000 nt, about 4,000 nt to about 11,000 nt, about 4,000 nt to about 12,000 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,000 nt to about 1,000 nt, about 5,000 nt to about 11,000 nt, about 5,000 nt to about 12,000 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,000 nt to about 1,000 nt, about 6,000 nt to about 11,000 nt, about 6,000 nt to about 12,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,000 nt to about 1,000 nt, about 7,000 nt to about 11,000 nt, about 7,000 nt to about 12,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 7,500 nt to about 1,000 nt, about 7,500 nt to about 11,000 nt, about 7,500 nt to about 12,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,000 nt to about 1,000 nt, about 8,000 nt to about 11,000 nt, about 8,000 nt to about 12,000 nt, about 8,500 nt to about 9,000 nt, about 8,500 nt to about 1,000 nt, about 8,500 nt to about 11,000 nt, about 8,500 nt to about 12,000 nt, about 9,000 nt to about 1,000 nt, about 9,000 nt to about 11,000 nt, about 9,000 nt to about 12,000 nt, about 1,000 nt to about 11,000 nt, about 1,000 nt to about 12,000 nt, about 11,000 nt to about 12,000 nt, about 3,000 nt, about 4,000 nt, about 5,000 nt, about 6,000 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, about 9,000 nt, about 1,000 nt, about 11,000 nt, or about 12,000 nt; and/or

the summed size of the three RNA molecules is about 7500 nt to about 13,500 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 7,500 nt to about 9,500 nt, about 7,500 nt to about 10,000 nt, about 7,500 nt to about 10,500 nt, about 7,500 nt to about 11,000 nt, about 7,500 nt to about 11,500 nt, about 7,500 nt to about 12,000 nt, about 7,500 nt to about 12,500 nt, about 7,500 nt to about 13,000 nt, about 7,500 nt to about 13,500 nt, about 8,500 nt to about 9,000 nt, about 8,500 nt to about 9,500 nt, about 8,500 nt to about 10,000 nt, about 8,500 nt to about 10,500 nt, about 8,500 nt to about 11,000 nt, about 8,500 nt to about 11,500 nt, about 8,500 nt to about 12,000 nt, about 8,500 nt to

about 12,500 nt, about 8,500 nt to about 13,000 nt, about 8,500 nt to about 13,500 nt, about 9,000 nt to about 9,500 nt, about 9,000 nt to about 10,000 nt, about 9,000 nt to about 10,500 nt, about 9,000 nt to about 11,000 nt, about 9,000 nt to about 11,500 nt, about 9,000 nt to about 12,000 nt, about 9,000 nt to about 12,500 nt, about 9,000 nt to about 13,000 nt, about 9,000 nt to about 13,500 nt, about 9,500 nt to about 10,000 nt, about 9,500 nt to about 10,500 nt, about 9,500 nt to about 11,000 nt, about 9,500 nt to about 11,500 nt, about 9,500 nt to about 12,000 nt, about 9,500 nt to about 12,500 nt, about 9,500 nt to about 13,000 nt, about 9,500 nt to about 13,500 nt, about 10,000 nt to about 10,500 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 11,500 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 12,500 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 13,500 nt, about 10,500 nt to about 11,000 nt, about 10,500 nt to about 11,500 nt, about 10,500 nt to about 12,000 nt, about 10,500 nt to about 12,500 nt, about 10,500 nt to about 13,000 nt, about 10,500 nt to about 13,500 nt, about 11,000 nt to about 11,500 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 12,500 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 13,500 nt, about 11,500 nt to about 12,000 nt, about 11,500 nt to about 12,500 nt, about 11,500 nt to about 13,000 nt, about 11,500 nt to about 13,500 nt, about 12,000 nt to about 12,500 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 13,500 nt, about 12,500 nt to about 13,000 nt, about 12,500 nt to about 13,500 nt, about 13,000 nt to about 13,500 nt about 7,500 nt, about 8,500 nt, about 9,000 nt, about 9,500 nt, about 10,000 nt, about 10,500 nt, about 11,000 nt, about 11,500 nt, about 12,000 nt, about 12,500 nt, about 13,000 nt, or about 13,500 nt.

60. The composition of any one of embodiments 3 and 4-24, wherein:

the total target protein coding sequence size is about 4000 nt to about 16,000 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 9,000 nt, about 5,000 nt to about 10,000 nt, about 5,000 nt to about 11,000 nt, about 5,000 nt to about 12,000 nt, about 5,000 nt to about 13,000 nt, about 5,000 nt to about 14,000 nt, about 5,000 nt to about 15,000 nt, about 5,000 nt to about 16,000 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 9,000 nt, about 6,000 nt to about 10,000 nt, about 6,000 nt to about 11,000 nt, about 6,000 nt to about 12,000 nt, about 6,000 nt to about 13,000 nt, about 6,000 nt to about 14,000 nt, about 6,000 nt to about 15,000 nt, about 6,000 nt to about 16,000 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 9,000 nt, about 7,000 nt to about 10,000 nt, about 7,000 nt to about 11,000 nt, about 7,000 nt to about 12,000 nt, about 7,000 nt to about 13,000 nt, about 7,000 nt to about 14,000 nt, about 7,000 nt to about 15,000 nt, about 7,000 nt to about 16,000 nt, about 8,000 nt to about 9,000 nt, about 8,000 nt to about 10,000 nt, about 8,000 nt to about 11,000 nt, about 8,000 nt to about 12,000 nt, about 8,000 nt to about 13,000 nt, about 8,000 nt to about 14,000 nt, about 8,000 nt to about 15,000 nt, about 8,000 nt to about 16,000 nt, about 9,000 nt to about 10,000 nt, about 9,000 nt to about 11,000 nt, about 9,000 nt to about 12,000 nt, about 9,000 nt to about 13,000 nt, about 9,000 nt to

about 14,000 nt, about 9,000 nt to about 15,000 nt, about 9,000 nt to about 16,000 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,000 nt to about 16,000 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 14,000 nt, about 11,000 nt to about 15,000 nt, about 11,000 nt to about 16,000 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,000 nt to about 16,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, about 13,000 nt to about 16,000 nt, about 14,000 nt to about 15,000 nt, about 14,000 nt to about 16,000 nt, or about 15,000 nt to about 16,000 nt. the total target protein coding sequence is about 5,000 nt, about 6,000 nt, about 7,000 nt, about 8,000 nt, about 9,000 nt, about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, about 15,000 nt, about 16,000 nt about 5,000 nt, about 6,000 nt, about 7,000 nt, about 8,000 nt, about 9,000 nt, about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, or about 15,000 nt; and/or

the summed size of the RNA molecules encoded by the four synthetic DNA molecules is about 10,000 nt to about 18,000 nt, about 10,000 nt to about 11,000 nt, about 10,000 nt to about 12,000 nt, about 10,000 nt to about 13,000 nt, about 10,000 nt to about 14,000 nt, about 10,000 nt to about 15,000 nt, about 10,000 nt to about 16,000 nt, about 10,000 nt to about 17,000 nt, about 10,000 nt to about 18,000 nt, about 11,000 nt to about 12,000 nt, about 11,000 nt to about 13,000 nt, about 11,000 nt to about 14,000 nt, about 11,000 nt to about 15,000 nt, about 11,000 nt to about 16,000 nt, about 11,000 nt to about 17,000 nt, about 11,000 nt to about 18,000 nt, about 12,000 nt to about 13,000 nt, about 12,000 nt to about 14,000 nt, about 12,000 nt to about 15,000 nt, about 12,000 nt to about 16,000 nt, about 12,000 nt to about 17,000 nt, about 12,000 nt to about 18,000 nt, about 13,000 nt to about 14,000 nt, about 13,000 nt to about 15,000 nt, about 13,000 nt to about 16,000 nt, about 13,000 nt to about 17,000 nt, about 13,000 nt to about 18,000 nt, about 14,000 nt to about 15,000 nt, about 14,000 nt to about 16,000 nt, about 14,000 nt to about 17,000 nt, about 14,000 nt to about 18,000 nt, about 15,000 nt to about 16,000 nt, about 15,000 nt to about 17,000 nt, about 15,000 nt to about 18,000 nt, about 16,000 nt to about 17,000 nt, about 16,000 nt to about 18,000 nt, about 17,000 nt to about 18,000 nt, about 10,000 nt, about 11,000 nt, about 12,000 nt, about 13,000 nt, about 14,000 nt, about 15,000 nt, about 16,000 nt, about 17,000 nt, or about 18,000 nt.

61. The system of any one of embodiments 1 to 24 and 57 to 60, wherein the first dimerization domain and the second dimerization domain, the third dimerization domain and the fourth dimerization domain, and/or the fifth dimerization domain and the sixth dimerization domain, are each no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50 to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least about 20%, at least about 25%, at least about

30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, or at least about 95%.

62. The system of any one of embodiments 1 to 24 and 57 to 61, wherein each dimerization domain is no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50 to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, or at least 90%.

63. The composition of any one of embodiments 1 to 24 and 57 to 62, wherein the RNA recombination efficiency is about 10% to about 100%, about 10% to about 20%, about 10% to about 30%, about 10% to about 35%, about 10% to about 40%, about 10% to about 45%, about 10% to about 50%, about 10% to about 55%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 20% to about 30%, about 20% to about 35%, about 20% to about 40%, about 20% to about 45%, about 20% to about 50%, about 20% to about 55%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 30% to about 35%, about 30% to about 40%, about 30% to about 45%, about 30% to about 50%, about 30% to about 55%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 35% to about 40%, about 35% to about 45%, about 35% to about 50%, about 35% to about 55%, about 35% to about 60%, about 35% to about 70%, about 35% to about 80%, about 35% to about 90%, about 40% to about 45%, about 40% to about 50%, about 40% to about 55%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about 40% to about 90%, about 45% to about 50%, about 45% to about 55%, about 45% to about 60%, about 45% to about 70%, about 45% to about 80%, about 45% to about 90%, about 50% to about 55%, about 50% to about 60%, about 50% to about 70%, about 50% to about 80%, about 50% to about 90%, about 55% to about 60%, about 55% to about 70%, about 55% to about 80%, about 55% to about 90%, about 60% to about 70%, about 60% to about 80%, about 60% to about 90%, about 70% to about 80%, about 70% to about 90%, about 80% to about 90%, about 10%, about 20%, about 30%, about 35%, about 40%, about 45%, about 50%, about 55%, about 60%, about 70%, about 80%, about 90%, about 95%, or about 100%.

64. The composition of any one of embodiments 25 to 30 and 44 to 45, the system of any one of embodiments 31 to 43, or the method of any one of embodiments 47-50, wherein the synthetic DNA is produced by transcription of an RNA virus genome by reverse transcriptase.

### Example 1

#### Synthetic RNA Dimerization and Recombination Domains



FIG. 1A depicts a schematic of vector designs (left) and RNA interactions and splicing (right). Left: 5' trans-splice (trsp) DNA vector: Open arrows are two opposing promoters. RFP coding domain and 3'UTR with poly adenylation elements are expressed opposite from the N-terminal portion of YFP (n-yfp), followed by a splice donor sequence (SD), a downstream intronic splicing enhancer (DISE), and two intronic splicing enhancers (2xISE), a binding domain (BD, also referred to as dimerization domain), and a stable stem loop BoxB element (boxB), a self-cleaving hammerhead ribozyme (HHRz), ending with a 3' UTR containing poly adenylation elements. The n-yfp segment has a small intron inserted (white segment within n-yfp). 3' trsp DNA vector: Open arrows are two opposing promoters. BFP coding domain and 3'UTR with poly adenylation elements are expressed opposite from complementary binding domain (anti-BD, also referred to as dimerization domain), followed by three intronic splicing enhancer sequences (3xISE), a branch point (BP), a polypyrimidine tract (PPT), a splice acceptor sequence (SA), the c-terminal portion of the YFP coding sequence, ending with a 3' UTR containing poly adenylation elements. Right: pre-mRNA interactions (5' trsp-RNA + 3' trsp-RNA) and trans-splicing to generate an mRNA encoding YFP protein are shown.

FIG. 1B depicts transfection of only the N-terminal expression plasmid does not lead to YFP fluorescence. Flow cytometry displaying 20k RFP+ cells.

FIG. 1C depicts transfection of only the C-terminal expression plasmid does not lead to YFP fluorescence. Flow cytometry displaying 20k BFP+ cells.

FIG. 1D depicts expression of N-terminal and C-terminal fragments without binding domains shows low levels of YFP induction. Flow cytometry displaying red and green fluorescence values for 20k BFP+ cells.

FIG. 1E depicts rationally designed dimerization/binding domain in a looped configuration. Segments of hypodiverse exclusively pyrimidine or exclusively purine containing sequences are interspaced with stable stem sequences. RNA folding predictions shows 6 stretches of open sequence (numbered 1-6) available for base pairing between the binding domain and its complementary sequence.

FIG. 1F depicts 3D rendering of the "looped" dimerization domain configuration showing the 6 stretches of open sequence (numbered 1-6).

FIG. 1G depicts negative control with no binding domain on the C-terminal half. Flow cytometry displaying red and green fluorescence values for 20k BFP+ cells.

FIG. 1H depicts negative control with no binding domain on the N-terminal half. Flow cytometry displaying red and green fluorescence values for 20k BFP+ cells.

FIG. 1I depicts matching binding domains on both N- and C-terminal half shows strong YFP induction in 90% of the cells. Flow cytometry displaying red and green fluorescence values for 20k BFP+ cells.

FIGS. 1J-1N depict data equivalent to that in FIGS. 1E-1I for a configuration of a binding domain with a stretch of 150 hypodiverse exclusively pyrimidine or exclusively purine containing sequence resulting in a fully open configuration.

FIG. 1O depicts representative fluorescence images for cells shown in FIG. 1G.

FIG. 1P depicts representative fluorescence images for cells shown in FIG. 1L.

FIG. 1Q depicts a comparison of conditions shown in FIG. 1D, FIGS. 1G-1I, and FIGS. 1L-1N. YFP induction coefficient is calculated:  $(\#R+Y+ \div \#R+Y-) \times 100 \times \text{med. Y-fluor}(R+Y+)$ . For comparison the recombination efficiency of a native intron (intron I of the mouse parvalbumin gene) on the N-terminus and an optimized binding domain for that intron on the C-terminal fragment are shown (white bar). This illustrates the benefits of the optimized synthetic RNA dimerization and recombination domains.

## Example 2

### Reconstitution of Protein from Three Synthetic Fragments

FIG. 2A depicts an exemplary schematic of vector designs. The protein coding sequence of a YFP is split into an N-terminal fragment, a middle fragment (m-yfp) and a C-terminal fragment. The junction of the n and m fragments is joined by a looped design binding domain (BD1) and the junction between m and c fragments is joined by a looped binding domain (BD2). The pyrimidine (Y) and purine (R) sequences are arranged to avoid self-circularization of the m-fragment and avoid direct recombination of the N- and C-fragment. The N-terminal fragment is co-expressed with red fluorescent protein as a transfection control, the C-terminal fragment is coexpressed with blue fluorescent protein as a transfection control.

FIG. 2B depicts matching binding domains on all three fragments shows strong YFP induction in 80% of the cells. Flow cytometry displaying red and green fluorescence values for 20k BFP+ cells.

FIG. 2C depicts representative fluorescent image of expression of the n and m fragment only shows no YFP fluorescence (negative control).

FIG. 2D depicts representative fluorescent image of expression of the m and c fragment only shows no YFP fluorescence (negative control).

FIG. 2E depicts representative fluorescent image showing that strong YFP fluorescence is induced by co-transfection of all three fragments.

## Example 3

### In vivo Delivery of Reconstituted full-length YFP Divided into Two Portions

Reconstitution of a YFP coding sequence from two fragments is achieved by using two synthetic RNA sequences, wherein one included the n-terminal coding half fragment of YFP, and one

included the c-terminal coding half fragment (FIG. 3A) (SEQ ID NOS 1 and 2). Each fragment was expressed from AAV2/8 after systemic (iv) administration in newborn (P3) mouse pups. A total of 1.88E11 viral genomes for each of the two fragments were administered per mouse. Expression of YFP was detected 3 weeks later in the liver, heart muscle, and skeletal muscle using fluorescence microscopy.

As shown in FIG. 3B, expression of full-length YFP was detected in the liver of the juvenile mouse, while uninjected liver showed no YFP expression.

As shown in FIG. 3C, expression of full-length YFP was detected in the heart muscle of the juvenile mouse, while uninjected heart muscle showed no YFP expression.

As shown in FIG. 3D, expression of full-length YFP was detected in the skeletal muscles of the leg, while uninjected liver showed no YFP expression.

Thus, the disclosed systems can be used to express full-length proteins *in vivo*, from two or more separate synthetic RNA molecules.

#### Example 4

##### **In vivo Delivery of Reconstituted full-length YFP Divided into Three Portions**

Reconstitution of a YFP coding sequence from three fragments is achieved by using three synthetic RNA sequences, wherein one included the n-terminal fragment of YFP, one included a middle fragment of YFP, and one included the c-terminal fragment (FIG. 4A) (SEQ ID NOS: 145, 146 and 2 respectively).

Each fragment was expressed from AAV2/8 after intramuscular injection into the tibialis anterior muscle of newborn (P3) mouse pups. A total of 1E11 viral genomes for each of the fragments was administered intramuscularly. Expression of YFP was detected 3 weeks later in the skeletal muscle using fluorescence microscopy.

As shown in FIG. 4B, expression of full-length YFP fluorescence was observed in the tibialis anterior muscle.

Thus, the disclosed systems can be used to express full-length proteins *in vivo*, from three or more separate synthetic RNA molecules.

#### Example 5

##### **in vivo Delivery of Reconstituted Full-Length Protein**

To demonstrate the feasibility of a three-part sRdR system *in vivo*, a combination of either two or three AAV-transfer plasmids (the DNA precursor plasmids of AAV) containing fragments of the YFP were transcutaneously electroporated into the tibialis anterior (TA) hindlimb muscle of adult mice.

Efficient reconstitution of both the two part split-YFP system as well as the three part split-YFP system was observed five days after intramuscular electroporation (FIGS. 5A-5F).

FIGS. 5A-5F depict efficient reconstitution of YFP from two and from three fragments in adult mouse tibialis anterior muscle. FIG. 5A depicts N-terminal and C-terminal halves of YFP coding sequences are equipped with synthetic RNA-dimerization and recombination domains. FIG. 5B depicts two AAV transfer plasmids expressing these two fragments were electroporated transcutaneously into adult mouse tibialis anterior (TA) muscle and strong fluorescence was detected at 5 days post electroporation. FIG. 5C shows no fluorescence was detectable in contralateral non-injected TA. FIG. 5D depicts N-terminal, middle, and C-terminal YFP coding sequence are equipped with synthetic RNA-dimerization and recombination domains linking each fragment to its adjacent fragment(s). FIG. 5E depicts transcutaneous electroporation of three AAV transfer plasmids expressing these three fragments. Strong YFP fluorescence is detected indicating efficient reconstitution of YFP from three fragments. FIG. 5F depicts fluorescence in contralateral non-injected TA. Fluorescent channel is overlaid onto grey scale photographs for context.

Two or three vectors were used to successfully express YFP in liver, cardiac muscle and skeletal muscle (two AAV vectors), and in skeletal muscle (three AAV vectors).

Hence the synthetic RNA-dimerization and recombination system provided herein can be deployed in the muscle. Based on these results, one can substitute the YFP coding sequence with a dystrophin (or other gene) coding sequence to achieve therapeutic full-length dystrophin (or other gene) expression from AAVs into a desired subject and/or tissue.

### Example 6

#### Delivery of Reconstituted full-length Dystrophin to treat DMD

An effective gene therapy using full-length dystrophin for patients who suffer from Duchenne muscular dystrophy (DMD) has remained challenging, because the coding sequence of this large protein exceeds the capacity of most viral vectors. Adeno-associated viruses (AAVs) are a common and the preferred method of gene delivery in gene replacement therapy. AAVs are non-toxic, well tolerated, and lead to long term expression of the replacement gene without random integration into the genome. However, the dystrophin gene is too large to be delivered by a single virus. If broken down into fragments, full-length dystrophin can only be delivered using a minimum of three viruses. Smaller versions of dystrophin called “micro-Dystrophin” or “mini-Dystrophin” are currently being tested for dystrophin gene replacement therapy, but these truncated versions of dystrophin are not expected to have full functionality as they are missing key domains in the rod and hinge section of the protein. To date, past attempts to overcome this limitation have not yielded the efficiency required for treating DMD.

Provided herein is a novel technology that can be used to efficiently reconstitute the coding sequence of large genes, including dystrophin, from multiple serial fragments. Using this technology in combination with AAV as a delivery vector, full-length dystrophin will be expressed in a murine model (as well as pig and canine models) for DMD. In one example the subject is a human adult, juvenile, or infant with DMD. For example, the disclosed methods and systems can be used to deliver synthetic RNA-dimerization and recombination domains encoding full-length dystrophin over two or three AAVs (*e.g.*, each AAV delivering a half or a third of the full-length coding sequence). In one example, the AAVs are myotropic AAVs (*e.g.*, those that preferentially infect muscles). This approach can be used to ameliorate or prevent the onset of dystrophy symptoms in a mouse or canine model for DMD, as well as human subjects.

Part 1: Construct efficiently reconstituted three-way split dystrophin expression cassettes. Three expression cassettes are constructed that efficiently reconstitute the full-length dystrophin coding sequence *in vitro* while each individual expression cassette is within the packaging limit of conventional AAV vectors. To achieve therapeutically effective levels of dystrophin, the expression system can be optimized to achieve roughly physiological levels of dystrophin or moderately supraphysiological levels. Up to 50-fold overexpression of dystrophin is tolerated without adverse effects. The dystrophin coding sequence can be split at a number of different points along its length. Efficiency of reconstitution, however, is affected by the local RNA microenvironment and maximization of reconstitution efficiency is done empirically by comparing efficiency of several possible split points. The natural dystrophin coding sequence can be codon optimized for optimal expression and modified to accommodate maximal reconstitution efficiency. It is expected that the full-length dystrophin coding sequence can be reconstituted from a three-way split precursor using the synthetic RNA-dimerization and recombination approach herein disclosed. In screening different configurations, the set of three expression cassettes that lead to the most efficient reconstitution of dystrophin (*e.g.*, approximately physiological or moderately supraphysiological levels) are selected. Experiments can be performed in HEK293T or Human Skeletal Muscle Cells (HSkMC, either primary or trans-differentiated). Using endogenous vs. exogenous specific quantitative RT-PCR probes, and by epitope tag detection in the exogenous dystrophin protein and Western blot analysis, reconstitution efficiencies will be determined different configurations of the split/reconstituted dystrophin.

Part 2: Maximize full-length dystrophin expression over non-reconstituted fragments. Suppression of fragmented background expression of non-reconstituted dystrophin can be achieved by modification of the synthetic RNA-dimerization and recombination domains. Non-reconstituted fragment expression caused by inefficiencies in RNA-recombination may lead to background

expression of dystrophin fragments. Further, suppression of this fragmented background expression may be achieved by modification of the synthetic RNA-dimerization and recombination domains. With the disclosed approach, each fragment of dystrophin is transcribed separately. Reconstitution occurs on the RNA level. Each individual fragment can therefore potentially be translated without being reconstituted. In a western blot, with full-length dystrophin running at roughly 430kDa, these fragments would run at sizes of about  $\frac{2}{3}$  (~290kDa) and  $\frac{1}{3}$  (~140kDa) of that. The synthetic RNA-dimerization and recombination domains can be optimized to avoid non-reconstituted fragment expression and favor full length expression of dystrophin. This can for example be achieved by strategically placing degon sequences, disrupting RNA nuclear export of non-recombined fragments, and introducing decoy translation initiation points. Experiments are carried out in HEK293T and HSkMC. The dystrophin coding sequence can be bookended with epitope tags that allow for identification and quantification of not fully reconstituted fragments of dystrophin using western blot analysis. Cellular distribution of these dystrophin fragments will be assessed using immunohistochemistry in skeletal human muscle cells. Additionally, quantitative assessment of fragment suppression will be done using conventional molecular biology techniques, including quantitative RT PCR across the recombination junctions will be used to determine how efficient the reconstitution on an RNA level occurs. It is expected that low levels of fragmented dystrophin expression will be observed. By modifying the synthetic RNA-dimerization and recombination domains, these fragments can be suppressed.

Part 3. Create high-titer AAV stocks of full-length dystrophin modules for *in vitro* and *in vivo* expression. Dystrophin expressing AAVs will be produced with high purity and viral genome counts higher than 3E13 GC/ml. Three myotropic AAV serotypes will be produced: AAV2/8, AAV2/9, and AAV2/rh10. A tripartite split fluorescent protein, a tripartite split of a full-length dystrophin bookended with epitope tags (see Part 2 above), and a non-tagged tripartite split of full-length dystrophin will be produced, resulting in 27 high-titer AAV preparations. Systemic delivery of therapeutic AAV particles requires high concentration large virus preparations. To achieve reconstituted expression of dystrophin from three separate viruses, repeated administration of the virus may be performed. AAV production in HEK293T cells. Iodixanol or CsCl purification. All batches will be tested *in vitro* in HEK293T and human skeletal muscle cells. As outlined in Part 1 and 2, reconstitution efficiency and unwanted fragment expression will be assessed.

Part 4. Measure expression/reconstitution levels of FLD-AAV modules *in vivo* and tissue distribution *in vivo* of full-length dystrophin expressing AAV modules. The same are assessed for a tripartite split fluorescent protein, as surrogate indicator. For *in vivo* delivery, direct intramuscular

(cardiac and skeletal muscles) and systemic intravenous delivery in newborn and juvenile mice will be compared. Direct muscle injection of FLD-AAV may result in efficient expression of full-length dystrophin as indicated in the Examples above. Systemic delivery of FLD-AAV will be examined using immunohistochemistry and western blot analysis. Different routes of administration, including direct intramuscular and systemic intravenous delivery, in newborn and juvenile mice will be compared. The analysis will focus on: (1) skeletal muscles (major forelimb, hindlimb, shoulder, abdominal and, face muscles) and differential infectivity of fast vs. slow twitch muscles, will be assessed by comparing tibialis anterior and soleus muscles, (2) cardiac muscle expression, and (3) liver expression. This cohort of animals will be monitored for possible adverse effects of the high-titer AAV injections.

Although direct muscular injection of AAVs represents an approach to delivering the FLD-AAV modules (which in light of the results in FIGS. 5A-5F is likely to be successful), it is nonetheless desirable from a clinical perspective to achieve full-length dystrophin expression using systemic i.v. delivery of the virus. In vitro FLD-AAV testing will be used to determine how AAV copy number and reconstituted dystrophin levels correlate. Tissue distribution and efficiency of reconstitution will be assessed in vivo, and different delivery paradigms (e.g., serotype, viral titer, route of application, number of repeat applications) will be examined to achieve optimal tissue distribution. Tissue coverage and expression levels will be assessed. Beneficial outcomes can be achieved even if only a portion of muscle fibers express dystrophin (e.g., normal heart function with only about 50% of cardiomyocytes being dystrophin deficient under non-stress conditions). Both, physiological and supraphysiological levels of dystrophin are of therapeutic value. Quantitative assessment will be performed as outlined in Part 1 & 2. In vivo intramuscular and systemic virus application will be performed in neonatal or juvenile mice under aseptic condition.

Part 5. Treat DMD mouse model (mdx) with FLD-AAV and assess disease onset/progression. FLD-AAV delivery in neonatal mdx mice may prevent the onset and progression of myopathy and cardiomyopathy. After optimization of the viral delivery of reconstituted full-length dystrophin (Parts 1-4) FLD-AAV treatment will be administered to a mouse model of DMD. These mice, depending on the genetic background they are bred, present with myopathy that is notably less pronounced than human DMD. Mice with the genetic background that presents with a more severe phenotype (D2.B10-Dmdmdx) show increased hind-limb weakness, lower muscle weight, fewer myofibers, and increased fat and fibrosis. These parameters can be compared between wild-type controls, treated mdx, and untreated mdx mice. The desired outcome is an amelioration or prevention of disease onset/progression.

Two mouse lines, C57BL/10ScSn-Dmdmdx/J, and D2.B10-Dmdmdx/J, which carry a mutation in the dystrophin gene are used. FLD-AAV is delivered according to parameters established as

described under Part 4. Animals are injected in the first postnatal week, in a time window before onset of myonecrosis in mdx mice. Wild-type, treated-mdx and vehicle/sham-treated-mdx mice are assessed for behavioral and anatomical signs of skeletal and cardiac myopathy. Using kinematic and electromyographic testing equipment, performance of these mice in a variety of motor tasks is assessed, such as balance beam, grip strength, horizontal ladder, treadmill speed challenge, over ground locomotor kinematic assessment, and swimming kinematic assessment (ambient temperature and cold water challenge). It will be determined whether FLD-AAV therapy can prevent the presentation of cardiomyopathy in mdx mice following chemical challenge.

The desired outcome of these experiments would be an amelioration or prevention of disease onset/progression.

### **Example 7**

#### **Delivery of reconstituted full-length MYO7A treat Usher Syndrome**

A first half of the MYO7A coding sequence is appended with a synthetic RNA dimerization and recombination domain and expressed from a first vector/plasmid. The second half of MYO7A is appended to the complementary RNA dimerization and recombination domain and expressed from a second vector/plasmid. If expressed together in the same cell the two halves of MYO7A are recombined to form the full-length MYO7A transcript which is then translated into protein.

### **Example 8**

#### **Transcriptional/expressional logic gate**

Breaking a target gene into two nonfunctional halves that get expressed from either two different promoters or using two different delivery vehicles can result in an intersectional expression pattern.

For example, promoter 1 of a first synthetic nucleic acid molecule provided herein can drive expression of the N-terminal half of the coding sequence in for example cell types A, B, and C, while promoter 2 of a second synthetic nucleic acid molecule provided herein drives expression of the C-terminal half in a subset of cells A, D, E, and F. In such an example, the effector gene encoding the target protein is only expressed in the overlapping area (in this example in cell population A).

A similar intersectionality can be used by making the two halves conditionally expressed, for example, under the condition of the presence of a recombinase. Another level at which intersectionality can be achieved is by delivering the two halves with two viruses that have different tropisms.

### **Example 9**

#### **Complementation**



The disclosed methods and systems can be used to make any gene (and corresponding target protein) into complementation parts (similar to the principle of alpha complementation of LacZ), by encoding two non-functional halves on separate plasmids that only become active when both plasmids are present.

5

### **Example 10**

#### **Trigger RNA**

The disclosed systems and methods can be configured such that reconstitution of the two or more portions of the coding sequences of the target protein depends on the presence of a specific “trigger” RNA molecule. As shown in FIG. 7B, in this example, the dimerization domains of each synthetic nucleic acid molecule are not reverse complements of one another, but instead specifically hybridize to adjacent regions of a third RNA molecule, a “trigger RNA”, which serves as a bridge to bring two synthetic nucleic acid molecules together. In this example, the system can “report” the presence of a specific RNA molecule which allows for “cell type specific triggering” of a reporter/effector protein.

15

### **Example 11**

#### **Inclusion of Stabilizing Element in 3'-UTR**

This example describes methods used to evaluate recombination of split coding sequences in the presence of a sequence in the 3'-UTR that stabilizes RNA. Woodchuck hepatitis posttranscriptional regulatory element 3 (WPRE3) was used as an exemplary stabilizing sequence. One skilled in the art will appreciate that other RNA sequence stabilizers can be used in place of WPRE3.

20

Median YFP fluorescence was measured by flow-cytometry for a two-way split YFP that is reconstituted using the disclosed synthetic RNA dimerization and recombination approach. The C-terminal YFP coding fragment is followed by a poly adenylation signal only (w/o WPRE3) or by a truncated version of the woodchuck hepatitis posttranscriptional regulatory element, WPRE3 followed by a poly adenylation signal (labelled w/WPRE3). The N-terminal YFP coding fragment is coexpressed with a red fluorescent protein from a bidirectional promoter for transfection control. The C-terminal fragment is co-expressed with a blue fluorescent protein from a bidirectional promoter as transfection control. Cells with equal red and blue fluorescent control values between conditions are compared.

25

30

As shown in FIG. 8, inclusion of a stabilizing element in the 3'-UTR increased expression efficiency of the recombined full-length YFP by about 50-60%. This enhancement is observed even though WPRE sequences stimulate nuclear export of the RNA molecule they are contained in, which may have negatively impacted the RNA joining reaction (and thus gene expression) by

35

shuttling molecule 150 of FIG. 6A outside the nucleus before the spliceosome mediated RNA joining can occur and thus rendering it non-functional.

Thus, the disclosed synthetic molecules (such as any of SEQ ID NOS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 145, 146, 147, 148, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, and 166) can be modified to further include a RNA sequence stabilizer.

### Example 12

#### Effect of Binding Domain Length on Reconstitution Efficiency

Binding domain length was assessed as follows. YFP was split into two non-fluorescent halves (SEQ ID NOS: 1 and 2, but with different length binding domains). Reconstitution efficiency for different length binding domains (ranging from 50 to 500 nucleotides) was assessed in cultured HEK 293t cells. N-terminal YFP is expressed from a bidirectional CMV promoter with a Red Fluorescent Protein (RFP) as a transfection control. C-terminal YFP is expressed from a bidirectional CMV promoter with a Blue Fluorescent Protein (BFP) as a transfection control. For the different binding domain lengths, YFP median fluorescence intensity was compared. Cells with matching RFP and BFP transfection levels are compared between conditions.

As shown in FIG. 11, all of the molecules achieved some level of expression of the full-length YFP, with varying degrees of reconstitution efficiency. Although maximal performance was observed with binding domain lengths of 150 bp and below (*e.g.* 50-150 bp), binding domain lengths of up to 500 bp were still able to recombine and express full-length YFP.

### Example 13

#### Effect of Splicing Enhancer Sequences

This example describes methods used to assess the effect of including one or more intronic splicing enhancer sequences (*e.g.*, 118, 120, 156 in FIG. 6A) in the disclosed synthetic introns.

YFP was split into two non-fluorescent halves (FIG. 12A). Reconstitution efficiency for different intron configurations was assessed in cultured HEK 293t cells. N-terminal YFP was expressed from a bidirectional CMV promoter with a Red Fluorescent Protein (RFP) as a transfection control. C-terminal YFP was expressed from a bidirectional CMV promoter with a Blue Fluorescent Protein (BFP) as a transfection control. For the different intron configurations, YFP median fluorescence intensity is compared. Cells with matching RFP and BFP transfection levels are compared between conditions.

As shown in FIG. 12A, the 5' molecule (SEQ ID NO: 1) includes the coding region of the N-terminal portion of YFP (n-yfp), followed by a splice donor sequence (SD), a downstream intronic splicing enhancer (DISE), and two intronic splicing enhancers (2xISE), a binding domain (BD), a self-cleaving hammerhead ribozyme (HHRz), ending with a poly adenylation signal (pA). The 3' molecule

(SEQ ID NO: 2) includes the complementary binding domain (anti-BD), followed by three intronic splicing enhancer sequences (3xISE), a branch point (BP), a polypyrimidine tract (PPT), a splice acceptor sequence (SA), the c-terminal portion of the YFP coding sequence, ending with a polyadenylation signal (pA).

5 As shown in FIG. 12B, inclusion of splice enhancers to both the 5' and the 3' molecules increases reconstitution efficiency of the full-length YFP. Removal of the splice enhancers reduces the reconstitution efficiency of the two coding sequences by about 50-90%. In the first column, YFP is reconstituted using the reference configuration (SEQ ID NOS: 1 and 2), the second column shows the reconstitution efficiency with deletion of the ISE elements in the 5' fragment, the third column shows reconstitution efficiency after deletion of the ISE and the DISE in the 5' fragment. The fourth column shows the reconstitution efficiency after deletion of the HHrz in the 5' fragment. The fifth column shows reconstitution efficiency using the reference configuration. The sixth column shows reconstitution efficiency after deletion of the ISE elements in the 3' fragment. The seventh shows reconstitution efficiency after deletion of the ISE in both 5' and 3' fragment and the DISE in the 5' fragment.

### Example 14

#### Dual Projection Tracing

This example describes methods used to perform dual projection tracing by reconstitution of full-length flp recombinase (Flpo) from two fragments (SEQ ID NOS: 147 and 148). As shown in FIG. 13A, the Flp recombinase gene was split into two non-functional halves. The N-terminal half of the Flpo gene was joined at its 3' end with a synthetic intron sequence followed by a dimerization domain sequence (RNA end joining module, REJ). The C-terminal half of the Flpo gene was joined at its 5' end with a synthetic intron and a dimerization domain (REJ-module). Upon infection of a cell with both constructs, and expression of pre-mRNA from each construct, the pre-mRNAs bound at the dimerization domains (shown by dark parallel bars in FIG. 13A), and the resulting complex was spliced to produce the full length Flpo recombinase mRNA transcript. Thus, the functional recombinase protein was produced by reconstitution of the two fragments. FIG. 13B shows a schematic of a flp activity reporter mouse carrying a flpo dependent red fluorescent protein (tdTomato) (Rosa-CAG-frt-STOP-frt-tdTomato). The two synthetic nucleic acid (DNA) constructs were packaged into separate adeno-associated viruses (retrogradely transported serotype AAV2/retro). AAV2/retro-n-flpo, virus carrying the first construct, was injected in the left primary motor cortex of the mouse, and AAV2/retro-c-flpo, virus carrying the second construct, was injected in the right primary motor cortex of the mouse.

As shown in FIGS. 13C-13D, primary motor cortex cells with axons that cross the midline are labeled with the red fluorescent protein (and appear white in FIGS. 13C and 13D). Hoechst staining (nuclei) is shown for context.

5

### Example 15

#### Expression of Long Protein *in vivo*

This example describes methods used to achieve efficient expression of oversized cargo in cell culture and *in vivo* in the mouse primary motor cortex.

To simulate a large disease-causing gene that fills up the adeno-associated virus (AAV) cargo capacity of two viruses (*i.e.*, it exceeds single AAV packaging capacity), a split YFP coding sequence was embedded inside a large uninterrupted open reading frame. N-terminally (*i.e.*, on the 5' side) the first part of the YFP coding sequence is flanked with long stuffer sequences (*i.e.*, an uninterrupted open reading frame) followed by a sequence encoding a 2A self-cleaving peptide. On the C-terminus (*i.e.*, 3' side) the second part of the YFP coding sequence is followed by a 2A self-cleaving peptide coding sequence and then followed by a long stuffer sequence (*i.e.*, and uninterrupted open reading frame) (FIG. 14A). The first and second synthetic DNA molecules encoding pre-mRNA molecules are shown in SEQ ID NOS: 22 and 23, excluding promoter sequences. The resulting RNA molecules expressed are each about 4000nt between the transcriptional start site at position 1 of SEQ ID NO: 22 and the transcriptional start site at position 1 of SEQ ID NO: 23 and the polyA tail. The resulting transcribed pre-mRNA molecule (5' fragment; transcribed from SEQ ID NO: 22) contains a stuffer open reading frame which is followed by a self-cleaving 2A peptide encoding sequence, followed by a sequence encoding the N-terminal portion of YFP, followed by a synthetic intron and a dimerization domain (having kissing loop architecture), and a polyA tail. The C-terminal pre-mRNA molecule (3' fragment; transcribed from SEQ ID NO: 23) is composed of a complementary kissing loop dimerization domain, a synthetic intron sequence, followed by the C-terminal YFP coding sequence, followed by a self-cleaving 2A peptide coding sequence, followed by a stuffer open reading frame, followed by a polyA tail.

Following production of the pre-mRNA molecules, the dimerization domains bind, and splicing joins the pre-mRNAs to produce a full-length mRNA. During translation, the 2A cleavage sequences flanking the YFP result in the cleaving off of the N and C-terminal stuffer sequences and the production of functional YFP protein.

To determine reconstitution efficiency on an RNA level, two probe based (5'-hydrolysis) quantitative real-time PCR assays are used. The first assay spans a sequence fully contained in the 3' exonic YFP sequence (labelled 3' probe). The second assay spans the junction between the 5' and the

3' exonic YFP sequence (labelled junction probe). Reconstitution efficiency is calculated as the ratio of (junction probe count)/(3' probe count).

Quantitative real-time PCR analysis of reconstitution efficiency of the oversize YFP constructs in HEK 293t cells was performed. Full-length oversized YFP is used as reference. The full-length oversized YFP ratio is set to 1 (FIG. 14B). Ratio of reconstituted is expressed as fraction of full-length (labelled split-REJ (split RNA end joining)). Reconstitution efficiency is calculated as follows: junction/3'prime. As shown in FIG. 14B about 60% of the RNAs joined in the split-REJ system.

Reconstituted YFP protein expression from full-length oversized YFP expression and split-REJ expression is assessed by flow cytometry of transiently transfected HEK 293t cells. As shown in FIG. 14C, the split REJ system achieved about a 45% joining efficiency, even with the large cargo.

*in vivo* analysis of reconstitution of the large YFP protein was performed as follows. 60nl of adeno-associated virus 2/8, containing 3E9 vg/injection/fragment, was injected into the primary motor cortex of the mouse. Tissue was harvested 10 days post injection. As shown in FIG. 14D, YFP fluorescence is readily detectable in the bulk tissue (top left, top middle panel, macroscopic top view of the mouse brain, YFP fluorescence plus auto-fluorescence for context are shown). Strong YFP signal is detected at and around the virus injection site in layer 5 of the motor cortex (right panel, cortical layers are numbered 1 to 6, approximate injection depth is indicated by gray bar, scale bar = 100 micrometers). Thus, the disclosed system can be used to express large proteins *in vivo*.

## Example 16

### Expression of Factor VIII

This example describes methods used to achieve efficient reconstitution of full-length human coagulation factor VIII (FVIII).

A schematic of the 5' and 3' nucleic acid molecules used for the experiment is shown in FIG. 15A (DNA encoding the pre-RNA molecules are set forth in SEQ ID NOS: 24 and 25, respectively). Each half includes about 3.8 kb of FVIII coding sequence. The resulting RNA 5'-sequence containing the N-terminal half (*e.g.*, as shown schematically in 110 of FIG. 6A) of the FVIII coding sequence is followed by an efficient synthetic intron and a dimerization domain (kissing loop architecture), and a polyA tail. The 3'-sequence containing the C-terminal half (*e.g.*, 150 of FIG. 6A) of the FVIII coding sequence is preceded by the complementary kissing loop dimerization domain, and an efficient synthetic intron sequence. To determine reconstitution efficiency on an RNA level, two probe based (5'-hydrolysis) quantitative real-time PCR assays are used. The first assay spans a sequence fully contained in the 3' exonic FVIII sequence (labelled 3' probe). The second assay spans the junction between the 5' and the 3' exonic FVIII sequence (labelled junction probe). Reconstitution efficiency is calculated as the ratio of (junction probe count)/(3' probe count).

PCR quantification of reconstitution efficiency after two days of expression in HEK 293t cells was performed. Full-length FVIII is used as reference. Full-length FVIII ratio is set to one. Reconstituted FVIII assay ratios are expressed as fraction of full-length (labelled split-REJ). As shown in FIG. 15B, a reconstitution efficiency of about 40-60% was achieved (that is about 40-60% of the two RNAs joined in the split-REJ system).

To demonstrate expression of FVIII *in vitro*, Western blotting was used. FVIII was tagged with an HA-tag at the N-terminus. Constructs are expressed in HEK 293t cells for 2 days. As shown in FIG. 15C, the disclosed split-REJ system successfully expressed full-length FVIII *in vitro*.

Based on these observations, expression of a full-length FVIII protein *in vivo* can be achieved, for example to treat hemophilia A. For example, a first half of a FVIII coding sequence is appended with a synthetic RNA dimerization and recombination domain and expressed from a first vector/plasmid. The second half of FVIII is appended to the complementary RNA dimerization and recombination domain and expressed from a second vector/plasmid. If expressed together in the same cell the two halves of FVIII are recombined to form the full-length FVIII transcript which is then translated into protein. For example, a sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ ID NO: 24, which includes an N-terminal FVIII coding sequence, and SEQ ID NO: 25 which includes a C-terminal FVIII coding sequence, can be utilized for *in vivo* expression.

### Example 17

#### Expression of Abca4

This example describes methods used to achieve efficient reconstitution of full-length human ATP binding cassette subfamily A member 4 (Abca4).

A schematic of the 5' and 3' molecules used are shown in FIG. 16A (DNA encoding the pre-RNA molecules are set forth in SEQ ID NOS: 20 and 21, respectively). The 5' half includes about 3.6kb of Abca4 coding sequence, and the 3' half about 3.2kb of the Abca4 coding region plus a C-terminal 3xFLAG tag. The 5'-sequence contains the N-terminal half of the coding sequence followed by the an efficient synthetic intron sequence and the first dimerization domain (kissing loop). The 3'-sequence containing the C-terminal half of the coding sequence is preceded by the complementary (kissing loop) dimerization domain and an efficient synthetic intron sequence. A Sanger sequencing trace across the junction is shown.

As shown in FIG. 16B, PCR amplification of the junction demonstrates faithful joining of the two coding sequences. To determine reconstitution efficiency on an RNA level, two probe based (5'-hydrolysis) quantitative real-time PCR assays are used (FIG. 16C). The first assay spans a sequence fully contained in the 3' exonic Abca4 sequence (labelled 3' probe). The second assay spans the

junction between the 5' and the 3' exonic Abca4 sequence (labelled junction probe). Reconstitution efficiency is calculated as the ratio of (junction probe count)/(3' probe count). PCR quantification of reconstitution efficiency after two days of expression in HEK 293t cells is shown in FIG. 16D. Full-length Abca4 is used as reference. Average full-length Abca4 ratio is set to one. Reconstituted Abca4 assay ratios are expressed as fraction of full-length (labelled split-REJ). As shown in FIG. 16D, a reconstitution efficiency of about 35% was achieved (that is about 30-40% of the two RNAs joined in the split-REJ system).

To demonstrate expression of Abca4 *in vitro*, Western blotting was used. Abca4 is tagged with a 3xFLAG-tag at the C-terminus. Constructs are expressed in HEK 293t cells for 2 days. As shown in FIG. 16E, the disclosed split-REJ system successfully expressed full-length Abca4 *in vitro*.

Quantification of the western blot is shown in Fig. 16F. To normalize for differential transfection efficiency between conditions, the full-length plasmid and the C-terminal plasmid co-express a Blue Fluorescent Protein for transfection control. BFP concentration in each sample was determined by dot blot and used to normalize between conditions. As shown in FIG. 16F reconstituted Abca4 is expressed at approximately 40% of the levels when compared with direct full-length expression. Hence, the protein levels as determined by western blot, track well with the RNA reconstitution efficiency determined by qPCR.

Based on these observations, expression of a full-length ABCA4 protein *in vivo* can be achieved, for example to treat Stargardt's Disease. For example, a first half of the ABCA4 coding sequence is appended with a synthetic RNA dimerization and recombination domain and expressed from a first vector/plasmid. The second half of ABCA4 is appended to the complementary RNA dimerization and recombination domain and expressed from a second vector/plasmid. If expressed together in the same cell the two halves of ABCA4 are recombined to form the full-length ABCA4 transcript which is then translated into protein. For example, a sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ ID NO: 20 (FIGS. 10R-10U), which includes an N-terminal Abca4 coding sequence, and SEQ ID NO: 21 (FIGS. 10V-10Z) which includes a C-terminal Abca4 coding sequence, can be utilized for *in vivo* expression.

### Example 18

#### Expression of Otof

This example describes methods used to achieve efficient reconstitution of full-length murine Otoferlin (Otof).

The sequences of the 5' and 3' DNA molecules used are shown in SEQ ID NOS: 155 and 156, respectively. The 5' half includes about 3.5kb of Otof coding sequence, the 3' half about 2.5kb of the

Otof coding region plus a C-terminal 3xFLAG tag. The 3'-sequence containing the C-terminal half (e.g., 150 of FIG. 6A) is preceded by the complementary binding domain and an efficient synthetic intron sequence. One skilled in the art will appreciate that a human OTOF coding sequence (e.g., GenBank Accession No. NM\_001287489.2 or NM\_194248.3) can be substituted for the mouse coding sequence in SEQ ID NOS: 155 and 156.

To demonstrate expression of Otof *in vitro*, Western blotting was used. Otof is tagged with a 3xFLAG-tag at the C-terminus for Western blot detection. Constructs are expressed in HEK 293t cells for 2 days. As shown in FIG. 18A, the disclosed split-REJ system successfully expressed full-length Otof *in vitro*.

Quantification of the western blot is shown in FIGS. 18B-C. Raw quantification is shown in the left bar plot (FIG. 18B) as fraction of full-length control. To normalize for differential transfection efficiency between conditions, the full-length plasmid and the C-terminal plasmid co-express a Blue Fluorescent Protein for transfection control. BFP concentration in each sample was determined by confocal fluorescence microscopy prior to harvesting the cells and used to normalize between conditions. Normalized quantification is shown in the right bar plot (FIG 18C) as normalized fraction of the full-length control. As shown in FIG. 18C reconstituted Otof is expressed at approximately 30% of the levels when compared with direct full-length expression.

Based on these observations, expression of a full-length OTOF protein *in vivo* can be achieved, for example to treat autosomal recessive deafness 9. For example, a first half of the OTOF coding sequence is appended with a synthetic RNA dimerization and recombination domain (that is an intron and binding domain) and expressed from a first vector/plasmid. The second half of OTOF is appended to the complementary RNA dimerization and recombination domain and expressed from a second vector/plasmid. If expressed together in the same cell the two RNA molecules are expressed in a target cell and the two halves of OTOF coding transcript are recombined to form the full-length OTOF transcript which is then translated into protein. For example, a sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ ID NO: 155, which includes an N-terminal Otof coding sequence, and SEQ ID NO: 156 which includes a C-terminal Otof coding sequence, can be utilized for *in vivo* expression, for example to treat hearing loss.

### Example 19

#### Expression of Myo7a

This example describes methods used to achieve efficient reconstitution of full-length human MYOSIN VIIA (Myo7a).



The sequences of the 5' and 3' DNA molecules used are shown in SEQ ID NOS: 157 and 158, respectively. The 5' half includes about 3.6kb of Myo7a coding sequence, the 3' half about 3.1kb of the Myo7a coding region plus a C-terminal 3xFLAG tag. The 3'-sequence containing the C-terminal half (e.g., 150 of FIG. 6A) is preceded by the complementary binding domain and an efficient synthetic intron sequence.

To demonstrate expression of Myo7a *in vitro*, Western blotting was used. Myo7a is tagged with a 3xFLAG-tag at the C-terminus for Western blot detection. Constructs are expressed in HEK 293t cells for 2 days. As shown in FIG. 19A, the disclosed split-REJ system successfully expressed full-length Myo7a *in vitro*.

Quantification of the western blot is shown in FIGS. 19B-19C. Raw quantification is shown in the left bar plot (FIG. 19B) as fraction of full-length control. To normalize for differential transfection efficiency between conditions, the full-length plasmid and the C-terminal plasmid co-express a Blue Fluorescent Protein for transfection control. BFP concentration in each sample was determined by confocal fluorescence microscopy prior to harvesting the cells and used to normalize between conditions. Normalized quantification is shown in the right bar plot (FIG. 19C) as normalized fraction of the full-length control. As shown in FIG. 19C reconstituted Myo7a is expressed at approximately 60% of the levels when compared with direct full-length expression.

Based on these observations, expression of a full-length MYO7A protein *in vivo* can be achieved, for example to treat Usher syndrome, type 1B. For example, a first half of the MYO7A coding sequence is appended with a synthetic RNA dimerization and recombination domain (that is an intron and binding domain) and expressed from a first vector/plasmid. The second half of MYO7A is appended to the complementary RNA dimerization and recombination domain and expressed from a second vector/plasmid. If expressed together in the same cell the two RNA molecules are expressed in a target cell and the two halves of MYO7A coding transcript are recombined to form the full-length MYO7A transcript which is then translated into protein. For example, a sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ ID NO: 157, which includes an N-terminal Myo7a coding sequence, and SEQ ID NO: 158 which includes a C-terminal Myo7a coding sequence, can be utilized for *in vivo* expression.

## Example 20

### Expression of dCas9-VPR

This example describes methods used to achieve efficient reconstitution of full-length enzymatically dead Cas9 fused to a VPR transcriptional activator domain (dCas9-VPR).

The sequences of the 5' and 3' DNA molecules used are shown in SEQ ID NOS: 159 and 160, respectively. The 5' half includes about 3.3kb of DCas9-VPR coding sequence, the 3' half about 2.5kb

of the DCas9-VPR coding region. The 3'-sequence containing the C-terminal half (e.g., 150 of FIG. 6A) is preceded by the complementary binding domain and an efficient synthetic intron sequence.

To demonstrate expression of DCas9-VPR *in vitro*, Western blotting was used. Constructs are expressed in HEK 293t cells for 2 days. As shown in FIG. 20A, the disclosed split-REJ system successfully expressed full-length DCas9-VPR *in vitro*.

Quantification of the western blot is shown in FIGS. 20B-20C. Raw quantification is shown in the left bar plot (FIG. 20B) as fraction of full-length control. To normalize for differential transfection efficiency between conditions, the full-length plasmid and the C-terminal plasmid co-express a Blue Fluorescent Protein for transfection control. BFP concentration in each sample was determined by confocal fluorescence microscopy prior to harvesting the cells and used to normalize between conditions. Normalized quantification is shown in the right bar plot (FIG. 20C) as normalized fraction of the full-length control. As shown in FIG. 20C reconstituted DCas9-VPR is expressed at approximately 35% of the levels when compared with direct full-length expression. When expressed together with a UAS targeting guide RNA in HEK 293t cells (FIG. 20D), both full-length and two-way split reconstituted dCas9-VPR induce yellow fluorescent protein expression from a UAS-YFP plasmid, demonstrating functionality of the reconstituted dCas9-VPR.

Based on these observations, expression of a full-length DCAS9-VPR protein *in vivo* can be achieved, for example to activate or overexpress genes. For example, a first half of the DCAS9-VPR coding sequence is appended with a synthetic RNA dimerization and recombination domain (that is an intron and binding domain) and expressed from a first vector/plasmid. The second half of DCAS9-VPR is appended to the complementary RNA dimerization and recombination domain and expressed from a second vector/plasmid. If expressed together in the same cell the two RNA molecules are expressed in a target cell and the two halves of DCAS9-VPR coding transcript are recombined to form the full-length DCAS9-VPR transcript which is then translated into protein. For example, a sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ ID NO: 159, which includes an N-terminal DCas9-VPR coding sequence, and SEQ ID NO: 160 which includes a C-terminal DCas9-VPR coding sequence, can be utilized for *in vivo* expression.

## Example 21

### Expression of Prime Editor

This example describes methods used to achieve efficient reconstitution of full-length humanized Cas9 Prime Editor (Prime Editor).

The sequences of the 5' and 3' DNA molecules used are shown in SEQ ID NOS: 161 and 162, respectively. The 5' half includes about 3.3kb of Prime Editor coding sequence, the 3' half about 3.0kb

of the Prime Editor coding region. The 3'-sequence containing the C-terminal half (e.g., 150 of FIG. 6A) is preceded by the complementary binding domain and an efficient synthetic intron sequence.

To demonstrate expression of Prime Editor *in vitro*, Western blotting was used. Constructs are expressed in HEK 293t cells for 2 days. As shown in FIG. 21A, the disclosed split-REJ system successfully expressed full-length Prime Editor *in vitro*.

Quantification of the western blot is shown in FIGS. 21B-21C. Raw quantification is shown in the left bar plot (FIG. 21B) as fraction of full-length control. To normalize for differential transfection efficiency between conditions, the full-length plasmid and the C-terminal plasmid co-express a Blue Fluorescent Protein for transfection control. BFP concentration in each sample was determined by confocal fluorescence microscopy prior to harvesting the cells and used to normalize between conditions. Normalized quantification is shown in the right bar plot (FIG. 21C) as normalized fraction of the full-length control. As shown in FIG. 21C reconstituted Prime Editor is expressed at approximately 60% of the levels when compared with direct full-length expression. FIG. 21D shows that targeted G to T transversion mutations could be introduced using the full-length and the two-way split prime editors, demonstrating functionality of the two-way split prime editor constructs.

Based on these observations, expression of a full-length PRIME EDITOR protein *in vivo* can be achieved, for example to treat genomic point mutations. For example, a first half of the PRIME EDITOR coding sequence is appended with a synthetic RNA dimerization and recombination domain (that is an intron and binding domain) and expressed from a first vector/plasmid. The second half of PRIME EDITOR is appended to the complementary RNA dimerization and recombination domain and expressed from a second vector/plasmid. If expressed together in the same cell the two RNA molecules are expressed in a target cell and the two halves of PRIME EDITOR coding transcript are recombined to form the full-length PRIME EDITOR transcript which is then translated into protein. For example, a sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ ID NO: 161, which includes an N-terminal Prime Editor coding sequence, and SEQ ID NO: 162 which includes a C-terminal Prime Editor coding sequence, can be utilized for *in vivo* expression.

## Example 22

### Expression of AncBE4

This example describes methods used to achieve efficient reconstitution of full-length humanized Cas9 Cytosine Base Editor (AncBE4).

The sequences of the 5' and 3' DNA molecules used are shown in SEQ ID NOS: 163 and 164, respectively. The 5' half includes about 2.4kb of AncBE4 coding sequence, the 3' half about 3.2kb of

the AncBE4 coding region. The 3'-sequence containing the C-terminal half (*e.g.*, 150 of FIG. 6A) is preceded by the complementary binding domain and an efficient synthetic intron sequence.

To demonstrate expression of AncBE4 *in vitro*, Western blotting was used. Constructs are expressed in HEK 293t cells for 2 days. As shown in FIG. 22A, the disclosed split-REJ system successfully expressed full-length AncBE4 *in vitro*.

Quantification of the western blot is shown in FIG. 22B. Raw quantification is shown in the left bar plot (FIG. 22B) as fraction of full-length control. As shown in FIG. 22B reconstituted AncBE4 is expressed at approximately 40-50% of the levels when compared with direct full-length expression. FIG. 22C shows that targeted C to T transition mutations could be introduced using the full-length and the two-way split AncBE4s, demonstrating functionality of the two-way split AncBE4 constructs.

Based on these observations, expression of a full-length ANCB E4 protein *in vivo* can be achieved, for example to treat genomic point mutations. For example, a first half of the ANCB E4 coding sequence is appended with a synthetic RNA dimerization and recombination domain (that is an intron and binding domain) and expressed from a first vector/plasmid. The second half of ANCB E4 is appended to the complementary RNA dimerization and recombination domain and expressed from a second vector/plasmid. If expressed together in the same cell the two RNA molecules are expressed in a target cell and the two halves of ANCB E4 coding transcript are recombined to form the full-length ANCB E4 transcript which is then translated into protein. For example, a sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ ID NO: 163, which includes an N-terminal AncBE4 coding sequence, and SEQ ID NO: 164 which includes a C-terminal AncBE4 coding sequence, can be utilized for *in vivo* expression.

### Example 23

#### Expression of Abe8e

This example describes methods used to achieve efficient reconstitution of full-length humanized Cas9 Adenosine Base Editor (Abe8e).

The sequences of the 5' and 3' DNA molecules used are shown in SEQ ID NOS: 165 and 166, respectively. The 5' half includes about 2.4kb of Abe8e coding sequence, the 3' half about 3.2kb of the Abe8e coding region. The 3'-sequence containing the C-terminal half (*e.g.*, 150 of FIG. 6A) is preceded by the complementary binding domain and an efficient synthetic intron sequence.

To demonstrate expression of Abe8e *in vitro*, Western blotting was used. Constructs are expressed in HEK 293t cells for 2 days. As shown in FIG. 23A, the disclosed split-REJ system successfully expressed full-length Abe8e *in vitro*.

Quantification of the western blot is shown in FIG. 23B. Raw quantification is shown in the left bar plot (FIG. 23B) as fraction of full-length control. As shown in FIG. 23B reconstituted Abe8e is

expressed at approximately 70% of the levels when compared with direct full-length expression. FIG. 23C shows that targeted C to T transition mutations could be introduced using the full-length and the two-way split Abe8es, demonstrating functionality of the two-way split Abe8e constructs.

Based on these observations, expression of a full-length ABE8E protein *in vivo* can be achieved, for example to treat genomic point mutations. For example, a first half of the ABE8E coding sequence is appended with a synthetic RNA dimerization and recombination domain (that is an intron and binding domain) and expressed from a first vector/plasmid. The second half of ABE8E is appended to the complementary RNA dimerization and recombination domain and expressed from a second vector/plasmid. If expressed together in the same cell the two RNA molecules are expressed in a target cell and the two halves of ABE8E coding transcript are recombined to form the full-length ABE8E transcript which is then translated into protein. For example, a sequence having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to SEQ ID NO: 165, which includes an N-terminal Abe8e coding sequence, and SEQ ID NO: 166 which includes a C-terminal Abe8e coding sequence, can be utilized for *in vivo* expression.

#### Example 24

##### Increasing RNA fragment length results in decreasing two-way split gene reconstitution

The impact of length of the 5' fragment encoding RNA molecule and the 3' fragment encoding RNA molecule was assessed.

The yellow fluorescent protein (yfp) coding sequence was split into two fragments. To extend the RNA encoding molecules, stuffer open reading frames were installed at the 5' end of the 5' fragment and the 3' end of the 3' fragment, respectively. The 5' yfp coding sequence was fused to an extended stuffer open reading frame via a self-cleaving 2A sequence. The 3' yfp coding sequence of yfp was linked via a self-cleaving 2A sequence to an extended stuffer open reading frame. At the split point of the 5' fragment of yfp and the 3' fragment of yfp an RNA end joining module (synthetic intron plus binding domain) was installed. The self-cleaving 2A sequences allow for the YFP protein to separate from the respective stuffer open reading frames after translation. By incorporating different length stuffer open reading frames, four 5' fragment encoding constructs and four 3' fragment encoding constructs were assembled. The length of the RNA (protein coding sequence plus synthetic intron and binding domain) transcribed from these constructs were: 1000nt, 2000nt, 3000nt, and 4000nt for the 5' fragment and 1000nt, 2000nt, 3000nt, and 4000nt for the 3' fragment.

Efficiency of YFP reconstitution was compared between all sixteen 5' to 3' fragment pairs. In this comparison, YFP was most efficiently reconstituted when the shortest constructs (i.e., 5'-1000nt with 3'-1000nt) were paired. A decline of reconstitution efficiency was observed when fragments with

longer stuffer sequences were paired. As percent of the shortest pairing (5'-1000nt with 3'-1000nt) the following YFP reconstitution efficiencies were observed:

|    |                                 |
|----|---------------------------------|
|    | 5'-1000nt with 3'-1000nt: 100%  |
|    | 5'-1000nt with 3'-2000nt: ~40%  |
| 5  | 5'-1000nt with 3'-3000nt: ~20%  |
|    | 5'-1000nt with 3'-4000nt: ~16%  |
|    | 5'-2000nt with 3'-1000nt: ~55%  |
|    | 5'-2000nt with 3'-2000nt: ~30%  |
|    | 5'-2000nt with 3'-3000nt: ~20%  |
| 10 | 5'-2000nt with 3'-4000nt: ~15%  |
|    | 5'-3000nt with 3'-1000nt: ~60%  |
|    | 5'-3000nt with 3'-2000nt: ~40%  |
|    | 5'-3000nt with 3'-3000nt: ~25%  |
|    | 5'-3000nt with 3'-4000nt: ~20%  |
| 15 | 5'-4000nt with 3'-1000nt: ~40%  |
|    | 5'-4000nt with 3'-2000nt: ~35%  |
|    | 5'-4000nt with 3'-3000nt: ~20%  |
|    | 5'-4000nt with 3'-4000nt: ~15%. |

These data illustrate that increasing the length of the fragments that encode the 5' and 3' coding sequences of a split gene progressively lowers the efficiency of split gene reconstitution.

### Example 25

#### Enhancement of RNA end joining reaction with downstream intronic splicing enhancer and intronic splicing enhancer sequences.

This example describes methods used to achieve efficient joining of two RNA molecules by incorporation of specific splicing enhancer sequences.

Particular effectiveness of select intronic splicing enhancer sequences was investigated using a screening platform where a split yellow fluorescent protein (YFP) was reconstituted using RNA end joining modules that were composed of a trimodal kissing loop RNA dimerization domain and a variable library of intronic segments. The sequences of the 5' and 3' DNA molecules used are shown in SEQ ID NOS: 171 and 172, respectively (the string of Ns in the sequence indicates the site of intronic library placement, such as at least one of the sequences in Table 2 below, such as 1, 2, 3, 4 or 5 of these sequences).

To demonstrate expression of reconstituted yfp *in vitro*, flow cytometry was used to determine yfp fluorescence intensity in HEK293t cells that were transfected with the 5' and 3' DNA molecules.

As shown in FIG. 24A, the intronic portion of the disclosed split-REJ system was subdivided into individual segments to screen for efficient intronic splicing enhancer sequences that facilitate the RNA joining reaction. The sequences used in the three positions of the 5' and the three positions of the 3' intronic portion of the constructs are given in SEQ NOS: 173 to 204 (Table 2) and listed in FIG. 24C.

5

Table 2: Exemplary Intronic Splicing Enhancer Sequences

| label | name                  | Sequence (SEQ ID NO)                           |
|-------|-----------------------|------------------------------------------------|
| ds1   | FGFR-2 pre-mRNA, IAS1 | GTAAGTATTgctttcattttgtcttttttaa (173)          |
| ds2   | Fas URI6              | GTAagttcttgctttgttcaaactgtctat (174)           |
| ds3   | CFTR E9 PY1/2         | GTAAGTATTCTTTTGTCTTCACtcat (175)               |
| ds4   | TIA1-preferred        | GTAAGTATTTTTTTACTCCtcaTTTTTACTCC (176)         |
| ds5   | FAS intron5           | GTAAGTATTTTTTTACGGTTATATTCTCCTTTCCCC (177)     |
| ds6   | CD46-D1/9             | GTAAGTATTTTCTGTTGTTTATTtccag (178)             |
| ds7   | B19V ISE              | GTAAGTATTGGGGTTGATTATGTGTGGGACGGTGTAAAGG (179) |
| ds8   | ratFGFR2DISE          | GTAAGTATTtctcttttcttccatgggttggcct (180)       |
| ds9   | just donor scramble   | GTAAGTATTaccagagattcgtagacctgcttgac (181)      |
| m1    | 6xWGGG                | TGGGGCTGGGCAGAGGGTTGAGGGGAGAGGGTCCTGGGG (182)  |
| m2    | C9-E6-ISE             | tcaTGGGTGGGTtcatTGGGTGGGTtca (183)             |
| m3    | AdMLBPadj             | Tagggcgtagtagtccagggtt (184)                   |
| m4    | bcl2-I2-BPadj         | Ttctctgtgggtggcattctctctctct (185)             |
| m5    | M2                    | GGGttatGGGACCtcaGGGataaGGGACC (186)            |
| m6    | GH1ivs                | CGGGGATGGGGGtca (187)                          |
| m7    | WangGrich             | TGGGGGGAGGtcaTGGGGGGAGG (188)                  |
| m8    | WangISE2              | GTTGGTGGTTtcatGTTGGTGGTT (189)                 |
| m9    | WangA                 | GGGTTTCGGGTTTtcaGGTGGTCGTTGGT (190)            |
| m10   | WangB                 | GGTGGTCGTTGGTtcaTTTGGGCTATTGG (191)            |
| m11   | WangC                 | TTTGGGCTATTGGtcaAGGGGGCGAGGGG (192)            |
| m12   | WangD                 | AGGGGGCGAGGGGtcaGGTATTCGGTATT (193)            |
| m13   | WangE                 | GGTATTCGGTATTtcaaggtaaCaggtaa (194)            |
| m14   | WangFmod              | aggtaaCaggtaatcaGGGTTTCGGGTTT (195)            |
| m15   | SMN-URC2/3            | TCTTACTTTTGTaaacTTTATGGTTTGTg (196)            |
| m16   | just scramble         | Cacgtattctcggtacggacgttacaga (197)             |
| ddl   | scramble              | Taagctggtatcc (198)                            |

|     |               |                                                 |
|-----|---------------|-------------------------------------------------|
| ap2 | 4.1R-E16-uISE | CACTAACTCTTTTTCCCCCtttttttACAG (199)            |
| ap3 | P6-cons-to30  | TACTAACtctttcttttCCTTTCCTTCTTCACAG (200)        |
| ap4 | AdMLSA        | CACTAACTCTgtcatactatctgtccctttttccaCAG (201)    |
| ap5 | bcl2-I2-SA    | CACTAACTCTctttcttttctccctctctccccaactgCAG (202) |
| ap6 | perfectT      | CACTAACTCTtttttttttttttttACAGCAG (203)          |
| ad1 | scramble      | Taagctggtatcc (204)                             |

Quantification of the flow cytometry is shown in FIG. 24B. Incorporation of intronic sequences to stimulate recruitment of the 5' splice site selection promoting splicing factor TIA-1 (T-Cell-Restricted Intracellular Antigen-1) can increase RNA end joining. In some examples, sequences containing WGGG motifs enhance RNA end joining.

Based on these observations, expression of a full-length split proteins *in vivo* can be enhanced by incorporating specific intronic splicing enhancer sequences into the intronic portion of the RNA end joining modules. For example, one or more sequences (such as 1, 2, or 3 sequences) having at least 80%, at least 90%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or 100% sequence identity to any one of SEQ NOS: 173 to 180, 182-196, or 199 to 203, can be utilized for *in vivo* expression of RNA end joining reaction products (for example can be use as the ISE for any embodiment provided herein).

In view of the many possible embodiments to which the principles of the disclosure may be applied, it should be recognized that the illustrated embodiments are only examples of the invention and should not be taken as limiting the scope of the invention. Rather, the scope of the invention is defined by the following claims. We therefore claim as our invention all that comes within the scope and spirit of these claims.



We claim:

1. A composition for expressing a target protein comprising (a) a first RNA molecule, the RNA molecule comprising from 5' to 3': (i) a coding sequence for an N-terminal portion of the target protein; (ii) a splice donor; and (iii) a first dimerization domain; and (b) a second RNA molecule, the RNA molecule comprising from 5' to 3': (i) a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice acceptor; and (v) a coding sequence for a C-terminal portion of the target protein.

2. The composition of claim 1, wherein the first and second dimerization domains bind by direct binding, indirect binding, or a combination thereof.

3. The composition of claim 2, wherein direct binding or indirect binding comprises basepairing interactions, non-canonical base pairing interactions, non-base pairing interactions, or a combination thereof.

4. The composition of claim 2 or 3, wherein direct binding comprises base pairing interactions between kissing loops or hypodiverse regions.

5. The composition of claim 2 or 3, wherein direct binding comprises non-canonical basepairing interactions, non-canonical base pairing interactions, non-base pairing interactions, or a combination thereof, between aptamer regions.

6. The composition of claim 2 or 3, wherein indirect binding comprises basepairing interactions through a nucleic acid bridge.

7. The composition of claim 2, wherein indirect binding comprises non-base pairing interactions between an aptamer and an aptamer target, or between two aptamers.

8. The composition of any one of claims 1 to 7, wherein the first or second dimerization domain does not comprise a cryptic splice acceptor.

9. The composition of any one of claims 1 to 8, wherein the dimerization domains are directly binding or indirectly binding aptamer sequence dimerization domains.

10. The composition of any one of claims 1 to 9, wherein the dimerization domains are kissing loop interaction domains.

11. The composition of any one of claims 1 to 10, wherein the target protein is a protein associated with disease, or a therapeutic protein.

12. The composition of claim 11, wherein the disease is a monogenic disease.

13. The composition of claim 12, wherein the therapeutic protein is a toxin.

14. The composition of any one of claims 11 to 13, wherein the disease and the target protein are one listed in Table 1.

15. The composition of any one of claims 1 to 14, wherein the first RNA molecule further comprises one or both of a downstream intronic splice enhancer (DISE) 3' to the splice donor and 5' to

the first dimerization domain, an intronic splice enhancer (ISE) 3' to the splice donor and 5' to the first dimerization domain; and/or

the second RNA molecule further comprises one or both of an ISE 3' to the second dimerization domain and 5' to the branch point sequence, and a DISE 3' to the splice donor and 5' to the dimerization domain;

or any combination thereof.

16. The composition of any one of claims 1 to 15, wherein

the first RNA molecule further comprises a self-cleaving RNA sequence or an RNA-cleaving enzyme target sequence positioned anywhere 3' to the splice donor such that it cleaves off the 3' located polyadenylated tail to decrease or suppress protein fragment expression from a non-recombined RNA molecule;

the second RNA molecule further comprises a self-cleaving RNA sequence or an RNA-cleaving enzyme target sequence positioned anywhere 5' to the branch point sequence such that it cleaves off the 5' located RNA cap to decrease or suppress protein fragment expression from a non-recombined RNA molecule;

the second RNA molecule further comprises a start codon anywhere 5' to the branch point sequence that is shifted relative to the open reading frame 3' of the splice acceptor to decrease or suppress translation of a target protein fragment from a non-recombined RNA molecule;

the first RNA molecule further comprises a micro RNA target site anywhere 3' to the splice donor such that an un-joined RNA fragment undergoes micro RNA dependent degradation once outside the nucleus;

the second RNA molecule further comprises a micro RNA target site anywhere 3' to the coding sequence such that an un-joined RNA fragment undergoes micro RNA dependent degradation once outside the nucleus;

the first RNA molecule further comprises a sequence encoding a degron protein degradation tag anywhere 3' to the splice donor such that it is in frame with the target protein open reading frame 5' of the splice donor site such that an un-joined protein fragment is tagged for degradation;

the second RNA molecule further comprises a start codon and an in-frame degron protein degradation tag anywhere 5' to the branch point sequence such that it is in frame with the target protein open reading frame 3' of the splice acceptor site such that an un-joined protein fragment is tagged for degradation;

or any combination thereof.

17. A composition for expressing a target protein comprising: (a) a first synthetic DNA molecule that encodes the first RNA molecule of any one of claims 1 to 16, wherein the first synthetic DNA molecule comprises (i) a first promoter operably linked to a sequence encoding the first RNA

molecule; and (b) a second synthetic DNA molecule that encodes the second RNA molecule of any one of claims 1 to 16, wherein the second synthetic DNA molecule comprises (i) a second promoter operably linked to a sequence encoding the second RNA molecule.

18. The composition of claim 17, wherein each promoter is independently selected.

19. The composition of claim 18 or 19, wherein:

the first and second promoter are the same promoter; or

the first and second promoter are different promoters.

20. The composition of any one of claims 17 to 19, wherein each of the first and second promoters is independently selected from: a constitutive promoter; a tissue-specific promoter; and a promoter endogenous to the target protein.

21. A system for expressing a target protein comprising a composition of any one of claims 17 to 20.

22. The system of claim 21, wherein when the system is introduced into a cell the RNA molecules are produced and recombine in the proper order, resulting in a full-length coding sequence of the target protein.

23. The system of claim 21 or 22, wherein each of the synthetic first and second RNA molecules are transcribed from a separate viral vector.

24. The system of any one of claims 21 to 23, wherein the viral vector is AAV.

25. The system of any one of claims 21 to 24, wherein each of the synthetic DNA molecules has a size independently selected from: about 2500 nt to about 5000 nt, 2,500 nt to about 2,750 nt, about 2,500 nt to about 3,000 nt, about 2,500 nt to about 3,250 nt, about 2,500 nt to about 3,500 nt, about 2,500 nt to about 3,750 nt, about 2,500 nt to about 4,000 nt, about 2,500 nt to about 4,250 nt, about 2,500 nt to about 4,500 nt, about 2,500 nt to about 4,750 nt, about 2,500 nt to about 5,000 nt, about 2,750 nt to about 3,000 nt, about 2,750 nt to about 3,250 nt, about 2,750 nt to about 3,500 nt, about 2,750 nt to about 3,750 nt, about 2,750 nt to about 4,000 nt, about 2,750 nt to about 4,250 nt, about 2,750 nt to about 4,500 nt, about 2,750 nt to about 4,750 nt, about 2,750 nt to about 5,000 nt, about 3,000 nt to about 3,250 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 3,750 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,250 nt, about 3,000 nt to about 4,500 nt, about 3,000 nt to about 4,750 nt, about 3,000 nt to about 5,000 nt, about 3,250 nt to about 3,500 nt, about 3,250 nt to about 3,750 nt, about 3,250 nt to about 4,000 nt, about 3,250 nt to about 4,250 nt, about 3,250 nt to about 4,500 nt, about 3,250 nt to about 4,750 nt, about 3,250 nt to about 5,000 nt, about 3,500 nt to about 3,750 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,250 nt, about 3,500 nt to about 4,500 nt, about 3,500 nt to about 4,750 nt, about 3,500 nt to about 5,000 nt, about 3,750 nt to about 4,000 nt, about 3,750 nt to about 4,250 nt, about 3,750 nt to about 4,500 nt, about 3,750 nt to about 4,750 nt, about 3,750 nt to about 5,000 nt, about 4,000 nt to about 4,250 nt, about

4,000 nt to about 4,500 nt, about 4,000 nt to about 4,750 nt, about 4,000 nt to about 5,000 nt, about 4,250 nt to about 4,500 nt, about 4,250 nt to about 4,750 nt, about 4,250 nt to about 5,000 nt, about 4,500 nt to about 4,750 nt, about 4,500 nt to about 5,000 nt, about 4,750 nt to about 5,000 nt, about 2,500 nt, about 2,750 nt, about 3,000 nt, about 3,250 nt, about 3,500 nt, about 3,750 nt, about 4,000 nt, about 4,250 nt, about 4,500 nt, about 4,750 nt, and about 5,000 nt.

26. The system of any one of claims 21 to 25, wherein the coding sequence for an N-terminal portion of the target protein, or a C-terminal portion of the target protein encoded by a synthetic DNA molecule of the system each has a size independently selected from: about 2500 to 4500 nt, about 2,500 nt to about 2,750 nt, about 2,500 nt to about 3,000 nt, about 2,500 nt to about 3,250 nt, about 2,500 nt to about 3,500 nt, about 2,500 nt to about 3,750 nt, about 2,500 nt to about 4,000 nt, about 2,500 nt to about 4,250 nt, about 2,500 nt to about 4,500 nt, about 2,750 nt to about 3,000 nt, about 2,750 nt to about 3,250 nt, about 2,750 nt to about 3,500 nt, about 2,750 nt to about 3,750 nt, about 2,750 nt to about 4,000 nt, about 2,750 nt to about 4,250 nt, about 2,750 nt to about 4,500 nt, about 3,000 nt to about 3,250 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 3,750 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,250 nt, about 3,000 nt to about 4,500 nt, about 3,250 nt to about 3,500 nt, about 3,250 nt to about 3,750 nt, about 3,250 nt to about 4,000 nt, about 3,250 nt to about 4,250 nt, about 3,250 nt to about 4,500 nt, about 3,500 nt to about 3,750 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,250 nt, about 3,500 nt to about 4,500 nt, about 3,750 nt to about 4,000 nt, about 3,750 nt to about 4,250 nt, about 3,750 nt to about 4,500 nt, about 4,000 nt to about 4,250 nt, about 4,000 nt to about 4,500 nt, about 4,250 nt to about 4,500 nt, about 2,500 nt, about 2,750 nt, about 3,000 nt, about 3,250 nt, about 3,500 nt, about 3,750 nt, about 4,000 nt, about 4,250 nt, and about 4,500 nt.

27. The system of any one of claims 21 to 26, wherein any one or both of the RNA molecules encoded by the synthetic DNA molecules of the system, respectively, has a size independently selected from: about 2500 to 4500 nt, about 2,500 nt to about 2,750 nt, about 2,500 nt to about 3,000 nt, about 2,500 nt to about 3,250 nt, about 2,500 nt to about 3,500 nt, about 2,500 nt to about 3,750 nt, about 2,500 nt to about 4,000 nt, about 2,500 nt to about 4,250 nt, about 2,500 nt to about 4,500 nt, about 2,750 nt to about 3,000 nt, about 2,750 nt to about 3,250 nt, about 2,750 nt to about 3,500 nt, about 2,750 nt to about 3,750 nt, about 2,750 nt to about 4,000 nt, about 2,750 nt to about 4,250 nt, about 2,750 nt to about 4,500 nt, about 3,000 nt to about 3,250 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 3,750 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,250 nt, about 3,000 nt to about 4,500 nt, about 3,250 nt to about 3,500 nt, about 3,250 nt to about 3,750 nt, about 3,250 nt to about 4,000 nt, about 3,250 nt to about 4,250 nt, about 3,250 nt to about 4,500 nt, about 3,500 nt to about 3,750 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,250 nt, about 3,500 nt to about 4,500 nt, about 3,750 nt to about 4,000 nt, about 3,750 nt to about 4,250 nt, about

3,750 nt to about 4,500 nt, about 4,000 nt to about 4,250 nt, about 4,000 nt to about 4,500 nt, about 4,250 nt to about 4,500 nt, about 2,500 nt, about 2,750 nt, about 3,000 nt, about 3,250 nt, about 3,500 nt, about 3,750 nt, about 4,000 nt, about 4,250 nt, and about 4,500 nt.

28. The system of any one of claims 21 to 27, wherein the system comprises a composition of any one of claims 17 to 20:

the synthetic DNA molecules have a total size selected from about 5000 nt to about 10,000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,000 nt to about 9,500 nt, about 5,000 nt to about 10,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 5,500 nt to about 8,500 nt, about 5,500 nt to about 9,000 nt, about 5,500 nt to about 9,500 nt, about 5,500 nt to about 10,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,000 nt to about 9,500 nt, about 6,000 nt to about 10,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 6,500 nt to about 8,500 nt, about 6,500 nt to about 9,000 nt, about 6,500 nt to about 9,500 nt, about 6,500 nt to about 10,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,000 nt to about 9,500 nt, about 7,000 nt to about 10,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 7,500 nt to about 9,500 nt, about 7,500 nt to about 10,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,000 nt to about 9,500 nt, about 8,000 nt to about 10,000 nt, about 8,500 nt to about 9,000 nt, about 8,500 nt to about 9,500 nt, about 8,500 nt to about 10,000 nt, about 9,000 nt to about 9,500 nt, about 9,000 nt to about 10,000 nt, about 9,500 nt to about 10,000 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, about 9,000 nt, about 9,500 nt, and about 10,000 nt;

the total target protein coding sequence is selected from about 2000 nt to about 8000 nt, about 2,000 nt to about 3,000 nt, about 2,000 nt to about 3,500 nt, about 2,000 nt to about 4,000 nt, about 2,000 nt to about 4,500 nt, about 2,000 nt to about 5,000 nt, about 2,000 nt to about 5,500 nt, about 2,000 nt to about 6,000 nt, about 2,000 nt to about 6,500 nt, about 2,000 nt to about 7,000 nt, about 2,000 nt to about 7,500 nt, about 2,000 nt to about 8,000 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,500 nt, about 3,000 nt to about 5,000 nt, about 3,000 nt to about 5,500 nt, about 3,000 nt to about 6,000 nt, about 3,000 nt to about 6,500 nt, about 3,000 nt to about 7,000 nt, about 3,000 nt to about 7,500 nt, about 3,000 nt to about 8,000 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,500 nt, about 3,500 nt to about 5,000 nt, about

3,500 nt to about 5,500 nt, about 3,500 nt to about 6,000 nt, about 3,500 nt to about 6,500 nt, about 3,500 nt to about 7,000 nt, about 3,500 nt to about 7,500 nt, about 3,500 nt to about 8,000 nt, about 4,000 nt to about 4,500 nt, about 4,000 nt to about 5,000 nt, about 4,000 nt to about 5,500 nt, about 4,000 nt to about 6,000 nt, about 4,000 nt to about 6,500 nt, about 4,000 nt to about 7,000 nt, about 4,000 nt to about 7,500 nt, about 4,000 nt to about 8,000 nt, about 4,500 nt to about 5,000 nt, about 4,500 nt to about 5,500 nt, about 4,500 nt to about 6,000 nt, about 4,500 nt to about 6,500 nt, about 4,500 nt to about 7,000 nt, about 4,500 nt to about 7,500 nt, about 4,500 nt to about 8,000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, or about 7,500 nt to about 8,000 nt. the total target protein coding sequence is about 2,000 nt, about 3,000 nt, about 3,500 nt, about 4,000 nt, about 4,500 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, and about 8,000 nt; and/or

the summed size of the RNA molecules encoded by the two synthetic DNA molecules is selected from about 5,000 nt to about 9000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 5,500 nt to about 8,500 nt, about 5,500 nt to about 9,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 6,500 nt to about 8,500 nt, about 6,500 nt to about 9,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,500 nt to about 9,000 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, and about 9,000 nt.

29. The system of any one of claims 21 to 28, wherein the first dimerization domain and the second dimerization domain are each no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50

to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, at least about 95%, or about 100%.

30. The system of any one of claims 21 to 29, wherein each dimerization domain is no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50 to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, at least 90%, or about 100%.

31. The system of any one of claims 21 to 30, wherein the RNA recombination efficiency is about 10% to about 100%, about 10% to about 20%, about 10% to about 30%, about 10% to about 35%, about 10% to about 40%, about 10% to about 45%, about 10% to about 50%, about 10% to about 55%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 20% to about 30%, about 20% to about 35%, about 20% to about 40%, about 20% to about 45%, about 20% to about 50%, about 20% to about 55%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 30% to about 35%, about 30% to about 40%, about 30% to about 45%, about 30% to about 50%, about 30% to about 55%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 35% to about 40%, about 35% to about 45%, about 35% to about 50%, about 35% to about 55%, about 35% to about 60%, about 35% to about 70%, about 35% to about 80%, about 35% to about 90%, about 40% to about 45%, about 40% to about 50%, about 40% to about 55%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about 40% to about 90%, about 45% to about 50%, about 45% to about 55%, about 45% to about 60%, about 45% to about 70%, about 45% to about 80%, about 45% to about 90%, about 50% to about 55%, about 50% to about 60%, about 50% to about 70%, about 50% to about 80%, about 50% to about 90%, about 55% to about 60%, about 55% to about 70%, about 55% to about 80%, about 55% to about 90%, about 60% to about 70%, about 60% to about 80%, about 60% to about 90%, about 70% to about 80%, about 70% to about 90%, about 80% to about 90%, about 10%, about 20%, about 30%, about 35%, about 40%, about 45%, about 50%, about 55%, about 60%, about 70%, about 80%, about 90%, about 95%, or about 100%.

32. A composition comprising a system of any one of claims 21 to 31.

33. The composition of claim 32, wherein the composition comprises first, second, third and optionally fourth RNA molecules, each encoding at least a portion of dystrophin, factor 8, ABCA4, or MYO7A.

34. A kit comprising the system of any one of claims 21 to 31, or composition of any one of claims 44 and 45, wherein any of the synthetic first, second, third and fourth nucleic acid molecules can be in separate containers, and optionally further comprising a buffer such as a pharmaceutically acceptable carrier.

35. A method of expressing a target protein in a cell, comprising:

introducing the system of any one of claims 21 to 31, or a composition of claim 32 or 33, into a cell, and expressing the first and second RNA molecules in the cell, wherein the target protein is produced in the cell.

36. The method of claim 35, wherein the cell is in a subject, and introducing comprises administering a therapeutically effective amount the system to the subject.

37. The method of claim 36, wherein the method treats a genetic disease caused by a mutation in a gene encoding the target protein in the subject, wherein the method results in expression of functional target protein in the subject.

38. The method of claim 37, wherein

the genetic disease is Duchenne muscular dystrophy and the target protein is dystrophin;

the genetic disease is Hemophilia A and the target protein is F8;

the genetic disease is Stargardt disease and the target protein is ABCA4; or

the genetic disease is Usher syndrome and the target protein is MYO7A.

39. A system of any one of claims 21 to 31, a composition of any one of claims 1 to 16, 32 and 33, or a method of any one of claims 35 to 38, wherein one or both of the first and second RNA molecules comprise at least 80%, at least 85%, at least 90%, at least 95%, at least 98%, at least 99%, or 100% sequence identity to a synthetic intron provided in any one of SEQ ID NOS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 145, 146, 147, 148, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, and 166.

40. A system of any one of claims 21 to 31 and 39, a composition of any one of claims 1 to 16, 32 and 33, or a method of any one of claims 35 to 38, or a method of any one of claims 35 to 38, wherein one or both of the first and second RNA molecules comprise a synthetic intron selected from: nt 3703 to 3975 of SEQ ID NO: 20, nt 1 to 228 of SEQ ID NO: 21, nt 3703 to 3975 of SEQ ID NO: 22, nt 1 to 225 of SEQ ID NO: 23, nt 3560 to 3828 of SEQ ID NO: 24, and nt 1-225 of SEQ ID NO: 25.

41. A system of any of claims 21 to 31, 39, and 40, a composition of any one of claims 1 to 16, 32 and 33, or a method of any one of claims 35 to 38, or a method of any one of claims 35 to 38, wherein one or both of the first and second RNA molecules further comprise a portion of a protein coding sequence.

42. A system of any of claims 21 to 31 and 39 to 41, a composition of any one of claims 1 to 16, 32 and 33, or a method of any one of claims 35 to 38, or a method of any one of claims 35 to 38,



wherein the portion of the protein coding sequence comprises an N-terminal half, an N-terminal portion, a C-terminal half, or a C-terminal portion, of the protein coding sequence.

43. A system of any one of claims 21 to 31 and 39 to 42, a composition of any one of claims 1 to 16, 32 and 33, or a method of any one of claims 35 to 38, comprising: (a) a first RNA molecule, the RNA molecule comprising from 5' to 3': (i) a coding sequence for an N-terminal portion of the target protein; (ii) a splice donor; (ii-2) a DISE, an ISE, or both; and (iii) a first dimerization domain; and (b) a second RNA molecule, the RNA molecule comprising from 5' to 3': (i) a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; (i-2) at least one ISE sequences; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice acceptor; and (v) a coding sequence for a C-terminal portion of the target protein.

44. A system of any one of claims 21 to 31 and 39 to 43, a composition of any one of claims 1 to 16, 32 and 33, or a method of any one of claims 35 to 38, comprising: (a) a first RNA molecule, the RNA molecule comprising from 5' to 3': (i) a coding sequence for an N-terminal portion of the target protein; (ii) a splice donor; (ii-2) a DISE, an ISE, and an ISE; and (iii) a first dimerization domain; and (b) a second RNA molecule, the RNA molecule comprising from 5' to 3': (i) a second dimerization domain, wherein the second dimerization domain binds to the first dimerization domain; (i-2) three ISE sequences; (ii) a branch point sequence; (iii) a polypyrimidine tract; (iv) a splice acceptor; and (v) a coding sequence for a C-terminal portion of the target protein.

45. The composition of any one of claims 1 to 16, wherein any one or two of the first and second RNA molecules each has a size independently selected from: about 2500 to 4500 nt, about 2,500 nt to about 2,750 nt, about 2,500 nt to about 3,000 nt, about 2,500 nt to about 3,250 nt, about 2,500 nt to about 3,500 nt, about 2,500 nt to about 3,750 nt, about 2,500 nt to about 4,000 nt, about 2,500 nt to about 4,250 nt, about 2,500 nt to about 4,500 nt, about 2,750 nt to about 3,000 nt, about 2,750 nt to about 3,250 nt, about 2,750 nt to about 3,500 nt, about 2,750 nt to about 3,750 nt, about 2,750 nt to about 4,000 nt, about 2,750 nt to about 4,250 nt, about 2,750 nt to about 4,500 nt, about 3,000 nt to about 3,250 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 3,750 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,250 nt, about 3,000 nt to about 4,500 nt, about 3,250 nt to about 3,500 nt, about 3,250 nt to about 3,750 nt, about 3,250 nt to about 4,000 nt, about 3,250 nt to about 4,250 nt, about 3,250 nt to about 4,500 nt, about 3,500 nt to about 3,750 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,250 nt, about 3,500 nt to about 4,500 nt, about 3,750 nt to about 4,000 nt, about 3,750 nt to about 4,250 nt, about 3,750 nt to about 4,500 nt, about 4,000 nt to about 4,250 nt, about 4,000 nt to about 4,500 nt, about 4,250 nt to about 4,500 nt, about 2,500 nt, about 2,750 nt, about 3,000 nt, about 3,250 nt, about 3,500 nt, about 3,750 nt, about 4,000 nt, about 4,250 nt, and about 4,500 nt.

46. The composition of any one of claims 1 to 16, wherein:

the total target protein coding sequence size is about 2000 nt to about 8000 nt, about 2,000 nt to about 3,000 nt, about 2,000 nt to about 3,500 nt, about 2,000 nt to about 4,000 nt, about 2,000 nt to about 4,500 nt, about 2,000 nt to about 5,000 nt, about 2,000 nt to about 5,500 nt, about 2,000 nt to about 6,000 nt, about 2,000 nt to about 6,500 nt, about 2,000 nt to about 7,000 nt, about 2,000 nt to about 7,500 nt, about 2,000 nt to about 8,000 nt, about 3,000 nt to about 3,500 nt, about 3,000 nt to about 4,000 nt, about 3,000 nt to about 4,500 nt, about 3,000 nt to about 5,000 nt, about 3,000 nt to about 5,500 nt, about 3,000 nt to about 6,000 nt, about 3,000 nt to about 6,500 nt, about 3,000 nt to about 7,000 nt, about 3,000 nt to about 7,500 nt, about 3,000 nt to about 8,000 nt, about 3,500 nt to about 4,000 nt, about 3,500 nt to about 4,500 nt, about 3,500 nt to about 5,000 nt, about 3,500 nt to about 5,500 nt, about 3,500 nt to about 6,000 nt, about 3,500 nt to about 6,500 nt, about 3,500 nt to about 7,000 nt, about 3,500 nt to about 7,500 nt, about 3,500 nt to about 8,000 nt, about 4,000 nt to about 4,500 nt, about 4,000 nt to about 5,000 nt, about 4,000 nt to about 5,500 nt, about 4,000 nt to about 6,000 nt, about 4,000 nt to about 6,500 nt, about 4,000 nt to about 7,000 nt, about 4,000 nt to about 7,500 nt, about 4,000 nt to about 8,000 nt, about 4,500 nt to about 5,000 nt, about 4,500 nt to about 5,500 nt, about 4,500 nt to about 6,000 nt, about 4,500 nt to about 6,500 nt, about 4,500 nt to about 7,000 nt, about 4,500 nt to about 7,500 nt, about 4,500 nt to about 8,000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to about 8,000 nt, about 7,500 nt to about 8,000 nt, about 2,000 nt, about 3,000 nt, about 3,500 nt, about 4,000 nt, about 4,500 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, or about 8,000 nt; and/or

the summed size of the two RNA molecules is about 5,000 nt to about 9000 nt, about 5,000 nt to about 5,500 nt, about 5,000 nt to about 6,000 nt, about 5,000 nt to about 6,500 nt, about 5,000 nt to about 7,000 nt, about 5,000 nt to about 7,500 nt, about 5,000 nt to about 8,000 nt, about 5,000 nt to about 8,500 nt, about 5,000 nt to about 9,000 nt, about 5,500 nt to about 6,000 nt, about 5,500 nt to about 6,500 nt, about 5,500 nt to about 7,000 nt, about 5,500 nt to about 7,500 nt, about 5,500 nt to about 8,000 nt, about 5,500 nt to about 8,500 nt, about 5,500 nt to about 9,000 nt, about 6,000 nt to about 6,500 nt, about 6,000 nt to about 7,000 nt, about 6,000 nt to about 7,500 nt, about 6,000 nt to about 8,000 nt, about 6,000 nt to about 8,500 nt, about 6,000 nt to about 9,000 nt, about 6,500 nt to about 7,000 nt, about 6,500 nt to about 7,500 nt, about 6,500 nt to about 8,000 nt, about 6,500 nt to about 8,500 nt, about 6,500 nt to about 9,000 nt, about 7,000 nt to about 7,500 nt, about 7,000 nt to

about 8,000 nt, about 7,000 nt to about 8,500 nt, about 7,000 nt to about 9,000 nt, about 7,500 nt to about 8,000 nt, about 7,500 nt to about 8,500 nt, about 7,500 nt to about 9,000 nt, about 8,000 nt to about 8,500 nt, about 8,000 nt to about 9,000 nt, about 8,500 nt to about 9,000 nt, about 5,000 nt, about 5,500 nt, about 6,000 nt, about 6,500 nt, about 7,000 nt, about 7,500 nt, about 8,000 nt, about 8,500 nt, or about 9,000 nt.

47. The composition of any one of claims 1 to 16, wherein the first dimerization domain and the second dimerization domain are each no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50 to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least about 20%, at least about 25%, at least about 30%, at least about 35%, at least about 40%, at least about 45%, at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, or at least about 95%.

48. The composition of any one of claims 1 to 16, wherein each dimerization domain is no more than 1000 nt, such as at least 50 nt, at least 100 nt, at least 150 nt, at least 200 nt, at least 300 nt, at least 400 nt, at least 500 nt, 50 to 1000 nt, 50 to 500 nt, 50 to 150 nt, 50, 100, 150, 200, 250, 300, 400, or 500 nt; and the system has a recombination efficiency of at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, or at least 90%.

49. The composition of any one of claims 1 to 16, wherein the RNA recombination efficiency is about 10% to about 100%, about 10% to about 20%, about 10% to about 30%, about 10% to about 35%, about 10% to about 40%, about 10% to about 45%, about 10% to about 50%, about 10% to about 55%, about 10% to about 60%, about 10% to about 70%, about 10% to about 80%, about 10% to about 90%, about 20% to about 30%, about 20% to about 35%, about 20% to about 40%, about 20% to about 45%, about 20% to about 50%, about 20% to about 55%, about 20% to about 60%, about 20% to about 70%, about 20% to about 80%, about 20% to about 90%, about 30% to about 35%, about 30% to about 40%, about 30% to about 45%, about 30% to about 50%, about 30% to about 55%, about 30% to about 60%, about 30% to about 70%, about 30% to about 80%, about 30% to about 90%, about 35% to about 40%, about 35% to about 45%, about 35% to about 50%, about 35% to about 55%, about 35% to about 60%, about 35% to about 70%, about 35% to about 80%, about 35% to about 90%, about 40% to about 45%, about 40% to about 50%, about 40% to about 55%, about 40% to about 60%, about 40% to about 70%, about 40% to about 80%, about 40% to about 90%, about 45% to about 50%, about 45% to about 55%, about 45% to about 60%, about 45% to about 70%, about 45% to about 80%, about 45% to about 90%, about 50% to about 55%, about 50% to about 60%, about 50% to about 70%, about 50% to about 80%, about 50% to about 90%, about 55% to about 60%, about 55% to about 70%, about 55% to about 80%, about 55% to about 90%, about 60% to about 70%, about 60% to about 80%, about 60% to about

90%, about 70% to about 80%, about 70% to about 90%, about 80% to about 90%, about 10%, about 20%, about 30%, about 35%, about 40%, about 45%, about 50%, about 55%, about 60%, about 70%, about 80%, about 90%, about 95%, or about 100%.

50. The composition of any one of claims 1 to 16, wherein:

- 5 (a) the first and second RNA molecules are each about 2500 nt to 4500 nt;  
(b) the total target protein coding sequence size is about 2000 nt to about 8000 nt; and/or  
(c) the summed size of the two RNA molecules is about 5,000 nt to about 9000 nt;  
and the RNA recombination efficiency is about 10% to about 100%.

10

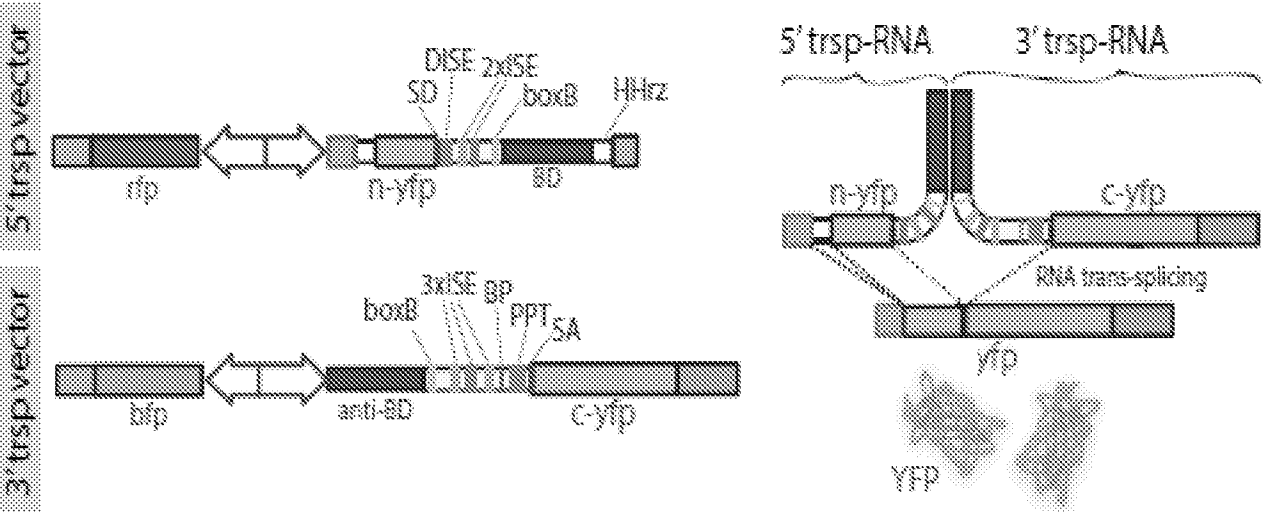


FIG. 1A

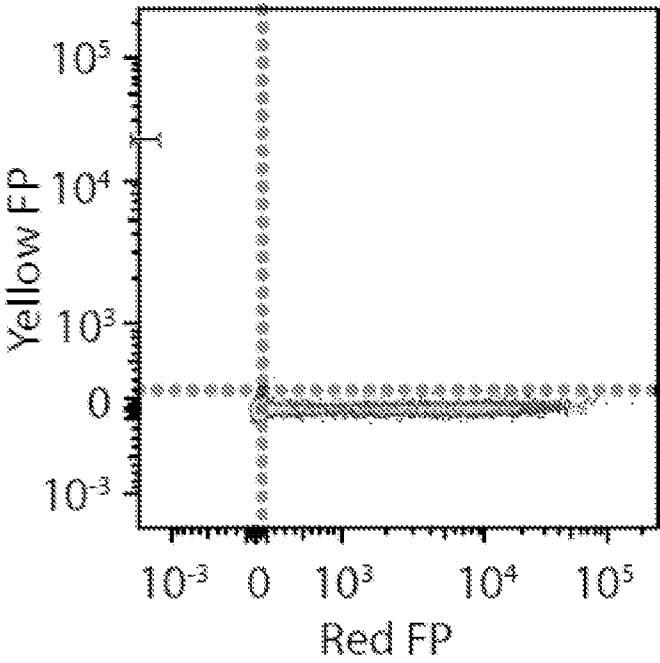


FIG. 1B

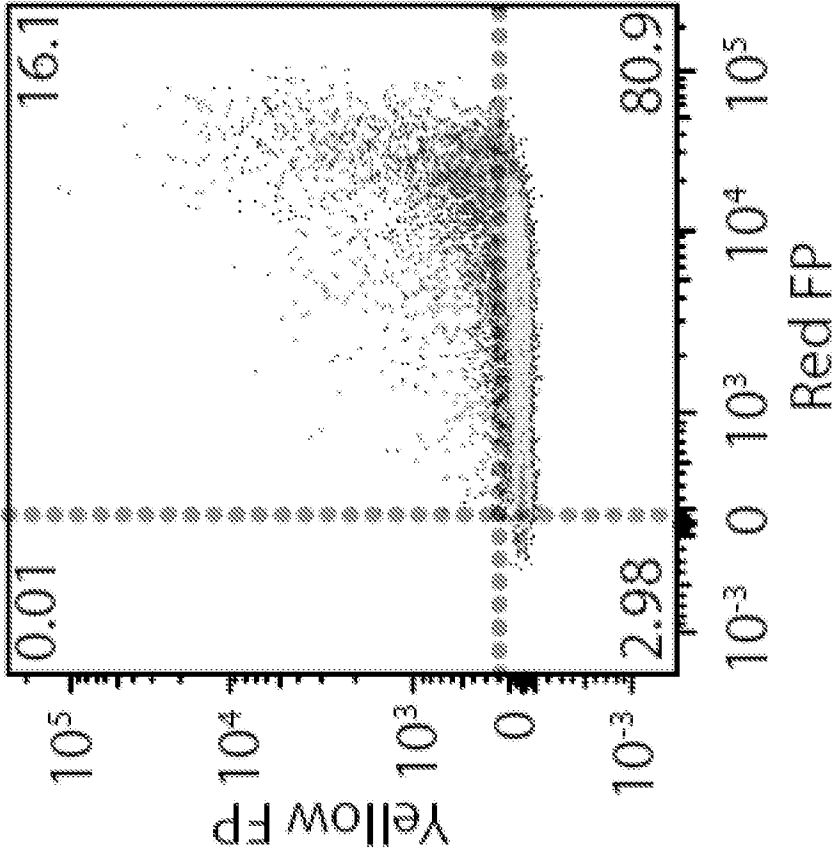


FIG. 1D

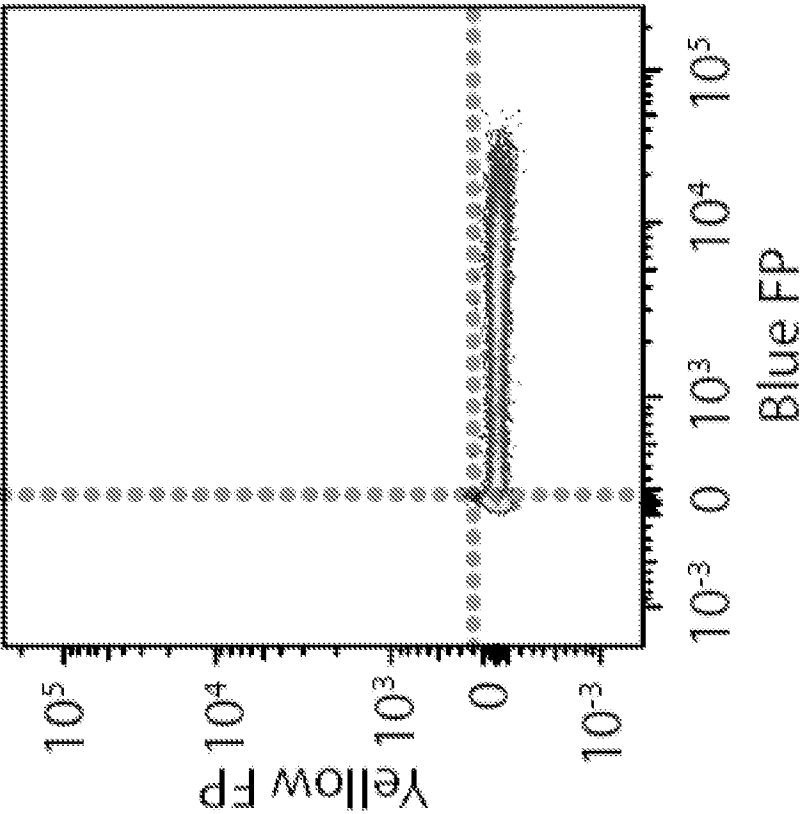


FIG. 1C

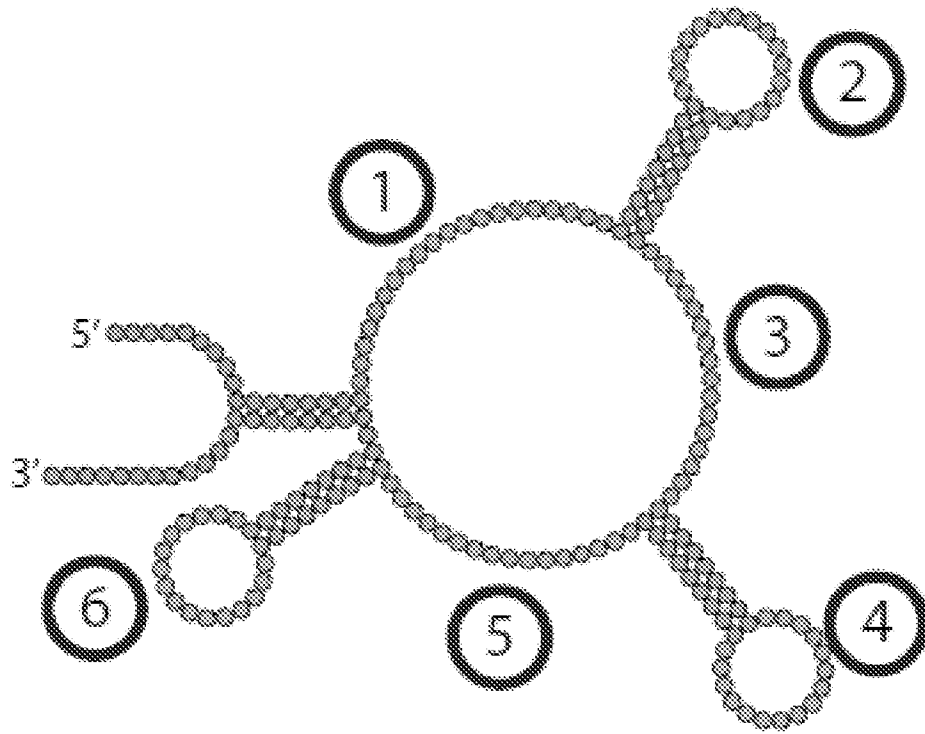


**FIG. 1E**

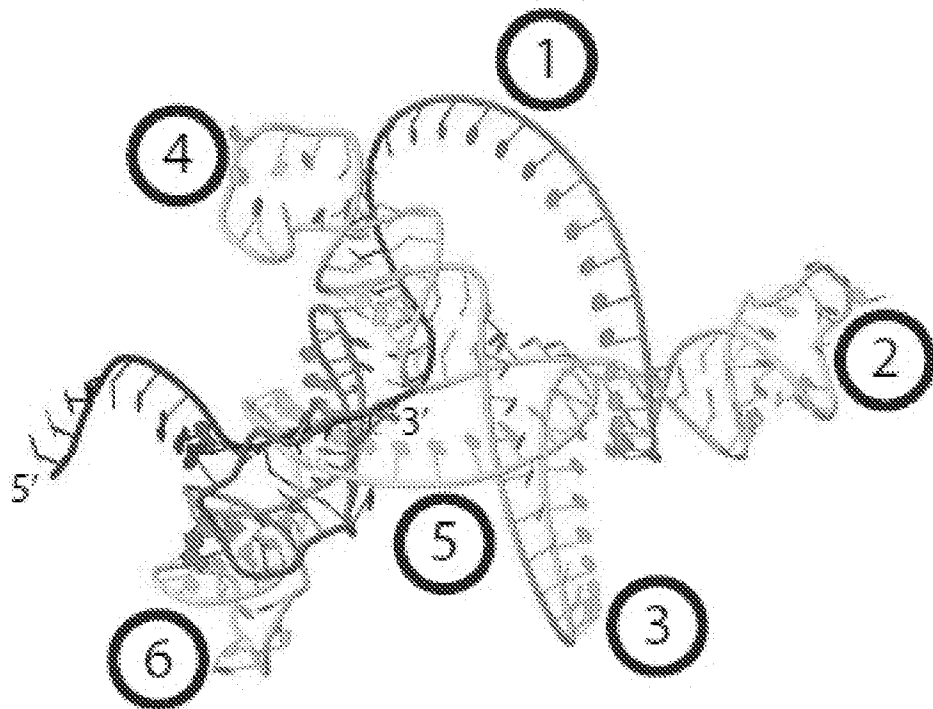


**FIG. 1F**

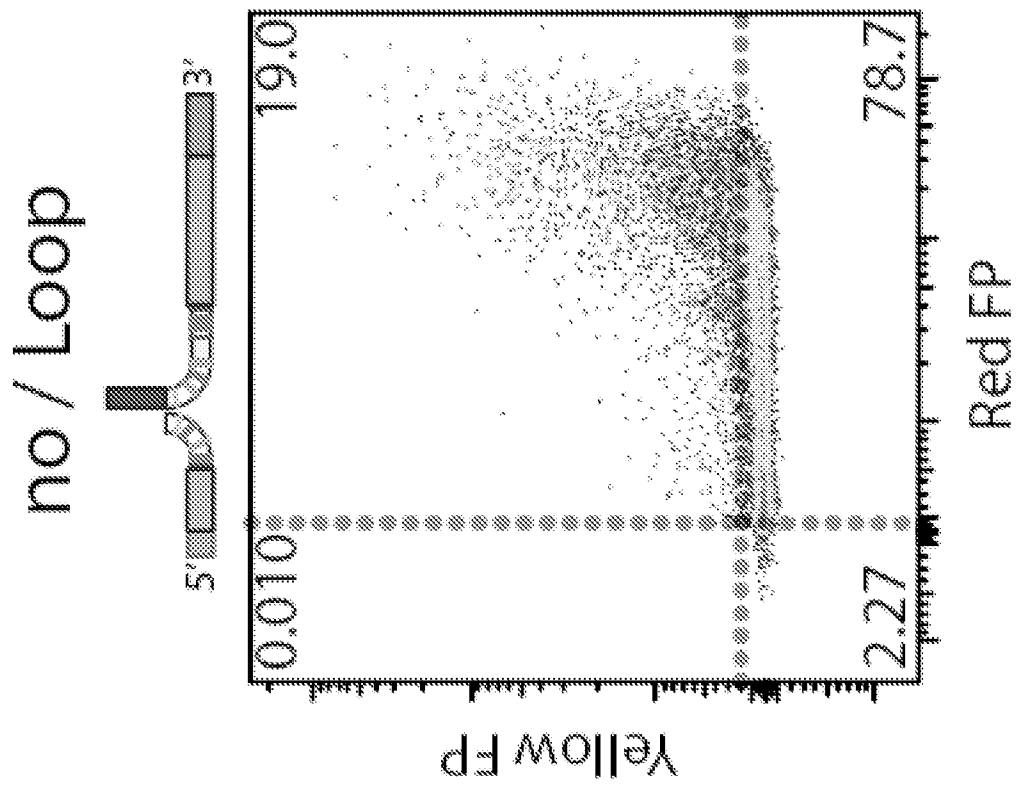


FIG. 1H

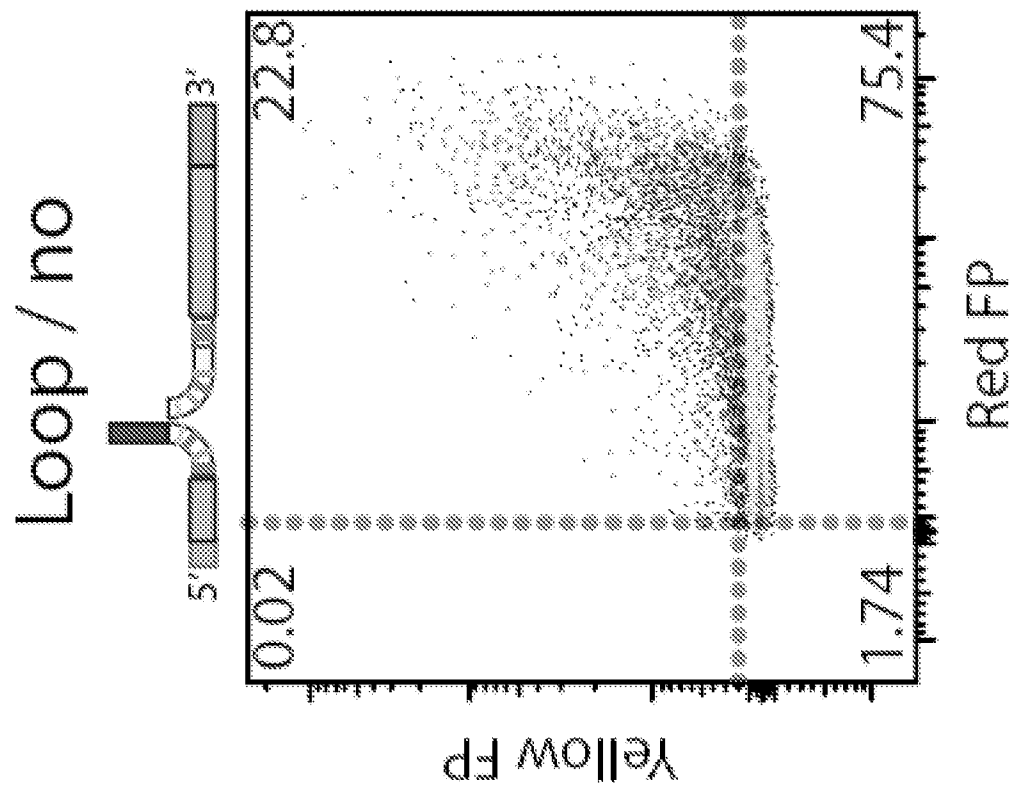


FIG. 1G



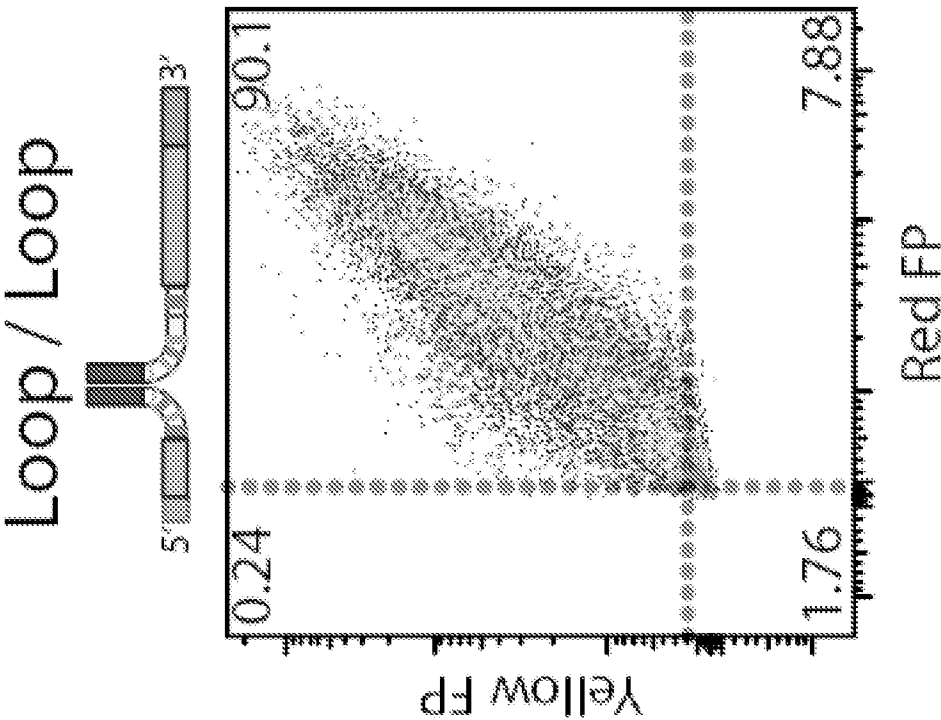


FIG. 1I

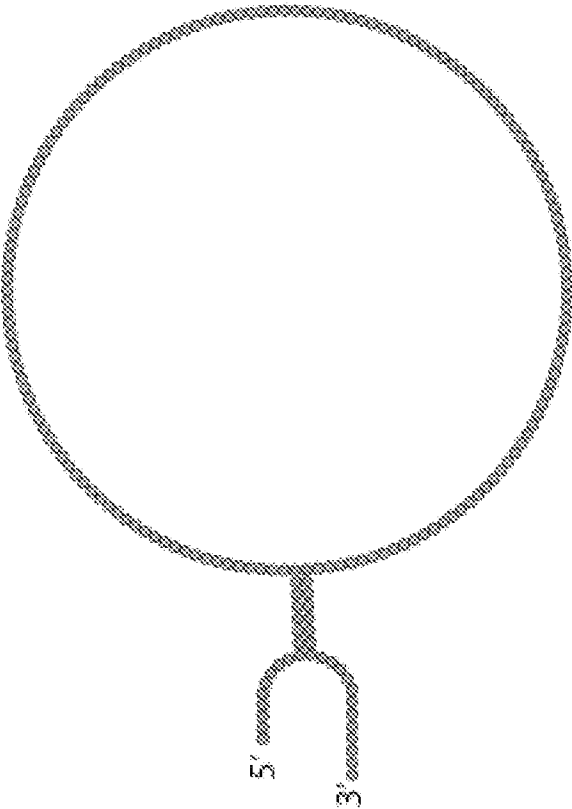


FIG. 1J

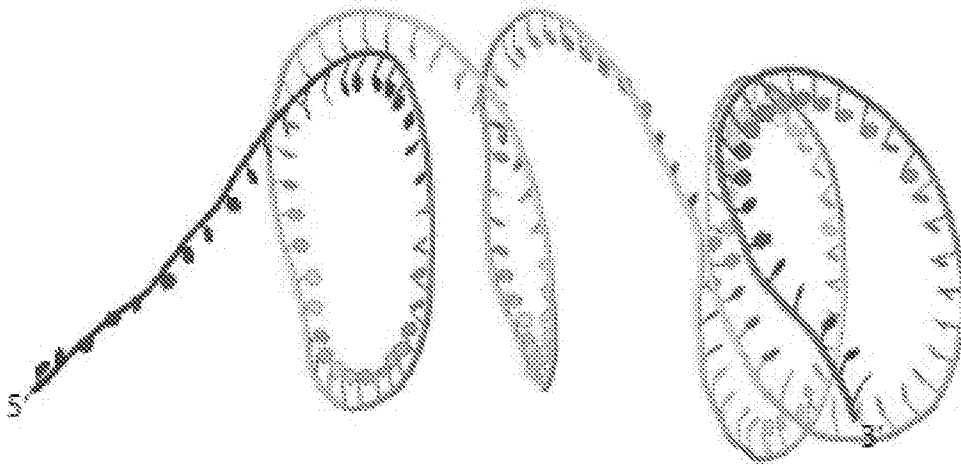


**FIG. 1K**

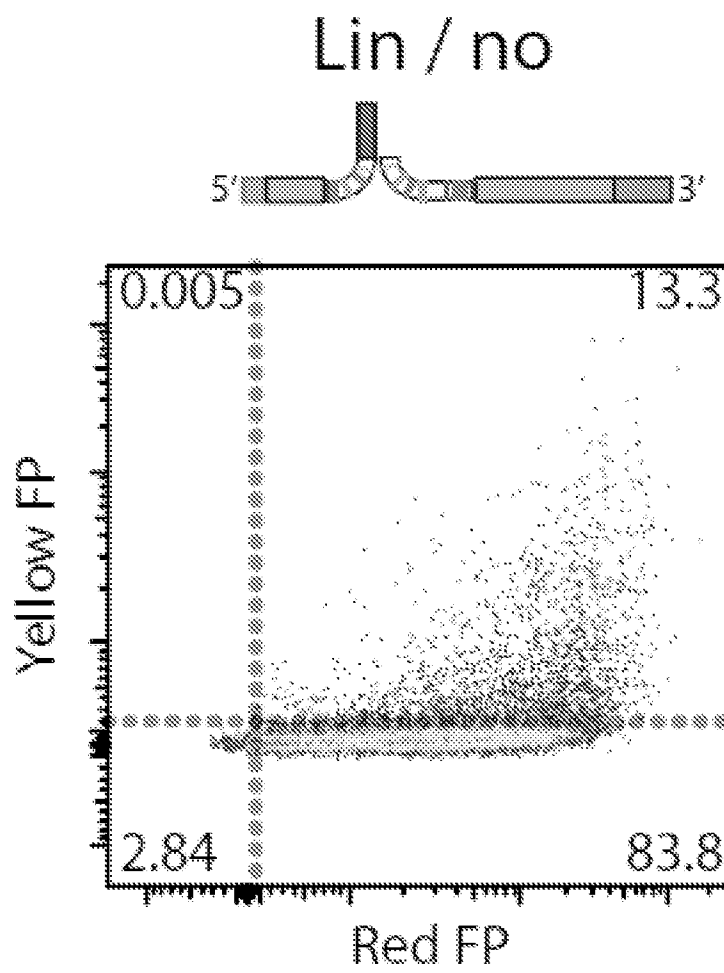


**FIG. 1L**

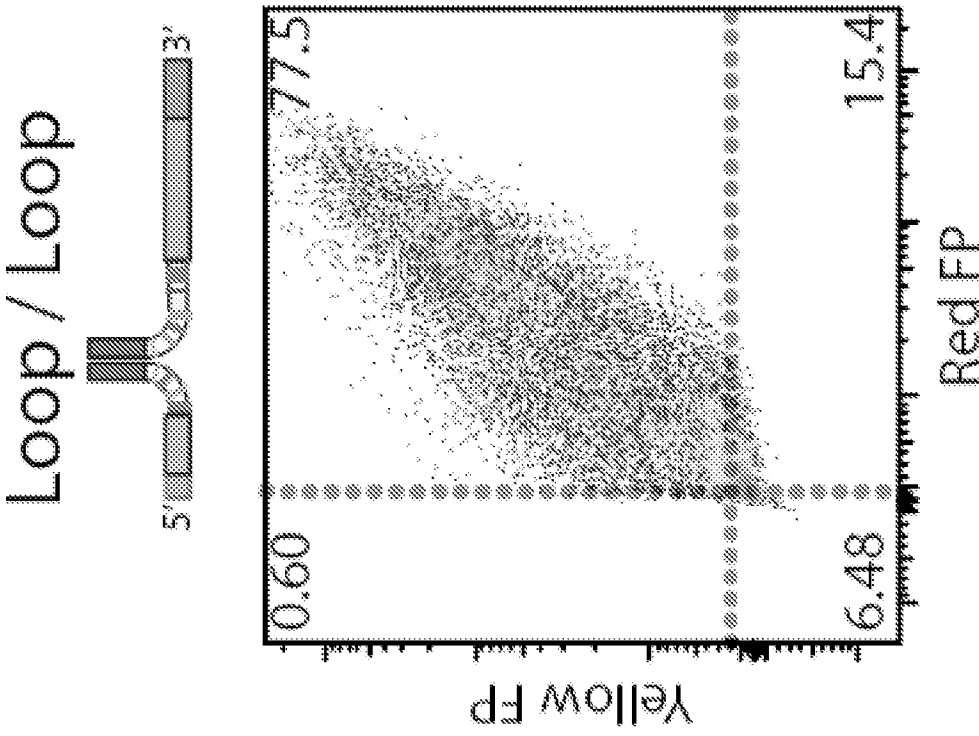


FIG. 1M

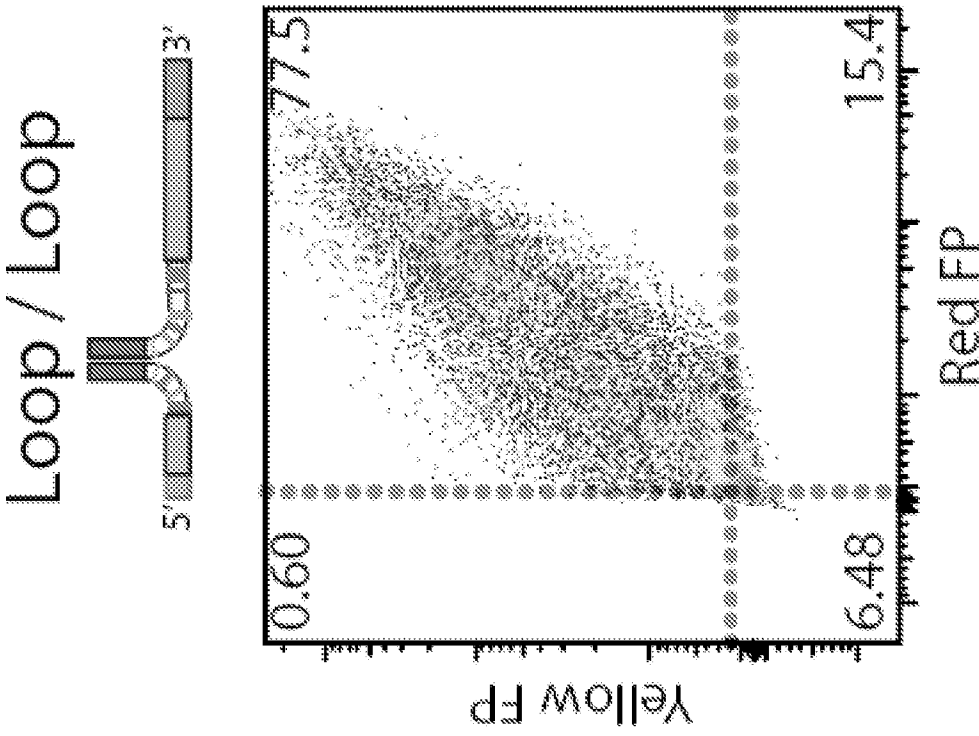


FIG. 1N

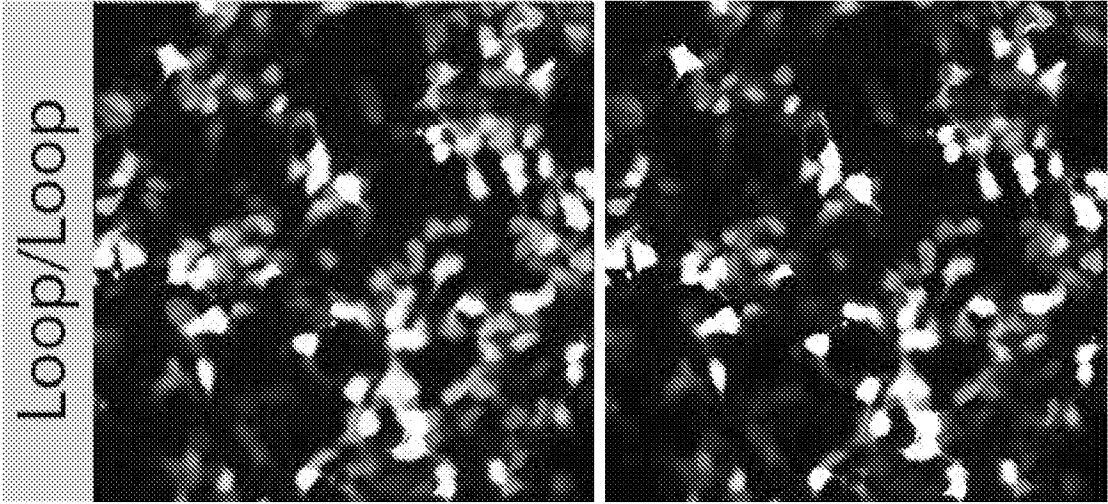


FIG. 1P

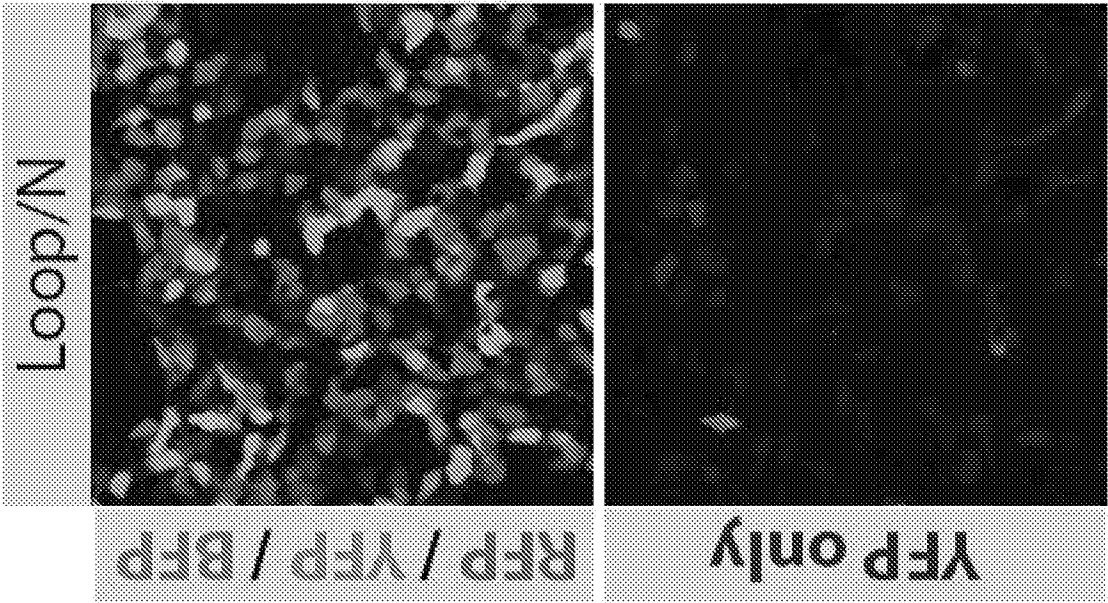


FIG. 1O

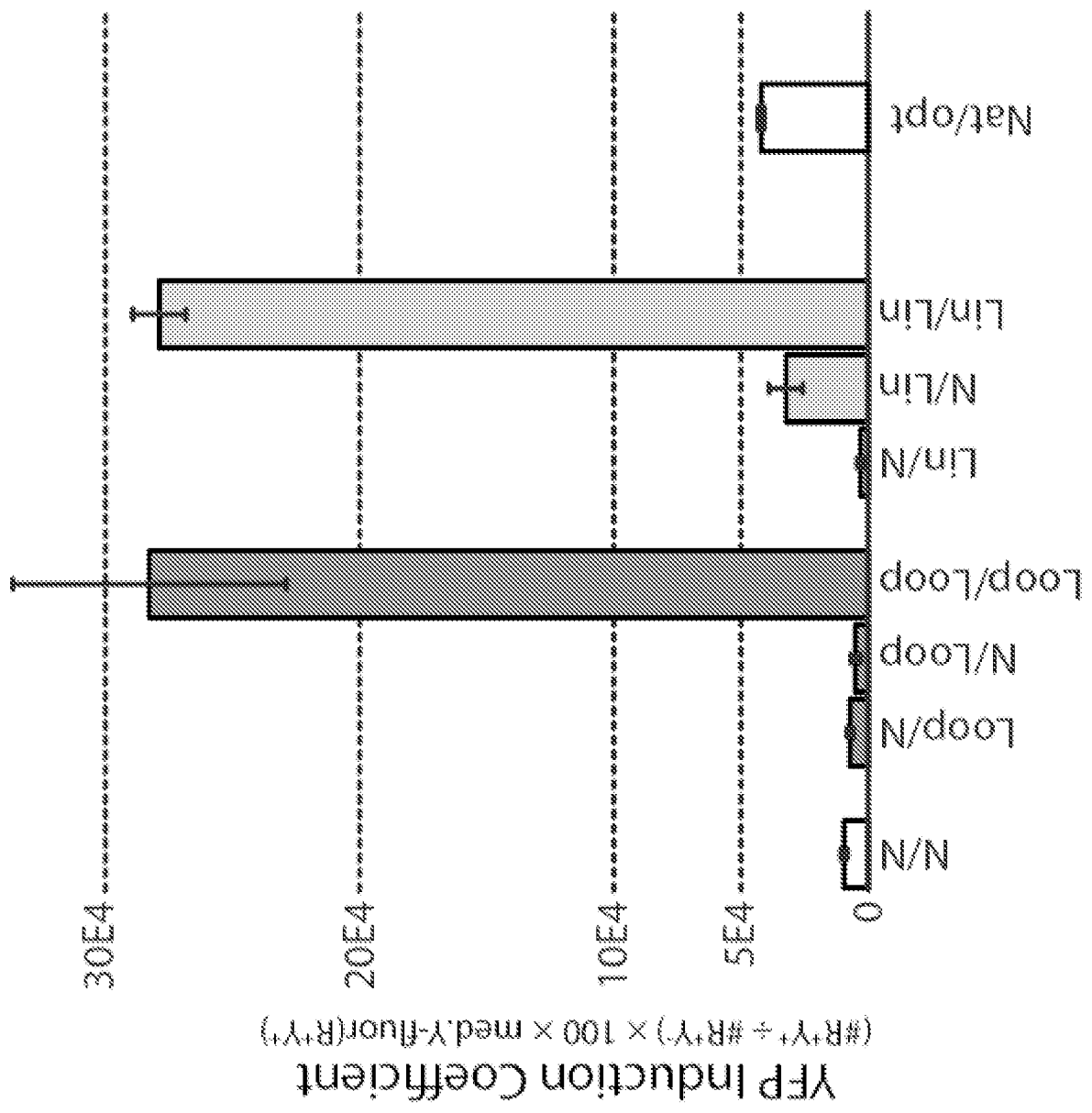


FIG. 1Q

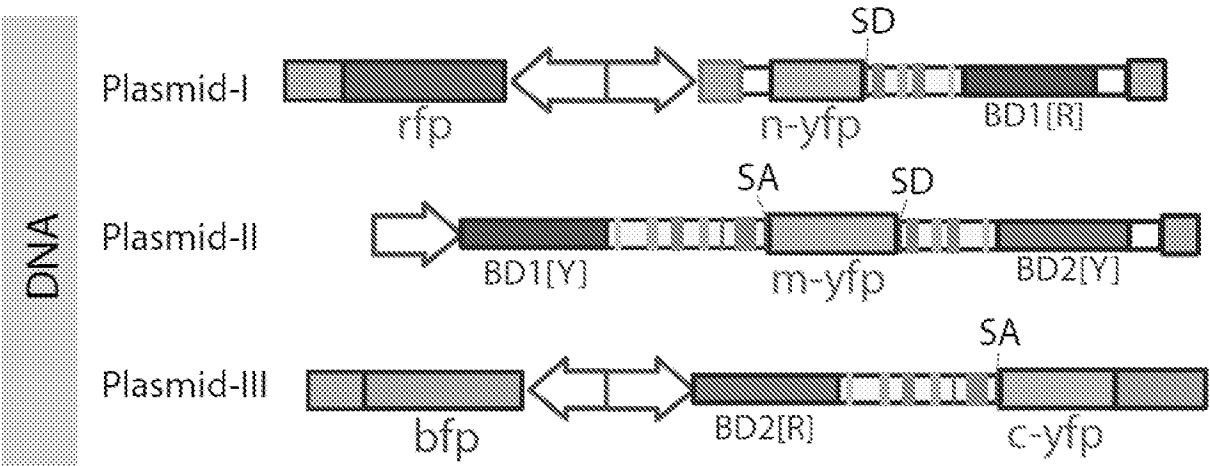


FIG. 2A

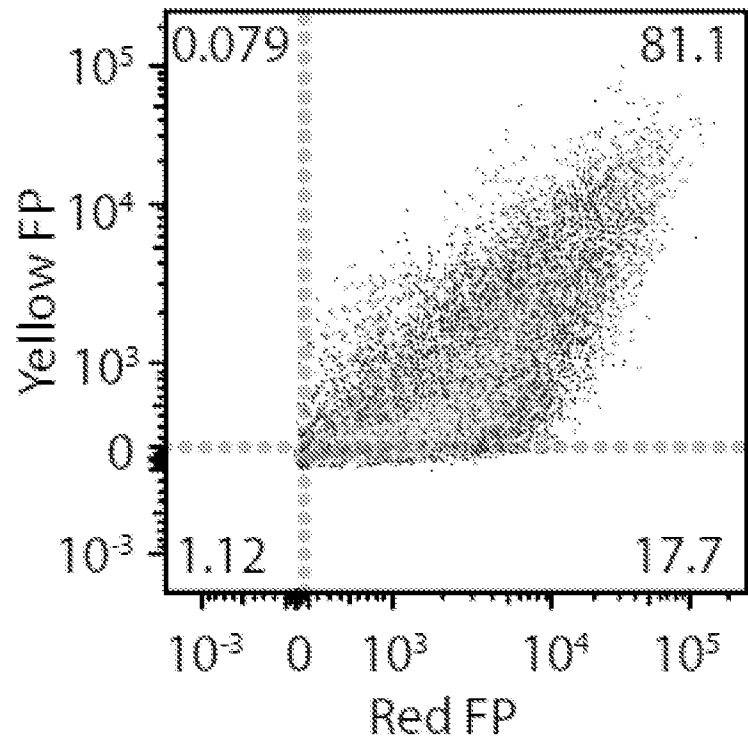


FIG. 2B

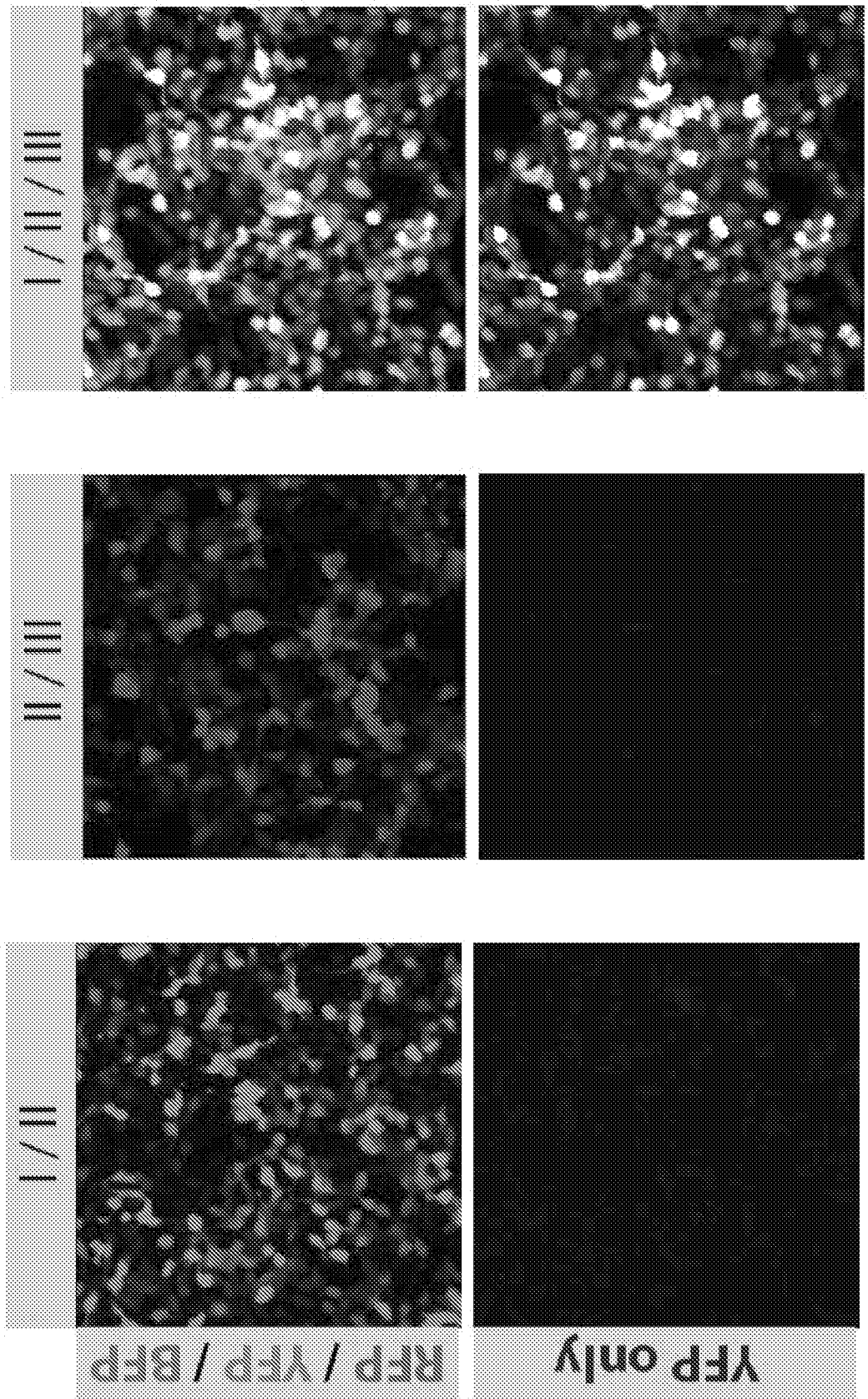


FIG. 2E

FIG. 2D

FIG. 2C

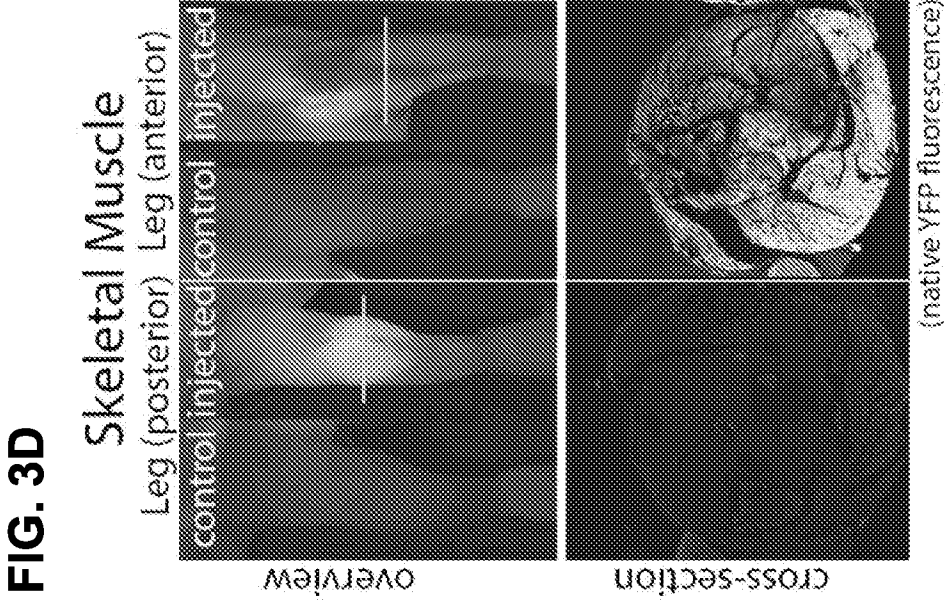
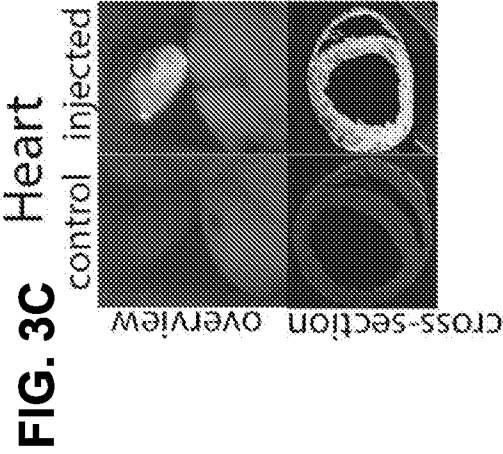
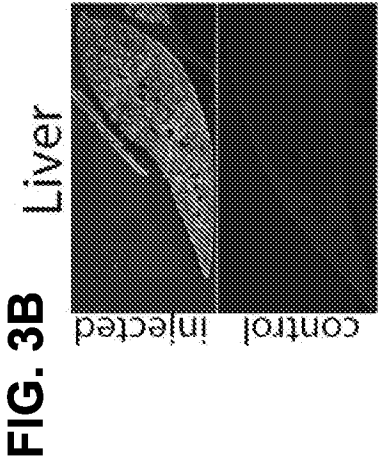
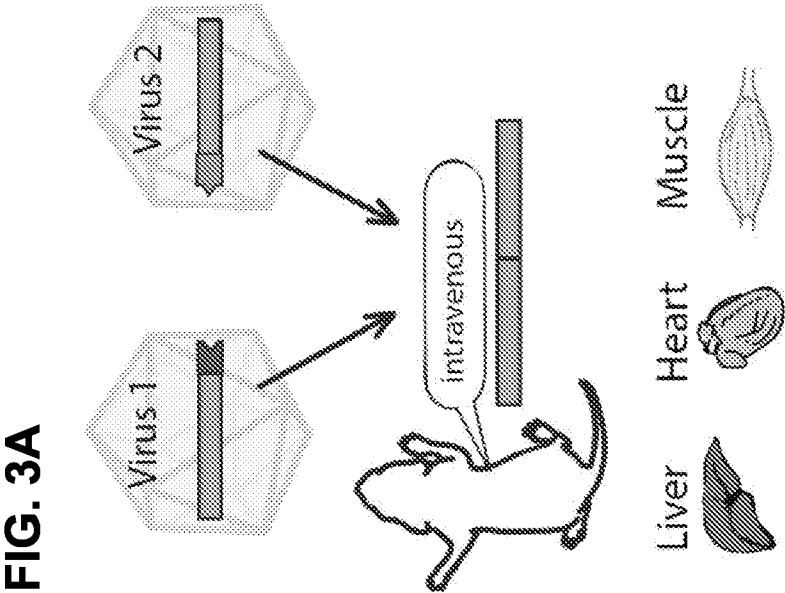




**FIG. 4B**

Mouse tibialis anterior

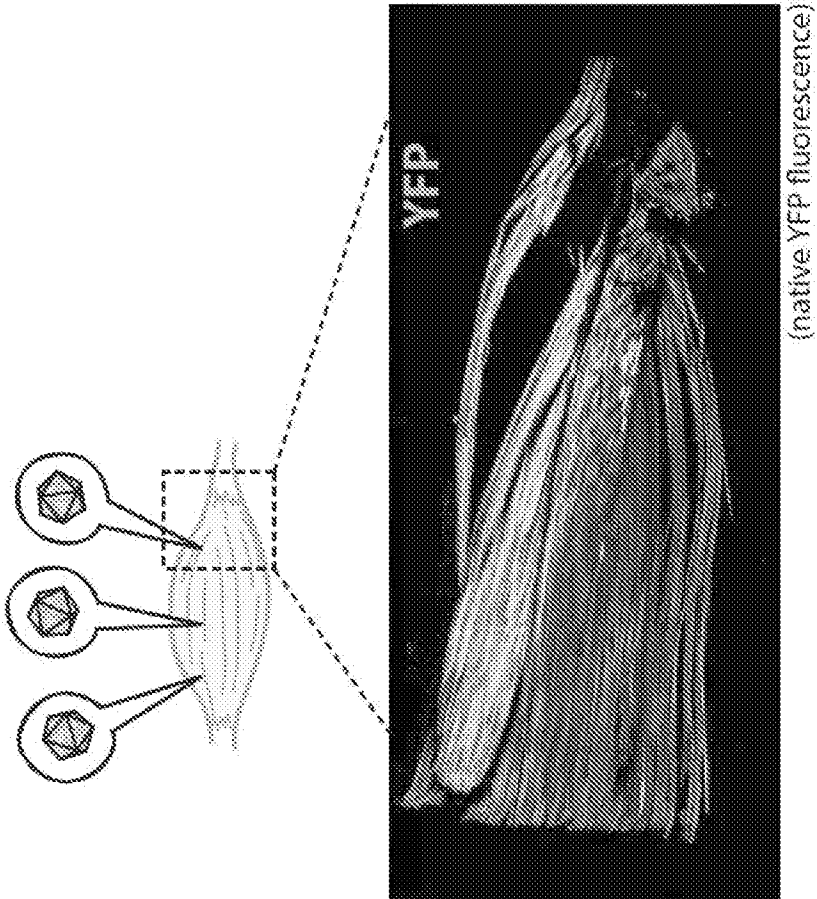
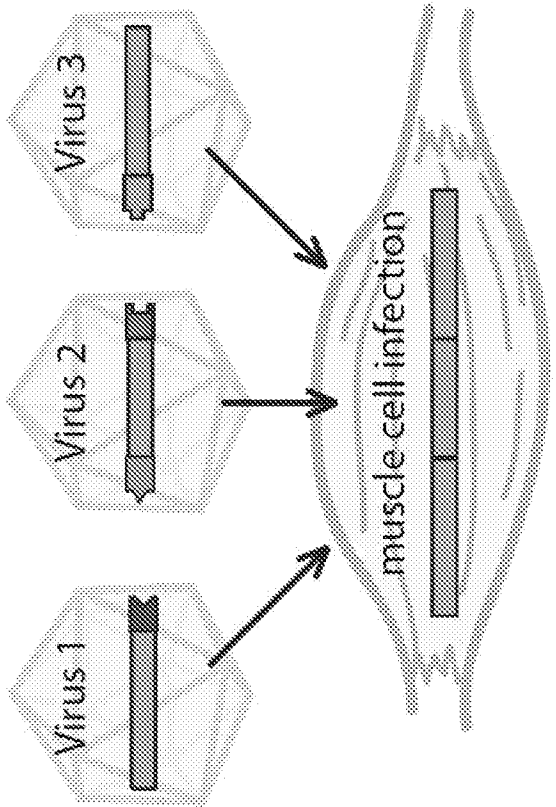


**FIG. 4A**



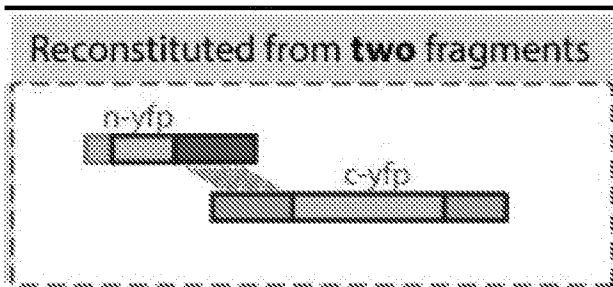


**FIG. 5A**

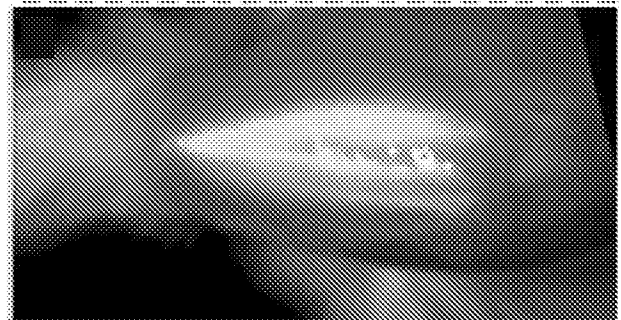


**FIG. 5B**

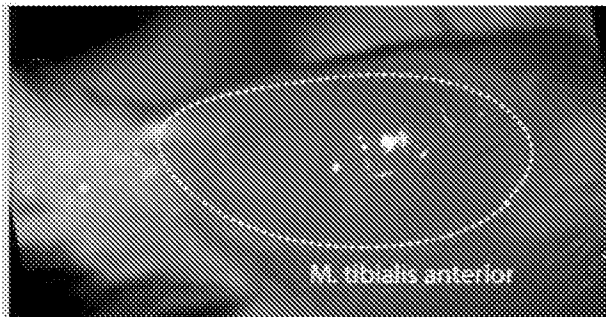


**FIG. 5C**

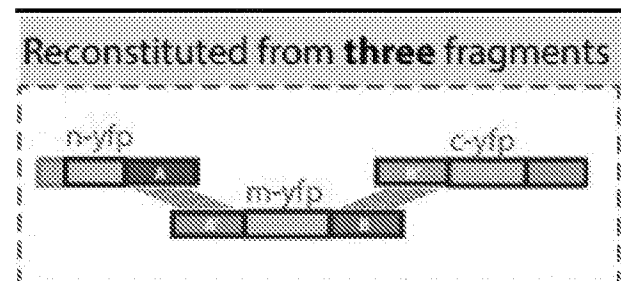


**FIG. 5D**

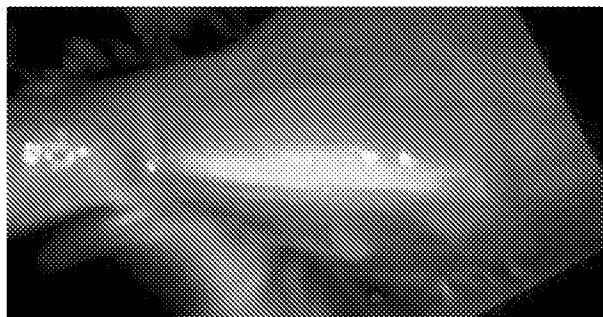


**FIG. 5E**

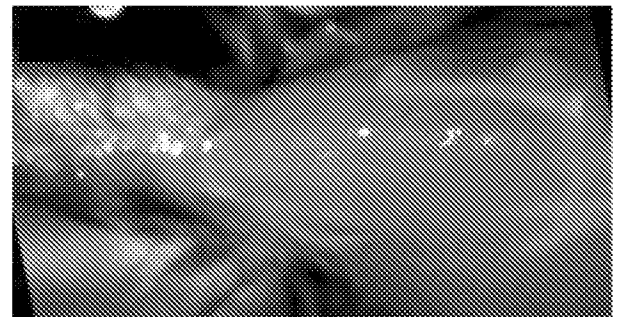


**FIG. 5F**

FIG. 6A

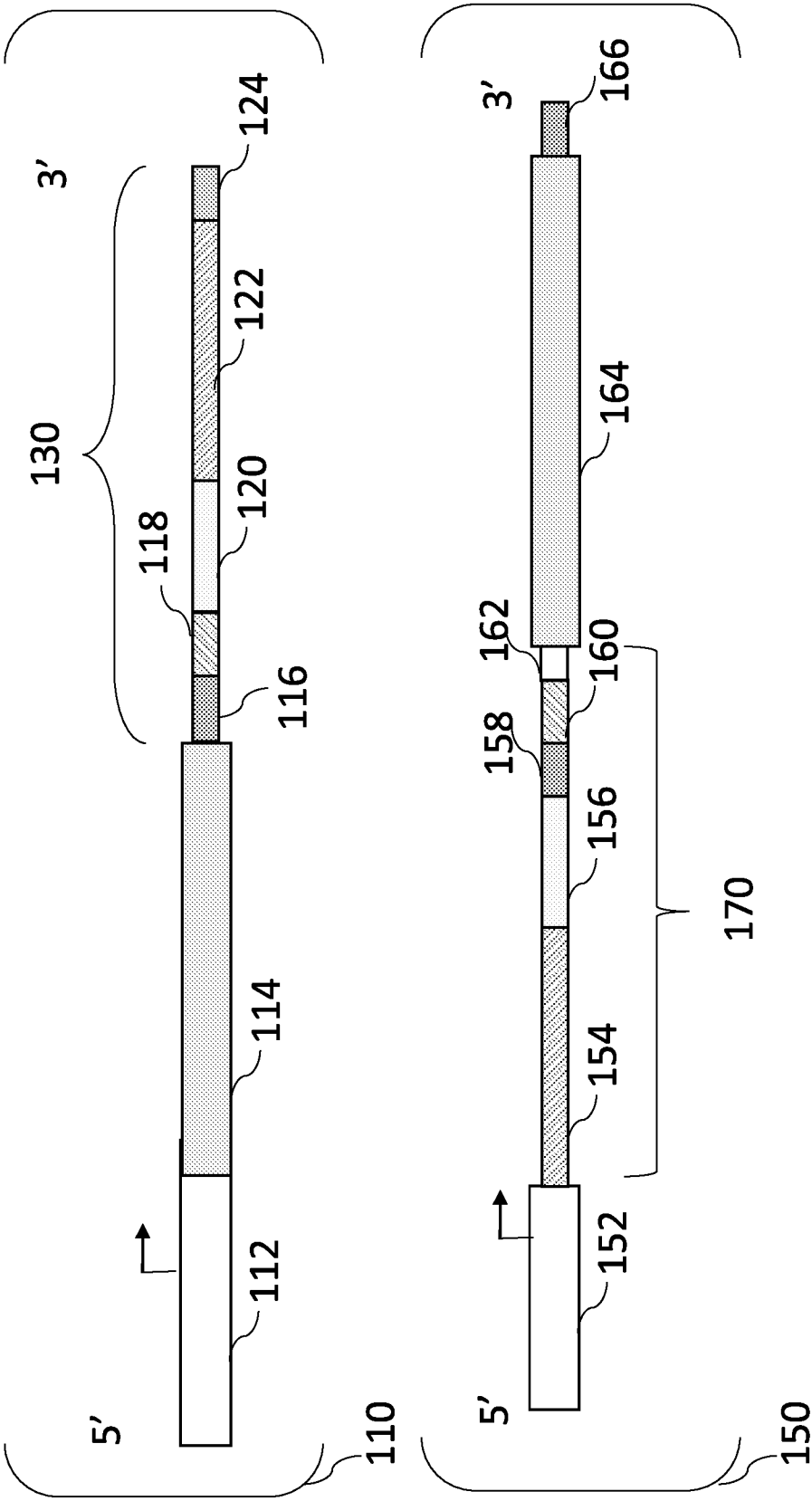


FIG. 6B

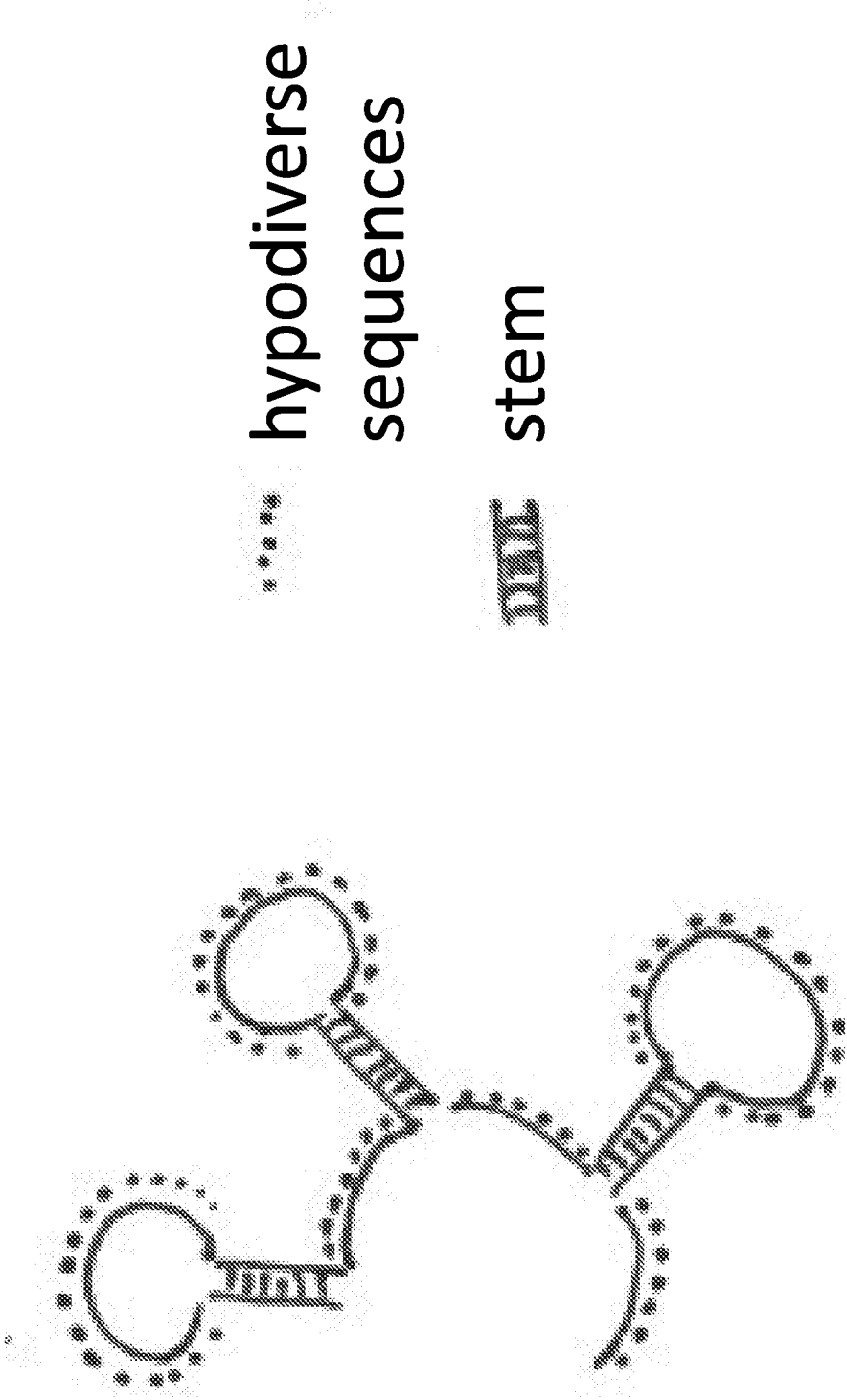


FIG. 6C

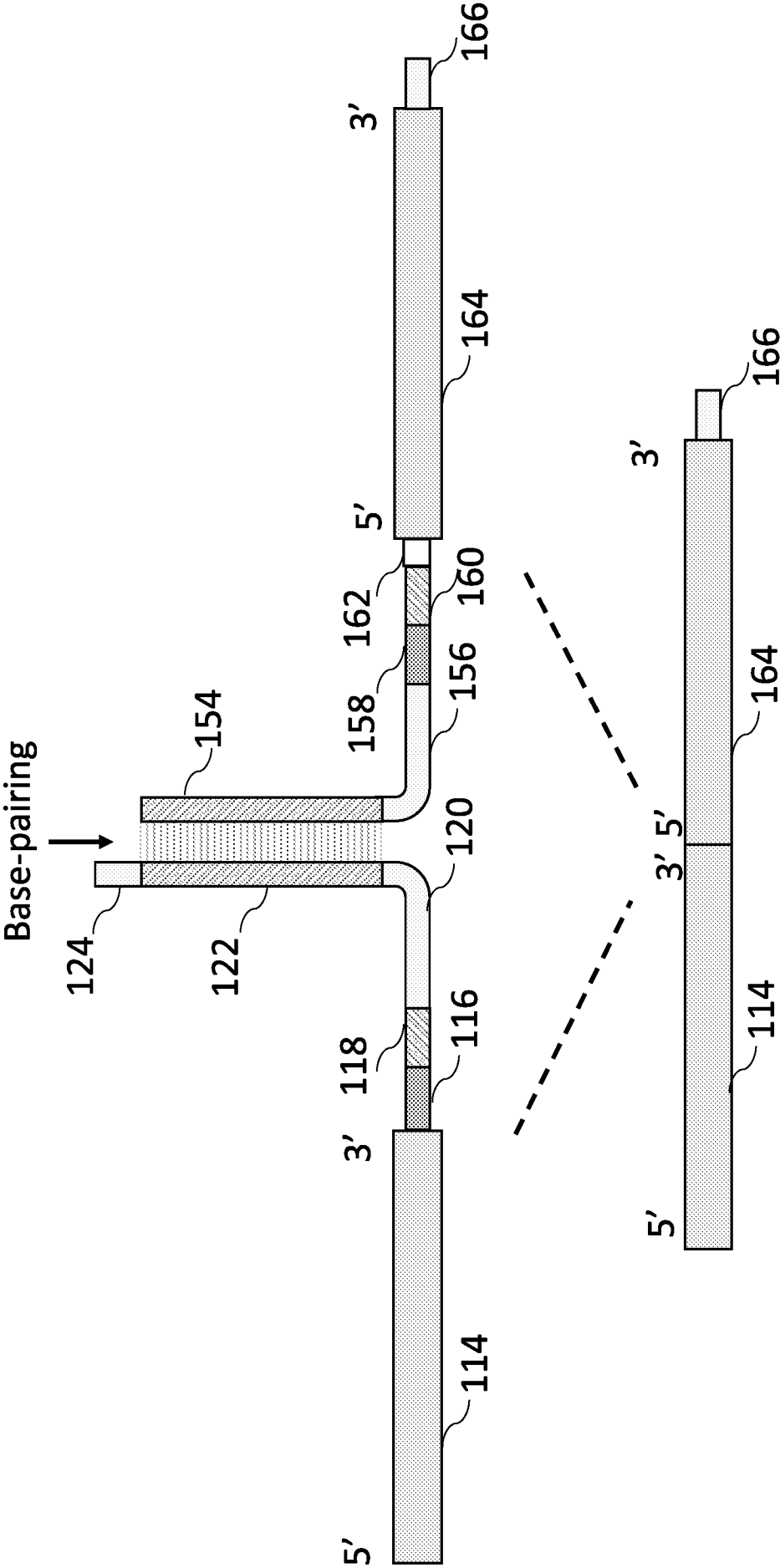


FIG. 6D

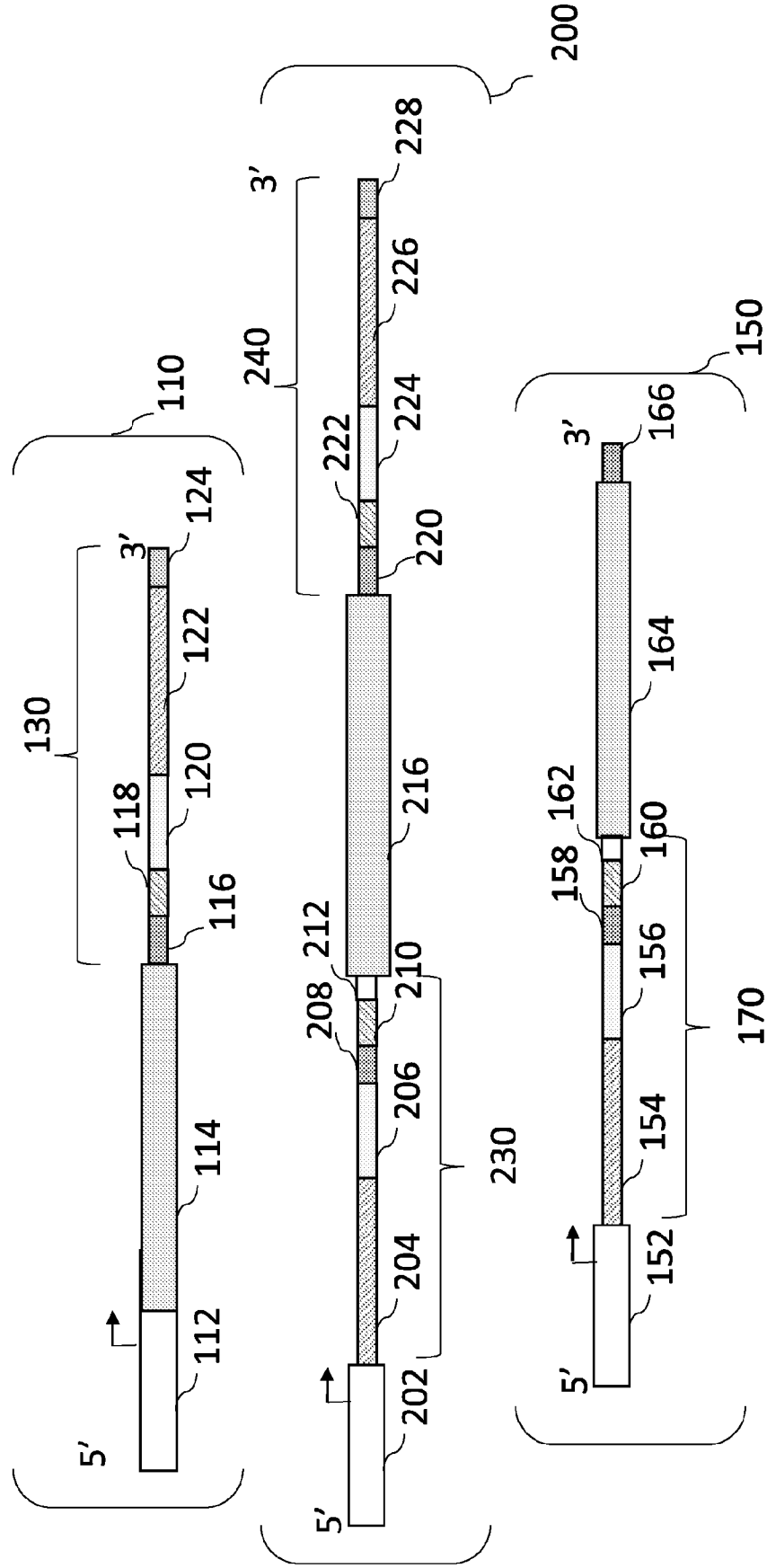


FIG. 6E

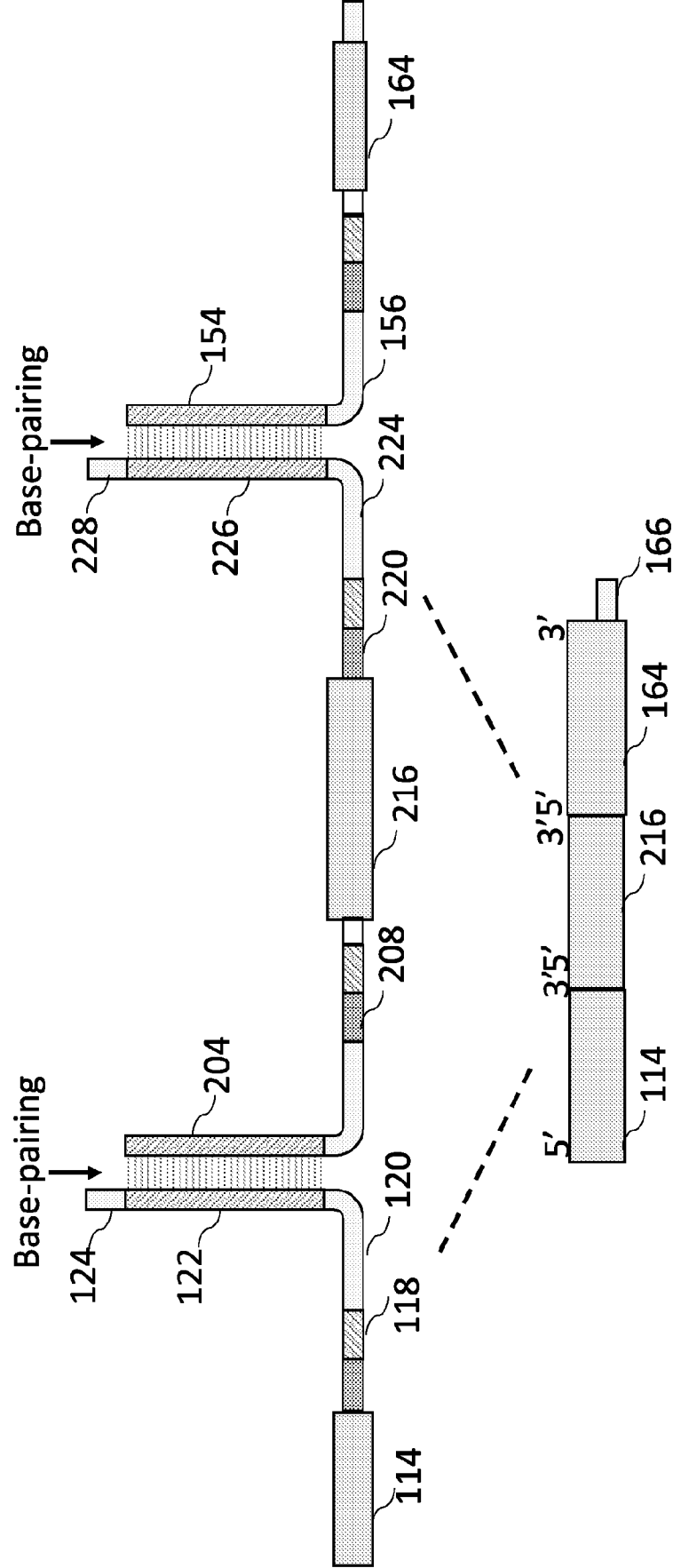


FIG. 6F

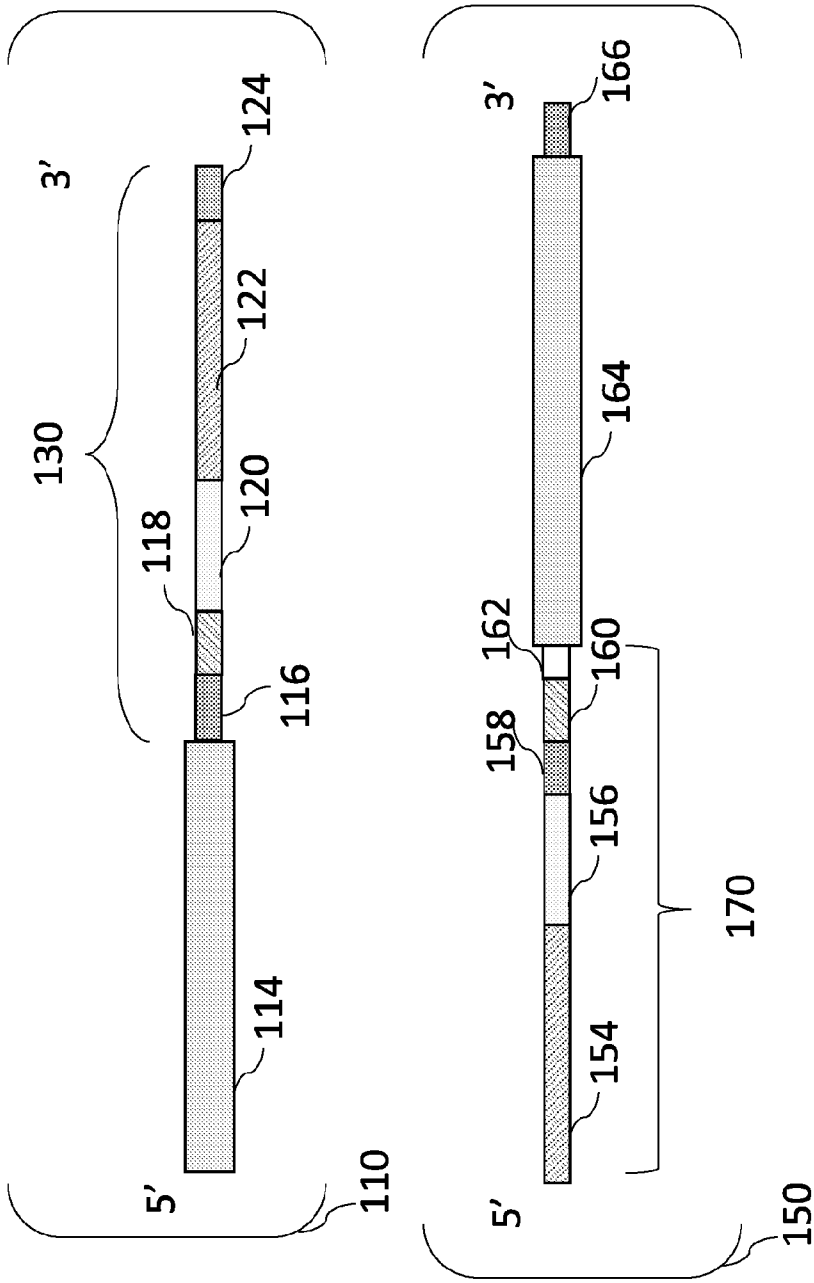




**FIG. 7A**

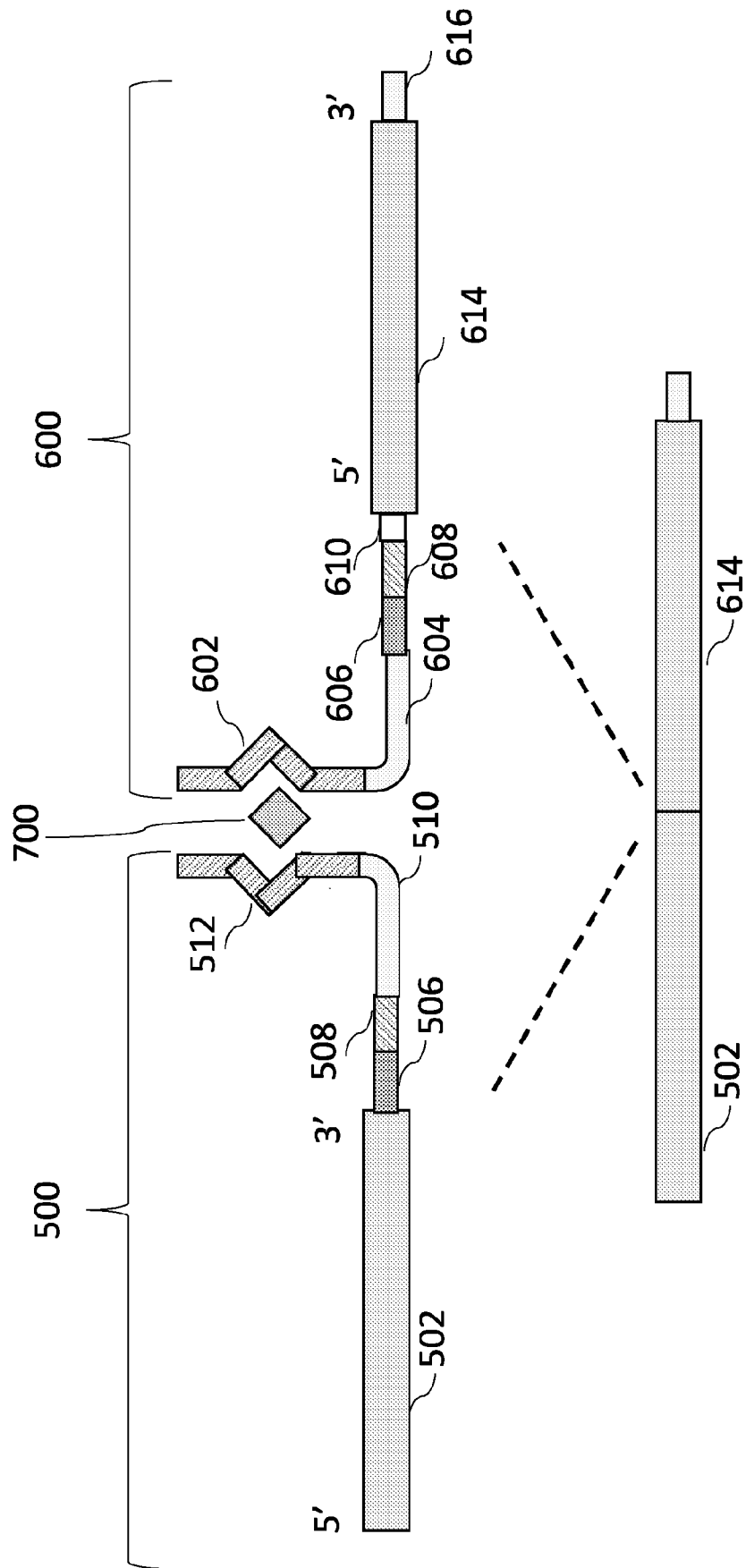
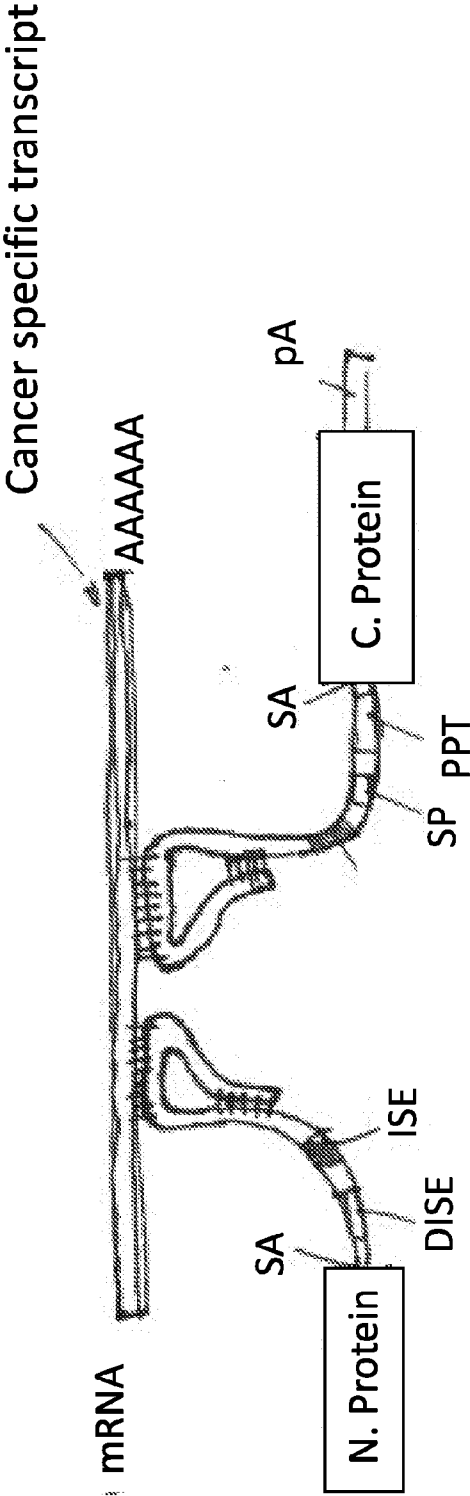


FIG. 7B



Recombination of RNA in presence of a specific transcript

FIG. 7C

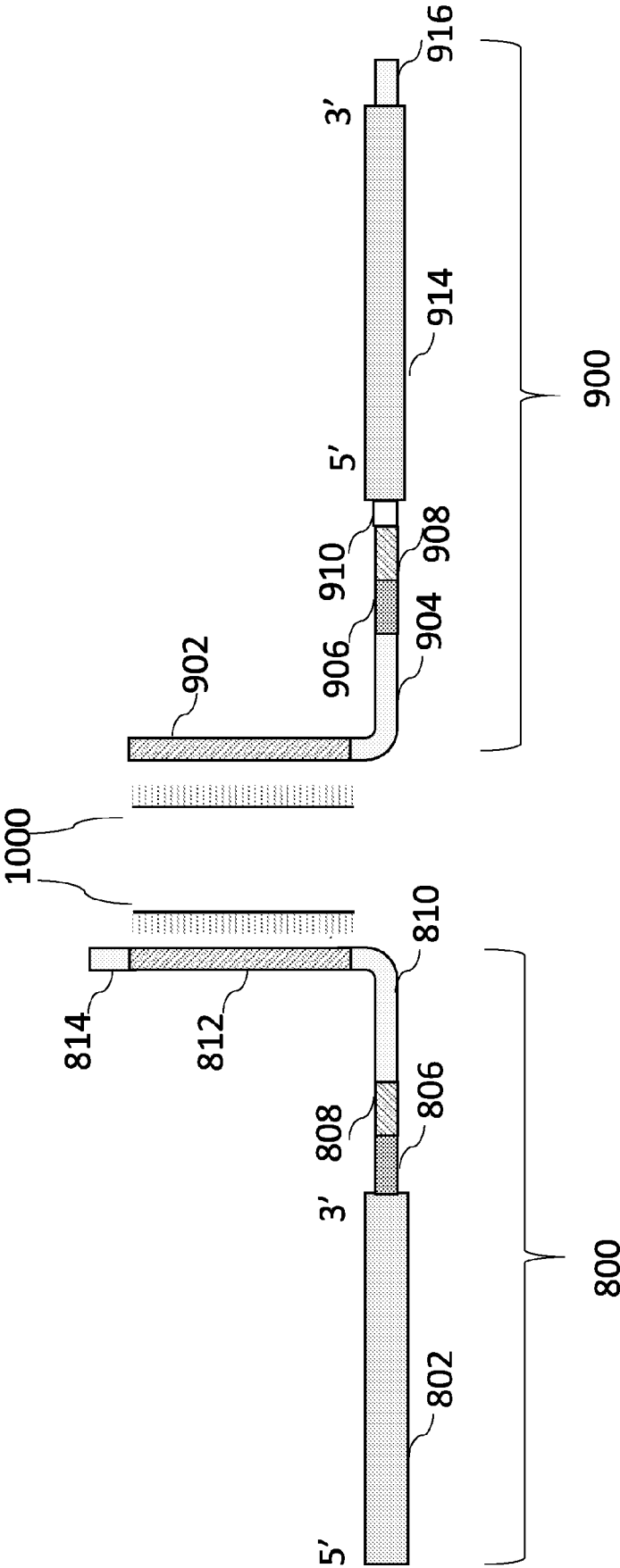


FIG. 8

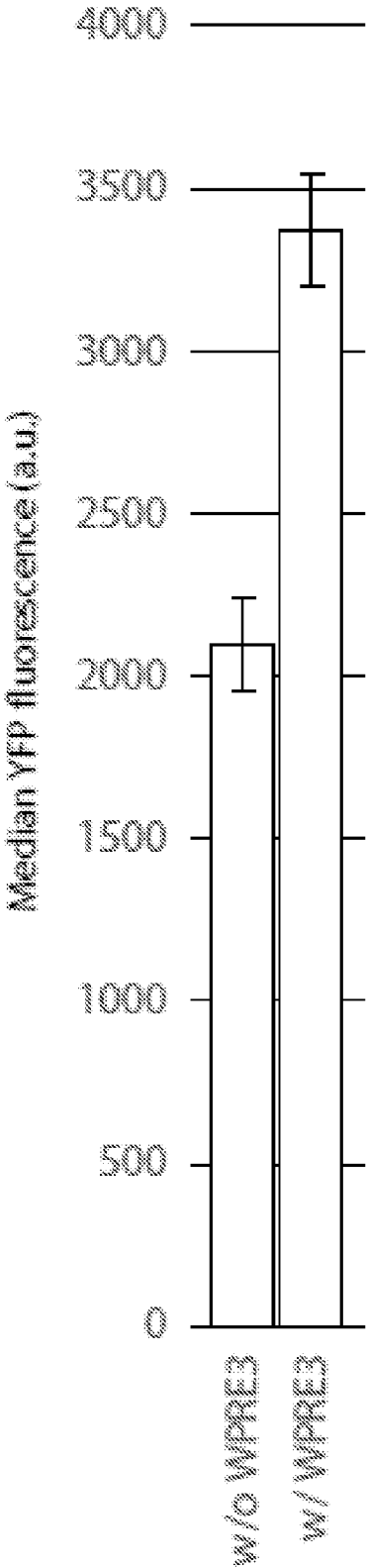
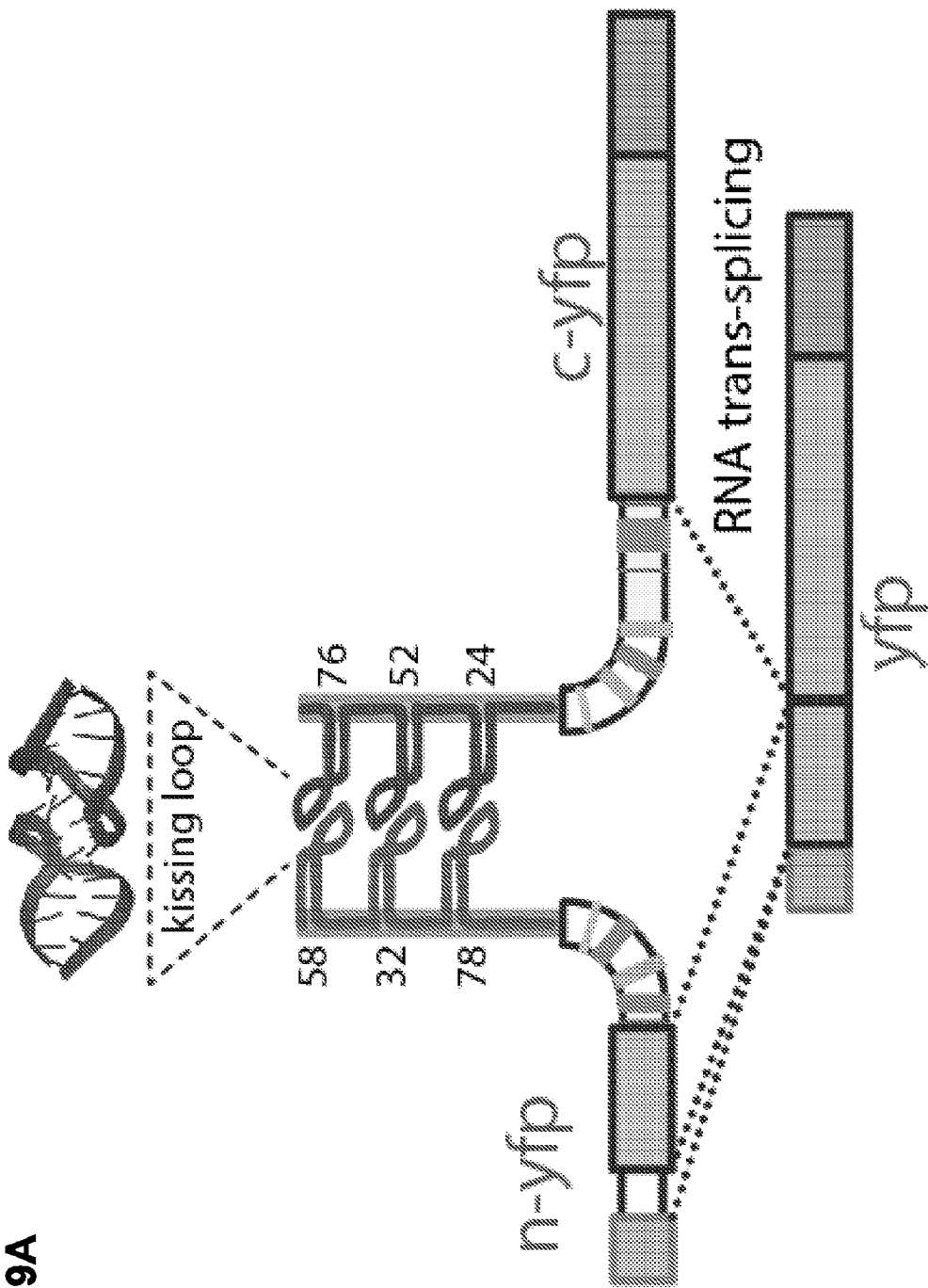


FIG. 9A



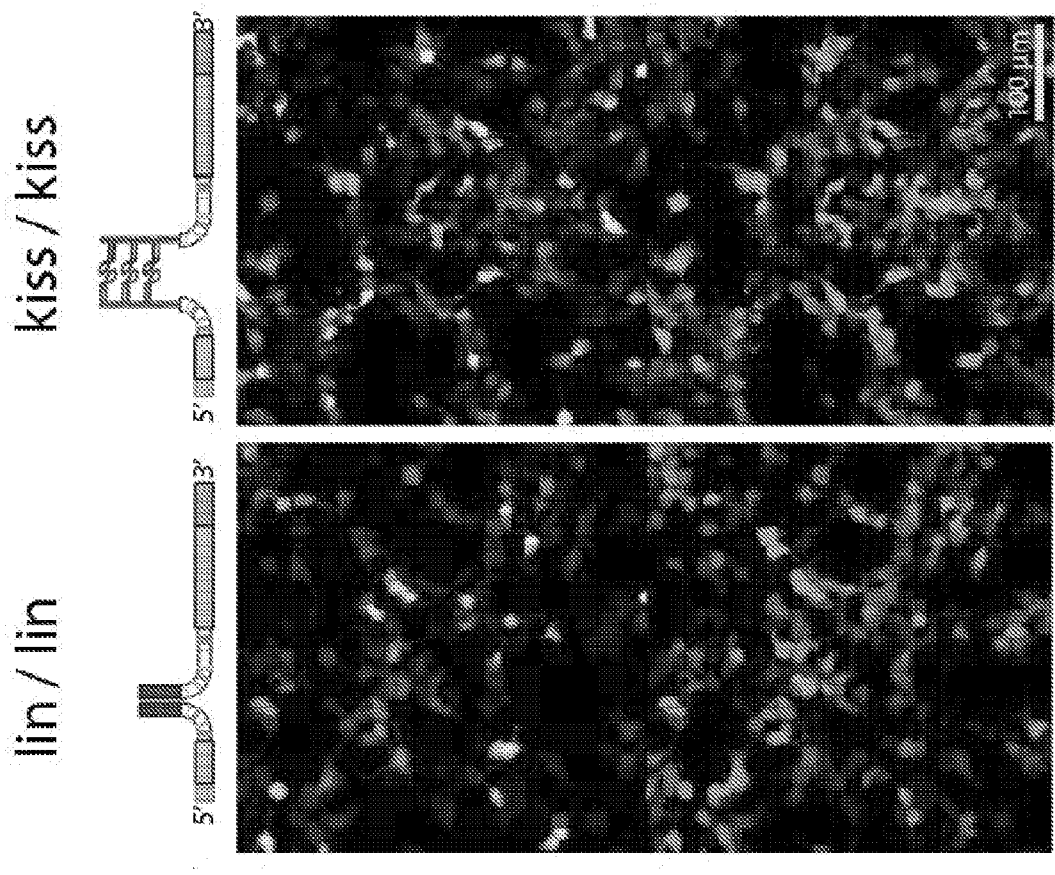


FIG. 9B

**FIG. 10A**

27/71

**FIG. 10B**

**FIG. 10B**



**FIG. 10C**

[illegible][illegible]

[illegible]

30/71

5'.Synthetic.Intron-DimerizationDomain(Lin)

[illegible]

**FIG. 10E**

## LCB290 - simpleSD-gfphybrid

[illegible]

**FIG. 10F**

## LCB291 - simpleSA-gfphybrid

**FIG. 10G**

**FIG. 10G**

## LCB300 - boxBloops with ISE

**FIG. 1**

**FIG. 10H**



**FIG. 10J**

10

20

30

40

50

60

70

80

90

100

110

120

130

140

150

160

170

180

190

200

210

220

230

240

250

260

270

280

290

300

310

320

330

340

350

360

370

380

390

400

410

420

430

440

450

460

470

480

490

500

510

520

530

540

550

560

570

580

590

600

610

620

630

640

650

660

670

680

690

700

710

720

730

740

750

760

770

780

790

800

810

820

830

840

850

860

870

880

890

900

910

920

930

940

950

960

970

980

990

1000

1010

1020

1030

1040

1050

1060

1070

1080

1090

1100

1110

1120

1130

1140

1150

1160

1170

1180

1190

1200

1210

1220

1230

1240

1250

1260

1270

1280

1290

1300

1310

1320

1330

1340

1350

1360

1370

1380

1390

1400

1410

1420

1430

1440

1450

1460

1470

1480

1490

1500

1510

1520

1530

1540

1550

1560

1570

1580

1590

1600

1610

1620

1630

1640

1650

1660

1670

1680

1690

1700

1710

1720

1730

1740

1750

1760

1770

1780

1790

1800

1810

1820

1830

1840

1850

1860

1870

1880

1890

1900

1910

1920

1930

1940

1950

1960

1970

1980

1990

2000

2010

2020

2030

2040

2050

2060

2070

2080

2090

2100

2110

2120

2130

2140

2150

2160

2170

2180

2190

2200

2210

2220

2230

2240

2250

2260

2270

2280

2290

2300

2310

2320

2330

2340

2350

2360

2370

2380

2390

2400

2410

2420

2430

2440

2450

2460

2470

2480

2490

2500

2510

2520

2530

2540

2550

2560

2570

2580

2590

2600

2610

2620

2630

2640

2650

2660

2670

2680

2690

2700

2710

2720

2730

2740

2750

2760

2770

2780

2790

2800

2810

2820

2830

2840

2850

2860

2870

2880

2890

2900

2910

2920

2930

2940

2950

2960

2970

2980

2990

3000

3010

3020

3030

3040

3050

3060

3070

3080

3090

3100

3110

3120

3130

3140

3150

3160

3170

3180

3190

3200

3210

3220

3230

3240

3250

3260

3270

3280

3290

3300

3310

3320

3330

3340

3350

3360

3370

3380

3390

3400

3410

3420

3430

3440

3450

3460

3470

3480

3490

3500

3510

3520

3530

3540

3550

3560

3570

3580

3590

3600

3610

3620

3630

3640

3650

3660

3670

3680

3690

3700

3710

3720

3730

3740

3750

3760

3770

3780

3790

3800

3810

3820

3830

3840

3850

3860

3870

3880

3890

3900

3910

3920

3930

3940

3950

3960

3970

3980

3990

4000

4010

4020

4030

4040

4050

4060

4070

4080

4090

4100

4110

4120

4130

4140

4150

4160

4170

4180

4190

4200

4210

4220

4230

4240

4250

4260

4270

4280

4290

4300

4310

4320

4330

4340

4350

4360

4370

4380

4390

4400

4410

4420

4430

4440

4450

4460

4470

4480

4490

4500

4510

4520

4530

4540

4550

4560

4570

4580

4590

4600

4610

4620

4630

4640

4650

4660

4670



**FIG. 10K**

37/71

**FIG. 10L**

38/71

**FIG. 10M**

**FIG. 10M**

## LCB572

[illegible]

**FIG. 10N**



**FIG. 10P**

42/71

## LCB606

**FIG. 10Q**

**FIG. 10Q**

**FIG. 10R**

[illegible]



LCB610 N.Abca4

[illegible]

**FIG. 10S**

N.Abca4-continued

46/71

**FIG. 10U**

**FIG. 10U**

## C.Abca4

**FIG 10V**

**FIG. 10V**

C.Abca4

49/71

**FIG. 10X**

50/71

C.Abca4 continued

[illegible]





FIG. 11

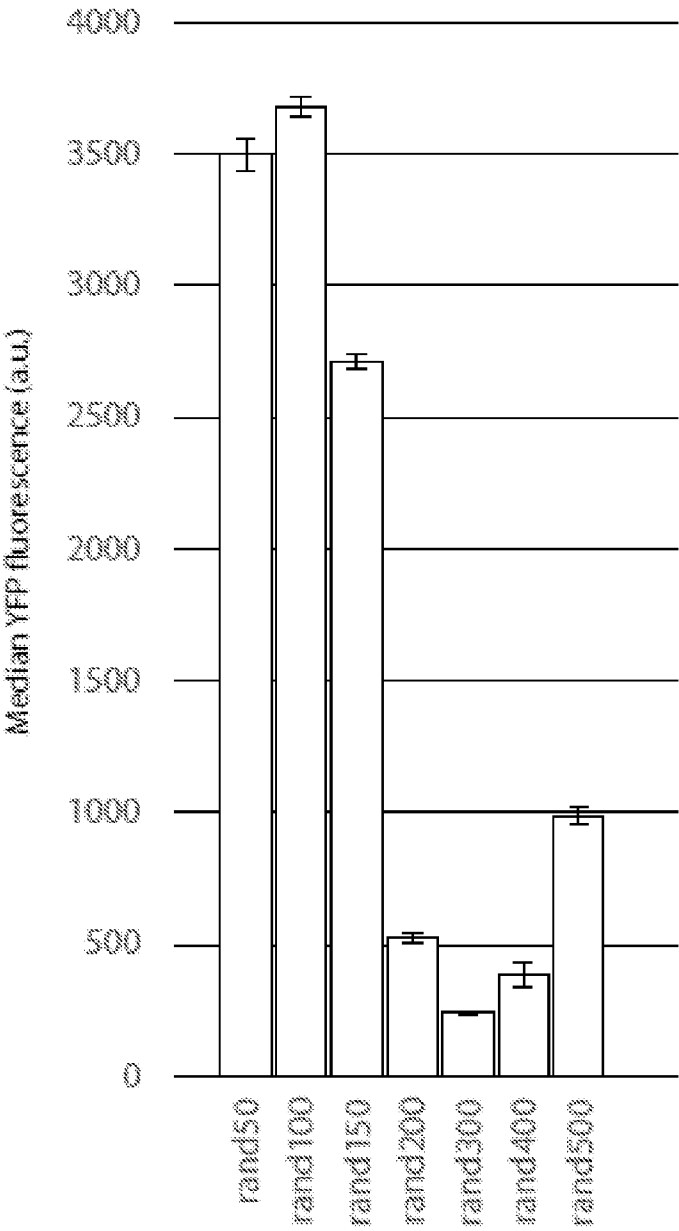


FIG. 12A

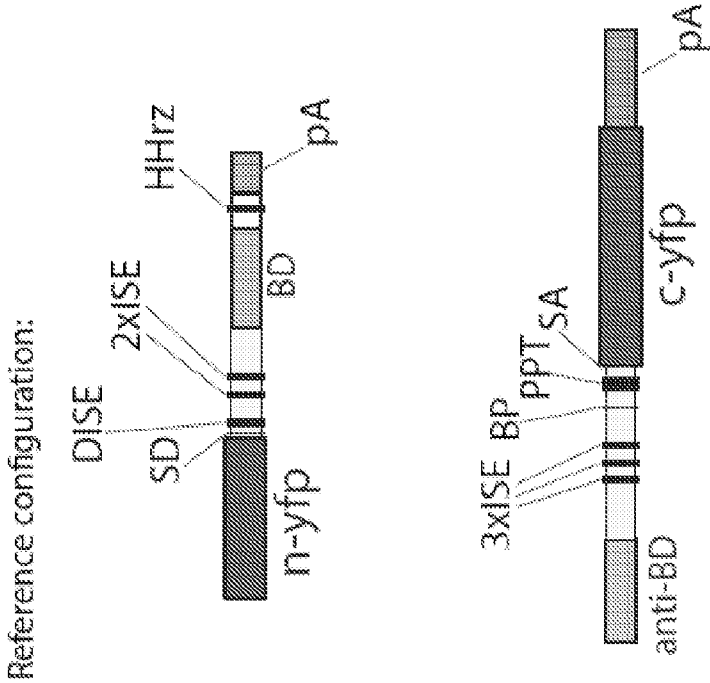


FIG. 12B

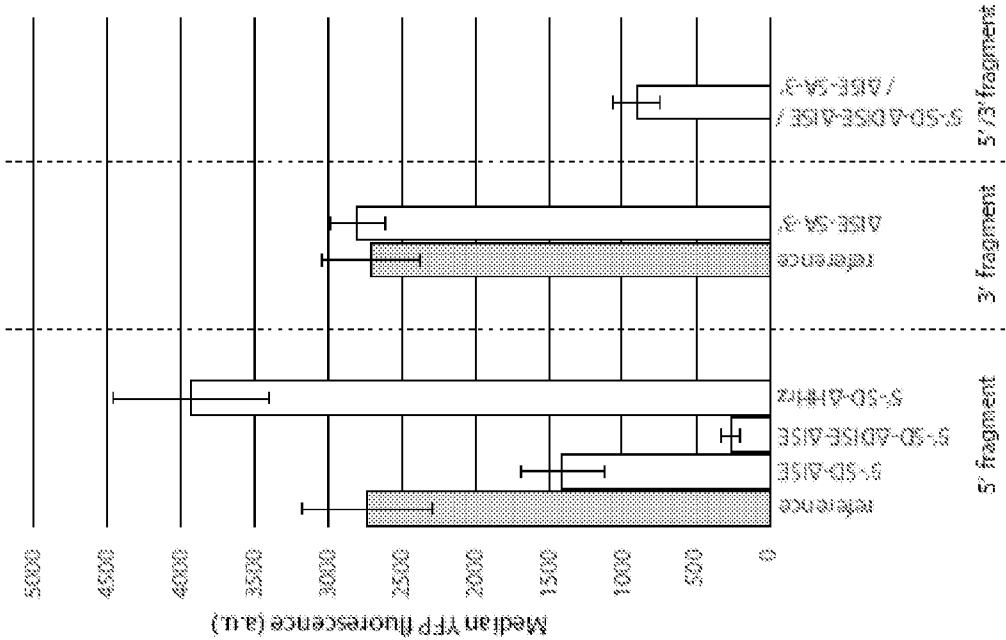


FIG. 13A

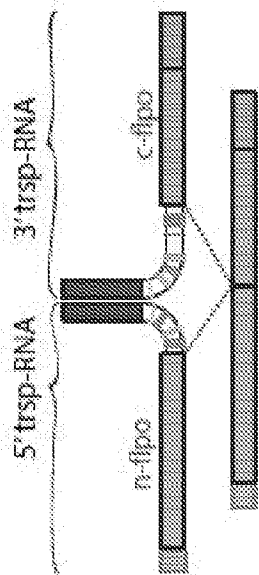


FIG. 13D

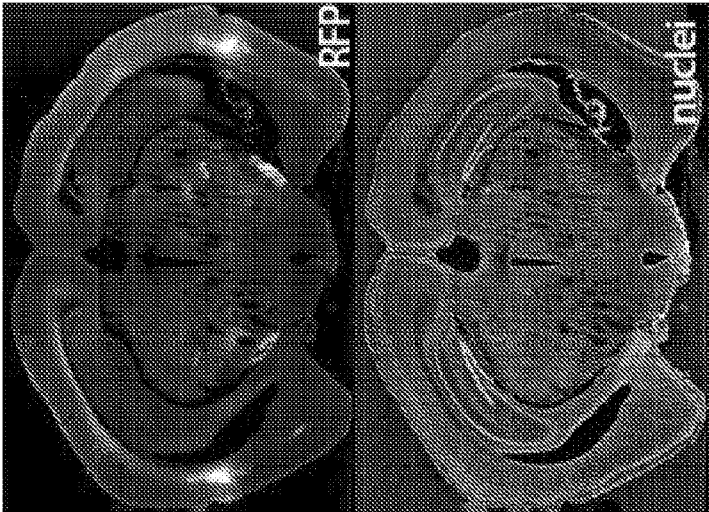


FIG. 13C

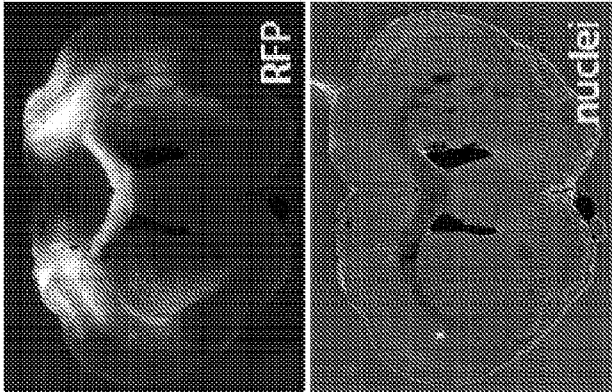
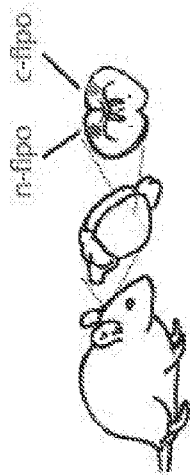
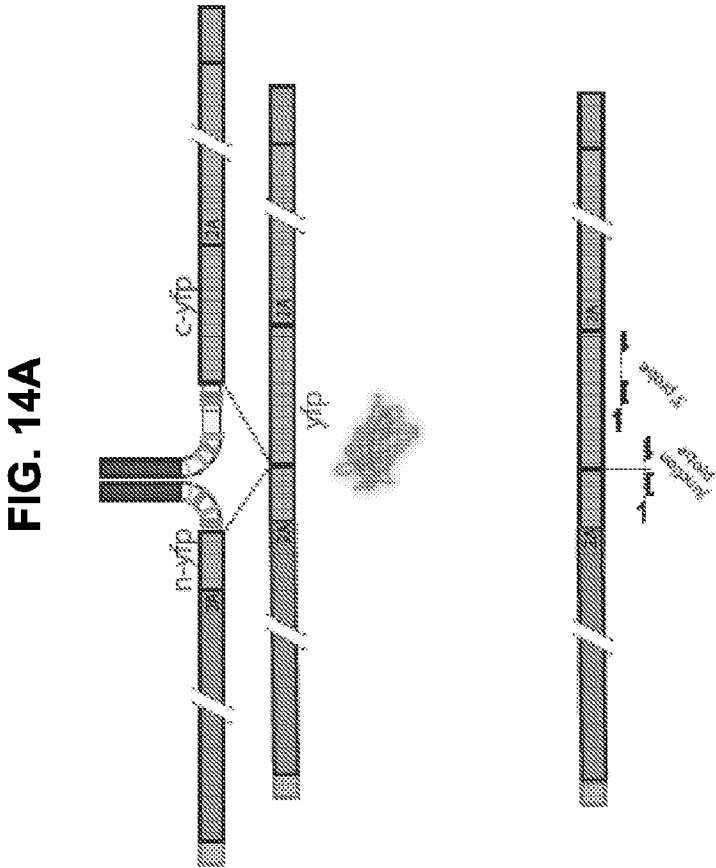
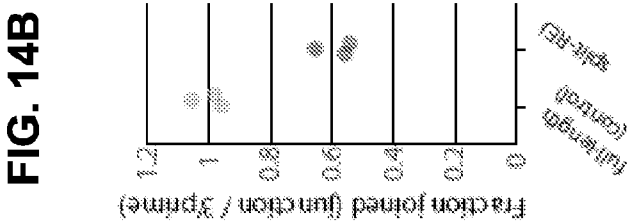
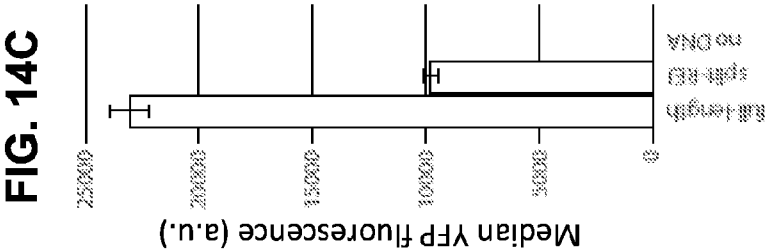
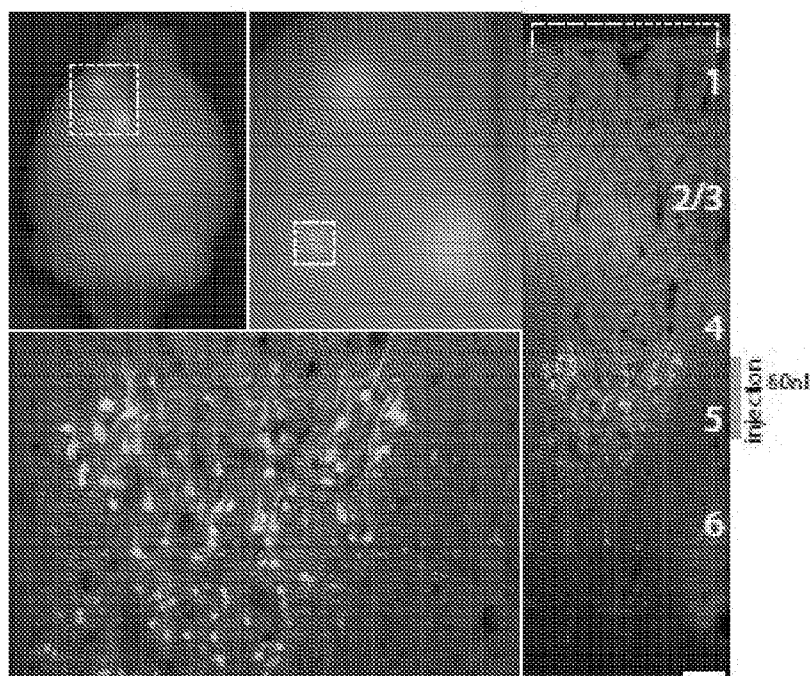
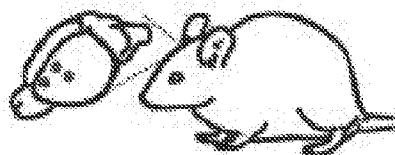


FIG. 13B







YFP (+ DRAQ5 for context)

**FIG. 14D**

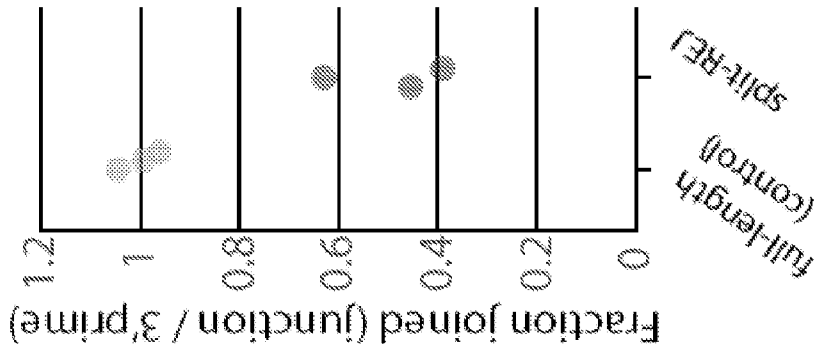


FIG. 15B

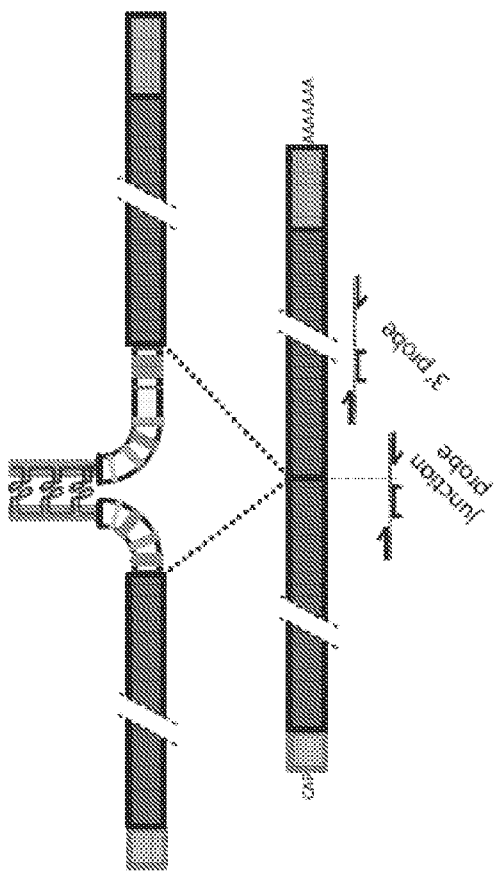
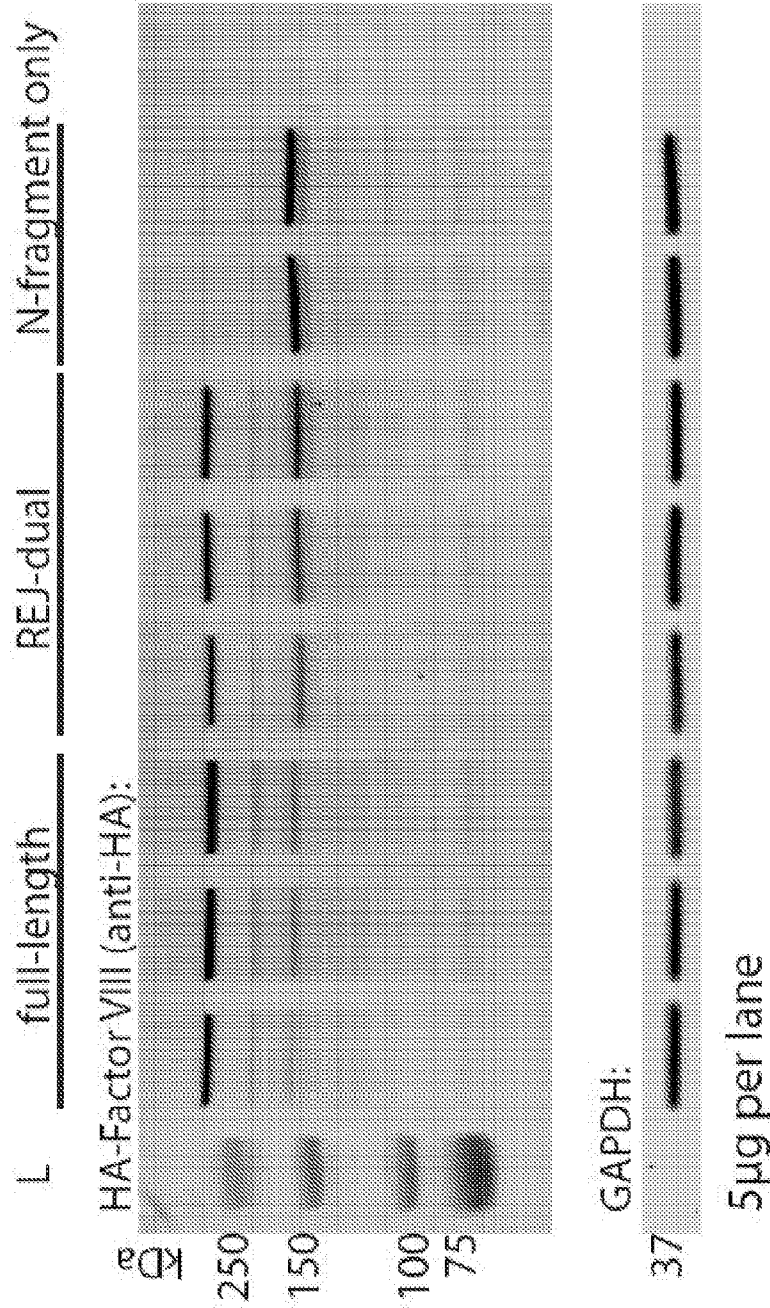


FIG. 15A

FIG. 15C



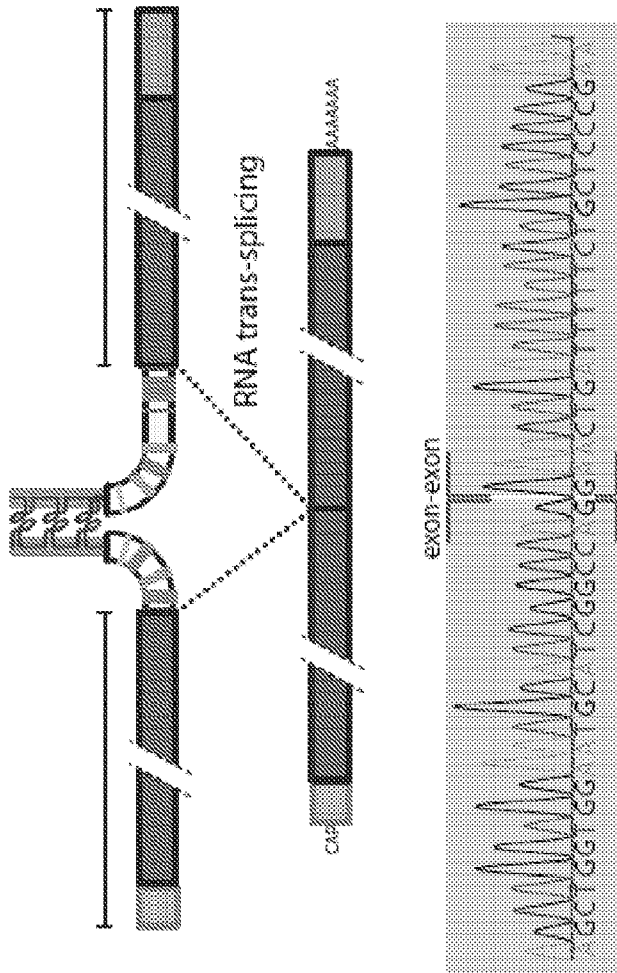


FIG. 16A

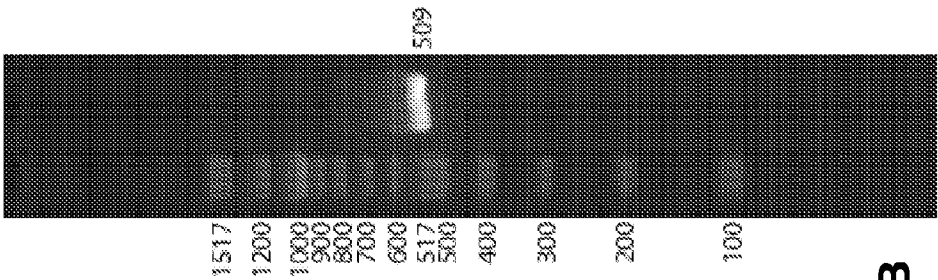
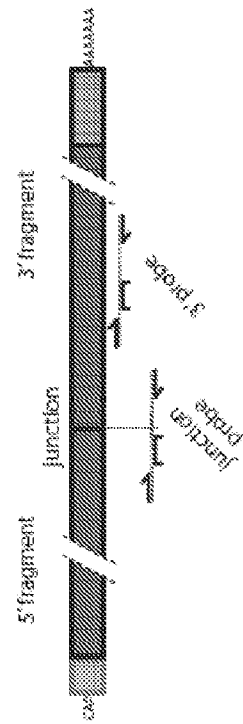


FIG. 16B



FIG. 16C



split-REJ AAV Abca4  
(prototype)

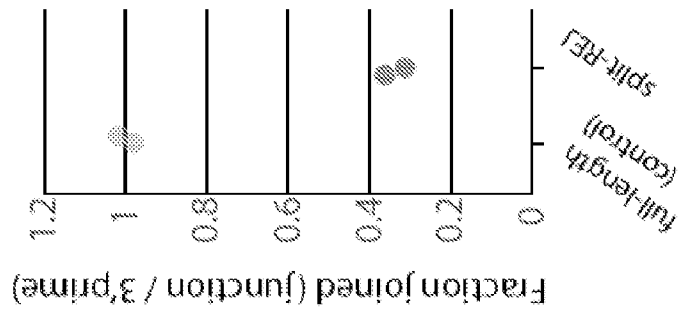
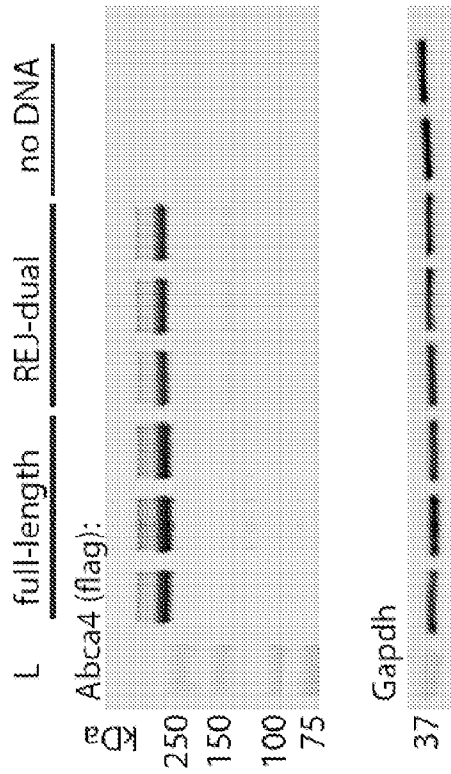


FIG. 16D

FIG. 16E



Transfection control (BFP)  
normalization factor:  
full-length: 1  
REJ-dual: 0.54

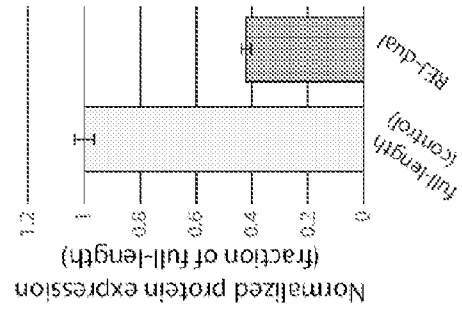


FIG. 16F

**FIG. 17A**

HIV-1 based kissing loop dimerization domain

```

5'~[...]agggcuuggagcAggaggGgcuuccgaagccu[...]-3' (N~fragment)
5'~[...]((((((((((((((((((((((((((((((((((((((((((((((((((((((((
//((((((((((((((((((((((((((((((((((((((((((((((((((((((((
3'~[...]((((((((((((((((((((((((((((((((((((((((((((((((((((((((
//((((((((((((((((((((((((((((((((((((((((((((((((((((((((
3'~[...]uccgaagccuugaccuccgaagcaggguccgga[...]-5' (C~fragment)

```

**FIG. 17B**

[illegible]

FIG. 18A

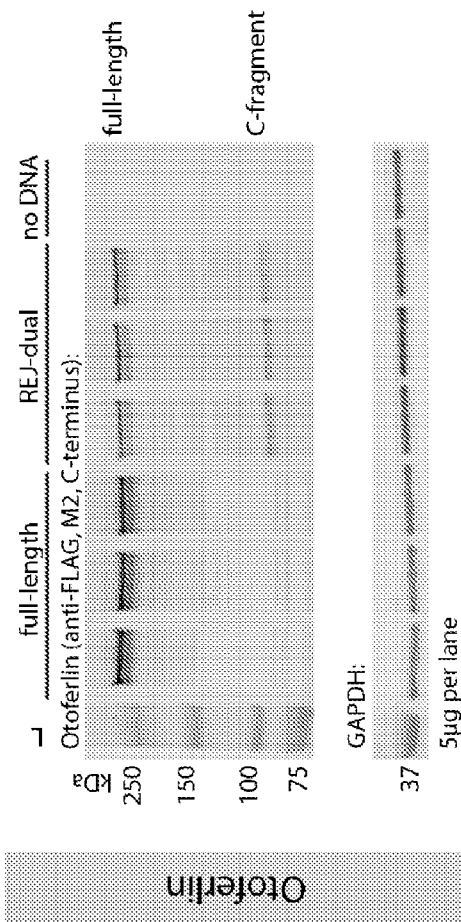


FIG. 18B

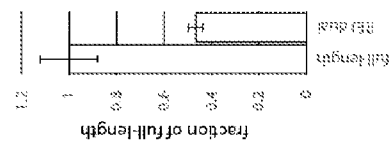


FIG. 18C

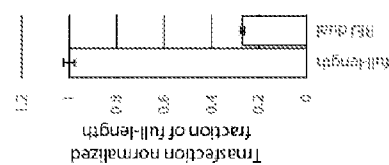


FIG. 19A

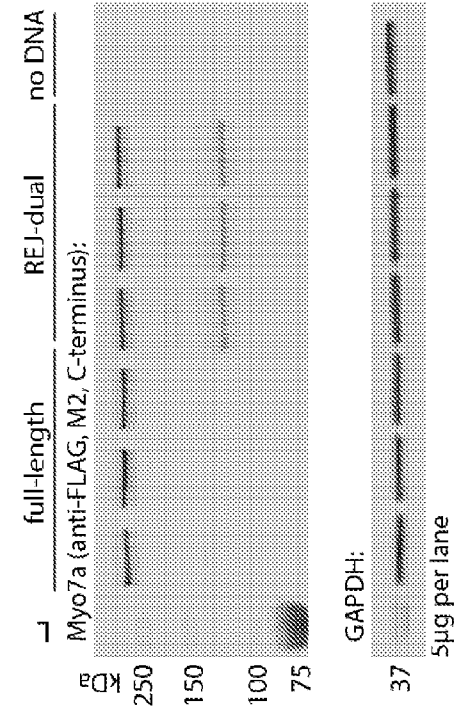


FIG. 19B

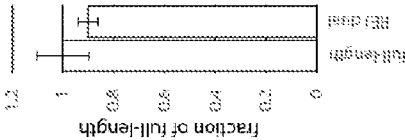


FIG. 19C

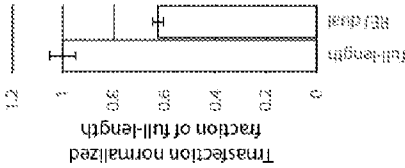


**FIG. 20A**

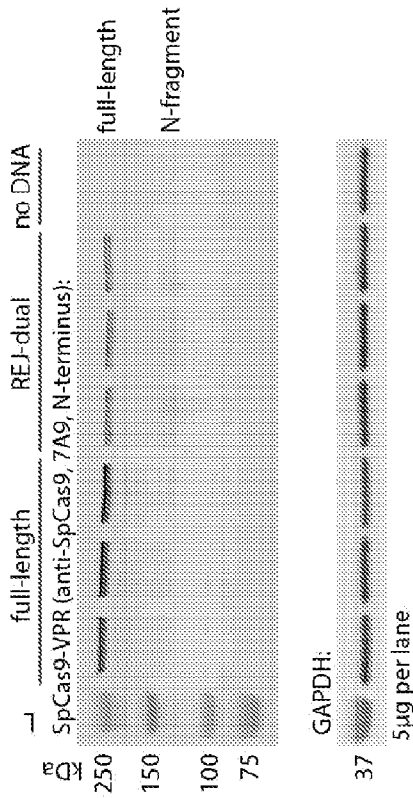


**FIG. 20B**

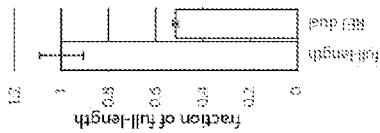
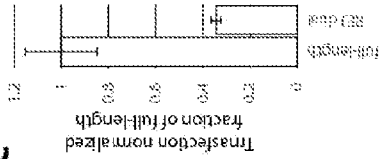


**FIG. 20C**



**SpCas9-VPR: transcriptional activation of UAS-YFP**

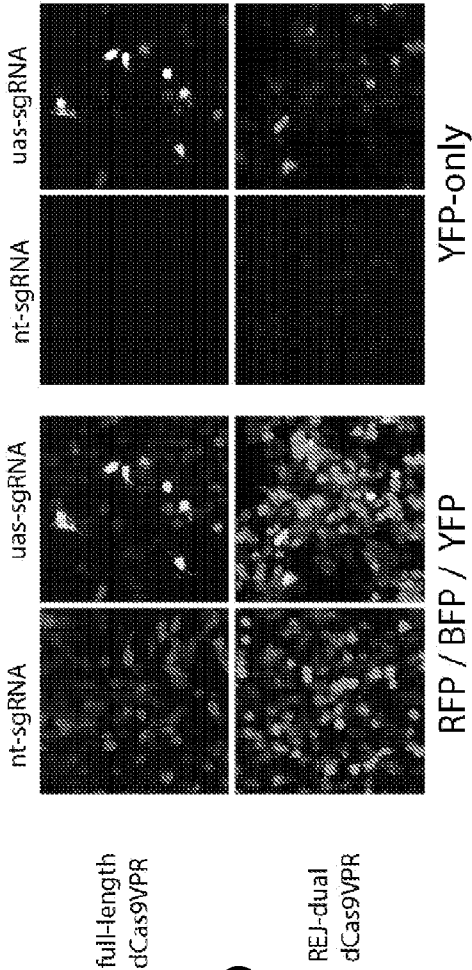


**FIG. 20D**

FIG. 21A

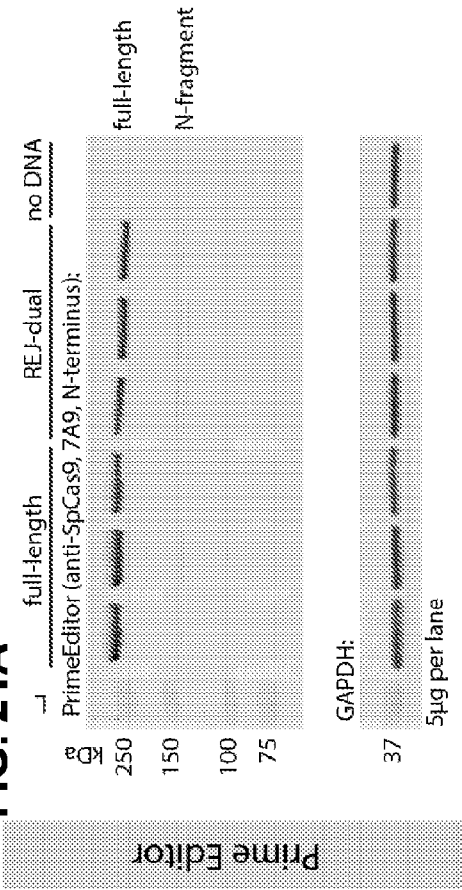


FIG. 21B

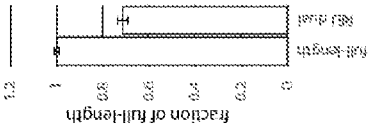
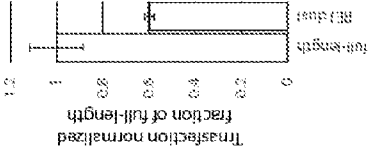


FIG. 21C



PrimeEditor: targeted G-to-T transversion mutation

FIG. 21D

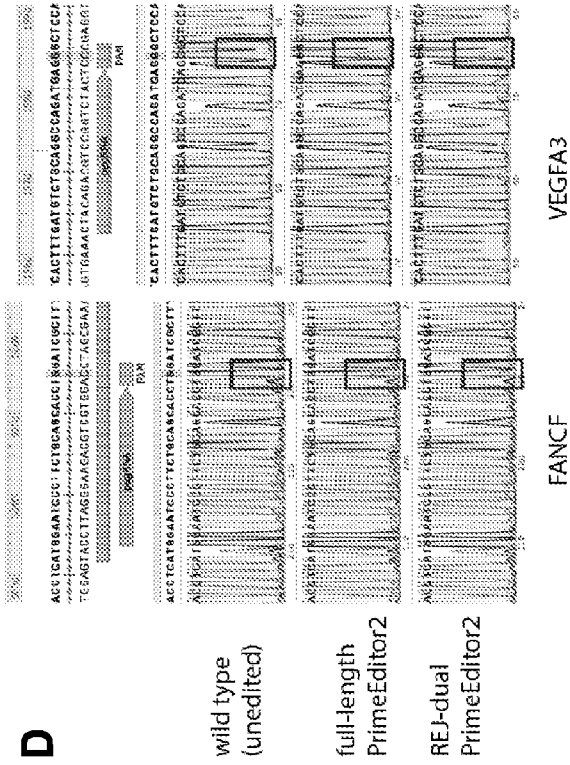


**FIG. 22A**

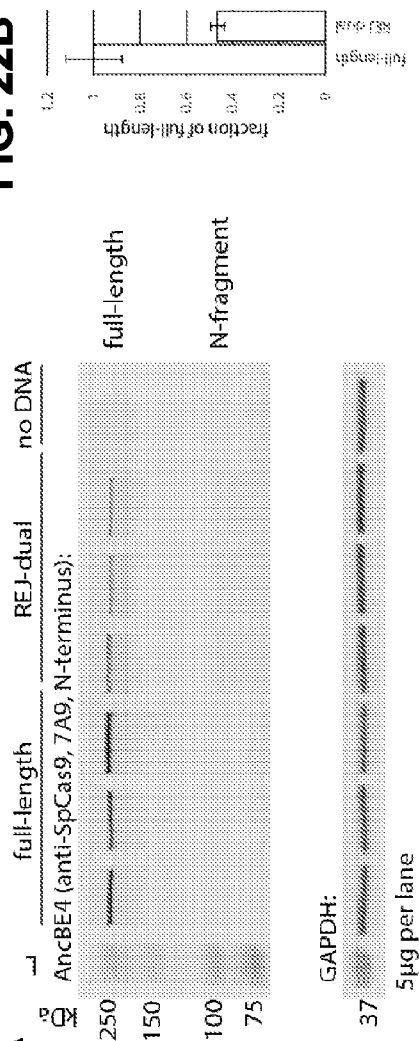


**FIG. 22B**

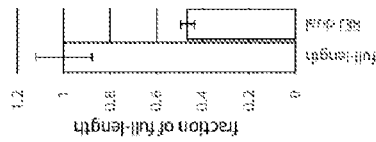


**FIG. 22C**

AncBE4: targeted C-to-T transition mutation

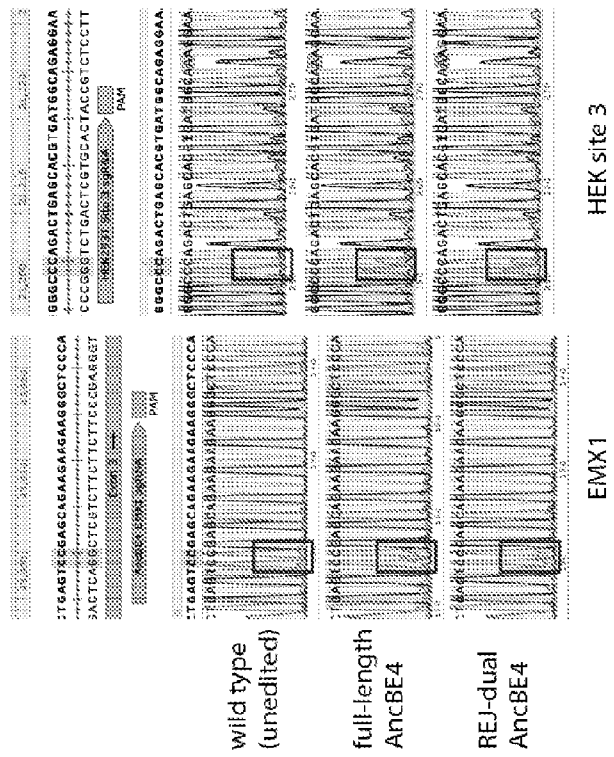


**FIG. 23A**

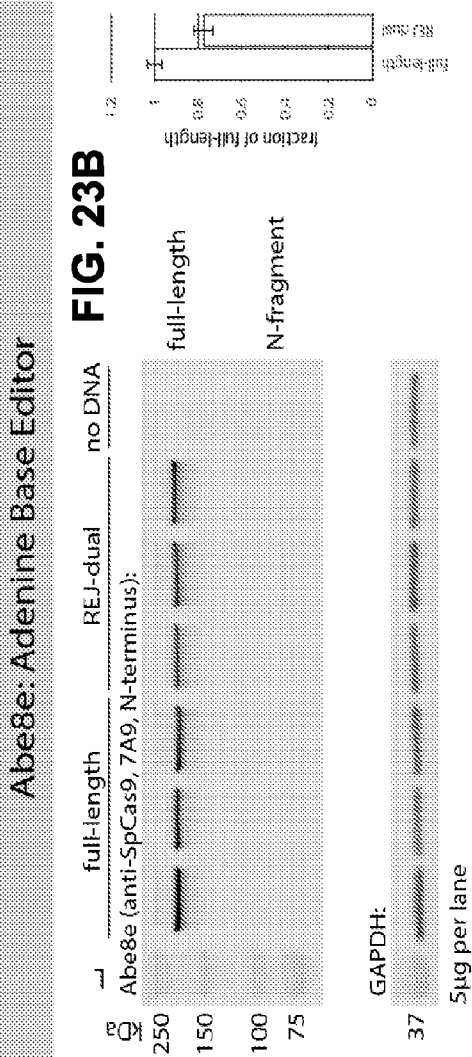


**FIG. 23C**

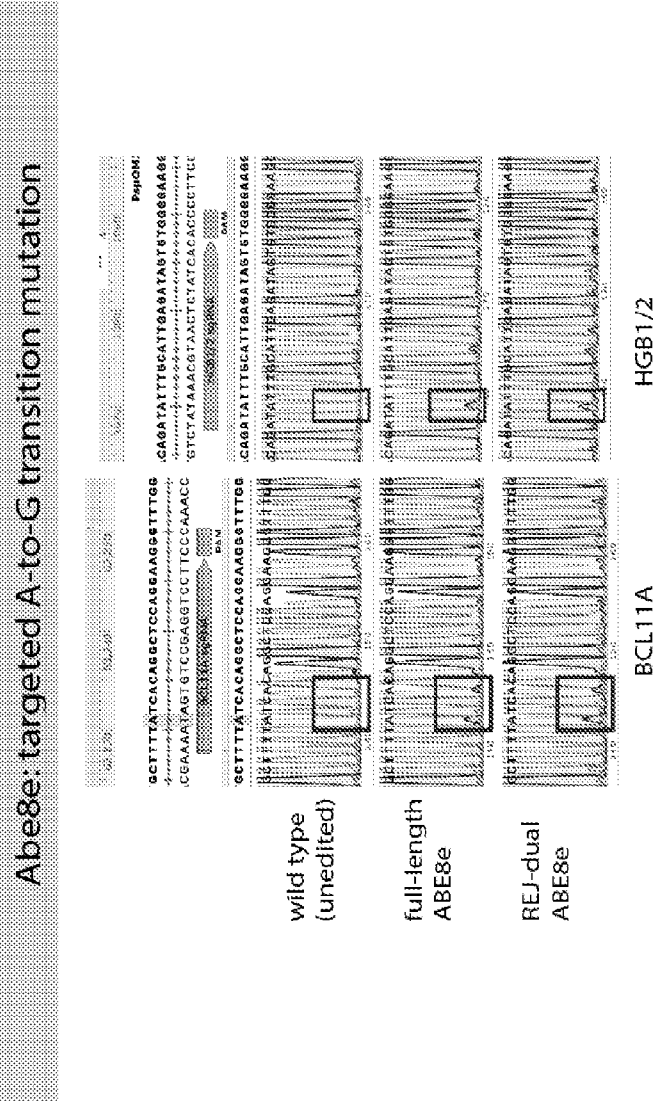




FIG. 24A

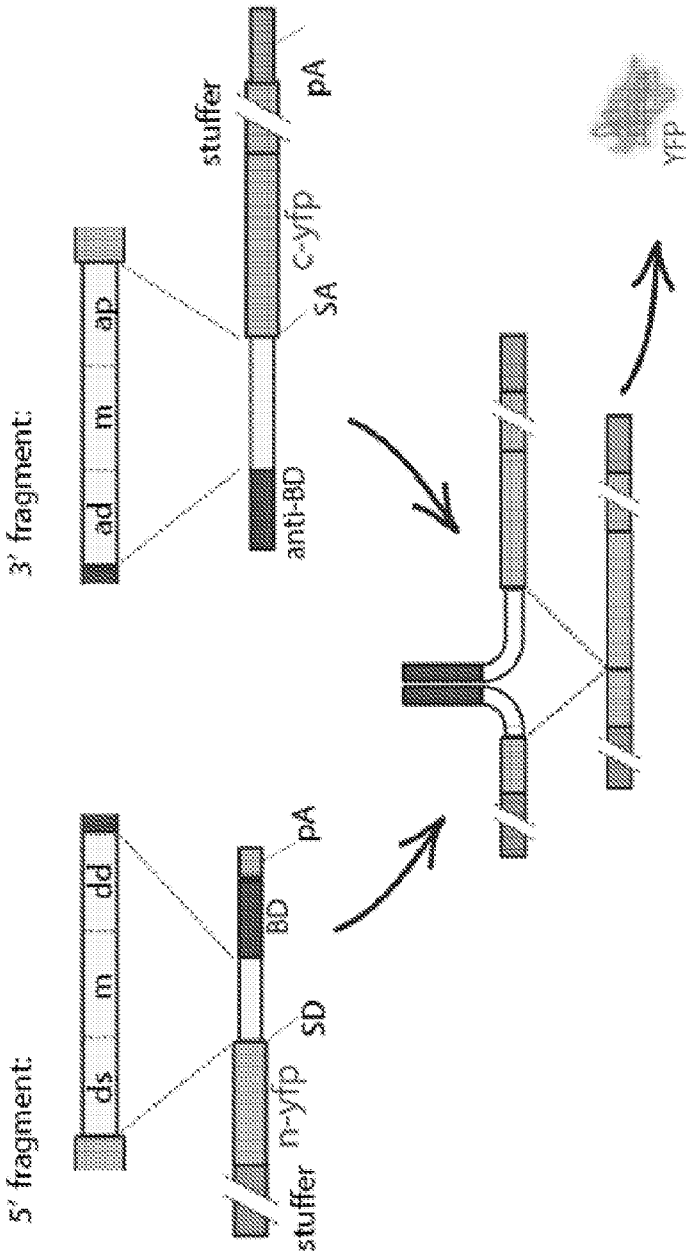


FIG. 24B

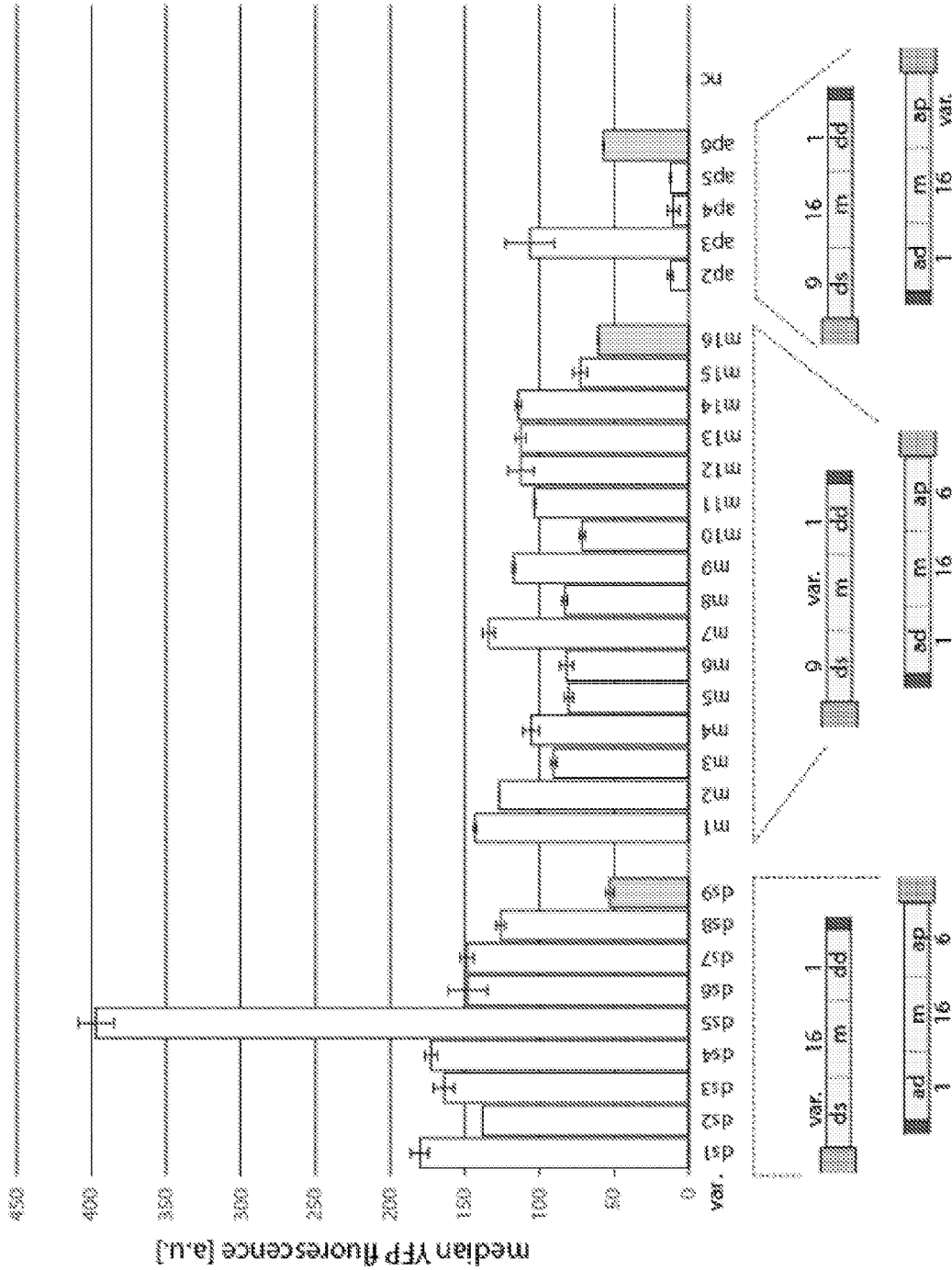


FIG. 24C

|                               |                              |                          |                      |                                |                      |
|-------------------------------|------------------------------|--------------------------|----------------------|--------------------------------|----------------------|
| Donor downstream segment (ds) |                              | Mid-intronic segment (m) |                      | Acceptor proximal segment (ap) |                      |
| ds1                           | FGFR-2 pre-mRNA, IAS1        | m1                       | 6xWGGG               | ap1                            | scramble             |
| ds2                           | Fas UR16                     | m2                       | C9-E6-ISE            | ap2                            | 4.1R-E16-uISE        |
| ds3                           | CFTR E9 PY1/2                | m3                       | AdMLBPadj            | ap3                            | P6-cons-to30         |
| ds4                           | TIA1-preferred               | m4                       | bcl2-l2-BPadj        | ap4                            | AdMLSA               |
| ds5                           | FAS intron5                  | m5                       | M2                   | ap5                            | bcl2-l2-SA           |
| ds6                           | CD46-D1/9                    | m6                       | GH11vs               | ap6                            | perfectT (reference) |
| ds7                           | B19V ISE                     | m7                       | WangGrich            | Acceptor distal segment (ad)   |                      |
| ds8                           | ratFGFR2DISE                 | m8                       | WangISE2             |                                |                      |
| ds9                           | donor & scramble (reference) | m9                       | WangA                | ad1                            | scramble (reference) |
| Donor distal segment (dd)     |                              | m10                      | WangB                |                                |                      |
|                               |                              | m11                      | WangC                |                                |                      |
|                               |                              | m12                      | WangD                |                                |                      |
|                               |                              | m13                      | WangE                |                                |                      |
|                               |                              | m14                      | WangFmod             |                                |                      |
|                               |                              | m15                      | SMN-URC2/3           |                                |                      |
|                               |                              | m16                      | scramble (reference) |                                |                      |

# SEQUENCE LISTING

<110> Salk Institute for Biological Studies

<120> COMPOSITIONS AND METHODS FOR HIGH-EFFICIENCY RECOMBINATION OF RNA MOLECULES

<130> 7158-102574-07

<150> 62/933,714

<151> 2019-11-11

<150> PCT/US2020/025430

<151> 2020-03-27

<160> 206

<170> PatentIn version 3.5

<210> 1

<211> 1491

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 1

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60  |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120 |
| atgggtggag tattttacggt aaactgcccc cttggcagta catcaagtgt atcatatgcc | 180 |
| aagtagcccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240 |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300 |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360 |
| atttccaagt ctccaccccc ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420 |
| ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt   | 480 |
| acggtgggag gtctatataa gcagagcttg gatgttgctt ttacttctag gcgcgccgcc  | 540 |
| accatggtga gcaagggcga ggagctgttc accgggggtg tgcccatcct ggtcgagctg  | 600 |
| gacggcgacg taaacggcca caagttcagc gtgtccggcg agggcgaggg cgatgccacc  | 660 |
| tacggcaagc tgaccctgaa gttcatctgc accaccggca agctgcccgt gccctggccc  | 720 |
| accctcgtga ccaccttcgg ctacggcctg atgtgcttcg cccgctaccc cgaccacatg  | 780 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| aagcagcacg acttcttcaa gtccgccatg cccgaaggct acgtccagga gcgccaccatc | 840  |
| ttcttcaagg acgacggcaa ctacaagacc cgcgccgagg tgaagttcga gggcgacacc  | 900  |
| ctggtgaacc gcatcgagct gaagggcatc gacttcaagg aggacggcaa catcctgggg  | 960  |
| cacaagctgg agtacaacta caacagccac aacgtctata tcatggccga caagcagaag  | 1020 |
| aacggcatca aggtaagtat tagctctttc tttccatggg ttggcctcgc cgcgtgggct  | 1080 |
| gaggggaagga ctgtcctggg actggacagg cgggttatgg gacctgaaaa gcggccctga | 1140 |
| aaaagggccg cgatgaaaac gaagcgagct aaagcctcct ctctcttctt cagaactcct  | 1200 |
| ctcttttctc tcctccagga gttcttcctc tctcccttct tctcaaatgc tttctccctc  | 1260 |
| tctcctgcat ttgagctcct tctttcctct ctcgacaatc cccttttctc cctcttgatt  | 1320 |
| gtcgactagc tcgcaatcat cgcggtatca aaaagcggtc aggcagctaa accaaaaggt  | 1380 |
| ttagcaattg cctctgatga gtcgctgaaa tgcgacgaaa accgcttttt ggtaccaata  | 1440 |
| aaatatcttt attttcatta catctgtgtg ttgggttttt gtgtgactag t           | 1491 |

<210> 2

<211> 1302

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 2

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgccatt  | 60  |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca | 120 |
| atgggtggag tattttacggt aaactgccc cttggcagta catcaagtgt atcatatgcc | 180 |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta | 240 |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac | 300 |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg | 360 |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg | 420 |
| ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt  | 480 |
| acggtgggag gtctatataa gcagagcttg gatgttgcct ttacttctag gcgcgccgcg | 540 |
| gaaaaccgcg ggataccgcg atgattgcga gctagtcgac aatcaagagg gagaaaaggg | 600 |

|                                                                     |      |
|---------------------------------------------------------------------|------|
| gattgtcgag agaggaaaga aggagctcaa atgcaggaga gagggagaaa gcatttgaga   | 660  |
| agaagggaga gaggaacaac tcgtggagga gagaaaagag acgagttgtg aagaagagag   | 720  |
| aggaggcttt agctcgcttc gttttcatca ttattgcggc cctgaaaaag ggccgcttat   | 780  |
| aacgttgctc gaattcgggt tatgggacca gtgaaggctg aggggaaggac tgtcctggga  | 840  |
| ctggacaggc gggttatggg acctgaaaat actaacaatc gatTTTTTTTT cccTTTTTTTT | 900  |
| ccaggtgaac ttcaagatcc gccacaacat cgaggacggc agcgtgcagc tcgccgacca   | 960  |
| ctaccagcag aacaccccca tcggcgacgg ccccgctgtg ctgcccgaca accactacct   | 1020 |
| gagctaccag tccgccctga gcaaagacct caacgagaag cgcgatcaca tggtcctgct   | 1080 |
| ggagttcgtg accgccgccg ggatcactct cggcatggac gagctgtaca aggaccttg    | 1140 |
| agaattcctc acctgcgac tcgatgcttt atttgtgaaa tttgtgatgc tattgcttta    | 1200 |
| tttgtaacca ttataagctg caataaaca gttacaaca acaattgcat tcattttatg     | 1260 |
| tttcaggttc agggggaggt gtgggaggtt ttttaaacta gt                      | 1302 |

<210> 3  
 <211> 404  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| <400> 3                                                           |     |
| gtaagtatta gctctttctt tccatgggtt ggcctcgccg cgtgggctga gggaaggact | 60  |
| gtcctgggac tggacaggcg ggttatggga cctgaaaagc ggccctgaaa aagggccgcg | 120 |
| atgaaaacga agcgagctaa agcctcctct ctcttcttca gaactcctct cttttctctc | 180 |
| ctccaggagt tcttcctctc tcccttcttc tcaaagtctt tctccctctc tcctgcattt | 240 |
| gagctccttc tttcctctct cgacaatccc cttttctccc tcttgattgt cgactagctc | 300 |
| gcaatcatcg cggtatcaaa aagcggtcag gcagctaaac caaaagggtt agcaattgcc | 360 |
| tctgatgagt cgctgaaatg cgacgaaaac cgcttttttg tacc                  | 404 |

<210> 4  
 <211> 382  
 <212> DNA  
 <213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 4

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| acttctaggc gcgccgcgga aaaccgcggg ataccgcgat gattgcgagc tagtcgacaa  | 60  |
| tcaagaggga gaaaagggga ttgtcgagag aggaaagaag gagctcaaat gcaggagaga  | 120 |
| gggagaaagc atttgagaag aaggagagaga ggaacaactc gtggaggaga gaaaagagac | 180 |
| gagttgtgaa gaagagagag gaggcttttag ctcgcttcgt tttcatcatt attgcggccc | 240 |
| tgaaaaaggg ccgcttataa cgttgctcga attcgggtta tgggaccagt gaaggctgag  | 300 |
| ggaaggactg tcctgggact ggacaggcgg gttatgggac ctgaaaatac taacaatcga  | 360 |
| ttttttttcc ctttttttcc ag                                           | 382 |

<210> 5

<211> 489

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 5

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| atggtgagca agggcgagga gctgttcacc ggggtggtgc ccatcctggt cgagctggac | 60  |
| ggcgacgtaa acggccacaa gttcagcgtg tccggcgagg gcgagggcga tgccacctac | 120 |
| ggcaagctga ccctgaagtt catctgcacc accggcaagc tgcccgtgcc ctggcccacc | 180 |
| ctcgtgacca ctttcggcta cggcctgatg tgcttcgccc gctaccccga ccacatgaag | 240 |
| cagcacgact tcttcaagtc cgccatgccc gaaggctacg tccaggagcg caccatcttc | 300 |
| ttcaaggacg acggcaacta caagaccgc gccgaggtga agttcgaggg cgacaccctg  | 360 |
| gtgaaccgca tcgagctgaa gggcatcgac ttcaaggagg acggcaacat cctggggcac | 420 |
| aagctggagt acaactacaa cagccacaac gtctatatca tggccgacaa gcagaagaac | 480 |
| ggcatcaag                                                         | 489 |

<210> 6

<211> 237

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 6  
gtgaacttca agatccgccca caacatcgag gacggcagcg tgcagctcgc cgaccactac 60  
cagcagaaca ccccatcgga cgacggcccc gtgctgctgc ccgacaacca ctacctgagc 120  
taccagtccg ccctgagcaa agaccccaac gagaagcgcg atcacatggg cctgctggag 180  
ttcgtgaccg ccgccgggat cactctcggc atggacgagc tgtacaagga cctttga 237

<210> 7  
<211> 382  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 7  
acttctaggc gcgccgcgga aaaccgcggg ataccgcgat gattgcgagc tagggagaga 60  
gaggggaaaag aaaagagaaa gaggaggagg aaagagggga gagaggggag ggaaaggaga 120  
gaaggaggga aggggaagaaa gaaagaagag gaaaagaggg gaggaggagg agaaaggaga 180  
aaaaaagaag ggaagggaga aaggcttttag ctcgcttcgt tttcatcatt attgcggccc 240  
tgaaaaaggg ccgcttataa cgttgctcga attcgggtta tgggaccagt gaaggctgag 300  
ggaaggactg tcctgggact ggacaggcgg gttatgggac ctgaaaatac taacaatcga 360  
ttttttttcc ctttttttcc ag 382

<210> 8  
<211> 301  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 8  
gtaagtgtcc cgcggaacat tattataacg ttgctcgaag atatcagatg gtgcgctcct 60  
ggacgtagcc ttcgggcatg gcggacttga agaagtcgtg ctgcttcatt tggtcggggg 120  
agcggctgaa gcaactgcac ccgtaggtca ggggtgggtcac gaggggtgggc caggggcacgg 180  
gcagcttgcc ggtggtgcag atgaacttca gggtcagctt gccgtaggtg gcatcgccct 240  
cgccctcgcc ggacacgctg aacttgtggc cgtttacgtc gccgtccagc tcgactctag 300



<210> 9  
<211> 326  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 9  
gctagcgtcg agctggacgg cgacgtaaac ggccacaagt tcagcgtgtc cggcgagggc 60  
gagggcgatg ccacctacgg caagctgacc ctgaagttca tctgcaccac cggcaagctg 120  
cccgtgccct ggcccaccct cgtgaccacc ctgacctacg gcgtgcagtg cttcagccgc 180  
taccgacgacc acatgaagca gcacgacttc ttcaagtccg ccatgcccga aggctacgtc 240  
caggagcgca ccatctccgc ggaacattat tataacgttg ctcgaatact aactggtacc 300  
tcttcttttt tttttgatat ctgcag 326

<210> 10  
<211> 278  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 10  
gttgcccttta cttctggcgc gccaaaaggc gtgccagaag taccgggcta ataatgtttc 60  
gcggctcctt taaatctgcc taaatacgta taaatttgat cgccctgaaa aaggcgcatc 120  
aaagccctga aaaaggcat acgtagccct gaaaaagggc aggagagcc ctgaaaaagg 180  
gcaagaggac gcggaacat tattagccgc caccatggac aggcgggtta tgggacctga 240  
aaatactaac aatcgatttt ttttcccttt ttttccag 278

<210> 11  
<211> 190  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 11

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| acttctaggc gcgccgcgga aaaccgcggg atatcattat tgcggccctg aaaaagggcc | 60  |
| gcttataacg ttgctcgaat tcgggttatg ggaccagtga aggctgaggg aaggactgtc | 120 |
| ctgggactgg acaggcgggt tatgggacct gaaaatacta acaatcgatt ttttttcct  | 180 |
| ttttttccag                                                        | 190 |

<210> 12  
 <211> 459  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| <400> 12                                                           |     |
| gtaagtatta gctctttctt tccatgggtt ggcctcgccg cgtgggctga gggaaggact  | 60  |
| gtcctgggac tggacaggcg gggttatggga cctgaaaagc ggccctgaaa aagggccgcg | 120 |
| atgaaaacga agcgagctaa agcctcctct ctcttcttca gaactcctct cttttctctc  | 180 |
| ctccaggagt tcttcctctc tcccttcttc tcaaattgctt tctccctctc tcctgcattt | 240 |
| gagctccttc tttcctctct cgacaatccc cttttctccc tcttgattgt cgactagctc  | 300 |
| gcaatcatcg cggtatcaaa aagcggtcag gcagctaaac caaaagggtt agcaattgcc  | 360 |
| tctgatgagt cgctgaaatg cgacgaaaac cgcttttttg taccaataaa atatctttat  | 420 |
| tttcattaca tctgtgtgtt ggttttttgt gtgactagt                         | 459 |

<210> 13  
 <211> 382  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| <400> 13                                                           |     |
| acttctaggc gcgccgcgga aaaccgcggg ataccgcgat gattgcgagc tagtcgacaa  | 60  |
| tcaagaggga gaaaagggga ttgtcgagag aggaaagaag gagctcaaatt gcaggagaga | 120 |
| gggagaaagc atttgagaag aaggagagaga ggaagaactc ctggaggaga gaaaagagag | 180 |
| gagttctgaa gaagagagag gaggcttttag ctcgcttcgt tttcatcatt attgcggccc | 240 |
| tgaaaaaggg ccgcttataa cgttgctcga attcgggtta tgggaccagt gaaggctgag  | 300 |

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| ggaaggactg tcctgggact ggacaggcgg gttatgggac ctgaaaatac taacaatcga | 360 |
| ttttttttcc ctttttttcc ag                                          | 382 |

<210> 14  
 <211> 372  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| <400> 14                                                           |     |
| gtaagtatta agcggccctg aaaaagggcc gcgatgaaaa cgaagcgagc taaagcctcc  | 60  |
| tctctcttct tcagaactcc tctcttttct ctctccagg agttcttcct ctctcccttc   | 120 |
| ttctcaaattg ctttctccct ctctcctgca tttagactcc ttctttcctc tctcgacaat | 180 |
| ccccctttct ccctcttgat tgtcgactag ctcgcaatca tcgcggtatc aaaaagcggc  | 240 |
| caggcagcta aaccaaaagg ttttagcaatt gcctctgatg agtcgctgaa atgcgacgaa | 300 |
| aaccgctttt tggtaccaat aaaatatctt tattttcatt acatctgtgt gttggttttt  | 360 |
| tgtgtgacta gt                                                      | 372 |

<210> 15  
 <211> 407  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| <400> 15                                                          |     |
| gtaagtatta gctcttttct tccatgggtt ggcctcgccg cgtgaagcgg ccctgaaaaa | 60  |
| gggccgcgat gaaaacgaag cgagctaaag cctcctctct cttcttcaga actcctctct | 120 |
| tttctctcct ccaggagttc ttctctctc cttctttctc aaatgctttc tccctctctc  | 180 |
| ctgcatttga gtcctttctt tcctctctcg acaatcccct tttctccctc ttgattgtcg | 240 |
| actagctcgc aatcatcgcg gtatcaaaaa gcggtcaggc agctaaacca aaaggtttag | 300 |
| caattgcctc tgatgagtcg ctgaaatgcg acgaaaaccg ctttttggtg ccaataaaat | 360 |
| atctttatct tcattacatc tgtgtgttgg ttttttgtgt gactagt               | 407 |

<210> 16

<211> 378  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 16  
gtaagtatta gctctttctt tccatgggtt ggcctcgccg cgtgggctga gggaaggact 60  
gtcctgggac tggacaggcg ggttatggga cctgaaaagc ggccctgaaa aagggccgcg 120  
atgaaaacga agcgagctaa agcctcctct ctcttcttca gaactcctct cttttctctc 180  
ctccaggagt tcttcctctc tcccttcttc tcaaagtctt tctccctctc tcctgcattt 240  
gagctccttc tttcctctct cgacaatccc cttttctccc tcttgattgt cgactagctc 300  
gcaatcatcg cggtatcggt accaataaaa tatctttatt ttcattacat ctgtgtgttg 360  
gttttttgtg tgactagt 378

<210> 17  
<211> 309  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 17  
acttctaggc gcgccgcgga aaaccgcggg ataccgcgat gattgcgagc tagtcgacaa 60  
tcaagaggga gaaaagggga ttgtcgagag aggaaagaag gagctcaaat gcaggagaga 120  
gggagaaagc atttgagaag aaggagagaga ggaagaactc ctggaggaga gaaaagagag 180  
gagttctgaa gaagagagag gaggcttttag ctcgcttcgt tttcatcatt attgcggccc 240  
tgaaaaaggg ccgcttataa cgttgctcga attctactaa caatcgattt tttttccctt 300  
tttttccag 309

<210> 18  
<211> 419  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 18

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| atatcctttt agggcagagt gaagagttag gaggaagggtg gttgggagag ggatttccag | 60  |
| gccttaggac atcatgacag atgaaaacga agcgagctaa agcctcctct ctcttcttca  | 120 |
| gaactcctct cttttctctc ctccaggagt tcttcctctc tcccttcttc tcaaagtgtt  | 180 |
| tctccctctc tcctgcattt gagctccttc tttcctctct cgacaatccc cttttctccc  | 240 |
| tcttgattgt cgactagctc gcaatcatcg cggtatcaaa aagcggtcag gcagctaaac  | 300 |
| caaaagggtt agcaattgcc tctgatgagt cgctgaaatg cgacgaaaac cgctttttgg  | 360 |
| taccaataaa atatctttat tttcattaca tctgtgtgtt ggttttttgt gtgactagt   | 419 |

<210> 19  
 <211> 275  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| <400> 19                                                           |     |
| acttctaggc gcgccgcgga aaaccgcggg ataccgcgat gattgcgagc tagtcgacaa  | 60  |
| tcaagaggga gaaaagggga ttgtcgagag aggaaagaag gagctcaaat gcaggagaga  | 120 |
| gggagaaagc atttgagaag aaggagagaga ggaacaactc gtggaggaga gaaaagagac | 180 |
| gagttgtgaa gaagagagag gaggccttag ctgccttcgt tttcatcatt tccaggcctt  | 240 |
| aggacatcat gacatttttc cttacttttg ctcac                             | 275 |

<210> 20  
 <211> 3975  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| <400> 20                                                           |     |
| acttctaggc gcgccccac catgggattc gtgcggcaga ttcagctgct gctgtggaag   | 60  |
| aactggaccc tgcggaagcg gcagaaaatc agattcgtgg tggaactcgt gtggcccctg  | 120 |
| agcctgtttc tgggtgctgat ctggctgcgg aacgccaatc ctctgtacag ccaccacgag | 180 |
| tgtcacttcc ccaacaaggc catgccttct gccggaatgc tgccttggct gcagggcac   | 240 |
| ttctgcaacg tgaacaaccc ctgctttcaa agccccacac ctggcgaaag ccctggcatc  | 300 |

|            |             |            |             |             |            |      |
|------------|-------------|------------|-------------|-------------|------------|------|
| gtgtccaact | acaacaacag  | catcctggcc | agagtgtacc  | gggacttcca  | agagctgctg | 360  |
| atgaacgccc | ctgagtctca  | gcacctgggc | agaatctgga  | ccgagctgca  | catcctgagc | 420  |
| cagttcatgg | acaccctgag  | aacacacccc | gagagaatcg  | ccggcagggg  | catcagaatc | 480  |
| cgggacatcc | tgaaggacga  | ggaaaccctg | acactgttcc  | tcatcaagaa  | catcggcctg | 540  |
| agcgacagcg | tggtgtacct  | gctgatcaac | agccaagtgc  | ggcccagagca | gtttgctcat | 600  |
| ggcgtgccag | atctcgccct  | gaaggatatc | gcctgttctg  | aggccctgct  | ggaacggttc | 660  |
| atcatcttca | gccagcggag  | aggcgccaag | accgtcagat  | atgccctgtg  | cagtctgagc | 720  |
| cagggaaccc | tgcaagtggat | cgaggatacc | ctgtacgcca  | acgtggactt  | cttcaagctg | 780  |
| ttccgggtgc | tgcccacact  | gctggattct | cgggtcccaag | gcatcaacct  | gagaagctgg | 840  |
| ggcggcatcc | tgtccgacat  | gagcccaaga | atccaagagt  | tcatccaccg  | gcctagcatg | 900  |
| caggacctgc | tgtgggttac  | cagacctctg | atgcagaacg  | gcggacccga  | gacattcacc | 960  |
| aagctgatgg | gcattctgag  | cgatctgctg | tgcggctacc  | ctgaaggcgg  | aggatctaga | 1020 |
| gtgctgagct | tcaattggta  | cgaggacaac | aactacaagg  | ccttcctggg  | catcgactcc | 1080 |
| accagaaagg | accccatcta  | cagctacgac | cggcggacaa  | ccagcttctg  | caatgccctg | 1140 |
| atccagagcc | tggaagcaa   | ccctctgacc | aagatcgctt  | ggagggccgc  | caaacctctg | 1200 |
| ctgatgggaa | agatcctgta  | caccctgac  | agccctgccg  | ccagaagaat  | cctgaagaac | 1260 |
| gccaacagca | ccttcgagga  | actggaacac | gtgcgcaagc  | tggtcaaggc  | ctgggaagaa | 1320 |
| gtgggacctc | agatctggta  | cttcttcgac | aatagcacc   | agatgaacat  | gatcagagac | 1380 |
| accctgggca | accctaccgt  | gaaggacttc | ctgaacagac  | agctgggcga  | agagggcatt | 1440 |
| accgccgagg | ccatcctgaa  | ctttctgtac | aaggggccca  | gagagtcca   | ggccgacgac | 1500 |
| atggccaact | tcgattggcg  | ggacatcttc | aacatcaccc  | acagaaccct  | gcggctggtc | 1560 |
| aaccagtacc | tggaatgcct  | ggtgctggac | aagtctgaga  | gctacaacga  | cgagacacag | 1620 |
| ctgaccaga  | gagccctgtc  | tctgctggaa | gagaatatgt  | tctgggctgg  | cgtggtgttc | 1680 |
| cccgacatgt | acccttggac  | aagcagcctg | cctcctcacg  | tgaagtacaa  | gatccggatg | 1740 |
| gacatcgacg | tggtcgaaaa  | gaccaacaag | atcaaggacc  | ggtactggga  | cagcggccct | 1800 |
| agagctgata | ccgtggaaga  | ttttcgctac | atctggggcg  | gattcgcata  | cctgcaggac | 1860 |
| atggtggaac | agggaatcac  | acggtcccag | gtgcaggctg  | aagctcctgt  | gggaatctac | 1920 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| ctgcagcaga tgccttatcc ttgcttcgtg gacgacagct tcatgatcat cctgaatcgg  | 1980 |
| tgcttcccca tcttcatggt gctggcctgg atctactccg tgtctatgac cgtgaagtcc  | 2040 |
| atcgtgctgg aaaaagagct gcggctgaaa gagacactga agaaccaggg cgtgtccaat  | 2100 |
| gccgtgatct ggtgcacctg gtttctggac agcttctcca ttatgagcat gagcatcttt  | 2160 |
| ctgctgacga tcttcatcat gcacggccgg atcctgcact acagcgaccc ctttatcctc  | 2220 |
| ttcctgttcc tgctggcctt ctccaccgct acaatcatgc tgtgttttct gctgtccacc  | 2280 |
| ttcttctcca aagcctctct ggccgctgct tgtagcggcg tgatctactt caccctgtac  | 2340 |
| ctgcctcaca tcctgtgctt cgcatggcag gacagaatga ccgccgagct gaagaaagct  | 2400 |
| gtgtccctgc tgagccctgt ggcctttggc tttggcaccg agtacctcgt cagatttgag  | 2460 |
| gaacaaggac tgggactgca gtggtccaac atcggcaata gccctacaga gggcgacgag  | 2520 |
| ttcagcttcc tgctgtctat gcaaattgatg ctgctggacg ccgccgtgta tggactgctg | 2580 |
| gcttggtatc tggaccaggt gttccctgcc gattacggca ctctctgcc ttggtatttc   | 2640 |
| ctgctgcaag agagctactg gctcggcggc gagggatgta gcaccagaga agaaagagcc  | 2700 |
| ctggaaaaga ccgagcctct gaccgaggaa acagaggacc ctgaacaccc agagggcac   | 2760 |
| cacgatagct ttttcgagag agaacacccc ggctgggtgc caggcgtgtg tgtgaagaat  | 2820 |
| ctggtcaaga tcttcgagcc ctgcggcaga cctgccgtgg acagactgaa catcaccttc  | 2880 |
| tacgagaacc agattaccgc ctttctgggc cacaacggcg ctggcaagac aaccacactg  | 2940 |
| agcatcctca ccggcctgct gcctccaaca agcggcacag ttctcgttgg cggcagagac  | 3000 |
| atcgagacaa gcctggatgc cgtcagacag tccctgggca tgtgccctca gcacaacatc  | 3060 |
| ctgtttcacc acctgaccgt ggccgagcac atgctgtttt atgccagct gaagggcaag   | 3120 |
| agccaagaag aggctcagct ggaaatggaa gccatgctcg aggacaccgg cctgcaccac  | 3180 |
| aagagaaatg aggaagccca ggatctgagc ggcggcatgc agagaaaact gagcgtggcc  | 3240 |
| attgccttcg tgggcgacgc caaggttgtg atcctggatg agcctacaag cggcgtggac  | 3300 |
| ccttacagca gaagatccat ctgggatctg ctgctgaagt acagaagcgg ccggaccatc  | 3360 |
| atcatgagca cccaccacat ggacgaggcc gatctgctcg gagacagaat cgccatcatt  | 3420 |
| gctcagggca gactgtactg cagcggcacc cactgtttc tgaagaactg tttcggcacc   | 3480 |
| ggactgtatc tgaccctcgt gcggaagatg aagaacatcc agtctcagcg gaagggcagc  | 3540 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| gagggcacct gtagctgttc tagcaagggc tttagcacca cctgtccagc tcacgtggac  | 3600 |
| gatctgaccc ctgaacaggt gctggatggc gacgtgaacg agctgatgga cgtgggtgctg | 3660 |
| caccatgtgc ctgaggccaa gctgggtggaa tgcacggcc aggtaagtat tagctctttc  | 3720 |
| tttccatggg ttggcctcgc cgcgtgggct gagggaagga ctgtcctggg actggacagg  | 3780 |
| cgggttatgg gacctgaagc gataaaaggc atgcacgttt gcggctacgt gcatgccaaa  | 3840 |
| aggagtcggg cttgcctccg tgcccgactc caaaagacct gctcgaggag gtggacgagc  | 3900 |
| aggtcaaaaa tccgggtacc aataaaatat ctttattttc attacatctg tgtgttggtt  | 3960 |
| ttttgtgtga ctagt                                                   | 3975 |

<210> 21

<211> 3611

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 21

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| aggatTTTTg acctgctcga ttgtccactg cgagcaggtc ttttggagtc gggcgaggcg | 60  |
| gaagcccgac tccttttggc atgcacgcta gccgcgtcgt gcatgccttt tatcgaattc | 120 |
| gggttatggg accagtgaag gctgagggaa ggactgtcct gggactggac aggcgggtta | 180 |
| tgggacctga aaatactaac aatcgatTTT ttttcccttt ttttccagga actgattttt | 240 |
| ctgctcccga acaagaactt caagcaccgg gcctacgcca gcctgttcag agagctggaa | 300 |
| gaaaccctgg ccgacctggg cctgtctagc tttggcatca gcgacacccc tctcgaagag | 360 |
| atcttcctga aagtgcaga ggacagcgat agcggccctc tgtttgctgg cggagcacag  | 420 |
| caaaagcgcg agaacgtgaa ccctagacac ccctgtctgg gcccaagaga gaaagccgga | 480 |
| cagaccctc aggacagcaa tgtgtgctct cctggtgctc ctgccgctca tcctgaggga  | 540 |
| caacctccac ctgaacctga gtgtcctgga cctcagctga acaccggaac acagctggtt | 600 |
| ctgcagcacg tgcaggctct gctcgtgaag agattccagc acaccatcag aagccacaag | 660 |
| gactttctgg ccagatcgt gctgcccgcc acctttgttt ttctggctct gatgctgagc  | 720 |
| atcgtgatcc ctccattcgg cgagtacccc gctctgacac tgcacccttg gatctacggc | 780 |
| cagcagtaca cttttttctc catggacgaa cccggcagcg agcagttcac agtgctggct | 840 |



|                                                                    |      |
|--------------------------------------------------------------------|------|
| gatgtcctgc tgaacaagcc cggcttcggc aaccggtgtc tgaaagaagg atggctgcct  | 900  |
| gagtaccctt gcggcaacag cacaccttgg aaaacccta gcgtgtcccc taacatcacc   | 960  |
| cagctgttcc aaaagcagaa atggacccaa gtgaaccctt ctccatcctg ccggtgctcc  | 1020 |
| acaagggaag agctgacat gctgcccag tgtccagaag gcgtggcgg acttcctcca     | 1080 |
| cctcagagaa cacagagatc caccgagatt ctccaggacc tgaccgaccg gaatatcagc  | 1140 |
| gacttcctgg ttaagacata cccgcactg atccgggtcca gcctgaagtc caagttctgg  | 1200 |
| gtcaacgaac agagatacgg cggcatcagc atcggcggaa aactgcctgt ggtgcctatc  | 1260 |
| acaggcgagg cccttggtggg ctttctgtcc gatctgggga gaatcatgaa cgtgtccggc | 1320 |
| ggacctatca ccagggaagc cagcaaagag atccccgatt tcctgaagca cctggaaacc  | 1380 |
| gaggacaata tcaaagtgtg gttcaacaac aaaggatggc acgccctcgt gtcttttctg  | 1440 |
| aacgtggccc acaatgccat cctgcgggct agcctgccta aggacagaag ccctgaggaa  | 1500 |
| tacggcatca ccgtgatctc ccagcctctg aatctgacca aagagcagct gagcgagatc  | 1560 |
| accgtgctga ccacctctgt ggatgctgtg gtggccatct gcgtgatctt cagcatgagc  | 1620 |
| ttcgtgcccc cctccttcgt gctgtacctg attcaagaga gagtgaacaa gagcaagcac  | 1680 |
| ctccagttca tctccggggt gtccccaacc acctactggg tcaccaatct tctgtgggac  | 1740 |
| atcatgaact acagcgtgtc agccggcctg gtcgtgggca tctttatcgg ctttcaaaag  | 1800 |
| aaggcctaca cgagccccga gaacctgcct gctttggttg ctctgctgct cctgtatggc  | 1860 |
| tgggccgtga ttcccatgat gtaccccgcc agctttctgt ttgacgtgcc cagcacagcc  | 1920 |
| tacgtggccc tgtcttgccg caatctgttc atcggcatca acagcagcg catcacattc   | 1980 |
| atcctggaac tgttcgagaa caacaggacc ctgctgcggg tcaacgccgt gctgcggaaa  | 2040 |
| ctgctgatcg tgttccctca cttctgtctc ggccggggcc tgatcgacct ggctctgtct  | 2100 |
| caagccgtga ccgatgtgta cgccagattt ggcgaggaa actccgcaa tccattccac    | 2160 |
| tgggacctga tcggcaagaa cctgttcgcc atggtggtgg aaggcgtcgt gtacttcctg  | 2220 |
| ctcactctgc tggatgcagag acactttttt ctgtcccaat ggatcgccga gcctacaaa  | 2280 |
| gaaccattg tggacgagga cgacgatgtg gccgaggaaa gacagagaat catcaccggc   | 2340 |
| ggcaacaaga ccgatatcct gagactgcac gagctgacaa agatctaccc cggcacaagc  | 2400 |
| tccccagccg tggataggct ttgtgtggga gttagaccg gcgagtgtt tggcctgctg    | 2460 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| ggagttaatg gcgccgaaa gaccaccacc ttcaagatgc tgaccggcga caccacagt    | 2520 |
| acaagcggag atgctacagt ggccggcaag agcatcctga ccaacatcag cgaagtgc    | 2580 |
| cagaacatgg gctactgccc tcagttcgac gccatcgacg aactgctgac aggccgcgaa  | 2640 |
| cacctgtatc tgtatgccag actgagaggc gtgcccgtg aagagatcga gaaggtggcc   | 2700 |
| aactggtcca tcaagtctct gggcctgaca gtgtacgccg actgtctggc cggaacatac  | 2760 |
| agcggaggaa acaagcggaa gctgagcacc gccattgctc tgatcggatg cccacctctg  | 2820 |
| gtcctgctgg atgaaccac caccggaatg gatccccagg ctagaagaat gctctggaac   | 2880 |
| gtgatcgtgt ctatcatccg cgagggcaga gctgtggtgc tgacctctca ctccatggaa  | 2940 |
| gagtgcgagg ctctgtgtac ccggctggcc attatggtca agggcgctt cagatgcatg   | 3000 |
| ggcaccattc agcatctgaa aagcaagttc ggcgacggct acatcgtgac aatgaagatc  | 3060 |
| aagagcccca aggacgacct cctgcctgat ctgaaccccg tggaacagtt ttttcagggc  | 3120 |
| aacttccccg gctccgtgca gcgggaaaga cactataaca tgctgcagtt tcaggtgtcc  | 3180 |
| tcctccagcc tggctcggat ctttcaactg ctgctctctc acaaggacag cctgctgatt  | 3240 |
| gaagagtaca gcgtgacaca gaccacactc gaccaggttt tcgtgaactt cgccaagcag  | 3300 |
| cagaccgaga gccacgacct gcctctgcat cctcgggccg ctggtgcctc tagacaagct  | 3360 |
| caggacggcg ctcgggctga ctacaaagac catgacggtg attataaaga tcatgacatc  | 3420 |
| gactataagg atgacgatga caaatgaggt accaattcct cacctgcgat ctcgagcttt  | 3480 |
| atttgtgaaa tttgtgatgc tattgcttta tttgtaacca ttataagctg caataaaca   | 3540 |
| gttaacaaca acaattgcat tcattttatg tttcaggttc agggggagggt gtgggaggtt | 3600 |
| ttttaaacta g                                                       | 3611 |

<210> 22

<211> 3975

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 22

|                                                                  |    |
|------------------------------------------------------------------|----|
| acttctaggc gcgccccac catggcccca aagaagaagc ggaaggctcg tatccacgga | 60 |
|------------------------------------------------------------------|----|

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| gtcccagcag ccaagcggaa ctacatcctg ggcctggaca tcggcatcac cagcgtgggc | 120 |
|-------------------------------------------------------------------|-----|

|                                                                    |      |
|--------------------------------------------------------------------|------|
| tacggcatca tcgactacga gacacgggac gtgatcgatg ccggcgtgcg gctgttcaaa  | 180  |
| gaggccaacg tggaaaacaa cgagggcagg cggagcaaga gaggcgccag aaggctgaag  | 240  |
| cggcggaggc ggcatagaat ccagagagtg aagaagctgc tgttcgacta caacctgctg  | 300  |
| accgaccaca gcgagctgag cggcatcaac ccctacgagg ccagagtga gggcctgagc   | 360  |
| cagaagctga gcgaggaaga gttctctgcc gccctgctgc acctggccaa gagaagaggc  | 420  |
| gtgcacaacg tgaacgaggt ggaagaggac accggcaacg agctgtccac caaagagcag  | 480  |
| atcagccgga acagcaaggc cctggaagag aaatacgtgg ccgaactgca gctggaacgg  | 540  |
| ctgaagaaag acggcgaagt gcggggcagc atcaacagat tcaagaccag cgactacgtg  | 600  |
| aaagaagcca aacagctgct gaagggtgcag aaggcctacc accagctgga ccagagcttc | 660  |
| atcgacacct acatcgacct gctggaaacc cggcggacct actatgaggg acctggcgag  | 720  |
| ggcagcccct tcggctggaa ggacatcaaa gaatggtacg agatgctgat gggccactgc  | 780  |
| acctacttcc ccgaggaact gcggagcgtg aagtacgcct acaacgccga cctgtacaac  | 840  |
| gccctgaacg acctgaacaa tctcgtgatc accagggacg agaacgagaa gctggaatat  | 900  |
| tacgagaagt tccagatcat cgagaacgtg ttcaagcaga agaagaagcc caccctgaag  | 960  |
| cagatcgcca aagaaatcct cgtgaacgaa gaggatatta agggctacag agtgaccagc  | 1020 |
| accggcaagc ccgagttcac caacctgaag gtgtaccacg acatcaagga cattaccgcc  | 1080 |
| cggaaagaga ttattgagaa cgccgagctg ctggatcaga ttgccaagat cctgaccatc  | 1140 |
| taccagagca gcgaggacat ccaggaagaa ctgaccaatc tgaactccga gctgaccag   | 1200 |
| gaagagatcg agcagatctc taatctgaag ggctataccg gcacccacaa cctgagcctg  | 1260 |
| aaggccatca acctgatcct ggacgagctg tggcacacca acgacaacca gatcgctatc  | 1320 |
| ttcaaccggc tgaagctggg gcccaagaag gtggacctgt cccagcagaa agagatcccc  | 1380 |
| accaccctgg tggacgactt catcctgagc cccgtcgtga agagaagctt catccagagc  | 1440 |
| atcaaagtga tcaacgccat catcaagaag tacggcctgc ccaacgacat cattatcgag  | 1500 |
| ctggcccgcg agaagaactc caaggacgcc cagaaaatga tcaacgagat gcagaagcgg  | 1560 |
| aaccggcaga ccaacgagcg gatcgaggaa atcatccgga ccaccggcaa agagaacgcc  | 1620 |
| aagtacctga tcgagaagat caagctgcac gacatgcagg aaggcaagt cctgtacagc   | 1680 |
| ctggaagcca tccctctgga agatctgctg aacaaccct tcaactatga ggtggaccac   | 1740 |

|            |            |            |            |            |            |      |
|------------|------------|------------|------------|------------|------------|------|
| atcatcccca | gaagcgtgtc | cttcgacaac | agcttcaaca | acaaggtgct | cgtgaagcag | 1800 |
| gaagaaaaca | gcaagaaggg | caaccggacc | ccattccagt | acctgagcag | cagcgacagc | 1860 |
| aagatcagct | acgaaacctt | caagaagcac | atcctgaatc | tggccaaggg | caagggcaga | 1920 |
| atcagcaaga | ccaagaaaga | gtatctgctg | gaagaacggg | acatcaacag | gttctccgtg | 1980 |
| cagaaagact | tcatcaaccg | gaacctggtg | gataccagat | acgccaccag | aggcctgatg | 2040 |
| aacctgctgc | ggagctactt | cagagtgaac | aacctggacg | tgaaagtgaa | gtccatcaat | 2100 |
| ggcggcttca | ccagctttct | gcggcggaag | tggaagttta | agaaagagcg | gaacaagggg | 2160 |
| tacaagcacc | acgccgagga | cgccctgata | attgccaacg | ccgatttcat | cttcaaagag | 2220 |
| tggaagaaac | tggacaaggc | caaaaaagtg | atggaaaacc | agatgttcga | ggaaaagcag | 2280 |
| gccgagagca | tgcccgatag | cgaaaccgag | caggagtaca | aagagatctt | catcaccccc | 2340 |
| caccagatca | agcacattaa | ggacttcaag | gactacaagt | acagccaccg | ggtggacaag | 2400 |
| aagcctaata | gagagctgat | taacgacacc | ctgtactcca | cccggaagga | cgacaagggc | 2460 |
| aacaccctga | tcgtgaacaa | tctgaacggc | ctgtacgaca | aggacaatga | caagctgaaa | 2520 |
| aagctgatca | acaagagccc | cgaaaagctg | ctgatgtacc | accacgaccc | ccagacctac | 2580 |
| cagaaactga | agctgattat | ggaacagtac | ggcgacgaga | agaatcccct | gtacaagtac | 2640 |
| tacgaggaaa | ccgggaacta | cctgaccaag | tactccaaaa | aggacaacgg | ccccgtgata | 2700 |
| aagaagatta | agtattacgg | caacaaactg | aacgcccata | tggacataca | cgacgactac | 2760 |
| ccaacagca  | gaaacaaggt | cgtgaagctg | tccctgaagc | cctacagatt | cgacgtgtac | 2820 |
| ctggacaatg | gcgtgtacaa | gttcgtgacc | gtgaagaatc | tggatgtgat | caaaaaagaa | 2880 |
| aactactacg | aagtgaatag | caagtgctat | gaggaagcta | agaagctgaa | gaagatcagc | 2940 |
| aaccaggccg | agtttatcgc | ctccttctac | aacaacgata | tgatcaagat | caacggcgag | 3000 |
| ctgtatagag | tgatcggcgt | gaacaacgac | ctgctgaacc | ggatcgaagt | gaacatgata | 3060 |
| gacatacct  | accgcgagta | cctggaaaac | atgaacgaca | agaggcccc  | caggatcatt | 3120 |
| aagacaatcg | ccggaagcgg | agctactaac | ttcagcctgc | tgaagcaggc | tggagacgtg | 3180 |
| gaggagaacc | ctggacctag | gcgcgccgcc | accatggtga | gcaagggcga | ggagctgttc | 3240 |
| accgggggtg | tgcccatact | ggtcgagctg | gacggcgacg | taaacggcca | caagttcagc | 3300 |
| gtgtccggcg | agggcgaggg | cgatgccacc | tacggcaagc | tgaccctgaa | gttcatctgc | 3360 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| accaccggca agctgcccgt gccctggccc accctcgtga ccaccttcgg ctacggcctg  | 3420 |
| atgtgcttcg cccgctaccc cgaccacatg aagcagcacg acttcttcaa gtccgccatg  | 3480 |
| cccgaaggct acgtccagga ggcgaccatc ttcttcaagg acgacggcaa ctacaagacc  | 3540 |
| cgcgccgagg tgaagttcga gggcgacacc ctggtgaacc gcatcgagct gaagggcac   | 3600 |
| gacttcaagg aggacggcaa catcctgggg cacaagctgg agtacaacta caacagccac  | 3660 |
| aacgtctata tcatggccga caagcagaag aacggcatca aggtaagtat tagctctttc  | 3720 |
| tttccatggg ttggcctcgc cgcgtgggct gaggggaagga ctgtcctggg actggacagg | 3780 |
| cgggttatgg gacctgaagc gataaaaggc atgcacgttt gcggctacgt gcatgccaaa  | 3840 |
| aggagtcggg cttgcctccg tgcccgactc caaaagacct gctcgaggag gtggacgagc  | 3900 |
| aggtcaaaaa tccgggtacc aataaaatat ctttattttc attacatctg tgtgttggtt  | 3960 |
| ttttgtgtga ctagt                                                   | 3975 |

<210> 23

<211> 3912

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 23

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| aggatTTTTg acctgctcga ttgtccactg cgagcaggtc ttttggagtc gggcgaggcg  | 60  |
| gaagcccgac tccttttggc atgcacgcta gccgcgtcgt gcatgccttt tatcttcggg  | 120 |
| ttatgggacc agtgaaggct gaggggaagga ctgtcctggg actggacagg cgggttatgg | 180 |
| gacctgaaaa tactaacaat cgattttttt tccctttttt tccaggtgaa cttcaagatc  | 240 |
| cgccacaaca tcgaggacgg cagcgtgcag ctccgccgacc actaccagca gaacaccccc | 300 |
| atcggcgacg gccccgtgct gctgcccgaac aaccactacc tgagctacca gtccgccctg | 360 |
| agcaaagacc ccaacgagaa gcgcgatcac atggtcctgc tggagtctgt gaccgccgcc  | 420 |
| gggatcactc tcggcatgga cgagctgtac aaggaccttg gaagcggagc tactaacttc  | 480 |
| agcctgctga agcaggctgg agacgtggag gagaaccctg gacctatcac aaagaagcac  | 540 |
| acagcccact tctccaagaa gggcgaagag gaaaacctgg aaggcctggg caatcagacc  | 600 |
| aagcagatcg tcgagaagta cgcctgcacc accagaatca gcccacacac aagccagcag  | 660 |

|             |            |            |             |            |            |      |
|-------------|------------|------------|-------------|------------|------------|------|
| aacttcgtga  | cccagcggag | caaaagagcc | ctgaagcagt  | ttcggctgcc | cctggaagaa | 720  |
| accgagctgg  | aaaagcggat | catcgtggac | gacaccagca  | cacagtggtc | caagaacatg | 780  |
| aagcacttga  | cccctagcac | actgaccag  | atcgactaca  | acgagaaaga | gaagggcgct | 840  |
| atcacacaga  | gcccactgag | cgactgtctg | accagaagcc  | acagcatccc | tcaggccaac | 900  |
| agatccccctc | tgccaatcgc | caaagtgtct | agcttcccca  | gcatcagacc | catctacctg | 960  |
| accagagtgc  | tgttccagga | caacagcagc | catctgccag  | ccgccagcta | ccggaagaaa | 1020 |
| gatagcggcg  | tgcaagagtc | cagccacttt | ctgcaaggcg  | ctaagaagaa | caatctgagc | 1080 |
| ctggctattc  | tgaccctgga | aatgaccggc | gatcagagag  | aagtcggctc | tctgggcacc | 1140 |
| agcgccacaa  | atagcgtgac | ctacaaaaag | gtggaaaaca  | ccgtgctgcc | taagcctgac | 1200 |
| ctgccaaaga  | caagcggcaa | ggtggaactg | ctgccaaagg  | tgcacatcta | ccagaaggac | 1260 |
| ctgtttccta  | ccgagacaag | caacggctct | cccggccatc  | tggatctggt | ggaaggatct | 1320 |
| ctgctgcagg  | gaaccgaggg | cgccatcaag | tggaacgagg  | ccaatagacc | tggcaaggtg | 1380 |
| cctttcctga  | gagtggccac | agagtctagc | gccaagacac  | cctccaaact | gctggatccc | 1440 |
| ctggcctggg  | ataaccacta | cggcactcag | atccccaag   | aggaatggaa | gtcccaagag | 1500 |
| aagtccccctg | aaaagaccgc | cttcaagaag | aaggacacca  | ttctgtccct | gaatgcctgc | 1560 |
| gagagcaacc  | acgccattgc | cgccatcaat | gagggccaga  | acaagcccga | gatcgaagtg | 1620 |
| acctggggcca | agcagggaag | aaccgagaga | ctgtgctccc  | agaatcctcc | tgtgctgaag | 1680 |
| cggcaccaga  | gagaaatcac | ccggaccaca | ctgcagagcg  | accaagaaga | gatcgattac | 1740 |
| gacgatacca  | tcagcgtcga | gatgaagaaa | gaagatttcg  | acatctacga | cgaggacgag | 1800 |
| aatcagagcc  | ctcggagctt | ccagaagaaa | accaggcact  | actttattgc | cgccgtcgag | 1860 |
| cggctgtggg  | actacggaat | gtctagctct | cctcacgtgc  | tgcggaatag | agcccagtct | 1920 |
| ggtagcgtgc  | cccagttcaa | aaaggctctg | ttccaagagt  | tcaccgacgg | cagcttcacc | 1980 |
| cagccactgt  | atagaggcga | gctgaacgag | catctgggcc  | tgctggggcc | ttatatcaga | 2040 |
| gccgaagtgg  | aagataacat | catggtcacc | ttccggaatc  | aggcctctcg | gccctacagc | 2100 |
| ttctacagct  | ccctgatctc | ctacgaagag | gaccagagac  | agggcgagca | gccccggaag | 2160 |
| aatttcgtga  | agcccaacga | gactaagacc | tactttttgga | aggtgcagca | ccatatggcc | 2220 |
| cctacaaagg  | acgagttcga | ctgcaaagcc | tgggcctact  | tctccgatgt | ggacctcgag | 2280 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| aaggatgtgc acagcggact catcggccca ctgcttgtgt gccacaccaa cacactgaac  | 2340 |
| cccgctcacg gcagacaagt gacagtgcaa gaattcgccc tgtttttcac catcttcgac  | 2400 |
| gaaacgaagt cctggtactt caccgaaaac atggaaagaa actgcagggc cccttgcaac  | 2460 |
| attcagatgg aagatccac cttcaaagag aactaccggt tccacgccat caacggctac   | 2520 |
| atcatggaca cactgcccgg cctggttatg gctcaggatc agagaatccg gtggtatctg  | 2580 |
| ctgtccatgg gtcceaagc gaatatccac tccatccact tctccggcca cgtgttcacc   | 2640 |
| gtgcggaaaa aagaagagta caaatggcc ctgtacaatc tgtaccctgg ggtgttcgaa   | 2700 |
| accgttgaga tgctgcctag caaggccgga atttggagag tggaatgtct gattggagag  | 2760 |
| cacctccacg ccgggatgag caccctgttt ctggtgtact ccaacaagtg tcagaccctt  | 2820 |
| ctcggcatgg cctctggcca cattagagac ttccagatca ccgccagcgg acagtatgga  | 2880 |
| cagtgggccc ctaaactggc cagactgcac tactccggca gcatcaatgc ctggtccacc  | 2940 |
| aaagagcctt tcagctggat caaagtggac ctgctggctc ccatgatcat ccacggaatc  | 3000 |
| aagaccaggg gcgccagaca aaagttcagc agcctgtaca tcagccagtt catcatcatg  | 3060 |
| tacagcctgg acggaaagaa gtggcagacc taccggggca atagcaccgg cacactgatg  | 3120 |
| gtgttcttcg gcaacgtgga ctccagcggc attaagcaca acatcttcaa ccctccaatc  | 3180 |
| attgcccgat acatccggct gcaccccaca cactacagca tcaggtctac cctgagaatg  | 3240 |
| gaactgatgg gctgcgacct gaacagctgc agcatgcccc tcggaatgga aagcaaggcc  | 3300 |
| atcagcgacg cccagatcac agcctctagc tacttcacca acatgttcgc cacttgagac  | 3360 |
| ccctctaagg cccggcttca tctgcaaggc agaagcaacg cttggaggcc ccaagtgaac  | 3420 |
| aaccccaaag aatggctgca ggtcgacttt cagaaaacca tgaaagtgaac aggcgtgacc | 3480 |
| acacagggcg tcaagtccct gctgacctct atgtacgtga aagagtttct gatcagctcc  | 3540 |
| agccaggacg gccaccagtg gaccctgttc ttccaaaacg gcaaagtgaac agtgttccag | 3600 |
| ggaaatcagg acagcttcac acccgtggtc aactccctgg atcctccact gctgacaaga  | 3660 |
| tacctgcgga ttaccctca gtcttgggtg caccagattg ccctgcggat ggaagtgctg   | 3720 |
| ggctgtgaag ctcaggacct ctactgaggt accaattcct cacctgcgat ctgatgctt   | 3780 |
| tatttgtaag atttgatg ctattgcttt atttgtaacc attataagct gcaataaaca    | 3840 |
| agttaacaac aacaattgca ttcattttat gtttcagggt cagggggagg tgtgggaggt  | 3900 |

tttttaaact ag

3912

<210> 24

<211> 3828

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 24

|                                                                    |      |
|--------------------------------------------------------------------|------|
| acttctaggc gcgccgccac catgtacca tacgatgttc cagattacgc ttatccttat   | 60   |
| gacgtgcctg actacgccta tccctacgac gtccccgact atgcagtgta caagaaaacc  | 120  |
| ctgttcgtgg aattcaccga ccacctgttc aatatcgcca agcctcggcc tccttggatg  | 180  |
| ggactgctgg gacctacaat tcaggccgag gtgtacgaca ccgtgggtcat caccctgaag | 240  |
| aacatggcca gccatcctgt gtctctgcac gccgtggggag tgtcttactg gaaggcttct | 300  |
| gagggcgccg agtacgacga tcagacaagc cagagagaga aagaggacga caaggttttc  | 360  |
| cctggcggca gccacaccta tgtctggcaa gtcctgaaag aaaacggccc tatggcctcc  | 420  |
| gacacctgtg gcctgacata cagctacctg agccacgtgg acctgggtcaa ggacctgaat | 480  |
| tctggcctga tcggagccct gctcgtgtgt agagaaggca gcctggccaa agagaaaacc  | 540  |
| cagacactgc acaagtcat cctgctgttc gccgtgttcg acgagggcaa gagctggcac   | 600  |
| agcgagacaa agaacagcct gatgcaggac agggatgccg cctctgctcg ggcttggcct  | 660  |
| aagatgcaca ccgtgaacgg ctacgtgaac agaagcctgc ctggactgat cggctgccac  | 720  |
| agaaagtccg tgtactggca cgtgatcggc atgggcacaa cacctgaggt gcacagcatc  | 780  |
| tttctggaag gacacacctt cctcgtgcgg aaccatagac aggccagcct ggaaatcagc  | 840  |
| cctatcacct tcctgaccgc tcagaccctg ctgatggatc tgggccagtt tctgctgttc  | 900  |
| tgccacatca gtccccacca gcacgatggc atggaagcct acgtgaaggt ggacagctgc  | 960  |
| cccgaagaac cccagctgcg gatgaagaac aacgaggaag ccgaggacta cgacgacgac  | 1020 |
| ctgaccgact ctgagatgga cgtcgtcaga ttcgacgacg ataacagccc cagcttcatc  | 1080 |
| caaatcagaa gcgtggccaa gaagcacccc aagacctggg tgcactatat cgccgccgag  | 1140 |
| gaagaggact gggattacgc tcctctggtg ctggcccctg acgacagaag ctacaagagc  | 1200 |
| cagtacctga acaacggccc tcagcggatc ggccggaagt ataagaaagt gcggttcatg  | 1260 |



|                                                                   |      |
|-------------------------------------------------------------------|------|
| gcctacaccg acgagacatt caagaccaga gaggccatcc agcacgagag cggaattctg | 1320 |
| ggccctctgc tgtatggcga agtgggcgat acactgctga tcatcttcaa gaaccaggcc | 1380 |
| agcagaccct acaacatcta ccctcacggc atcaccgatg tgcggcccct gtattctaga | 1440 |
| aggctgcca agggcgtgaa gcacctgaag gacttccta tcctgcctgg cgagatcttc   | 1500 |
| aagtacaagt ggaccgtgac cgtggaagat ggccccacca agagcgaccc tagatgtctg | 1560 |
| acacggtact acagcagctt cgtgaacatg gaacgcgacc tggccagcgg cctgattgga | 1620 |
| cctctgctga tctgctacaa agaaagcgtg gaccagcggg gcaaccagat catgagcgac | 1680 |
| aagcggaacg tgatcctgtt tagcgtgttc gatgagaacc ggtcctggta tctgaccgag | 1740 |
| aacatccagc ggtttctgcc caatcctgct ggcgtgcagc tggaagatcc tgagttccag | 1800 |
| gcctccaaca tcatgcactc catcaatggc tatgtgttcg acagcctgca gctgagcgtg | 1860 |
| tgctgcacg aagtggccta ctggtacatc ctgagcattg gcgcccagac cgacttcctg  | 1920 |
| tccgtgttct tttccggcta caccttcaag cacaagatgg tgtacgagga taccctgaca | 1980 |
| ctgttcccat tctccggcga gacagtgttc atgagcatgg aaaaccccgg cctgtggatc | 2040 |
| ctgggctgtc acaacagcga cttccggaac agaggcatga cagccctgct gaaggtgtcc | 2100 |
| agctgcgaca agaacaccgg cgactactac gaggacagct atgaggacat cagcgcctac | 2160 |
| ctgctgagca agaacaatgc catcgagccc agaagcttca gccagaatag cagacacccc | 2220 |
| tccaccagac agaagcagtt caacgccaca acaatccccg agaacgacat cgagaaaacc | 2280 |
| gacacctggg ttgcccaccg gaccctatg cctaagatcc agaacgtgtc ctccagcgat  | 2340 |
| ctgctgatgc tcctgagaca gagccctaca cctcacggac tgagcctgtc cgatctgcaa | 2400 |
| gaggccaaat acgaaacctt cagcgacgac ctttctcctg gcgccatcga cagcaacaat | 2460 |
| agcctgagcg agatgaccca cttcagacca cagctgcacc acagcggcga catggtgttt | 2520 |
| acacctgaga gcggcctcca gctgagactg aatgagaagc tgggaaccac cgccgccacc | 2580 |
| gagctgaaga aactggactt caaggtgtcc tctaccagca acaacctgat cagcacaatc | 2640 |
| ccctccgaca acctggctgc cggcaccgac aacacatctt ctctggggcc acctagcatg | 2700 |
| cccgtgcact acgatagcca gctggatacc acactgttcg gcaagaagtc tagccctctg | 2760 |
| acagagtctg gcggccctct gtctctgagc gaggaaaaca acgacagcaa gctgctggaa | 2820 |
| tccggcctga tgaacagcca agagtcctcc tggggcaaga atgtgtccag caccgagtcc | 2880 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| ggcagactgt tcaagggaaa gagagcccac ggacctgctc tgctgaccaa ggataacgcc  | 2940 |
| ctgttcaaag tgtccatcag cctgctcaag accaacaaga cctccaacaa ctccgccacc  | 3000 |
| aacagaaaga cccacatcga cggccctagc ctgctgatcg agaataagccc tagcgtctgg | 3060 |
| cagaatatcc tggaaagcga caccgagttc aagaaagtga cccctctgat ccacgaccgg  | 3120 |
| atgctcatgg acaagaacgc caccgctctg cggctgaacc acatgagcaa caagacaacc  | 3180 |
| agcagcaaga atatggaaat ggtgcagcag aagaaagagg gccccattcc tccagacgct  | 3240 |
| cagaaccccg atatgagctt cttcaagatg ctctttctgc ccgagagcgc ccggtggatc  | 3300 |
| cagagaacac acggcaagaa ctccctgaac tccggccagg gaccttctcc aaagcagctg  | 3360 |
| gtttccctgg gacctgagaa gtccgtggaa ggccagaact tcctgagcga aaagaacaaa  | 3420 |
| gtggtcgtcg gcaagggcga gttcaccaag gatgtgggcc tgaaagagat ggtctttccc  | 3480 |
| agcagccgga acctgttcct gaccaacctg gacaacctgc acgagaacaa caccacaat   | 3540 |
| caagagaaga agatccaaga ggtaagtatt agctctttct ttccatgggt tggcctcgcc  | 3600 |
| gcgtgggctg aggggaaggac tgtcctggga ctggacaggc gggttatggg acctgaagcg | 3660 |
| ataaaaggca tgcacgtttg cggctacgtg catgccaaaa ggagtcgggc ttgcctccgt  | 3720 |
| gcccgactcc aaaagacctg ctcgaggagg tggacgagca ggtcaaaaat ccgggtacca  | 3780 |
| ataaaatatc tttattttca ttacatctgt gtgttggttt tttgtgtg               | 3828 |

<210> 25

<211> 3802

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 25

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| aggatTTTTg acctgctcga ttgtccactg cgagcaggtc ttttggagtc gggcgaggcg  | 60  |
| gaagcccgac tccttttggc atgcacgcta gccgcgtcgt gcatgccttt tatcttcggg  | 120 |
| ttatgggacc agtgaaggct gaggggaagga ctgtcctggg actggacagg cgggttatgg | 180 |
| gacctgaaaa tactaacaat cgattttttt tccctttttt tccaggaaat cgaaaagaaa  | 240 |
| gagacactca tccaagagaa cgtgggtgctg cctcagatcc acacagtgac cggcaccaag | 300 |
| aactttatga agaattctgtt cctgctgagt acccggcaga atgtggaagg cagctacgac | 360 |

|             |            |            |             |            |             |      |
|-------------|------------|------------|-------------|------------|-------------|------|
| ggcgcttatg  | cccctgtgct | gcaagacttc | agatccctga  | acgactccac | caatcggaca  | 420  |
| aagaagcaca  | cagcccactt | ctccaagaag | ggcgaagagg  | aaaacctgga | aggcctgggc  | 480  |
| aatcagacca  | agcagatcgt | cgagaagtac | gcctgcacca  | ccagaatcag | ccccaacaca  | 540  |
| agccagcaga  | acttcgtgac | ccagcggagc | aaaagagccc  | tgaagcagtt | tcggctgccc  | 600  |
| ctggaagaaa  | ccgagctgga | aaagcggatc | atcgtggacg  | acaccagcac | acagtgggtcc | 660  |
| aagaacatga  | agcacttgac | ccctagcaca | ctgacctaga  | tcgactacaa | cgagaaagag  | 720  |
| aagggcgcta  | tcacacagag | cccactgagc | gactgtctga  | ccagaagcca | cagcatccct  | 780  |
| caggccaaca  | gatccccctt | gccaatcgcc | aaagtgtcta  | gcttccccag | catcagacct  | 840  |
| atctacctga  | ccagagtgtt | gttccaggac | aacagcagcc  | atctgccagc | cgccagctac  | 900  |
| cgggaagaaag | atagcggcgt | gcaagagtcc | agccactttc  | tgcaaggcgc | taagaagaac  | 960  |
| aatctgagcc  | tggctattct | gaccctggaa | atgaccggcg  | atcagagaga | agtcggctct  | 1020 |
| ctgggcacca  | gcgccacaaa | tagcgtgacc | tacaaaaagg  | tggaaaacac | cgtgctgcct  | 1080 |
| aagcctgacc  | tgccaaagac | aagcggcaag | gtggaactgc  | tgccaaaggt | gcacatctac  | 1140 |
| cagaaggacc  | tgtttcttac | cgagacaagc | aacggctctc  | ccggccatct | ggatctgggtg | 1200 |
| gaaggatctc  | tgctgcaggg | aaccgagggc | gccatcaagt  | ggaacgaggc | caatagacct  | 1260 |
| ggcaagggtgc | ccttcctgag | agtggccaca | gagtctagcg  | ccaagacacc | ctccaaactg  | 1320 |
| ctggatcccc  | tggcctggga | taaccactac | ggcactcaga  | tccccaaaga | ggaatggaag  | 1380 |
| tccaagaga   | agtcccctga | aaagaccgcc | ttcaagaaga  | aggacaccat | tctgtccctg  | 1440 |
| aatgcctgcg  | agagcaacca | cgccattgcc | gccatcaatg  | agggccagaa | caagcccagag | 1500 |
| atcgaagtga  | cctgggccaa | gcagggaaga | accgagagac  | tgtgctccca | gaatcctcct  | 1560 |
| gtgctgaagc  | ggcaccagag | agaaatcacc | cggaccacac  | tgcaagcga  | ccaagaagag  | 1620 |
| atcgattacg  | acgataccat | cagcgtcgag | atgaagaaag  | aagatttcga | catctacgac  | 1680 |
| gaggacgaga  | atcagagccc | tcggagcttc | cagaagaaaa  | ccaggcacta | ctttattgcc  | 1740 |
| gccgtcgagc  | ggctgtggga | ctacggaatg | tctagctctc  | ctcacgtgct | gcggaataga  | 1800 |
| gcccagtctg  | gtagcgtgcc | ccagttcaaa | aaggctcgtgt | tccaagagtt | caccgacggc  | 1860 |
| agcttcaccc  | agccactgta | tagaggcgag | ctgaacgagc  | atctgggcct | gctgggcccct | 1920 |
| tatatcagag  | ccgaagtgga | agataacatc | atggtcacct  | tccggaatca | ggcctctcgg  | 1980 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| ccctacagct tctacagctc cctgatctcc tacgaagagg accagagaca gggcgcagag  | 2040 |
| ccccggaaga atttcgtgaa gcccaacgag actaagacct acttttggaa ggtgcagcac  | 2100 |
| catatggccc ctacaaagga cgagttcgac tgcaaagcct gggcctactt ctccgatgtg  | 2160 |
| gacctcgaga aggatgtgca cagcggactc atcggcccac tgcttgtgtg ccacaccaac  | 2220 |
| acactgaacc ccgctcacgg cagacaagtg acagtgaag aattcgccct gtttttcacc   | 2280 |
| atcttcgacg aaacgaagtc ctggtacttc accgaaaaca tggaaagaaa ctgcagggcc  | 2340 |
| ccttgcaaca ttcagatgga agatcccacc ttcaaagaga actaccggtt ccacgccatc  | 2400 |
| aacggctaca tcatggacac actgcccggc ctggttatgg ctcaggatca gagaatccgg  | 2460 |
| tggtatctgc tgtccatggg ctccaacgag aatatccact ccatccactt ctccggccac  | 2520 |
| gtgttcaccg tgcggaaaaa agaagagtac aaaatggccc tgtacaatct gtaccctggg  | 2580 |
| gtgttcgaaa ccgttgagat gctgcctagc aaggccggaa tttggagagt ggaatgtctg  | 2640 |
| attggagagc acctccacgc cgggatgagc accctgtttc tgggtgtactc caacaagtgt | 2700 |
| cagaccctc tcggcatggc ctctggccac attagagact tccagatcac cgccagcggg   | 2760 |
| cagtatggac agtgggcccc taaactggcc agactgcact actccggcag catcaatgcc  | 2820 |
| tggtccacca aagagccttt cagctggatc aaagtggacc tgctggctcc catgatcatc  | 2880 |
| cacggaatca agaccaggg cgccagacaa aagttcagca gcctgtacat cagccagttc   | 2940 |
| atcatcatgt acagcctgga cggaaagaag tggcagacct accggggcaa tagcaccggc  | 3000 |
| acactgatgg tgttcttcgg caacgtggac tccagcggca ttaagcaca catcttcaac   | 3060 |
| cctccaatca ttgcccgata catccggctg caccacacac actacagcat caggtctacc  | 3120 |
| ctgagaatgg aactgatggg ctgcgacctg aacagctgca gcatgcccct cggaatggaa  | 3180 |
| agcaaggcca tcagcgacgc ccagatcaca gcctctagct acttcaccaa catgttcgcc  | 3240 |
| acttgagacc cctctaaggc ccggcttcat ctgcaaggca gaagcaacgc ttggaggccc  | 3300 |
| caagtgaaca accccaaaga atggctgcag gtcgactttc agaaaaccat gaaagtgaca  | 3360 |
| ggcgtgacca cacagggcgt caagtccctg ctgacctcta tgtacgtgaa agagtttctg  | 3420 |
| atcagctcca gccaggacgg ccaccagtgg accctgttct tccaaaacgg caaagtgaaa  | 3480 |
| gtgttcagg gaaatcagga cagcttcaca cccgtggtca actccctgga tcctccactg   | 3540 |
| ctgacaagat acctgcggat tcaccctcag tcttgggtgc accagattgc cctgcggatg  | 3600 |

|                                                                   |      |
|-------------------------------------------------------------------|------|
| gaagtgctgg gctgtgaagc tcaggacctc tactgaggta ccaattcctc acctgcgatc | 3660 |
| tcgatgcttt atttgtgaaa tttgtgatgc tattgcttta tttgtaacca ttataagctg | 3720 |
| caataaaca gttaacaaca acaattgcat tcattttatg tttcagggtc aggggggaggt | 3780 |
| gtgggaggtt ttttaaacta gt                                          | 3802 |

<210> 26  
 <211> 10  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|            |    |
|------------|----|
| <400> 26   |    |
| tggggggagg | 10 |

<210> 27  
 <211> 10  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|            |    |
|------------|----|
| <400> 27   |    |
| gtagtgaggg | 10 |

<210> 28  
 <211> 10  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|            |    |
|------------|----|
| <400> 28   |    |
| gttggtggtt | 10 |

<210> 29  
 <211> 10  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 29

agttgtggtt 10

<210> 30  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 30  
gtattgggtc 10

<210> 31  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 31  
agtgtgaggg 10

<210> 32  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 32  
gggtaatggg 10

<210> 33  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 33  
tcattggggt 10

<210> 34  
<211> 10  
<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 34

ggtgggggtc

10

<210> 35

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 35

ggttttgttg

10

<210> 36

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 36

tatactcccg

10

<210> 37

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 37

gtattcgatc

10

<210> 38

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 38

gtagttccct 10

<210> 39  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 39  
gttaatagta 10

<210> 40  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 40  
tgctggttag 10

<210> 41  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 41  
ataggtaacg 10

<210> 42  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 42  
tctgaattgc 10

<210> 43  
<211> 10  
<212> DNA



<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 43

tctgggtttg

10

<210> 44

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 44

cattctcttt

10

<210> 45

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 45

gtattggtgt

10

<210> 46

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 46

tttagatttg

10

<210> 47

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 47

|                                       |    |
|---------------------------------------|----|
| ataagtactg                            | 10 |
|                                       |    |
| <210> 48                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
|                                       |    |
| <400> 48                              |    |
| tagtctatta                            | 10 |
|                                       |    |
| <210> 49                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
|                                       |    |
| <400> 49                              |    |
| aggtattgca                            | 10 |
|                                       |    |
| <210> 50                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
|                                       |    |
| <400> 50                              |    |
| gtagattacg                            | 10 |
|                                       |    |
| <210> 51                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
|                                       |    |
| <400> 51                              |    |
| gggcgggtgc                            | 10 |
|                                       |    |
| <210> 52                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |

|                                       |    |
|---------------------------------------|----|
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 52                              |    |
| cgtttacaat                            | 10 |
|                                       |    |
| <210> 53                              |    |
| <211> 11                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 53                              |    |
| gtacagggat g                          | 11 |
|                                       |    |
| <210> 54                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 54                              |    |
| aatcagggga                            | 10 |
|                                       |    |
| <210> 55                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 55                              |    |
| ggaggttttg                            | 10 |
|                                       |    |
| <210> 56                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 56                              |    |

gtattccctg 10

<210> 57  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 57  
tggttaagatc 10

<210> 58  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 58  
gtagttaagt 10

<210> 59  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 59  
gttggtttgg 10

<210> 60  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 60  
gtatttactt 10

<210> 61  
<211> 10  
<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 61

gtaacggggt

10

<210> 62

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 62

tttttttctg

10

<210> 63

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 63

ggggaaggga

10

<210> 64

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 64

ttaccccgt

10

<210> 65

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 65

gtattctatg 10

<210> 66  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 66  
aggtattgtg 10

<210> 67  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 67  
tttggggggg 10

<210> 68  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 68  
gttgtagcg 10

<210> 69  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 69  
ggtagttggg 10

<210> 70  
<211> 10  
<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 70

ctaagtactg

10

<210> 71

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 71

aaccatcttc

10

<210> 72

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 72

gtacctgggt

10

<210> 73

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 73

gtatctcatt

10

<210> 74

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 74

|                                       |    |
|---------------------------------------|----|
| aaataaaaatt                           | 10 |
|                                       |    |
| <210> 75                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
|                                       |    |
| <400> 75                              |    |
| ggtgggttat                            | 10 |
|                                       |    |
| <210> 76                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
|                                       |    |
| <400> 76                              |    |
| taaggaggagg                           | 10 |
|                                       |    |
| <210> 77                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
|                                       |    |
| <400> 77                              |    |
| tatgggaggagg                          | 10 |
|                                       |    |
| <210> 78                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
|                                       |    |
| <400> 78                              |    |
| gatgggaggagg                          | 10 |
|                                       |    |
| <210> 79                              |    |
| <211> 10                              |    |
| <212> DNA                             |    |



<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 79

tggggggggt

10

<210> 80

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 80

ggggaagggg

10

<210> 81

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 81

tggttaagagg

10

<210> 82

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 82

gggttagggt

10

<210> 83

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 83

gtatcggggg 10

<210> 84  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 84  
ggttttgctg 10

<210> 85  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 85  
tgggggtgga 10

<210> 86  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 86  
acttttagag 10

<210> 87  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 87  
gtaacgggtt 10

<210> 88  
<211> 10  
<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 88

gtttggggga

10

<210> 89

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 89

atttttagag

10

<210> 90

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 90

ttaaagtagg

10

<210> 91

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 91

gtattaatat

10

<210> 92

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 92

ggttttgggtg 10

<210> 93  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 93  
tatgggaaag 10

<210> 94  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 94  
ggttgggagg 10

<210> 95  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 95  
gtatttagtg 10

<210> 96  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 96  
gagttaaag 10

<210> 97  
<211> 10  
<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 97

ttgtaagttg

10

<210> 98

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 98

tgggggtagg

10

<210> 99

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 99

gttcttaggg

10

<210> 100

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 100

gtattctaag

10

<210> 101

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 101

ggaggttttg 10

<210> 102  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 102  
agaatatgta 10

<210> 103  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 103  
atctttcggg 10

<210> 104  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 104  
ttgcattgaa 10

<210> 105  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 105  
ggtgggattt 10

<210> 106  
<211> 10  
<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 106

tttatcta

10

<210> 107

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 107

gcgggtggtg

10

<210> 108

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 108

ggttagata

10

<210> 109

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 109

tttatgcgtt

10

<210> 110

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 110

tgggtaaggc 10

<210> 111  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 111  
gggggtgggc 10

<210> 112  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 112  
gtagtatatt 10

<210> 113  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 113  
ggaggtattt 10

<210> 114  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 114  
gtattgtaag 10

<210> 115  
<211> 10  
<212> DNA



<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 115

tttacgggag

10

<210> 116

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 116

tagttctggg

10

<210> 117

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 117

ccacgtctat

10

<210> 118

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 118

agtgggtagg

10

<210> 119

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 119

|                                       |    |
|---------------------------------------|----|
| caatttttac                            | 10 |
|                                       |    |
| <210> 120                             |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 120                             |    |
| ggtctggggg                            | 10 |
|                                       |    |
| <210> 121                             |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 121                             |    |
| atcaagattg                            | 10 |
|                                       |    |
| <210> 122                             |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 122                             |    |
| gttagctaaa                            | 10 |
|                                       |    |
| <210> 123                             |    |
| <211> 10                              |    |
| <212> DNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 123                             |    |
| agtgtggggt                            | 10 |
|                                       |    |
| <210> 124                             |    |
| <211> 10                              |    |
| <212> DNA                             |    |

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 124

ggtatgtggg

10

<210> 125

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 125

gtagtgtggg

10

<210> 126

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 126

aggaggtgtt

10

<210> 127

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 127

gttggtaggt

10

<210> 128

<211> 10

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 128

gtaggtggtt 10

<210> 129  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 129  
aggtgttggt 10

<210> 130  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 130  
tatggttggtg 10

<210> 131  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 131  
ttaggttagt 10

<210> 132  
<211> 10  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 132  
gattggagtt 10

<210> 133  
<211> 10  
<212> DNA

|                                       |    |
|---------------------------------------|----|
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 133                             |    |
| gtagagtgga                            | 10 |
| <210> 134                             |    |
| <211> 24                              |    |
| <212> RNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 134                             |    |
| cucuuucuuu uccauggguu ggcu            | 24 |
| <210> 135                             |    |
| <211> 24                              |    |
| <212> RNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 135                             |    |
| ggcugaggga aggacugucc ugga            | 24 |
| <210> 136                             |    |
| <211> 13                              |    |
| <212> RNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 136                             |    |
| ggguuauagg acc                        | 13 |
| <210> 137                             |    |
| <211> 12                              |    |
| <212> RNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 137                             |    |

|                                       |    |
|---------------------------------------|----|
| aauuccuuuu ua                         | 12 |
|                                       |    |
| <210> 138                             |    |
| <211> 12                              |    |
| <212> RNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 138                             |    |
| guauccuuuu ua                         | 12 |
|                                       |    |
| <210> 139                             |    |
| <211> 33                              |    |
| <212> RNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 139                             |    |
| aggcuucgga gcaaggaggc agcuccgaag ccu  | 33 |
|                                       |    |
| <210> 140                             |    |
| <211> 33                              |    |
| <212> RNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 140                             |    |
| aggcuucgga gcaagccucc agcuccgaag ccu  | 33 |
|                                       |    |
| <210> 141                             |    |
| <211> 29                              |    |
| <212> RNA                             |    |
| <213> Artificial Sequence             |    |
| <220>                                 |    |
| <223> Synthetic nucleic acid sequence |    |
| <400> 141                             |    |
| gucgaggccg agcgggcaaa ggccucgac       | 29 |
|                                       |    |
| <210> 142                             |    |
| <211> 29                              |    |
| <212> RNA                             |    |

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 142

gucgaggccg agcccgcaaa ggccucgac

29

<210> 143

<211> 10

<212> RNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<220>

<221> misc\_feature

<222> (1)..(3)

<223> n is a, c, g, or u

<220>

<221> misc\_feature

<222> (8)..(10)

<223> n is a, c, g, or u

<400> 143

nnnaggunnn

10

<210> 144

<211> 12

<212> RNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 144

uuuuccuuaa cu

12

<210> 145

<211> 1305

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 145

cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt

60

|                                                                     |      |
|---------------------------------------------------------------------|------|
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca   | 120  |
| atgggtggag tatttacggt aaactgcca cttggcagta catcaagtgt atcatatgcc    | 180  |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta   | 240  |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac   | 300  |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg   | 360  |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg   | 420  |
| ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt    | 480  |
| acggtgggag gtctatataa gcagagcttg gatgttgctt ttacttctag gcgcgccgcc   | 540  |
| accatggtga gcaagggcga ggagctgttc accgggggtg tgcccatcct ggtcgagctg   | 600  |
| gacggcgacg taaacggcca caagttcagc gtgtccggcg agggcgaggg cgatgccacc   | 660  |
| tacggcaagc tgaccctgaa gttcatctgc accaccggca agctgcccgt gccctggccc   | 720  |
| accctcgtga ccaccttcgg ctacggcctg atgtgcttcg cccgctaccc cgaccacatg   | 780  |
| aagcagcacg acttcttcaa gtccgccatg cccgaaggct acgtccagga gcgcaccatc   | 840  |
| ttcttcaagg taagtattag ctctttcttt ccatgggttg gcctcgccgc gtgggctgag   | 900  |
| ggaaggactg tcctgggact ggacaggcgg gttatgggac ctgaaaagcg gccctgaaaa   | 960  |
| agggccgcga tctgtagaaa gcgagctagt gccggacagt tagaggaaaa ggggaagaac   | 1020 |
| tgtccgaaaa aaggggggga agacagtgac tagaaaggga agggagaagt cactgtagag   | 1080 |
| gggaaggaaa aggctagcta gaggagaagg aaagaggcta gctagcagag gagaaggaaa   | 1140 |
| ggcgccagca gttcgggtgct atcaaaaagc ggtcaggcag ctaaaccaaa aggttttagca | 1200 |
| attgcctctg atgagtcgct gaaatgcgac gaaaaccgct ttttggtagc aataaaatat   | 1260 |
| ctttattttc attacatctg tgtgttggtt ttttgtgtga ctagt                   | 1305 |

<210> 146  
 <211> 1543  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                  |    |
|------------------------------------------------------------------|----|
| <400> 146                                                        |    |
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgccatt | 60 |



|                                                                    |      |
|--------------------------------------------------------------------|------|
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120  |
| atgggtggag tatttacggt aaactgcca cttggcagta catcaagtgt atcatatgcc   | 180  |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240  |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300  |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360  |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420  |
| ggactttcca aaatgtcgta acaactccgc cccattgacg caaatgggcg gtaggcgtgt  | 480  |
| acggtgggag gtctatataa gcagagcttg gatgttgcct ttacttctag gcgcgccgcg  | 540  |
| gaaaaccgcg ggatagcacc gaactgctgg cgcctttcct tctcctctgc tagctagcct  | 600  |
| ctttccttct cctctagcta gccttttcct tcccctctac agtgacttct cccttcctt   | 660  |
| tctagtcact gtcttcccc ccttttttcg gacagttcct ccccttttcc tctaactgtc   | 720  |
| cggcactagc tcgctttcta cagatcatta ttgcggccct gaaaaagggc cgcttataac  | 780  |
| gttgctcgaa ttcgggttat gggaccagtg aaggctgagg gaaggactgt cctgggactg  | 840  |
| gacaggcggg ttatgggacc tgaaaatact aacaatcgat tttttttccc tttttttcca  | 900  |
| ggacgacggc aactacaaga cccgcgccga ggtgaagttc gagggcgaca ccctggtgaa  | 960  |
| ccgcatcgag ctgaagggca tcgacttcaa ggaggacggc aacatcctgg ggcacaagct  | 1020 |
| ggagtacaac tacaacagcc acaacgtcta tatcatggcc gacaagcaga agaacggcat  | 1080 |
| caaggtaagt attagctctt tctttccatg gggttggcctc gccgcgtggg ctgagggaag | 1140 |
| gactgtcctg ggactggaca ggcgggttat gggacctgaa aagcggccct gaaaaagggc  | 1200 |
| cgcagcgaac acgaagcgag ctaaagcctc ctctctcttc ttcagaactc ctctcttttc  | 1260 |
| tctcctccag gagttcttcc tctctccctt cttctcaaat gctttctccc tctctcctgc  | 1320 |
| atttgagctc cttctttcct ctctcgacaa tccccttttc tccctcttga ttgtcgacta  | 1380 |
| gctcgcaatc atcgcggtgc taaaaagcgg tcaggcagct aaacaaaaag gtttagcaat  | 1440 |
| tgcctctgat gagtcgctga aatgcgacga aaaccgcttt ttggtaccaa taaaatatct  | 1500 |
| ttattttcat tacatctgtg tgttggtttt ttgtgtgact agt                    | 1543 |

<210> 147  
<211> 1571

<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 147  
cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt 60  
gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca 120  
atgggtggag tatttacggt aaactgccca cttggcagta catcaagtgt atcatatgcc 180  
aagtacgcc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta 240  
catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac 300  
catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg 360  
atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg 420  
ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt 480  
acggtgggag gtctatataa gcagagcttg gatgttgctt ttacttctag gcgcgccacc 540  
atgagccagt tcgacatcct gtgcaagacc cccccaagg tgctgggtgcg gcagttcgtg 600  
gagagattcg agaggccag cggcgagaag atcgccagct gtgccgccga gctgacctac 660  
ctgtgctgga tgatcaccca caacggcacc gccatcaaga gggccacctt catgagctac 720  
aacaccatca tcagcaacag cctgagcttc gacatcgtga acaagagcct gcagttcaag 780  
tacaagacc agaaggccac catcctggag gccagcctga agaagctgat cccgcctgg 840  
gagttacca tcatccctta caacggccag aagcaccaga gcgacatcac cgacatcgtg 900  
tccagcctgc agctgcagtt cgagagcagc gaggaggccg acaagggcaa cagccacagc 960  
aagaagatgc tgaaggccct gctgtccgag ggcgagagca tctgggagat caccgagaag 1020  
atcctgaaca gcttcgagta caccagcagg ttcaccaaga ccaagaccct gtaccagttc 1080  
ctgttcctgg ccacattcat caactgcggc aggttaagtat tagctctttc tttccatggg 1140  
ttggcctcgc cgcgtgggct gagggaagga ctgtcctggg actggacagg cgggttatgg 1200  
gacctgaaaa gcggccctga aaaagggccg cgatgaaaac gaagcgagct aaagcctcct 1260  
ctctcttctt cagaactcct ctcttttctc tcctccagga gttcttcctc tctcccttct 1320  
tctcaaatgc tttctccctc tctcctgcat ttgagctcct tctttcctct ctcgacaatc 1380  
cccttttctc cctcttgatt gtcgactagc tcgcaatcat cgcggtatca aaaagcggtc 1440

|                                                                   |      |
|-------------------------------------------------------------------|------|
| aggcagctaa accaaaaggt ttagcaattg cctctgatga gtcgctgaaa tgcgacgaaa | 1500 |
| accgcttttt ggtaccaata aaatatcttt attttcatta catctgtgtg ttggtttttt | 1560 |
| gtgtgactag t                                                      | 1571 |

<210> 148

<211> 1765

<212> DNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 148

|                                                                    |      |
|--------------------------------------------------------------------|------|
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60   |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120  |
| atgggtggag tattttacggt aaactgccc cttggcagta catcaagtgt atcatatgcc  | 180  |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240  |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300  |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360  |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420  |
| ggactttcca aaatgtcgta acaactccgc cccattgacg caaatgggcg gtaggcgtgt  | 480  |
| acggtgggag gtctatataa gcagagcttg gatgttgcct ttacttctag gcgcgccgcg  | 540  |
| gaaaaccgcg ggataccgcg atgattgcga gctagtcgac aatcaagagg gagaaaaggg  | 600  |
| gattgtcgag agaggaaaga aggagctcaa atgcaggaga gagggagaaa gcatttgaga  | 660  |
| agaagggaga gaggaagaac tcctggagga gagaaaagag aggagttctg aagaagagag  | 720  |
| aggaggcttt agctcgcttc gttttcatca ttattgcggc cctgaaaaag ggccgcttat  | 780  |
| aacgttgctc gaattcgggt tatgggacca gtgaaggctg agggaaggac tgtcctggga  | 840  |
| ctggacaggc gggttatggg acctgaaaat actaacaatc gatttttttt cctttttttt  | 900  |
| ccaggttcag cgacatcaag aacgtggacc ccaagagctt caagctggtg cagaacaagt  | 960  |
| acctgggctg gatcattcag tgcctggtga ccgagaccaa gacaagcgtg tccaggcaca  | 1020 |
| tctacttttt cagcgccaga ggcaggatcg accccctggt gtacctggac gagttcctga  | 1080 |
| ggaacagcga gcccgctgctg aagagagtga acaggaccgg caacagcagc agcaacaagc | 1140 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| aggagtacca gctgctgaag gacaacctgg tgcgcagcta caacaaggcc ctgaagaaga  | 1200 |
| acgcccccta ccccatcttc gctatcaaga acggccctaa gagccacatc ggccaggcacc | 1260 |
| tgatgaccag ctttctgagc atgaagggcc tgaccgagct gacaaacgtg gtgggcaact  | 1320 |
| ggagcgacaa gagggcctcc gccgtggcca ggaccaccta caccaccag atcaccgcca   | 1380 |
| tccccgacca ctacttcgcc ctggtgtcca ggtactacgc ctacgacccc atcagcaagg  | 1440 |
| agatgatcgc cctgaaggac gagaccaacc ccatcgagga gtggcagcac atcgagcagc  | 1500 |
| tgaagggcag cgccgagggc agcatcagat accccgcctg gaacggcatc atcagccagg  | 1560 |
| aggtgctgga ctacctgagc agctacatca acaggcggat ctgagaattc ctcacctgcg  | 1620 |
| atctcgatgc tttatttgtg aaatttgtga tgctattgct ttatttgtaa ccattataag  | 1680 |
| ctgcaataaa caagttaaca acaacaattg cattcatttt atgtttcagg ttcaggggga  | 1740 |
| ggtgtgggag gttttttaaa ctagt                                        | 1765 |

<210> 149  
 <211> 10  
 <212> RNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|            |    |
|------------|----|
| <400> 149  |    |
| aaagaaggaa | 10 |

<210> 150  
 <211> 12  
 <212> RNA  
 <213> Artificial Sequence

<220>  
 <223> Synthetic nucleic acid sequence

|               |    |
|---------------|----|
| <400> 150     |    |
| cuuucuuuuc uu | 12 |

<210> 151  
 <211> 11  
 <212> RNA  
 <213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<220>

<221> misc\_feature

<222> (1)..(3)

<223> n is a, c, g, or u

<220>

<221> misc\_feature

<222> (8)..(11)

<223> n is a, c, g, or u

<400> 151

nnnaggunnn n

11

<210> 152

<211> 11

<212> RNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<220>

<221> misc\_feature

<222> (1)..(3)

<223> n is a, c, g, or u

<220>

<221> misc\_feature

<222> (8)..(11)

<223> n is a, c, g, or u

<400> 152

nnnuggunnn n

11

<210> 153

<211> 11

<212> RNA

<213> Artificial Sequence

<220>

<223> Synthetic nucleic acid sequence

<220>

<221> misc\_feature

<222> (3)..(8)

<223> n is a, c, g, or u

<400> 153  
gannnnnnnaa a 11

<210> 154  
<211> 12  
<212> DNA  
<213> Artificial Sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 154  
gccgccacca tg 12

<210> 155  
<211> 4311  
<212> DNA  
<213> Artificial sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 155  
cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt 60  
gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca 120  
atgggtggag tatttacggt aaactgccc cttggcagta catcaagtgt atcatatgcc 180  
aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta 240  
catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac 300  
catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg 360  
atttccaagt ctccaccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg 420  
ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt 480  
acggtgggag gtctatataa gcagagcttg gatgttgccct ttacttctag gcgcgccgcc 540  
gccaccatgg ctctgatcgt gcacctgaaa accgtgtccg agctgagagg caagggcgac 600  
agaatcgcca aagtacctt cagaggccag agcttctaca gcagagtgtt ggaaaactgc 660  
gaaggcgtgg ccgacttcga cgagacattc agatggcctg tggccagcag catcgacaga 720  
aacgaggtgc tcgagatcca gatcttcaac tacagcaagg tgttcagcaa caagctgatc 780  
gggaccttct gcatggtgct gcagaaagtg gtggaagaga accgcgtgga agtgaccgac 840  
acactgatgg acgacagcaa cgccatcatc aagaccagcc tgagcatgga agtgcgctac 900

|                                                                    |      |
|--------------------------------------------------------------------|------|
| caggccacag atggcacagt cggaccttgg gacgatggcg atttcctggg agatgagagc  | 960  |
| ctgcaagagg aaaaggacag ccaagagaca gacggcctgc tgcctggctc tcggcctagc  | 1020 |
| acaagaatca gcggcgagaa gtccttcaga agcaagggca gagaaaagac caaaggcggc  | 1080 |
| agagatggcg agcacaaggc tggcagatct gtgttcagcg ccatgaagct gggcaagacc  | 1140 |
| agaagccaca aagaggaacc ccagagacag gacgagccag ccgttctgga aatggaagat  | 1200 |
| ctcgaccatc tggccatcca gctcggcgac ggacttgacc ctgattctgt gtctctggcc  | 1260 |
| agcgtgacag ccctgacaag caacgtgtcc aacaagagaa gcaagcccga catcaagatg  | 1320 |
| gaacccagcg ccggcagacc catggattac cagggtgtcca tcaccgtgat cgaggccaga | 1380 |
| cagctcgtgg gcctgaacat ggatcctgtc gtgtgtgtgg aagtgggcga cgacaaaaag  | 1440 |
| tacaccagca tgaaggaaag caccaactgt ccctactaca acgagtactt cgtgttcgac  | 1500 |
| ttccacgtgt cccagacgt gatgttcgac aagatcatta agatcagcgt gatccacagc   | 1560 |
| aagaacctgc tgagaagcgg cacactcgtg ggcagcttta agatggacgt gggcaccgtg  | 1620 |
| tacagccagc cagagcacca gtttcaccac aagtgggcca tcctgagcga ccccgatgat  | 1680 |
| atctctgctg gcctgaaggg ctacgtgaag tgtgatgtgg ctgtcgtcgg caaaggcgac  | 1740 |
| aacatcaaga caccacaa ggccaacgag actgacgagg acgatatcga gggcaacctg    | 1800 |
| ctgctgccag aaggcgtgcc accagaaaga cagtgggcca gattctatgt gaagatctac  | 1860 |
| agagccgagg gcctgcctag aatgaacaca agcctgatgg ccaacgtgaa gaaggctttc  | 1920 |
| atcggcgaga acaaggacct ggtggacccc tacgtccagg tgttcttcgc tggacagaaa  | 1980 |
| ggcaagacct ccgtgcagaa gtccagctac gagcccctgt ggaacgaaca ggtggtgttc  | 2040 |
| accgatctgt tccctccact gtgcaagaga atgaaggtgc agatccggga cagcgacaaa  | 2100 |
| gtgaacgatg tggccatcgg caccacttc atcgacctga gaaagatcag caacgacggc   | 2160 |
| gacaagggct tcctgcctac acttggaact gcctgggtca acatgtacgg cagcaccaga  | 2220 |
| aactacaccc tgctggacga gcaccaggac ctgaacgaag gactcggaga gggcgtgtcc  | 2280 |
| ttccgggcta gactgatgct gggactcgcc gtggaaatcc tggacacaag caaccctgag  | 2340 |
| ctgaccagca gcacagaggt gcaggttgaa caggccacac ctgtgtctga gagctgcacc  | 2400 |
| ggcagaatgg aagagttctt cctgttcggc gccttcctgg aagcctccat gatcgataga  | 2460 |
| aagaacggcg ataagcccat caccttcgaa gtgaccatcg gcaactacgg caacgaggtg  | 2520 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| gacggcatgt ctagaccctt ccggcctaga ccaagaaaag agcccggcga cgaggaagag  | 2580 |
| gtggacctga tccagaacag cagcgacgat gagggcgacg aagctggcga tctggcaagc  | 2640 |
| gtagcagca cccctcctat gagggcccag atcaccgacc ggaactactt tcatctgccc   | 2700 |
| tacctgaaa gaaagccctg catctacatc aagagctggg ggccctgacca gagaaggcgg  | 2760 |
| ctgtacaacg ctaacatcat ggaccatata gccgacaagc tggaagaggg actgaacgac  | 2820 |
| gtccaagaga tgatcaagac cgagaagtct taccctcgaga gaaggctgag gggcgtgctc | 2880 |
| gaggaactga gctgtggatg ccacagattt ctgagcctgt ccgacaagga ccagggcaga  | 2940 |
| agcagcagaa ccagactgga tagagagcgg ctgaagtcct gcatgcgcga gctggaatct  | 3000 |
| atggggccagc aggccaagag cctgagagcc caagtgaaga gacacaccgt gcgggacaag | 3060 |
| ctgagatcct gccagaactt cctgcagaag ctgcggttcc tggccgatga gcctcagcac  | 3120 |
| tctatccccg acgtgttcat ctggatgatg agcaacaaca agaggatcgc ctacgccaga  | 3180 |
| gtgcccagca aggatctgct gtttagcatc gtggaagagg aactcggcaa ggactgcgcc  | 3240 |
| aaagtcaaga ccctgttcct gaagctgcca ggcaagagag gcttcggctc tgctggatgg  | 3300 |
| acagtgcagg ctaagctgga actgtacctg tggctgggcc tgagcaagca gagaaaggac  | 3360 |
| ttcctgtgcg gcctgccttg cggcttcgaa gaagtgaagg ctgctcaagg cctgggcctg  | 3420 |
| cacagcttcc ctccaatctc tctggtgtac acaaagaagc aggccttcca gctgagggcc  | 3480 |
| cacatgtacc aggctagatc tctgttcgcc gccgactcta gcggcctgtc tgatcctttc  | 3540 |
| gctcgggtgt tcttcatcaa ccagagccag tgcaccgagg tgctgaacga gacactgtgt  | 3600 |
| cctacctggg accagatgct ggtctttgac aacctcgagc tgtacggcga ggctcacgaa  | 3660 |
| ctgagagatg accctcctat catcgtcatc gagatctacg accaggacag catgggcaaa  | 3720 |
| gccgacttca tgggcagAAC cttcgccaag cctctggtca agatggccga cgaggcttac  | 3780 |
| tgccctctc gggtcccacc tcagctcgag tactaccaga tctaccgggg ctctgctaca   | 3840 |
| gccggcgatc tgctggctgc ttttgagctg ctgcaaatcg gccctagcgg caaggctgat  | 3900 |
| ctgcctcaa tcaacggccc tgtggacatg gacagaggcc ccattatgcc tgtgcctgtg   | 3960 |
| ggcatcagac ccgtgctgag caagtacaga gtggaagtgc tgttttgggg cctgcgcgac  | 4020 |
| ctgaagagag tgaacctggc tcaggtaagt attagctctt tctttccatg ggttggcctc  | 4080 |
| gccgcgtggg ctgagggaag gactgtcctg ggactggaca ggcgggttat gggacctgaa  | 4140 |



|                                                                     |      |
|---------------------------------------------------------------------|------|
| gcgataaaag gcatgcacgt ttgcggctac gtgcatgcca aaaggagtcg ggcttgccctc  | 4200 |
| cgtgccccgac tccaaaagac ctgctcgagg aggtggacga gcagggtcaaa aatccgggta | 4260 |
| ccaataaaaat atctttatatt tcattacatc tgtgtgtttgg ttttttgtgt g         | 4311 |

<210> 156  
 <211> 3467  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |      |
|--------------------------------------------------------------------|------|
| <400> 156                                                          |      |
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60   |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120  |
| atgggtggag tattttacggt aaactgcccc cttggcagta catcaagtgt atcatatgcc | 180  |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240  |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300  |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360  |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420  |
| ggactttcca aaatgtcgta acaactccgc cccattgacg caaatgggcg gtaggcgtgt  | 480  |
| acgggtgggag gtctatataa gcagagcttg gatgttgctt ttaggatttt tgacctgctc | 540  |
| gattgtccac tgcgagcagg tcttttggag tcgggcgagg cggaagcccc actccttttg  | 600  |
| gcatgcacgc tagccgcgtc gtgcatgcct tttatcttcg ggttatggga ccagtgaagg  | 660  |
| ctgaggggaag gactgtcctg ggactggaca ggcgggttat gggacctgaa aatactaaca | 720  |
| atcgatTTTT tttccctttt tttccaggtg gacagaccca gagtggatat cgagtgtgct  | 780  |
| ggcaaagggg tgcagagcag cctgatccat aactacaaga agaaccccaa cttcaacacc  | 840  |
| ctggtcaagt ggttcgaagt ggatctgccc gagaacgaac tgctgcaccc acctctgaac  | 900  |
| atcagagtgg tggactgcag agccttcggc agatacacc tcgtgggatc tcacgccgtg   | 960  |
| tctagcctga gaagattcat ctacagacct ccagacagaa gcgcccctaa ctggaacaca  | 1020 |
| acaggcgagg tgggtggtgtc catggaaccc gaggaacccg tgaagaaact ggaaaccatg | 1080 |
| gtcaagctgg acgccacctc cgatgctgtc gtgaaagtgg acgtggccga ggacgagaaa  | 1140 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| gagcgcaaga agaagaaaaa gaagggcccc agcgaggaac ctgaagagga agaacctgac  | 1200 |
| gagagcatgc tggactggtg gtccaagtac ttgcctcca tcgacacaat gaaggaacag   | 1260 |
| ctgagacagc acgagacaag cggcaccgac ctcgaagaga aagaagagat ggaatccgcc  | 1320 |
| gaaggactga agggccctat gaagtccaaa gagaagtcta gggccgcaa agaagagaaa   | 1380 |
| aaaaagaaga accagtctcc tggaccaggc cagggatctg aggctcccga aaagaaaaag  | 1440 |
| gccaagatcg acgagctgaa ggtgtacccc aaagagctgg aaagcgagtt cgacagcttc  | 1500 |
| gaggactggc tgcacacctt caatctgctg agaggaaaga caggcgacga cgaggatggc  | 1560 |
| agcactgaag aagagagaat cgtcggcaga ttcaagggca gcctgtgcgt gtacaagggtg | 1620 |
| ccactgcctg aggacgtgtc cagagaggct ggctacgac ctacctacgg catgttccaa   | 1680 |
| ggcatcccta gcaacgaccc catcaatgtg ctctgtcgga tctatgtcgt gcggggccact | 1740 |
| gatctgcatc ccgccgatat caacggcaag gcagaccctt atatcgctat caagctgggg  | 1800 |
| aaaaccgaca tcagggacaa agagaactac atcagcaagc agctgaaccc cgtgttcggc  | 1860 |
| aagagcttcg acatcgaggc tagcttcccc atggaatcca tgctgaccgt ggccgtgtac  | 1920 |
| gactgggatc tcgtgggaac agacgacctg atcggagaga caaagattga cctggaaaac  | 1980 |
| cggttctact ccaagcaccg ggccacctgt ggaatcgccc agacctactc tatccacggc  | 2040 |
| tacaacatct ggcgggaccc catgaagcct agccagatcc tgaccaggct gtgcaaagaa  | 2100 |
| ggcaaggctg acggccctca ctttggacct cacggccggg tcagagtggc caacagagtg  | 2160 |
| ttcacaggcc cctccgagat cgaggatgag aacggccaga gaaagcccac cgatgagcat  | 2220 |
| gtggctctga gcgctctgag aactgggaa gatatcccta gagtgggctg cagactggtg   | 2280 |
| cccgagcacg tggaaacaag acccctgctg aaccagaca agcccggaat cgaacagggc   | 2340 |
| agactcgaac tgtgggtcga catgttcctt atggacatgc ccgcacctgg cacaccactg  | 2400 |
| gacatcagcc ctaggaagcc caagaaatac gagctgcgcg tgatcgtgtg gaacaccgac  | 2460 |
| gaagtgggtg tggaagatga cgacttcttc accggcgaaa agtccagcga catcttcgtc  | 2520 |
| agaggatggc tgaagggaca gcaagaggat aagcaggaca ccgacgtgca ctaccacagc  | 2580 |
| cttacaggcg aaggcaactt taactggcgc tacctgtttc ctttcgacta cctggccgcc  | 2640 |
| gaagagaaga tcgtgatgtc caagaaagaa tctatgttca gctgggacga gacagagtac  | 2700 |
| aagatccccg ccagactgac cctgcagatc tgggatgccg atcacttcag cgccgacgac  | 2760 |

|                                                                   |      |
|-------------------------------------------------------------------|------|
| tttctgggag ccatcgagct ggacctgaat agattcccca gaggcgcca gaccgccaag  | 2820 |
| cagtgcacaa tggaaatggc cactggcgag gtcgacgtgc cactggtgtc tatcttcaag | 2880 |
| cagaagcgcg tcaaaggctg gtggcccctg ctggctagaa acgagaacga cgagttcgag | 2940 |
| ctgaccggaa aggtggaagc cgagctgcat ctgctgacag ctgaagaggc cgagaagaat | 3000 |
| cctgtgggcc tcgctaggaa tgagcccgat cctctggaaa agcccaacag acccgatacc | 3060 |
| gccttcgtgt ggtttctgaa cccactgaag tccatcaagt acctgatctg taccgggtac | 3120 |
| aagtggctga ttatcaagat cgtgctggcc ctgctggggc tgctgatgct tgctctgttc | 3180 |
| ctgtactccc tgcctggcta tatggtcaag aagctgctgg gcgccggcgc tcgggctgac | 3240 |
| tacaaagacc atgacggtga ttataaagat catgacatcg actataagga tgacgatgac | 3300 |
| aaatgaggta ccaattcctc acctgcgac tcgatgcttt atttgtgaaa tttgtgatgc  | 3360 |
| tattgcttta tttgtaacca ttataagctg caataaaca gttaacaaca acaattgcat  | 3420 |
| tcattttatg tttcaggttc agggggaggt gtgggaggtt ttttaaa               | 3467 |

<210> 157

<211> 4392

<212> DNA

<213> Artificial sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 157

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60  |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120 |
| atgggtggag tattttacggt aaactgccca cttggcagta catcaagtgt atcatatgcc | 180 |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240 |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300 |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360 |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420 |
| ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt   | 480 |
| acggtgggag gtctatataa gcagagcttg gatgttgctt ttacttctag gcgcgccgcc  | 540 |
| accatggtca ttctgcagca gggcgaccac gtgtggatgg atctgagact gggccaagag  | 600 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| ttcgacgtgc caatcggcgc cgtggtcaag ctgtgtgatt ctggccaggt gcaagtcgtg  | 660  |
| gacgacgagg ataatgagca ctggatcagc cctcagaacg ccacacacat caagcctatg  | 720  |
| cacccacat ctgtgcacgg cgtggaagat atgatccggc tgggcgatct gaacgaggcc   | 780  |
| ggcatcctga gaaacctgct gatcagatac cgggaccacc tgatctacac ctacaccggc  | 840  |
| tctatcctgg tggccgtgaa tccctaccag ctgctgagca tctacagccc cgagcacatc  | 900  |
| cggcagtaca ccaacaagaa aatcggcgag atgcctcctc acatcttcgc cattgccgac  | 960  |
| aactgctact tcaacatgaa gcggaacagc cgggaccagt gctgcatcat ctctggcgaa  | 1020 |
| tctggcgccg gaaagaccga gagcacaaag ctgatcctgc agttcctggc cgccatcagc  | 1080 |
| ggacagcact cttggattga gcagcaggtc ctggaagcca cacctattct ggaagccttc  | 1140 |
| ggcaacgcca agaccatccg gaacgacaac agcagcagat tcggcaaata catcgacatc  | 1200 |
| cacttcaaca agagaggcgc cattgagggc gccaaagatcg agcagtacct gctggaaaag | 1260 |
| tccagagtgt gcagacaggc cctggacgag agaaactacc acgtgttcta ctgcatgctg  | 1320 |
| gaaggcatga gcgaggacca gaagaagaag ctcggactcg gccaggccag cgactacaat  | 1380 |
| tatctggcca tgggcaactg catcacatgc gagggcagag tggacagcca agagtacgcc  | 1440 |
| aacatccgca gcgccatgaa ggtgctgatg ttcaccgaca ccgagaactg ggagatcagc  | 1500 |
| aaactgctgg ccgctatcct gcctctgggc aacctgcagt acgaggccag aaccttcgag  | 1560 |
| aacctggatg cctgcgaggt gctgttctct ccttccctgg ctaccgccgc ctctctgctg  | 1620 |
| gaagtgaacc ctctgatct gatgagctgc ctgaccagca gaacctgat caccagaggc    | 1680 |
| gagacagtgt ctaccctct gagcagagaa caggctctgg atgtgcggga cgccttcgtg   | 1740 |
| aagggcattc acggcagact gttcgtgtgg atcgtggaca agatcaacgc cgccatctac  | 1800 |
| aagcctcaa gccaggacgt gaagaacagc agaagatcca tcggcctgct ggacatcttc   | 1860 |
| ggcttcgaga atttcgccgt gaacagcttc gagcagctgt gcatcaactt cgccaacgag  | 1920 |
| cacctccagc agttcttcgt gcggcacgtg ttcaagctgg aacaagagga atacgacctg  | 1980 |
| gaatccatcg actggctgca catcgagttc accgataacc aggacgccct ggacatgac   | 2040 |
| gccaacaagc ccatgaacat catcagcctg atcgacgagg aaagcaagtt cccaagggc   | 2100 |
| accgatacca ccatgctgca caagctgaac agccagcaca aactgaatgc caactacatc  | 2160 |
| ccgcctaaga acaaccacga gacacagttc ggcatcaacc acttcgccgg catcgtgtac  | 2220 |

|            |            |            |            |            |             |      |
|------------|------------|------------|------------|------------|-------------|------|
| tacgaaaccc | agggctttct | ggaaaagaac | cgggacaccc | tgcacggcga | catcattcag  | 2280 |
| ctggtgcaca | gcagccggaa | caagttcatc | aagcagatct | tccaggccga | cgctgccatg  | 2340 |
| ggagccgaga | caagaaagag | aagccccaca | ctgagcagcc | agttcaagcg | gagtctggaa  | 2400 |
| ctgctgatga | gaaccctggg | agcctgccag | cctttctttg | tgcggtgcat | caagcccaac  | 2460 |
| gagttcaaga | aacccatgct | gttcgaccgg | cacctgtgtg | tgcggcagct | gagatacagc  | 2520 |
| ggcatgatgg | aaaccatcag | gattcggaga | gccggctatc | ccatccggta | cagcttcgtg  | 2580 |
| gaattcgtcg | agcggtacag | agtgtgtgtg | cctggcgtga | agcctgccta | caaacagggc  | 2640 |
| gatctcagag | gcacctgtca | gagaatggcc | gaagccgtgc | tgggcaccca | tgacgattgg  | 2700 |
| cagatcggaa | agacaaagat | cttcctgaag | gaccaccacg | acatgctgct | cgaggtggaa  | 2760 |
| agagacaagg | ccatcaccga | cagagtgatc | ctgctccaga | aagtgatccg | gggcttcaag  | 2820 |
| gacagaagca | atttcctgaa | gctgaagaat | gccgccactc | tgatccagag | acactggcgg  | 2880 |
| ggacacaact | gccggaagaa | ctacggcctg | atgaggctgg | gcttcctgag | actgcaggcc  | 2940 |
| ctgcacagaa | gcagaaagct | gcaccagcag | tacagactgg | cccggcagcg | gatcatccag  | 3000 |
| tttcaagcca | gatgtcgggc | ctacctgtg  | cgcaaggcct | tcagacatag | actgtgggcc  | 3060 |
| gtgctgaccg | tgcaggccta | tgccagagga | atgattgccc | gcagactgca | ccagagactg  | 3120 |
| agagccgagt | atctgtggcg | gctggaagcc | gagaaaatgc | ggctggccga | ggaagagaag  | 3180 |
| ctgcggaaag | agatgagcgc | caagaaggcc | aaagaagagg | ccgagcggaa | gcaccaagag  | 3240 |
| agactggctc | aactggccag | agaggacgcc | gagagagagc | tgaaagagaa | agaggccgcc  | 3300 |
| agacggaaga | aagaactcct | ggaacagatg | gaacgggcca | gacacgagcc | cgtgaaccac  | 3360 |
| agcgatatgg | tggataagat | gttcggcttc | ctgggcacct | ctggcggact | gcctggacaa  | 3420 |
| gaaggacagg | cccctagcgg | ctttgaggac | ctggaacgtg | ggagaagaga | aatggtggaa  | 3480 |
| gaggatctgg | acgccgctct | gcctctgcct | gacgaggatg | aagaagatct | gagcgagtac  | 3540 |
| aagttcgcca | agtttgccgc | cacctacttt | caaggcacca | ccacacacag | ctacaccaga  | 3600 |
| aggcctctga | agcagcccct | gctgtaccac | gatgatgagg | gcgatcaact | ggcagcccctg | 3660 |
| gccgtgtgga | ttaccatcct | cagattcatg | ggcgacctgc | ctgagcctaa | gtaccacacc  | 3720 |
| gcatgtctg  | acggctccga | gaagatcccc | gtgatgacca | agatctacga | gactctgggc  | 3780 |
| aagaaaacct | acaagcgcga | gctgcaggct | ctccaaggcg | aaggcgaagc | tcaactgcct  | 3840 |

|                                                                   |      |
|-------------------------------------------------------------------|------|
| gagggccaga aaaagtcctc tgtgcgccac aaactggtgc acctgacact gaagaagaaa | 3900 |
| agcaagctga cagaggaagt gaccaagcgg ctgcacgatg gcgagtctac agtgcagggc | 3960 |
| aacagcatgc tcgaggacag acccaccagc aacctggaaa aactgcactt catcatcggc | 4020 |
| aacggaatcc tgcggcctgc tctgagggat gagatctact gccagatctc caagcagctg | 4080 |
| acacacaacc ccagcaagag cagctacgcc agaggctgga ttctggtaag tattagctct | 4140 |
| ttctttccat gggttggcct cgccgcgtgg gctgagggaa ggactgtcct gggactggac | 4200 |
| aggcgggtta tgggacctga agcgataaaa ggcatgcacg tttgcggcta cgtgcatgcc | 4260 |
| aaaaggagtc gggcttgcct ccgtgcccga ctccaaaaga cctgctcgag gaggtggacg | 4320 |
| agcaggtcaa aaatccgggt accaataaaa tatctttatt ttcattacat ctgtgtgttg | 4380 |
| gttttttgtg tg                                                     | 4392 |

<210> 158  
 <211> 4055  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| <400> 158                                                          |     |
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60  |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120 |
| atgggtggag tattttacgtt aaactgccc cttggcagta catcaagtgt atcatatgcc  | 180 |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240 |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300 |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360 |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420 |
| ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt   | 480 |
| acggtgggag gtctatataa gcagagcttg gatgttgctt ttaggatttt tgacctgctc  | 540 |
| gattgtccac tgcgagcagg tcttttggag tcgggcgagg cggaagcccg actccttttg  | 600 |
| gcatgcacgc tagccgcgtc gtgcatgcct tttatcttcg ggttatggga ccagtgaagg  | 660 |
| ctgaggggaag gactgtcctg ggactggaca ggcgggttat gggacctgaa aatactaaca | 720 |

|            |            |             |            |            |             |      |
|------------|------------|-------------|------------|------------|-------------|------|
| atcgatTTTT | tttccTTTT  | tttccaggtg  | tctctgtgcg | tgggctgttt | cgccccaagc  | 780  |
| gagaagtTCG | tgaagtacct | gaggaacttc  | atccacggcg | gacctccagg | ctacgcccct  | 840  |
| tactgtgaag | agaggctgag | aaggaccttt  | gtgaacggca | cccggacaca | gcctccatcc  | 900  |
| tggctggaac | tccaggccac | caagagcaaa  | aagcccatca | tgctgcccgt | gacctttatg  | 960  |
| gatggcacca | caaagaccct | gctgaccgat  | agcgccacca | ccgccaaaga | gctgtgtaac  | 1020 |
| gccctggctg | acaagattag | cctgaaggat  | agattcggct | tcagcctgta | cattgccctg  | 1080 |
| ttcgacaagg | tgtccagcct | cggctctggc  | tctgaccatg | tgatggatgc | catcagccag  | 1140 |
| tgcgagcagt | atgccaaaga | acagggcgcc  | caagagagga | acgctccttg | gcggctgttc  | 1200 |
| tttcggaaag | aggtgttcac | cccttggcac  | agccccagcg | aagataacgt | ggccaccaat  | 1260 |
| ctgatctacc | agcaagttgt | gcggggcgctg | aagttcggcg | agtacagatg | cgaaaaagag  | 1320 |
| gacgatctgg | ccgagctggc | ctctcagcag  | tactttgtgg | actacggcag | cgagatgatc  | 1380 |
| ctggaacggc | tgctgaatct | ggtgcccacc  | tacattcccc | atcgggagat | cacccccactg | 1440 |
| aaaaccctcg | agaagtgggc | ccagctggcc  | attgctgccc | acaagaaagg | catctatgcc  | 1500 |
| cagcggagaa | cagacgccc  | gaaagtcaaa  | gaggatgtcg | ttagctacgc | ccggttcaag  | 1560 |
| tggcctctgc | tgtttagccg | gttctacgag  | gcctacaagt | tcagcggccc | cagtctgccc  | 1620 |
| aagaacgatg | tgatcgtggc | tgtgaactgg  | accggcgtgt | acttcgtgga | tgagcaagaa  | 1680 |
| caagtgctgc | ttgagctgag | cttccccgag  | atcatggccg | tgtccagctc | cagagaatgc  | 1740 |
| agagtgtggc | tgagcctggg | ctgtagcgat  | ctgggatgtg | ccgctcctca | ttctggatgg  | 1800 |
| gctggactga | caccagccgg | accttgtagc  | ccttgttggt | cttgccgggg | ggccaagaca  | 1860 |
| acagccccta | gctttaccct | ggccaccatt  | aagggcgacg | agtacacctt | caccagcagc  | 1920 |
| aacgccgagg | acatcagaga | tctggtcgtg  | accttccttg | aaggcctgcg | gaagcggagc  | 1980 |
| aaatatgtgg | tggccctgca | ggacaacccc  | aatcctgctg | gcgaggaatc | cggctttctg  | 2040 |
| agctttgcca | aaggcgacct | gatcatcctg  | gaccacgaca | ccggcgagca | agtgatgaat  | 2100 |
| agcggctggg | ccaacggcat | caatgagcgg  | acaaagcagc | ggggcgactt | ccctaccgat  | 2160 |
| agcgtgtacg | tgatgcccac | cgtgaccatg  | cctccaaggg | aaatcgtggc | cctggtcacc  | 2220 |
| atgacacccg | accagagaca | ggatgtttgtg | cggctgctgc | agctgaggac | agccgaacca  | 2280 |
| gaagtgcggg | ccaagcctta | cacactggaa  | gagttcagct | acgactactt | ccggcctcct  | 2340 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| ccaaagcaca ccctgtctag agtgatggtg tccaaggcca gaggcaagga taggctgtgg  | 2400 |
| tcccacacaa gagagcccct gaaacaggca ctgctgaaaa agctgctggg cagcgaggaa  | 2460 |
| ctgagccaag aagcctgtct ggcctttatc gccgtgctga agtacatggg cgattacccc  | 2520 |
| tccaagcgga ccagatccgt gaacgaactg accgaccaga ttttcgaggg cccactgaag  | 2580 |
| gccgagcctc tgaaagatga ggcctacgtg cagattctga aacagctgac cgacaaccac  | 2640 |
| atccgctaca gcgaggaacg cggatgggaa ctgctgtggc tgtgtaccgg actgttccca  | 2700 |
| cctagcaaca ttctgctgcc ccacgtgcag cggtttctgc agtctagaaa gcactgccct  | 2760 |
| ctggccatcg attgcctgca gaggctgcaa aaggccctga gaaatggctc ccggaagtac  | 2820 |
| cctcctcacc tgggtggaagt ggaagccatc cagcacaaga ccacacagat ctttcacaag | 2880 |
| gtctacttcc ccgacgacac agacgaggcc tttgaggtgg aatcctctac caaggccaag  | 2940 |
| gacttctgcc agaatatcgc caccaggctg ctgctgaagt ccagcgaagg ctttagcctg  | 3000 |
| tttgtgaaga tcgccgacaa agtgctgagc gtgcccgaga acgacttctt tttcgatttt  | 3060 |
| gtgcgccatc tgaccgactg gattaagaag gctagacca tcaaggatgg catcgtgccc   | 3120 |
| agcctgacct atcaggtgtt ctttatgaag aagctgtgga cgaccaccgt gcctggcaag  | 3180 |
| gatcctatgg ccgacagcat cttccactac taccaagagc tgcccaagta cctgcggggc  | 3240 |
| taccacaagt gtaccagaga agaggctctg cagctgggag ccctgatcta tagagtgaag  | 3300 |
| tttgaagagg acaagagcta cttccctagc atccccaagc tgctgcgcga actggttccc  | 3360 |
| caggatctga tccggcaagt gtcccctgat gactggaagc ggtctatcgt ggcctacttt  | 3420 |
| aacaagcacg ccggcaagag taaagaggaa gccaaagctgg cttttctgaa gctcatcttt | 3480 |
| aagtggccta ctttcggctc cgccttcttc gaagtgaagc agaccaccga gcctaacttc  | 3540 |
| cctgagattc tgctgatcgc catcaacaaa tacggcgtgt ccctgatcga tcccaagaca  | 3600 |
| aaggacatcc tgacaacaca ccccttcacc aaaatcagca actggtccag cggcaacacc  | 3660 |
| tacttccaca tcaccatcgg caatctcgtg cggggctcta agctgctgtg tgaaaccagc  | 3720 |
| ctgggataca agatggacga cctgctgaca agctacatct cccagatgct gaccgccatg  | 3780 |
| agcaaacaga gaggctctcg gagcggcaag tggggcgctc gggctgacta caaagaccat  | 3840 |
| gacggtgatt ataaagatca tgacatcgac tataaggatg acgatgacaa atgaggtacc  | 3900 |
| aattcctcac ctgcgatctc gatgctttat ttgtgaaatt tgtgatgcta ttgctttatt  | 3960 |



|                                                                   |      |
|-------------------------------------------------------------------|------|
| tgtaaccatt ataagctgca ataaacaagt taacaacaac aattgcattc attttatgtt | 4020 |
| tcaggttcag ggggaggtgt gggaggtttt ttaaa                            | 4055 |

<210> 159  
 <211> 4161  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |      |
|--------------------------------------------------------------------|------|
| <400> 159                                                          |      |
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60   |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120  |
| atgggtggag tatttacggt aaactgccc cttggcagta catcaagtgt atcatatgcc   | 180  |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240  |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300  |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360  |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420  |
| ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt   | 480  |
| acggtgggag gtctatataa gcagagcttg gatgttgctt ttacttctag gcgcgccgcc  | 540  |
| accatgaaga gaaccgccga cggcagcgag ttcgagagcc ctaagaaaaa gcggaagggtg | 600  |
| gacaagaagt acagcatcgg cctggctatc ggcaccaatt ctgttggctg ggccgtgatc  | 660  |
| accgacgagt acaaggtgcc cagcaagaaa ttcaaggtgc tgggcaacac cgaccggcac  | 720  |
| agcatcaaga agaatctgat cggcgccctg ctgttcgact ctggcgaaac agccgaagcc  | 780  |
| accagactga agaggacagc cagacggcgg tacaccagaa gaaagaaccg gatctgctac  | 840  |
| ctgcaagaga tcttcagcaa cgagatggcc aaggtggacg acagcttctt ccaccggctg  | 900  |
| gaagagtcct tcctggtgga agaggataag aagcacgagc ggcaccccat cttcggcaac  | 960  |
| atcgtggatg aggtggccta ccacgagaag taccacacca tctaccacct gagaaagaaa  | 1020 |
| ctggtggaca gcaccgaaa ggccgacctg agactgatct atctggccct ggctcacatg   | 1080 |
| atcaagttcc ggggccactt cctgatcgag ggcgacctga atcctgacaa cagcgacgtg  | 1140 |
| gacaagctgt tcatccagct ggtgcagacc tacaaccagc tgttcgagga aaaccccatc  | 1200 |

|            |            |             |            |            |            |      |
|------------|------------|-------------|------------|------------|------------|------|
| aacgccagcg | gagtggatgc | caaggccatc  | ctgtctgccc | ggctgagcaa | gagcagacgg | 1260 |
| ctggaaaacc | tgatcgctca | gctgcccggc  | gagaagaaga | atggcctgtt | cggcaacctg | 1320 |
| attgccctga | gcctgggcct | gacacctaac  | ttcaagagca | acttcgacct | ggccgaggac | 1380 |
| gccaaactgc | agctgtccaa | ggacacctac  | gacgacgacc | tggacaatct | gctggcccag | 1440 |
| atcggcgatc | agtacgccga | cttgtttctg  | gccgccaaga | acctgtccga | cgccatcctg | 1500 |
| ctgagcgaca | tcctgagagt | gaacaccgag  | atcacaaagg | cccctctgag | cgcctctatg | 1560 |
| atcaagagat | acgacgagca | ccaccaggat  | ctgaccctgc | tgaaggccct | cgtagacag  | 1620 |
| cagctgcctg | agaagtacaa | agagattttc  | ttcgaccaga | gcaagaacgg | ctacgccggc | 1680 |
| tacattgatg | gcggagccag | ccaagaggaa  | ttctacaagt | tcatcaagcc | catcctcgag | 1740 |
| aagatggacg | gcaccgagga | actgctggtc  | aagctgaaca | gagaggacct | gctgcggaag | 1800 |
| cagcggacct | tcgacaatgg | ctctatccct  | caccaaattc | acctgggaga | gctgcacgcc | 1860 |
| attctgcgga | gacaagagga | cttttaccba  | ttcctgaagg | acaaccggga | aaagattgag | 1920 |
| aagatcctga | ccttcaggat | cccctactac  | gtgggaccac | tggccagagg | caatagcaga | 1980 |
| ttgcctgga  | tgaccagaaa | gagcgaggaa  | accatcacac | cctggaactt | cgaggaagtg | 2040 |
| gtggataagg | gcgccagcgc | tcagtccttc  | atcgagcgga | tgaccaactt | cgataagaac | 2100 |
| ctgcctaacg | agaagggtgt | gcccgaagcac | agcctgctgt | acgagtactt | caccgtgtac | 2160 |
| aacgagctga | ccaaagtga  | atacgtgacc  | gagggaatga | gaaagcccgc | ctttctgagc | 2220 |
| ggcgagcaga | aaaaggccat | tgtggatctg  | ctgttcaaga | ccaaccggaa | agtgaccgtg | 2280 |
| aagcagctga | aagaggacta | cttcaagaaa  | atcgagtgtc | tcgacagcgt | ggaaatcagc | 2340 |
| ggcgtggaag | atcggttcaa | tgccagcctg  | ggcacatacc | acgacctgtc | gaaaattatc | 2400 |
| aaggacaagg | acttcctgga | caacgaagag  | aacgaggaca | tcctggaaga | tatcgtgctg | 2460 |
| accctgacac | tgtttgagga | cagagagatg  | atcgaggaac | ggctgaaaac | atacggccac | 2520 |
| ctgttcgacg | acaaagtgat | gaagcaactg  | aagcggcgga | gatacaccgg | ctggggcaga | 2580 |
| ctgtctcgga | agctgatcaa | cggcatccgg  | gataagcagt | ccggcaagac | catcctggac | 2640 |
| tttctgaagt | ccgacggctt | cgccaatcgg  | aacttcatgc | agctgatcca | cgacgacagc | 2700 |
| ctgaccttta | aagaggatat | ccagaaagcc  | caggtgtccg | gccagggcga | ttctctgcat | 2760 |
| gagcacattg | ccaacctggc | cggctctccc  | gccattaaga | agggcattct | gcagacagtg | 2820 |

|                                                                     |      |
|---------------------------------------------------------------------|------|
| aaggtggtgg acgagctggt caaagtcattg ggcagacaca agcccagagaa catcgtgatc | 2880 |
| gaaatggcca gagagaacca gaccacacag aagggccaga agaacagccg cgagagaatg   | 2940 |
| aagcggatcg aagagggcat caaagagctg ggcagccaga tcctgaaaga acaccccgtg   | 3000 |
| gaaaacaccc agctgcagaa cgagaagctg tacctgtact acctccaaaa cggccgggat   | 3060 |
| atgtatgtgg accaagagct ggacatcaac cggctgtccg actacgatgt ggacgctatc   | 3120 |
| gtgccccagt cttttctgaa agacgactcc atcgacaaca aggtcctgac cagaagcgac   | 3180 |
| aagaaccggg gcaagagcga taacgtgccc tccgaagagg tcgtgaagaa gatgaagaac   | 3240 |
| tactggcgac agctgctgaa cgccaagctg attaccacgc ggaagttcga taacctgacc   | 3300 |
| aaggccgaga gaggcggcct gtctgaactg gataaggccg gcttcatcaa gagacagctg   | 3360 |
| gtggaaaccc ggcagatcac caaacacgtg gcacagattc tggactcccc gatgaacact   | 3420 |
| aagtacgacg agaatgacaa gctgatccgg gaagtgaaag tgatcacctt gaagtccaag   | 3480 |
| ctggtgtccg attttccgaa ggattttccag ttctacaaag tgcgcgagat caacaactac  | 3540 |
| catcacgccc acgacgccta cctgaatgcc gttgtttgaa cagccctgat caagaagtat   | 3600 |
| cccaagctgg aatccgagtt cgtgtacggc gactacaagg tgtacgacgt gcggaagatg   | 3660 |
| atcgccaaga gcgagcaaga gattggcaag gctaccgcca agtactttct ctacagcaac   | 3720 |
| atcatgaact ttttcaagac cgagattacc ctggccaacg gcgagatcag aaagcggcct   | 3780 |
| ctgatcgaga caaacggcga aaccggcgag attgtgtggg acaagggcag agattttgcc   | 3840 |
| accgtgcgga aagtgtctgag catgccccaa gtgaatatcg tgaaaaagac cgaggtaagt  | 3900 |
| attagctctt tctttccatg ggttggcctc gccgcgtggg ctgagggaag gactgtcctg   | 3960 |
| ggactggaca ggcgggttat gggacctgaa gcgataaaag gcatgcacgt ttgcggctac   | 4020 |
| gtgcatgcca aaaggagtcg ggcttgccctc cgtgcccgcac tccaaaagac ctgctcgagg | 4080 |
| aggtggacga gcaggtcaaa aatccgggta ccaataaaat atctttatct tcattacatc   | 4140 |
| tgtgtgttgg ttttttgtgt g                                             | 4161 |

<210> 160  
 <211> 3410  
 <212> DNA  
 <213> Artificial sequence  
 <220>

<223> Synthetic nucleic acid sequence

<400> 160

|                                                                    |      |
|--------------------------------------------------------------------|------|
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60   |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120  |
| atgggtggag tattttacggt aaactgccc cttggcagta catcaagtgt atcatatgcc  | 180  |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240  |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300  |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360  |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420  |
| ggactttcca aaatgtcgt acaactccgc ccattgacg caaatgggcg gtaggcgtgt    | 480  |
| acggtgggag gtctatataa gcagagcttg gatgttgcct ttaggatttt tgacctgctc  | 540  |
| gattgtccac tgcgagcagg tcttttggag tcgggcgagg cggaagcccg actccttttg  | 600  |
| gcatgcacgc tagccgcgtc gtgcatgcct tttatcttcg ggttatggga ccagtgaagg  | 660  |
| ctgagggaag gactgtcctg ggactggaca ggcgggttat gggacctgaa aatactaaca  | 720  |
| atcgatTTTT tttccctttt tttccaggtg cagacaggcg gcttcagcaa agagtccatt  | 780  |
| ctgcccaga gaaacagcga taagctgac gcccggaaga aggactggga ccctaagaag    | 840  |
| tacggcggct tcgatagccc taccgtggcc tattctgtgc tgggtggtggc caaagtggaa | 900  |
| aagggcaagt ccaagaaact caagagcgtg aaagagctgc tggggatcac catcatggaa  | 960  |
| agaagcagct tcgagaagaa tcctatcgat ttctctgagg ccaagggcta caaagaagt   | 1020 |
| aaaaaggacc tgatcatcaa gctccccaag tactccctgt tcgagctgga aaatggccgg  | 1080 |
| aagcggatgc tggcttctgc tggcgaactg cagaaggga acgaactggc cctgcctagc   | 1140 |
| aaatatgtga acttcctgta cctggccagc cactatgaga agctgaaggg cagccccgag  | 1200 |
| gacaatgacg aaaagcagct gtttgtggaa cagcacaagc actacctgga cgagatcatc  | 1260 |
| gagcagatct ccgagttctc caagagagt atcctggccg acgctaattt ggacaaagt    | 1320 |
| ctgtccgcct acaacaagca ccgggacaag cctatcagag agcaggccga gaatatcatc  | 1380 |
| cacctgttta ccctgaccaa tctgggagcc cctgccgcct tcaagtactt tgacaccacc  | 1440 |
| atcgaccgga agcgtacac cagcaccaaa gaggtgctgg acgccacact gatccaccag   | 1500 |
| tctatcaccg gcctgtacga gacacggatc gacctgtctc agctcgagg cgatagcagg   | 1560 |

|            |            |             |            |            |            |      |
|------------|------------|-------------|------------|------------|------------|------|
| gctgacccca | agaagaagag | gaaggtgtcg  | ccagggatcc | gtcgacttga | cgcgttgata | 1620 |
| tcaacaagtt | tgtacaaaaa | agcaggctac  | aaagaggcca | gcggttccgg | acgggctgac | 1680 |
| gcattggacg | attttgatct | ggatatgctg  | ggaagtgacg | ccctcgatga | ttttgacctt | 1740 |
| gacatgcttg | gttcggatgc | ccttgatgac  | tttgacctcg | acatgctcgg | cagtgacgcc | 1800 |
| cttgatgatt | tcgacctgga | catgctgatt  | aactctagaa | gttccggatc | tccgaaaaag | 1860 |
| aaacgcaaag | ttggtagcca | gtacctgccc  | gacaccgacg | accggcaccg | gatcgaggaa | 1920 |
| aagcggaaag | ggacctacga | gacattcaag  | agcatcatga | agaagtcccc | cttcagcggc | 1980 |
| cccaccgacc | ctagacctcc | acctagaaga  | atgcccgctg | ccagcagatc | cagcgccagc | 2040 |
| gtgccaaaac | ctgcccccca | gccttacccc  | ttcaccagca | gcctgagcac | catcaactac | 2100 |
| gacgagttcc | ctaccatggt | gttccccagc  | ggccagatct | ctcaggcctc | tgctctggct | 2160 |
| ccagcccctc | ctcaggtgct | gcctcaggct  | cctgctcctg | caccagctcc | agccatggtg | 2220 |
| tctgcactgg | ctcaggcacc | agcaccctg   | cctgtgctgg | ctcctggacc | tccacaggct | 2280 |
| gtggctccac | cagcccctaa | acctacacag  | gccggcgagg | gcacactgtc | tgaagctctg | 2340 |
| ctgcagctgc | agttcgacga | cgaggatctg  | ggagccctgc | tgggaaacag | caccgatcct | 2400 |
| gccgtgttca | ccgacctggc | cagcgtggac  | aacagcgagt | tccagcagct | gctgaaccag | 2460 |
| ggcatccctg | tggcccctca | caccaccgag  | cccatgctga | tggaataccc | cgaggccatc | 2520 |
| acccggctcg | tgacaggcgc | tcagaggcct  | cctgatccag | ctcctgcccc | tctgggagca | 2580 |
| ccaggcctgc | ctaatggact | gctgtctggc  | gacgaggact | tcagctctat | cgccgatatg | 2640 |
| gatttctcag | ccttgctggg | ctctggcagc  | ggcagccggg | attccaggga | agggatgttt | 2700 |
| ttgccgaagc | ctgaggccgg | ctccgctatt  | agtgacgtgt | ttgagggccg | cgagggtgtg | 2760 |
| cagccaaaac | gaatccggcc | atttcatacct | ccaggaagtc | catgggccaa | ccgcccactc | 2820 |
| cccgccagcc | tcgcaccaac | accaaccggt  | ccagtacatg | agccagtcgg | gtcactgacc | 2880 |
| ccggcaccag | tcctcagcc  | actggatcca  | gcgcccgcag | tgactcccga | ggccagtcac | 2940 |
| ctgttgaggg | atcccgatga | agagacgagc  | caggctgtca | aagcccttcg | ggagatggcc | 3000 |
| gatactgtga | ttccccagaa | ggaagaggct  | gcaatctgtg | gccaaatgga | cctttcccat | 3060 |
| ccgcccccaa | ggggccatct | ggatgagctg  | acaaccacac | ttgagtccat | gaccgaggat | 3120 |
| ctgaacctgg | actcaccctt | gaccccggaa  | ttgaacgaga | ttctggatac | cttcctgaac | 3180 |

|                                                                       |      |
|-----------------------------------------------------------------------|------|
| gacgagtgcc tcttgc atgc catgc atatc agcacaggac tgtccatctt cgacacatct   | 3240 |
| ctgtttttgag gtaccaattc ctcacctgcg atctc gatgc tttattttgtg aaattttgtga | 3300 |
| tgctattgct ttattttgtaa ccattataag ctgcaataaa caagttaaca acaacaattg    | 3360 |
| cattcatttt atgttttcagg ttcaggggga ggtgtgggag gtttttttaa               | 3410 |

<210> 161  
 <211> 4161  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |      |
|--------------------------------------------------------------------|------|
| <400> 161                                                          |      |
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60   |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120  |
| atgggtggag tattttacggt aaactgccc cttggcagta catcaagtgt atcatatgcc  | 180  |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240  |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300  |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360  |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420  |
| ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt   | 480  |
| acggtgggag gtctatataa gcagagcttg gatgttgcct ttacttctag gcgcgccgcc  | 540  |
| accatgaaga gaaccgccga cggcagcgag ttcgagagcc ctaagaaaaa gcggaaggtg  | 600  |
| gacaagaagt acagcatcgg cctggacatc ggcaccaatt ctgttggctg ggccgtgatc  | 660  |
| accgacgagt acaaggtgcc cagcaagaaa ttcaaggtgc tgggcaacac cgaccggcac  | 720  |
| agcatcaaga agaatctgat cggcgccctg ctgttcgact ctggcgaaac agccgaagcc  | 780  |
| accagactga agaggacagc cagacggcgg tacaccagaa gaaagaaccg gatctgctac  | 840  |
| ctgcaagaga tcttcagcaa cgagatggcc aaggtggacg acagcttctt ccaccggctg  | 900  |
| gaagagtcct tcctgggtgga agaggataag aagcacgagc ggcaccccat cttcggcaac | 960  |
| atcgtggatg aggtggccta ccacgagaag taccacacca tctaccacct gagaaagaaa  | 1020 |
| ctggtggaca gcaccgacaa ggccgacctg agactgatct atctggccct ggctcacatg  | 1080 |

|            |            |            |            |            |            |      |
|------------|------------|------------|------------|------------|------------|------|
| atcaagttcc | ggggccactt | cctgatcgag | ggcgacctga | atcctgacaa | cagcgacgtg | 1140 |
| gacaagctgt | tcatccagct | ggtgcagacc | tacaaccagc | tgttcgagga | aaaccccatc | 1200 |
| aacgccagcg | gagtggatgc | caaggccatc | ctgtctgccc | ggctgagcaa | gagcagacgg | 1260 |
| ctggaaaacc | tgatcgctca | gctgcccggc | gagaagaaga | atggcctgtt | cggcaacctg | 1320 |
| attgccctga | gcctgggcct | gacacctaac | ttcaagagca | acttcgacct | ggccgaggac | 1380 |
| gccaaactgc | agctgtccaa | ggacacctac | gacgacgacc | tggacaatct | gctggcccag | 1440 |
| atcggcgatc | agtacgccga | cttgtttctg | gccgccaaga | acctgtccga | cgccatcctg | 1500 |
| ctgagcgaca | tcctgagagt | gaacaccgag | atcacaaagg | cccctctgag | cgcctctatg | 1560 |
| atcaagagat | acgacgagca | ccaccaggat | ctgaccctgc | tgaaggccct | cgttagacag | 1620 |
| cagctgcctg | agaagtacaa | agagattttc | ttcgaccaga | gcaagaacgg | ctacgccggc | 1680 |
| tacattgatg | gcggagccag | ccaagaggaa | ttctacaagt | tcatcaagcc | catcctcgag | 1740 |
| aagatggacg | gcaccgagga | actgctggtc | aagctgaaca | gagaggacct | gctgcggaag | 1800 |
| cagcggacct | tcgacaatgg | ctctatccct | caccaaattc | acctgggaga | gctgcacgcc | 1860 |
| attctgcgga | gacaagagga | cttttaccba | ttcctgaagg | acaaccggga | aaagattgag | 1920 |
| aagatcctga | ccttcaggat | cccctactac | gtgggaccac | tggccagagg | caatagcaga | 1980 |
| ttcgcttga  | tgaccagaaa | gagcgaggaa | accatcacac | cctggaactt | cgaggaagtg | 2040 |
| gtggataagg | gcgccagcgc | tcagtccttc | atcgagcgga | tgaccaactt | cgataagaac | 2100 |
| ctgcctaacg | agaagggtgt | gcccagcac  | agcctgctgt | acgagtactt | caccgtgtac | 2160 |
| aacgagctga | ccaaagtga  | atacgtgacc | gagggaatga | gaaagccgc  | ctttctgagc | 2220 |
| ggcgagcaga | aaaaggccat | tgtggatctg | ctgttcaaga | ccaaccggaa | agtgaccgtg | 2280 |
| aagcagctga | aagaggacta | cttcaagaaa | atcgagtgtt | tcgacagcgt | ggaaatcagc | 2340 |
| ggcgtggaag | atcggttcaa | tgccagcctg | ggcacatacc | acgacctgct | gaaaattatc | 2400 |
| aaggacaagg | acttcctgga | caacgaagag | aacgaggaca | tcctggaaga | tatcgtgctg | 2460 |
| accctgacac | tgtttgagga | cagagagatg | atcgaggaa  | ggctgaaaac | atacgcccac | 2520 |
| ctgttcgacg | acaaagtgat | gaagcaactg | aagcggcgga | gatacaccgg | ctggggcaga | 2580 |
| ctgtctcgga | agctgatcaa | cggcatccgg | gataagcagt | ccggcaagac | catcctggac | 2640 |
| tttctgaagt | ccgacggctt | cgccaatcgg | aacttcatgc | agctgatcca | cgacgacagc | 2700 |

|                                                                     |      |
|---------------------------------------------------------------------|------|
| ctgaccttta aagaggatat ccagaaagcc cagggtgtccg gccagggcga ttctctgcat  | 2760 |
| gagcacattg ccaacctggc cggctctccc gccattaaga agggcattct gcagacagt    | 2820 |
| aagggtggtgg acgagctggt caaagtcattg ggcagacaca agcccagaaa catcgtgatc | 2880 |
| gaaatggcca gagagaacca gaccacacag aaggggccaga agaacagccg cgagagaatg  | 2940 |
| aagcggatcg aagagggcat caaagagctg ggcagccaga tcctgaaaga acaccccg     | 3000 |
| gaaaacaccc agctgcagaa cgagaagctg tacctgtact acctccaaaa cggccgggat   | 3060 |
| atgtatgtgg accaagagct ggacatcaac cggctgtccg actacgatgt ggacgctatc   | 3120 |
| gtgccccagt cttttctgaa agacgactcc atcgacaaca aggtcctgac cagaagcgac   | 3180 |
| aagaaccggg gcaagagcga taacgtgccc tccgaagagg tcgtgaagaa gatgaagaac   | 3240 |
| tactggcgac agctgctgaa cgccaagctg attaccagc ggaagttcga taacctgacc    | 3300 |
| aaggccgaga gaggcggcct gtctgaactg gataaggccg gcttcatcaa gagacagctg   | 3360 |
| gtggaaaccc ggcagatcac caaacacgtg gcacagattc tggactcccg gatgaacact   | 3420 |
| aagtacgacg agaatgacaa gctgatccgg gaagtgaag tgatcaccct gaagtccaag    | 3480 |
| ctggtgtccg atttccggaa ggatttccag ttctacaaag tgcgcgagat caacaactac   | 3540 |
| catcacgccc acgacgccta cctgaatgcc gttgttgga cagccctgat caagaagtat    | 3600 |
| cccaagctgg aatccgagtt cgtgtacggc gactacaagg tgtacgacgt gcggaagatg   | 3660 |
| atcgccaaga gcgagcaaga gattggcaag gctaccgcca agtacttctt ctacagcaac   | 3720 |
| atcatgaact ttttcaagac cgagattacc ctggccaacg gcgagatcag aaagcggcct   | 3780 |
| ctgatcgaga caaacggcga aaccggcgag attgtgtggg acaagggcag agattttgcc   | 3840 |
| accgtgcgga aagtgtgag catgccccaa gtgaatatcg tgaaaaagac cgaggtaagt    | 3900 |
| attagctctt tctttccatg gggttgccctc gccgcgtggg ctgagggaag gactgtcctg  | 3960 |
| ggactggaca ggcggttat gggacctgaa gcgataaaag gcatgcacgt ttgcggctac    | 4020 |
| gtgcatgcca aaaggagtcg ggcttgccctc cgtgcccgcac tccaaaagac ctgctcgagg | 4080 |
| aggtggacga gcaggtcaaa aatccgggta ccaataaaat atctttatct tcattacatc   | 4140 |
| tgtgtgttgg ttttttgtgt g                                             | 4161 |

<210> 162  
<211> 3911



<212> DNA  
<213> Artificial sequence

<220>  
<223> Synthetic nucleic acid sequence

<400> 162  
cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt 60  
gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca 120  
atgggtggag tatttacggt aaactgccca cttggcagta catcaagtgt atcatatgcc 180  
aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta 240  
catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac 300  
catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg 360  
atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg 420  
ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt 480  
acggtgggag gtctatataa gcagagcttg gatgttgctt ttaggatttt tgacctgctc 540  
gattgtccac tgcgagcagg tcttttggag tcgggcgagg cggaagcccg actccttttg 600  
gcatgcacgc tagccgcgtc gtgcatgcct tttatcttcg ggttatggga ccagtgaagg 660  
ctgagggaag gactgtcctg ggactggaca ggcgggttat gggacctgaa aatactaaca 720  
atcgattttt tttccctttt tttccaggtg cagacaggcg gcttcagcaa agagtccatt 780  
ctgccaaga gaaacagcga taagctgatc gcccggaaga aggactggga ccctaagaag 840  
tacggcggct tcgatagccc taccgtggcc tattctgtgc tgggtggtggc caaagtggaa 900  
aagggaagt ccaagaaact caagagcgtg aaagagctgc tggggatcac catcatggaa 960  
agaagcagct tcgagaagaa tcctatcgat ttcctcgagg ccaagggcta caaagaagtg 1020  
aaaaaggacc tgatcatcaa gctccccaag tactccctgt tcgagctgga aaatggccgg 1080  
aagcggatgc tggcttctgc tggcgaactg cagaaggga acgaactggc cctgcctagc 1140  
aaatatgtga acttcctgta cctggccagc cactatgaga agctgaaggg cagccccgag 1200  
gacaatgagc aaaagcagct gtttgtggaa cagcacaagc actacctgga cgagatcatc 1260  
gagcagatct ccgagttctc caagagagtg atcctggccg acgctaattt ggacaaagtg 1320  
ctgtccgcct acaacaagca ccgggacaag cctatcagag agcaggccga gaatatcatc 1380  
cacctgttta ccctgaccaa tctgggagcc cctgccgcct tcaagtactt tgacaccacc 1440

|                                                                    |      |
|--------------------------------------------------------------------|------|
| atcgaccgga agcgctacac cagcaccaaa gaggtgctgg acgccacact gatccaccag  | 1500 |
| tctatcaccg gcctgtacga gacacggatc gacctgtctc agctcggagg cgattctggc  | 1560 |
| ggatctagcg gtggaagctc tggctctgag acacctggca caagcgagtc tgccacacct  | 1620 |
| gagtctagcg gcggatcttc aggcggcagc agcacctga atatcgagga tgagtacaga   | 1680 |
| ctgcacgaga caagcaaaga acccgacgtg tccctgggct ctacctggct gtctgatttt  | 1740 |
| cctcaagcct gggccgaaac aggcggaatg ggacttgctg ttagacaggc tcccctgatc  | 1800 |
| attcccctga aggccacaag caccctgtg tccatcaagc agtaccatcat gtctcaagag  | 1860 |
| gcccggctgg gaatcaagcc ccacattcag agactgctgg accagggcac cctgggtgcct | 1920 |
| tgtcaaagcc cttggaatac ccctctgctg cctgtgaaga agcccggcac caacgactac  | 1980 |
| agaccctgac aggatctgac cgaagtgaac aagagagtcg aggacattca cccaccgtg   | 2040 |
| cctaatacctt acaacctgct gtctggcctg cctccttccc accaatggta cacagtgtg  | 2100 |
| gacctgaagg atgccttctt ctgcctgcgg ctgcacccta caagccagcc tctgtttgcc  | 2160 |
| ttcgagtggc gggatccaga gatgggcatt agcggacagc tgacctggac cagactgccc  | 2220 |
| cagggttca agaatagccc cacactgttc aacgaggccc tgcacaggga cctcgccgac   | 2280 |
| tttagaattc agcaccatga cctgattctg ctgcagtatg tggatgatct gctgctggcc  | 2340 |
| gctaccagcg agctggattg tcagcaggga acaagagccc tgctgcagac cctgggcaat  | 2400 |
| ctgggctata gagcctctgc caagaaggcc cagatttgcc agaagcaagt taagtacctg  | 2460 |
| ggctacctgc tcaaagaagg ccagcgttgg ctgaccgagg ccagaaaaga aaccgtgatg  | 2520 |
| ggccagccta cacctaagac acccagacag ctgagagagt tcctgggcaa agccggattc  | 2580 |
| tgcaggctgt ttatccctgg cttcgccgag atggctgccc ctctgtatcc tctgacaaag  | 2640 |
| cccggaactc tgttcaactg gggcccagac cagcagaaag cctaccaaga gatcaagcag  | 2700 |
| gctctgctga cagcccctgc tctgggactg cctgatctga ccaagccttt cgagctgttc  | 2760 |
| gtggacgaga agcagggtta tgccaagggc gtgctgacac agaaactcgg cccttgagga  | 2820 |
| aggcccgtgg cttacctgag caaaaagctg gatcctgtgg ccgctggctg gcctccttgt  | 2880 |
| ctgagaatgg tggccgctat cgccgtgctg actaaggatg ccggcaagct gacaatggga  | 2940 |
| cagcctctgg ttattctggc ccctcatgcc gtggaagccc tcgtgaaaca gcctcctgat  | 3000 |
| cggtggctga gcaacgccag aatgaccac taccaggcac tgctgctcga caccgacaga   | 3060 |

|                                                                   |      |
|-------------------------------------------------------------------|------|
| gtgcaatttg gccctgtggt ggccctgaat ccagccacat tgctgcctct gcctgaggag | 3120 |
| ggactgcagc acaactgcct cgatatcctg gctgaggccc acggcacaag acccgatctg | 3180 |
| acagatcagc cactgcctga cgccgaccac acctggtata cagatggcag ctctctgctg | 3240 |
| caagagggcc agagaaaagc tggcgccgct gtgaccacag agacagaagt gatttgggcc | 3300 |
| aaagctctgc ctgccggcac atctgccccaa agagccgaac tgatcgact gacacaggcc | 3360 |
| ctgaagatgg ccgagggcaa gaaactgaac gtgtacaccg actccagata cgccttcgcc | 3420 |
| accgctcaca tccacggcga aatctacaga cgagaggat ggctgaccag cgagggaaaa  | 3480 |
| gagattaaga acaaggacga gattctcgcc ctctcaagg ccctgttcct gcctaagcgg  | 3540 |
| ctgagcatca tccactgtcc tggccaccag aagggaact ctgccgaggc tagaggcaac  | 3600 |
| agaatggccg atcaggctgc cagaaaggcc gccattaccg agacacccga taccagcaca | 3660 |
| ctgctgattg agaacagcag cccttcggc ggctccaaaa gaacagctga cggctccgag  | 3720 |
| tttgagccca aaaagaaacg gaaagtgtga ggtaccaatt cctcacctgc gatctcgatg | 3780 |
| ctttatttgt gaaatttgtg atgctattgc tttatttgta accattataa gctgcaataa | 3840 |
| acaagttaac aacaacaatt gcattcattt tatgtttcag gttcaggggg aggtgtggga | 3900 |
| ggttttttta a                                                      | 3911 |

<210> 163  
 <211> 3159  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| <400> 163                                                         |     |
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt | 60  |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca | 120 |
| atgggtggag tatttacggt aaactgcccc cttggcagta catcaagtgt atcatatgcc | 180 |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta | 240 |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac | 300 |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg | 360 |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg | 420 |

|            |            |             |             |            |            |      |
|------------|------------|-------------|-------------|------------|------------|------|
| ggactttcca | aaatgtcgt  | acaactccgc  | cccattgacg  | caaatgggcg | gtaggcgtgt | 480  |
| acggtgggag | gtctatataa | gcagagcttg  | gatgttgctt  | ttacttctag | gcgcgccacc | 540  |
| atgaaacgga | cagccgacgg | aagcgagttc  | gagtcaccaa  | agaagaagcg | gaaagtcagc | 600  |
| agtgaaaccg | gaccagtggc | agtggaccca  | accctgagga  | gacggattga | gccccatgaa | 660  |
| tttgaagtgt | tctttgaccc | aaggagctg   | aggaaggaga  | catgcctgct | gtacgagatc | 720  |
| aagtggggca | caagccacaa | gatctggcgc  | cacagctcca  | agaacaccac | aaagcacgtg | 780  |
| gaagtgaatt | tcatcgagaa | gtttacctcc  | gagcggcact  | tctgcccctc | taccagctgt | 840  |
| tccatcacat | ggtttctgtc | ttggagccct  | tgcggcgagt  | gttccaaggc | catcaccgag | 900  |
| ttcctgtctc | agcaccctaa | cgtgaccctg  | gtcatctacg  | tggcccggct | gtatcaccac | 960  |
| atggaccagc | agaacaggca | gggcctgcgc  | gatctgggtga | attctggcgt | gaccatccag | 1020 |
| atcatgacag | ccccagagta | cgactattgc  | tggcgggaact | tcgtgaatta | tccacctggc | 1080 |
| aaggaggcac | actggccaag | atacccaccc  | ctgtggatga  | agctgtatgc | actggagctg | 1140 |
| cacgcaggaa | tcctgggcct | gcctccatgt  | ctgaatatcc  | tgcggagaaa | gcagccccag | 1200 |
| ctgacatfff | tcaccattgc | tctgcaatct  | tgtcactatc  | agcggctgcc | tcctcatatt | 1260 |
| ctgtgggcta | ccggcctgaa | gtctggagga  | tctagcggag  | gatcctctgg | cagcgagaca | 1320 |
| ccaggaacaa | gcgagtcagc | aacaccagag  | agcagtggcg  | gcagcagcgg | cggcagcgac | 1380 |
| aagaagtaca | gcatcggcct | ggccatcggc  | accaattctg  | ttggctgggc | cgtgatcacc | 1440 |
| gacgagtaca | aggtgcccag | caagaaattc  | aaggtgctgg  | gcaacaccga | ccggcacagc | 1500 |
| atcaagaaga | atctgatcgg | cggcctgctg  | ttcgactctg  | gcgaaacagc | cgaagccacc | 1560 |
| agactgaaga | ggacagccag | acggcgggtac | accagaagaa  | agaaccggat | ctgctacctg | 1620 |
| caagagatct | tcagcaacga | gatggccaag  | gtggacgaca  | gcttcttcca | ccggctggaa | 1680 |
| gagtccttcc | tggtggaaga | ggataagaag  | cacgagcggc  | accccatctt | cggcaacatc | 1740 |
| gtggatgagg | tggcctacca | cgagaagtac  | cccaccatct  | accacctgag | aaagaaactg | 1800 |
| gtggacagca | ccgacaaggc | cgacctgaga  | ctgatctatc  | tggccctggc | tcacatgatc | 1860 |
| aagttccggg | gccacttcct | gatcgagggc  | gacctgaatc  | ctgacaacag | cgacgtggac | 1920 |
| aagctgttca | tccagctggg | gcagacctac  | aaccagctgt  | tcgaggaaaa | ccccatcaac | 1980 |
| gccagcggag | tggatgccaa | ggccatcctg  | tctgcccggc  | tgagcaagag | cagacggctg | 2040 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| gaaaacctga tcgctcagct gcccggcgag aagaagaatg gcctgttcgg caacctgatt  | 2100 |
| gccctgagcc tgggcctgac acctaacttc aagagcaact tcgacctggc cgaggacgcc  | 2160 |
| aaactgcagc tgtccaagga cacctacgac gacgacctgg acaatctgct ggcccagatc  | 2220 |
| ggcgatcagt acgccgactt gtttctggcc gccaagaacc tgtccgacgc catcctgctg  | 2280 |
| agcgacatcc tgagagtga caccgagatc acaaaggccc ctctgagcgc ctctatgatc   | 2340 |
| aagagatacg acgagcacca ccaggatctg accctgctga aggccctcgt tagacagcag  | 2400 |
| ctgcctgaga agtaciaaaga gattttcttc gaccagagca agaacggcta cgccggctac | 2460 |
| attgatggcg gagccagcca agaggaattc tacaagttca tcaagcccat cctcgagaag  | 2520 |
| atggacggca ccgaggaact gctgggtcaag ctgaacagag aggacctgct gcggaagcag | 2580 |
| cggaccttcg acaatggctc tatccctcac caaatccacc tgggagagct gcacgccatt  | 2640 |
| ctgcggagac aagaggactt ttaccatttc ctgaaggaca accgggaaaa gattgagaag  | 2700 |
| atcctgacct tcaggatccc ctactacgtg ggaccactgg ccagaggcaa tagcagattc  | 2760 |
| gcctggatga ccagaaagag cgaggaaacc atcacaccct ggaacttcga ggaagtgggtg | 2820 |
| gataagggcg ccagcgctca gtccttcac gagcggatga ccaacttcga taagaacctg   | 2880 |
| cctaacgaga aggtaagtat tagctctttc tttccatggg ttggcctcgc cgcgtgggct  | 2940 |
| gaggggaagga ctgtcctggg actggacagg cgggttatgg gacctgaagc gataaaaggc | 3000 |
| atgcacgttt gcggctacgt gcatgccaaa aggagtcggg cttgcctccg tgcccgactc  | 3060 |
| caaaagacct gctcgaggag gtggacgagc aggtcaaaaa tccgggtacc aataaaatat  | 3120 |
| ctttattttc attacatctg tgtgttggtt ttttgtgtg                         | 3159 |

<210> 164

<211> 4115

<212> DNA

<213> Artificial sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 164

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt | 60  |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca | 120 |
| atgggtggag tattttacggt aaactgcca cttggcagta catcaagtg atcatatgcc  | 180 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240  |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300  |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360  |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420  |
| ggactttcca aaatgtcgta acaactccgc cccattgacg caaatgggcg gtaggcgtgt  | 480  |
| acggtgggag gtctatataa gcagagcttg gatgttgctt ttaggatttt tgacctgctc  | 540  |
| gattgtccac tgcgagcagg tcttttggag tcgggcgagg cggaagcccg actccttttg  | 600  |
| gcatgcacgc tagccgcgtc gtgcatgcct tttatcttcg ggttatggga ccagtgaagg  | 660  |
| ctgagggaag gactgtcctg ggactggaca ggcgggttat gggacctgaa aatactaaca  | 720  |
| atcgatTTTT tttccctttt tttccaggtg ctgcccgaagc acagcctgct gtacgagtac | 780  |
| ttcaccgtgt acaacgagct gaccaaagtg aaatacgtga ccgagggaat gagaaagccc  | 840  |
| gccttttctga gcggcgagca gaaaaaggcc attgtggatc tgctgttcaa gaccaaccgg | 900  |
| aaagtgaccg tgaagcagct gaaagaggac tacttcaaga aaatcgagtg cttcgacagc  | 960  |
| gtggaaatca gcggcgtgga agatcggttc aatgccagcc tgggcacata ccacgacctg  | 1020 |
| ctgaaaatta tcaaggacaa ggacttcttg gacaacgaag agaacgagga catcctggaa  | 1080 |
| gatatcgtgc tgaccctgac actgtttgag gacagagaga tgatcgagga acggctgaaa  | 1140 |
| acatacgccc acctgttcga cgacaaagtg atgaagcaac tgaagcggcg gagatacacc  | 1200 |
| ggctggggca gactgtctcg gaagctgata aacggcatcc gggataagca gtccggcaag  | 1260 |
| accatcctgg acttttctgaa gtccgacggc ttcgccaatc ggaacttcat gcagctgata | 1320 |
| cacgacgaca gcctgacctt taaagaggat atccagaaag cccaggtgtc cggccagggc  | 1380 |
| gatttctctg atgagcacat tgccaacctg gccggctctc ccgccattaa gaagggcatt  | 1440 |
| ctgcagacag tgaagggtgt ggacgagctg gtcaaagtca tgggcagaca caagcccgag  | 1500 |
| aacatcgtga tcgaaatggc cagagagaac cagaccacac agaagggccga gaagaacagc | 1560 |
| cgcgagagaa tgaagcggat cgaagagggc atcaaagagc tgggcagcca gatcctgaaa  | 1620 |
| gaacaccccg tggaaaacac ccagctgcag aacgagaagc tgtacctgta ctacctcaa   | 1680 |
| aacggccggg atatgtatgt ggaccaagag ctggacatca accggctgtc cgactacgat  | 1740 |
| gtggaccata tcgtgccccca gtcttttctg aaagacgact ccatcgacaa caaggtcctg | 1800 |

|             |            |             |            |            |            |      |
|-------------|------------|-------------|------------|------------|------------|------|
| accagaagcg  | acaagaaccg | gggcaagagc  | gataacgtgc | cctccgaaga | ggtcgtgaag | 1860 |
| aagatgaaga  | actactggcg | acagctgctg  | aacgccaagc | tgattaccca | gcggaagttc | 1920 |
| gataacctga  | ccaaggccga | gagaggcggc  | ctgtctgaac | tggataaggc | cggcttcac  | 1980 |
| aagagacagc  | tggtggaaac | ccggcagatc  | accaaacacg | tggcacagat | tctggactcc | 2040 |
| cggatgaaca  | ctaagtacga | cgagaatgac  | aagctgatcc | gggaagtga  | agtgatcacc | 2100 |
| ctgaagtcca  | agctggtgtc | cgatttccgg  | aaggatttcc | agttctacaa | agtgcgag   | 2160 |
| atcaacaact  | accatcacgc | ccacgacgcc  | tacctgaatg | ccgttggttg | aacagccctg | 2220 |
| atcaagaagt  | atcccaagct | ggaatccgag  | ttcgtgtacg | gcgactacaa | ggtgtacgac | 2280 |
| gtgcggaaga  | tgatcgccaa | gagcgagcaa  | gagattggca | aggctaccgc | caagtacttc | 2340 |
| ttctacagca  | acatcatgaa | ctttttcaag  | accgagatta | ccctggccaa | cggcgagatc | 2400 |
| agaaagcggc  | ctctgatcga | gacaaacggc  | gaaaccggcg | agattgtgtg | ggacaagggc | 2460 |
| agagattttg  | ccaccgtgcg | gaaagtgtctg | agcatgcccc | aagtgaatat | cgtgaaaaag | 2520 |
| accgaggtgc  | agacaggcgg | cttcagcaaa  | gagtccattc | tgcccaagag | aaacagcgat | 2580 |
| aagctgatcg  | cccggaagaa | ggactgggac  | cctaagaagt | acggcggctt | cgatagccct | 2640 |
| accgtggcct  | attctgtgct | ggtggtggcc  | aaagtggaaa | agggcaagtc | caagaaactc | 2700 |
| aagagcgtga  | aagagctgct | ggggatcacc  | atcatggaaa | gaagcagctt | cgagaagaat | 2760 |
| cctatcgatt  | tcctcaggcg | caagggttac  | aaagaagtga | aaaaggacct | gatcatcaag | 2820 |
| ctccccaagt  | actccctggt | cgagctggaa  | aatggccgga | agcggatgct | ggcttctgct | 2880 |
| ggcgaactgc  | agaaggga   | cgaactggcc  | ctgcctagca | aatatgtgaa | cttcctgtac | 2940 |
| ctggccagcc  | actatgagaa | gctgaagggc  | agccccgagg | acaatgagca | aaagcagctg | 3000 |
| tttgtggaac  | agcacaagca | ctacctggac  | gagatcatcg | agcagatctc | cgagttctcc | 3060 |
| aagagagtga  | tcctggccga | cgctaactctg | gacaaagtgc | tgtccgccta | caacaagcac | 3120 |
| cgggacaagc  | ctatcagaga | gcaggccgag  | aatatcatcc | acctgtttac | cctgaccaat | 3180 |
| ctggggagccc | ctgccgcctt | caagtacttt  | gacaccacca | tcgaccggaa | gcgctacacc | 3240 |
| agcaccaaag  | aggtgctgga | cgccacactg  | atccaccagt | ctatcaccgg | cctgtacgag | 3300 |
| acacggatcg  | acctgtctca | gctcggaggc  | gatagcggcg | ggagcggcgg | gagcgggggg | 3360 |
| agcactaatc  | tgagcgacat | cattgagaag  | gagactggga | aacagctggt | cattcaggag | 3420 |

|                                                                     |      |
|---------------------------------------------------------------------|------|
| tccatcctga tgctgcctga ggaggtggag gaagtgatcg gcaacaagcc agagtctgac   | 3480 |
| atcctggtgc acaccgccta cgacgagtcc acagatgaga atgtgatgct gctgacctct   | 3540 |
| gacgcccccg agtataagcc ttggggccctg gtcattccagg attctaacgg cgagaataag | 3600 |
| atcaagatgc tgagcggagg atccggagga tctggaggca gcaccaacct gtctgacatc   | 3660 |
| atcgagaagg agacaggcaa gcagctggtc atccaggaga gcatcctgat gctgcccga    | 3720 |
| gaagtcgaag aagtgatcgg aaacaagcct gagagcgata tcctggtcca taccgcctac   | 3780 |
| gacgagagta ccgacgaaaa tgtgatgctg ctgacatccg acgccccaga gtataagccc   | 3840 |
| tgggctctgg tcatccagga ttccaacgga gagaacaaaa tcaaaatgct gtctggcggc   | 3900 |
| tcaaaaagaa ccgccgacgg cagcgaattc gagcccaaga agaaggaggaa agtctaaacc  | 3960 |
| aattcctcac ctgcatctc gatgctttat ttgtgaaatt tgtgatgcta ttgctttatt    | 4020 |
| tgtaaccatt ataagctgca ataaacaagt taacaacaac aattgcattc attttatgtt   | 4080 |
| tcaggttcag ggggaggtgt gggaggtttt ttaaa                              | 4115 |

<210> 165  
 <211> 2973  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| <400> 165                                                          |     |
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60  |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120 |
| atgggtggag tattttacggt aaactgccca cttggcagta catcaagtgt atcatatgcc | 180 |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240 |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300 |
| catggtgatg cggtttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg | 360 |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420 |
| ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt   | 480 |
| acggtgggag gtctatataa gcagagcttg gatgttgctt ttacttctag gcgcgccacc  | 540 |
| atgaaacgga cagccgacgg aagcgagttc gagtcaccaa agaagaagcg gaaagtctct  | 600 |



|            |            |            |            |             |            |      |
|------------|------------|------------|------------|-------------|------------|------|
| gaggtggagt | tttccacga  | gtactggatg | agacatgccc | tgaccctggc  | caagagggca | 660  |
| cgggatgaga | gggaggtgcc | tgtgggagcc | gtgctggtgc | tgaacaatag  | agtgatcggc | 720  |
| gagggctgga | acagagccat | cggcctgcac | gacccaacag | cccatgccga  | aattatggcc | 780  |
| ctgagacagg | gcggcctggt | catgcagaac | tacagactga | ttgacgccac  | cctgtacgtg | 840  |
| acattcgagc | cttgcgtgat | gtgcgccggc | gccatgatcc | actctaggat  | cggccgcgtg | 900  |
| gtgtttggcg | tgaggaactc | aaaaagaggc | gccgcaggct | ccctgatgaa  | cgtgctgaac | 960  |
| taccccgga  | tgaatcaccg | cgtcgaaatt | accgagggaa | tcctggcaga  | tgaatgtgcc | 1020 |
| gccctgctgt | gcgatttcta | tcggatgcct | agacagggtg | tcaatgctca  | gaagaaggcc | 1080 |
| cagagctcca | tcaactccgg | aggatctagc | ggaggctcct | ctggctctga  | gacacctggc | 1140 |
| acaagcgaga | gcgcaacacc | tgaaagcagc | gggggcagca | gcgggggggtc | agacaagaag | 1200 |
| tacagcatcg | gcctggccat | cggcaccaat | tctgttggct | gggccgtgat  | caccgacgag | 1260 |
| tacaaggtgc | ccagcaagaa | attcaaggtg | ctgggcaaca | ccgaccggca  | cagcatcaag | 1320 |
| aagaatctga | tcggcgccct | gctgttcgac | tctggcgaaa | cagccgaagc  | caccagactg | 1380 |
| aagaggacag | ccagacggcg | gtacaccaga | agaaagaacc | ggatctgcta  | cctgcaagag | 1440 |
| atcttcagca | acgagatggc | caaggtggac | gacagcttct | tccaccggct  | ggaagagtcc | 1500 |
| ttcctggtgg | aagaggataa | gaagcacgag | cggcacccca | tcttcggcaa  | catcgtggat | 1560 |
| gaggtggcct | accacgagaa | gtaccccacc | atctaccacc | tgagaaagaa  | actggtggac | 1620 |
| agcaccgaca | aggccgacct | gagactgata | tatctggccc | tggctcacat  | gatcaagttc | 1680 |
| cggggccact | tcctgatcga | gggcgacctg | aatcctgaca | acagcgacgt  | ggacaagctg | 1740 |
| ttcatccagc | tggtgcagac | ctacaaccag | ctgttcgagg | aaaaccccat  | caacgccagc | 1800 |
| ggagtggatg | ccaaggccat | cctgtctgcc | cggctgagca | agagcagacg  | gctggaaaac | 1860 |
| ctgatcgctc | agctgcccgg | cgagaagaag | aatggcctgt | tcggcaacct  | gattgccctg | 1920 |
| agcctgggcc | tgacacctaa | cttcaagagc | aacttcgacc | tggccgagga  | cgccaaactg | 1980 |
| cagctgtcca | aggacaccta | cgacgacgac | ctggacaatc | tgctggccca  | gatcggcgat | 2040 |
| cagtacgccg | acttgtttct | ggccgccaag | aacctgtccg | acgccatcct  | gctgagcgac | 2100 |
| atcctgagag | tgaacaccga | gatcacaaag | gcccctctga | gcgctcttat  | gatcaagaga | 2160 |
| tacgacgagc | accaccagga | tctgaccctg | ctgaaggccc | tcgttagaca  | gcagctgcct | 2220 |

|                                                                       |      |
|-----------------------------------------------------------------------|------|
| gagaagtaca aagagatttt cttcgaccag agcaagaacg gctacgccgg ctacattgat     | 2280 |
| ggcgggagcca gccaaagagga attctacaag ttcacatcaagc ccatcctcga gaagatggac | 2340 |
| ggcaccgagg aactgctggt caagctgaac agagaggacc tgctgcggaa gcagcggacc     | 2400 |
| ttcgacaatg gctctatccc tcaccaaadc cacctgggag agctgcacgc cattctgcgg     | 2460 |
| agacaagagg actttttaccc attcctgaag gacaaccggg aaaagattga gaagatcctg    | 2520 |
| accttcagga tcccctacta cgtgggacca ctggccagag gcaatagcag attcgcctgg     | 2580 |
| atgaccagaa agagcgagga aaccatcaca ccctggaact tcgaggaagt ggtggataag     | 2640 |
| ggcgccagcg ctacgtcctt catcgagcgg atgaccaact tcgataagaa cctgcctaac     | 2700 |
| gagaaggtaa gtattagctc tttctttcca tgggttggcc tcgccgcgtg ggctgaggga     | 2760 |
| aggactgtcc tgggactgga caggcgggtt atgggacctg aagcgataaa aggcatgcac     | 2820 |
| gtttgcggct acgtgcatgc caaaaggagt cgggcttgcc tccgtgcccg actccaaaag     | 2880 |
| acctgctcga ggaggtggac gagcaggtca aaaatccggg taccaataaa atatctttat     | 2940 |
| tttcattaca tctgtgtgtt ggttttttgt gtg                                  | 2973 |

<210> 166

<211> 3560

<212> DNA

<213> Artificial sequence

<220>

<223> Synthetic nucleic acid sequence

<400> 166

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60  |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120 |
| atgggtggag tattttacggt aaactgccca cttggcagta catcaagtgt atcatatgcc | 180 |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240 |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300 |
| catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg  | 360 |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420 |
| ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt   | 480 |
| acggtgggag gtctatataa gcagagcttg gatgttgctt ttaggatttt tgacctgctc  | 540 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| gattgtccac tgcgagcagg tcttttggag tcgggcgagg cggaagcccg actccttttg  | 600  |
| gcatgcacgc tagccgcgtc gtgcatgcct tttatcttcg ggttatggga ccagtgaagg  | 660  |
| ctgaggggaag gactgtcctg ggactggaca ggcgggttat gggacctgaa aatactaaca | 720  |
| atcgatTTTT tttccctttt tttccaggtg ctgcccgaagc acagcctgct gtacgagtac | 780  |
| ttcaccgtgt acaacgagct gaccaaagtg aaatacgtga ccgagggaat gagaaagccc  | 840  |
| gccttttctga gcggcgagca gaaaaaggcc attgtggatc tgctgttcaa gaccaaccgg | 900  |
| aaagtgaccg tgaagcagct gaaagaggac tacttcaaga aaatcgagtg cttcgacagc  | 960  |
| gtggaaatca gcggcgtgga agatcggttc aatgccagcc tgggcacata ccacgacctg  | 1020 |
| ctgaaaatta tcaaggacaa ggacttcctg gacaacgaag agaacgagga catcctggaa  | 1080 |
| gatatcgtgc tgaccctgac actgtttgag gacagagaga tgatcgagga acggctgaaa  | 1140 |
| acatacgccc acctgttcga cgacaaagtg atgaagcaac tgaagcggcg gagatacacc  | 1200 |
| ggctggggca gactgtctcg gaagctgatc aacggcatcc gggataagca gtccggcaag  | 1260 |
| accatcctgg acttttctgaa gtccgacggc ttcgccaatc ggaacttcat gcagctgatc | 1320 |
| cacgacgaca gcctgacctt taaagaggat atccagaaag cccaggtgtc cggccagggc  | 1380 |
| gattctctgc atgagcacat tgccaacctg gccggctctc ccgccattaa gaagggcatt  | 1440 |
| ctgcagacag tgaaggtggt ggacgagctg gtcaaagtca tgggcagaca caagcccgag  | 1500 |
| aacatcgtga tcgaaatggc cagagagaac cagaccacac agaagggccca gaagaacagc | 1560 |
| cgcgagagaa tgaagcggat cgaagagggc atcaaagagc tgggcagcca gatcctgaaa  | 1620 |
| gaacaccccg tggaaaacac ccagctgcag aacgagaagc tgtacctgta ctacctcaa   | 1680 |
| aacggccggg atatgtatgt ggaccaagag ctggacatca accggctgtc cgactacgat  | 1740 |
| gtggaccata tcgtgcccc a gtcttttctg aaagacgact ccatcgacaa caaggtcctg | 1800 |
| accagaagcg acaagaaccg gggcaagagc gataacgtgc cctccgaaga ggtcgtgaag  | 1860 |
| aagatgaaga actactggcg acagctgctg aacgccaagc tgattacca gcggaagttc   | 1920 |
| gataacctga ccaaggccga gagaggcggc ctgtctgaac tggataaggc cggcttcac   | 1980 |
| aagagacagc tgggtggaac ccggcagatc accaaacacg tggcacagat tctggactcc  | 2040 |
| cggatgaaca ctaagtacga cgagaatgac aagctgatcc gggaagtga agtgatcacc   | 2100 |
| ctgaagtcca agctggtgtc cgatttccgg aaggatttcc agttctacaa agtgcgcgag  | 2160 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| atcaacaact accatcacgc ccacgacgcc tacctgaatg cgttggttg aacagccctg   | 2220 |
| atcaagaagt atcccaagct ggaatccgag ttcgtgtacg gcgactacaa ggtgtacgac  | 2280 |
| gtgcggaaga tgatcgccaa gagcgagcaa gagattggca aggctaccgc caagtacttc  | 2340 |
| ttctacagca acatcatgaa ctttttcaag accgagatta ccctggccaa cggcgagatc  | 2400 |
| agaaagcggc ctctgatcga gacaaacggc gaaaccggcg agattgtgtg ggacaagggc  | 2460 |
| agagattttg ccaccgtgcg gaaagtgtctg agcatgcccc aagtgaatat cgtgaaaaag | 2520 |
| accgaggtgc agacaggcgg cttcagcaaa gagtccattc tgcccaagag aaacagcgat  | 2580 |
| aagctgatcg cccggaagaa ggactgggac cctaagaagt acggcggctt cgatagccct  | 2640 |
| accgtggcct attctgtgct ggtggtggcc aaagtggaaa agggcaagtc caagaaactc  | 2700 |
| aagagcgtga aagagctgct ggggatcacc atcatggaaa gaagcagctt cgagaagaat  | 2760 |
| cctatcgatt tcctcgaggc caagggtac aaagaagtga aaaaggacct gatcatcaag   | 2820 |
| ctccccaagt actccctgtt cgagctggaa aatggccgga agcggatgct ggcttctgct  | 2880 |
| ggcgaactgc agaagggaaa cgaactggcc ctgcctagca aatatgtgaa cttcctgtac  | 2940 |
| ctggccagcc actatgagaa gctgaagggc agccccgagg acaatgagca aaagcagctg  | 3000 |
| tttgtggaac agcacaagca ctacctggac gagatcatcg agcagatctc cgagttctcc  | 3060 |
| aagagagtga tcctggccga cgctaattctg gacaaagtgc tgtccgccta caacaagcac | 3120 |
| cgggacaagc ctatcagaga gcaggccgag aatatcatcc acctgtttac cctgaccaat  | 3180 |
| ctgggagccc ctgccgcctt caagtacttt gacaccacca tcgaccggaa gcgctacacc  | 3240 |
| agcaccaaag aggtgctgga cgccacactg atccaccagt ctatcaccgg cctgtacgag  | 3300 |
| acacggatcg acctgtctca gctcggaggc gattctggcg gctcaaaaag aaccgccgac  | 3360 |
| ggcagcgaat tcgagcccaa gaagaagagg aaagtctaag gtaccaattc ctcacctgcg  | 3420 |
| atctcgatgc tttatttgtg aaatttgtga tgctattgct ttatttgtaa ccattataag  | 3480 |
| ctgcaataaa caagttaaca acaacaattg cattcatttt atgtttcagg ttcaggggga  | 3540 |
| ggtgtgggag gttttttaa                                               | 3560 |

<210> 167  
 <211> 112  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 167  
 gatttttgac ctgctcgatt gtccactgcg agcaggtcctt ttggagtcgg gcgaggcgga 60  
 agccccgactc cttttggcat gcacgctagc cgcgtcgtgc atgcctttta tc 112

<210> 168  
 <211> 13  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 168  
 gggttatggg acc 13

<210> 169  
 <211> 24  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 169  
 ggctgaggga aggactgtcc tggg 24

<210> 170  
 <211> 24  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 170  
 ctctttcttt ccatggggttg gcct 24

<210> 171  
 <211> 4463  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<220>  
<221> misc\_feature  
<222> (4225)..(4294)  
<223> n is a, c, g, or t

<400> 171  
cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt 60  
gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca 120  
atgggtggag tattttacggt aaactgccc a ttggcagta catcaagtgt atcatatgcc 180  
aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta 240  
catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac 300  
catggtgatg cggttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg 360  
atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg 420  
ggactttcca aaatgtcgta acaactccgc ccattgacg caaatgggcg gtaggcgtgt 480  
acggtgggag gtctatataa gcagagcttg gatgttgctt ttacttctag gcgcgccgcc 540  
accatggccc caaagaagaa gcggaaggtc ggtatccacg gagtcccagc agccaagcgg 600  
aactacatcc tgggcctgga catcggcatc accagcgtgg gctacggcat catcgactac 660  
gagacacggg acgtgatcga tgccggcgtg cggctgttca aagaggccaa cgtggaaaac 720  
aacgagggca ggcggagcaa gagaggcgcc agaaggctga agcggcggag gcggcataga 780  
atccagagag tgaagaagct gctgttcgac tacaacctgc tgaccgacca cagcgagctg 840  
agcggcatca acccctacga ggccagagtg aagggcctga gccagaagct gagcgaggaa 900  
gagttctctg ccgccctgct gcacctggcc aagagaagag gcgtgcacaa cgtgaacgag 960  
gtggaagagg acaccggcaa cgagctgtcc accaaagagc agatcagccg gaacagcaag 1020  
gccctggaag agaaatacgt ggccgaactg cagctggaac ggctgaagaa agacggcgaa 1080  
gtgcggggca gcatcaacag attcaagacc agcgactacg tgaaagaagc caaacagctg 1140  
ctgaagggtc agaaggccta ccaccagctg gaccagagct tcatcgacac ctacatcgac 1200  
ctgctggaaa cccggcggac ctactatgag ggacctggcg agggcagccc cttcggctgg 1260  
aaggacatca aagaatggta cgagatgctg atgggccact gcacctactt ccccgaggaa 1320  
ctgcggagcg tgaagtacgc ctacaacgcc gacctgtaca acgccctgaa cgacctgaac 1380  
aatctcgtga tcaccaggga cgagaacgag aagctggaat attacgagaa gttccagatc 1440

|             |             |            |             |             |             |      |
|-------------|-------------|------------|-------------|-------------|-------------|------|
| atcgagaacg  | tgttcaagca  | gaagaagaag | cccaccctga  | agcagatcgc  | caaagaaatc  | 1500 |
| ctcgtgaacg  | aagaggatat  | taagggctac | agagtgaacca | gcaccggcaa  | gcccgagttc  | 1560 |
| accaacctga  | aggtgtacca  | cgacatcaag | gacattaccg  | cccggaaaga  | gattattgag  | 1620 |
| aacgccgagc  | tgctggatca  | gattgccaag | atcctgaacca | tctaccagag  | cagcgaggac  | 1680 |
| atccaggaag  | aactgaccaa  | tctgaactcc | gagctgaccc  | aggaagagat  | cgagcagatc  | 1740 |
| tctaattctga | agggtctatac | cggcacccac | aacctgagcc  | tgaaggccat  | caacctgatc  | 1800 |
| ctggacgagc  | tgtggcacac  | caacgacaac | cagatcgcta  | tcttcaaccg  | gctgaagctg  | 1860 |
| gtgccaaga   | aggtggacct  | gtcccagcag | aaagagatcc  | ccaccaccct  | ggtggacgac  | 1920 |
| ttcatcctga  | gccccgtcgt  | gaagagaagc | ttcatccaga  | gcatcaaagt  | gatcaacgcc  | 1980 |
| atcatcaaga  | agtacggcct  | gccaacgac  | atcattatcg  | agctggccccg | cgagaagaac  | 2040 |
| tccaaggacg  | cccagaaaat  | gatcaacgag | atgcagaagc  | ggaaccggca  | gaccaacgag  | 2100 |
| cggatcgagg  | aaatcatccg  | gaccaccggc | aaagagaacg  | ccaagtacct  | gatcgagaag  | 2160 |
| atcaagctgc  | acgacatgca  | ggaaggcaag | tgcctgtaca  | gcctggaagc  | catccctctg  | 2220 |
| gaagatctgc  | tgaacaaccc  | cttcaactat | gaggtggacc  | acatcatccc  | cagaagcgtg  | 2280 |
| tccttcgaca  | acagcttcaa  | caacaaggtg | ctcgtgaagc  | aggaagaaaa  | cagcaagaag  | 2340 |
| ggcaaccgga  | ccccattcca  | gtacctgagc | agcagcgaca  | gcaagatcag  | ctacgaaacc  | 2400 |
| ttcaagaagc  | acatcctgaa  | tctggccaag | ggcaagggca  | gaatcagcaa  | gaccaagaaa  | 2460 |
| gagtatctgc  | tggaagaacg  | ggacatcaac | aggttctccg  | tgcagaaaga  | cttcatcaac  | 2520 |
| cggaacctgg  | tggataccag  | atacgccacc | agaggcctga  | tgaacctgct  | gcggagctac  | 2580 |
| ttcagagtga  | acaacctgga  | cgtgaaagtg | aagtccatca  | atggcggctt  | caccagcttt  | 2640 |
| ctgcggcgga  | agtggaagtt  | taagaaagag | cggaacaagg  | ggtacaagca  | ccacgccgag  | 2700 |
| gacgccctga  | tcattgccaa  | cgccgatttc | atcttcaaag  | agtggaagaa  | actggacaag  | 2760 |
| gccaaaaaag  | tgatggaaaa  | ccagatgttc | gaggaaaagc  | aggccgagag  | catgccccgag | 2820 |
| atcgaaaccg  | agcaggagta  | caaagagatc | ttcatcacc   | cccaccagat  | caagcacatt  | 2880 |
| aaggacttca  | aggactacaa  | gtacagccac | cgggtggaca  | agaagcctaa  | tagagagctg  | 2940 |
| attaacgaca  | ccctgtactc  | cacccggaag | gacgacaagg  | gcaacaccct  | gatcgtgaac  | 3000 |
| aatctgaacg  | gcctgtacga  | caaggacaat | gacaagctga  | aaaagctgat  | caacaagagc  | 3060 |

|                                                                   |      |
|-------------------------------------------------------------------|------|
| cccgaaaagc tgctgatgta ccaccacgac cccagacct accagaaact gaagctgatt  | 3120 |
| atggaacagt acggcgacga gaagaatccc ctgtacaagt actacgagga aaccgggaac | 3180 |
| tacctgacca agtactccaa aaaggacaac ggccccgtga tcaagaagat taagtattac | 3240 |
| ggcaacaaac tgaacgccc tctggacatc accgacgact accccaacag cagaaacaag  | 3300 |
| gtcgtgaagc tgtccctgaa gccctacaga ttcgacgtgt acctggacaa tggcgtgtac | 3360 |
| aagttcgtga ccgtgaagaa tctggatgtg atcaaaaaag aaaactacta cgaagtgaat | 3420 |
| agcaagtgct atgaggaagc taagaagctg aagaagatca gcaaccaggc cgagtttatc | 3480 |
| gcctccttct acaacaacga tctgatcaag atcaacggcg agctgtatag agtgatcggc | 3540 |
| gtgaacaacg acctgctgaa cgggatcgaa gtgaacatga tcgacatcac ctaccgag   | 3600 |
| tacctggaaa acatgaacga caagaggccc cccaggatca ttaagacaat cgccggaagc | 3660 |
| ggagctacta acttcagcct gctgaagcag gctggagacg tggaggagaa ccctggacct | 3720 |
| aggcgcgccg ccaccatggt gagcaagggc gaggagctgt tcaccggggt ggtgcccac  | 3780 |
| ctggtcgagc tggacggcga cgtaaaggc cacaagttca gcgtgtccgg cgagggcgag  | 3840 |
| ggcgatgcca cctacggcaa gctgaccctg aagttcatct gcaccaccgg caagctgccc | 3900 |
| gtgccctggc ccaccctcgt gaccaccttc ggctacggcc tgatgtgctt cgcccgctac | 3960 |
| cccgaccaca tgaagcagca cgacttcttc aagtccgcca tgcccgaagg ctacgtccag | 4020 |
| gagcgcacca tcttcttcaa ggacgacggc aactacaaga cccgcgccga ggtgaagttc | 4080 |
| gagggcgaca ccctggtgaa ccgcatcgag ctgaagggca tcgacttcaa ggaggacggc | 4140 |
| aacatcctgg ggcacaagct ggagtacaac tacaacagcc acaacgtcta tatcatggcc | 4200 |
| gacaagcaga agaacggcat caagnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn | 4260 |
| nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnngataaa aggcatgcac gtttgaggct | 4320 |
| acgtgcatgc caaaaggagt cgggcttgcc tccgtgcccg actccaaaag acctgctcga | 4380 |
| ggaggtggac gagcaggtca aaaatccggg taccaataaa atatctttat tttcattaca | 4440 |
| tctgtgtgtt ggttttttgt gtg                                         | 4463 |

<210> 172  
 <211> 3467  
 <212> DNA  
 <213> Artificial sequence



<220>

<223> Synthetic nucleic acid sequence

<400> 172

|                                                                    |      |
|--------------------------------------------------------------------|------|
| cgttacataa cttacggtaa atggcccgcc tggctgaccg cccaacgacc cccgcccatt  | 60   |
| gacgtcaata atgacgtatg ttcccatagt aacgccaata gggactttcc attgacgtca  | 120  |
| atgggtggag tattttacggt aaactgccc cttggcagta catcaagtgt atcatatgcc  | 180  |
| aagtacgccc cctattgacg tcaatgacgg taaatggccc gcctggcatt atgcccagta  | 240  |
| catgacctta tgggactttc ctacttggca gtacatctac gtattagtca tcgctattac  | 300  |
| catggtgatg cggtttttggc agtacatcaa tgggcgtgga tagcggtttg actcacgggg | 360  |
| atttccaagt ctccacccca ttgacgtcaa tgggagtttg ttttggcacc aaaatcaacg  | 420  |
| ggactttcca aaatgtcgt acaactccgc ccattgacg caaatgggcg gtaggcgtgt    | 480  |
| acggtgggag gtctatataa gcagagcttg gatgttgcct ttaggatttt tgacctgctc  | 540  |
| gattgtccac tgcgagcagg tcttttggag tcgggcgagg cggaagcccg actccttttg  | 600  |
| gcatgcacgc tagccgcgtc gtgcatgcct tttatcttcg ggttatggga ccagtgaagg  | 660  |
| ctgagggaag gactgtcctg ggactggaca ggcgggttat gggacctgaa aatactaaca  | 720  |
| atcgattttt tttccctttt tttccaggtg gacagaccca gagtggatat cgagtgtgct  | 780  |
| ggcaaagggg tgcagagcag cctgatccat aactacaaga agaaccctaa cttcaacacc  | 840  |
| ctgggtcaagt ggttcgaagt ggatctgccc gagaacgaac tgctgcaccc acctctgaac | 900  |
| atcagagtgg tggactgcag agccttcggc agatacacc tcgtgggatc tcacgccgtg   | 960  |
| tctagcctga gaagattcat ctacagacct ccagacagaa gcgcccctaa ctggaacaca  | 1020 |
| acaggcgagg tgggtggtgtc catggaaccc gaggaacccg tgaagaaact ggaaaccatg | 1080 |
| gtcaagctgg acgccacctc cgatgctgtc gtgaaagtgg acgtggccga ggacgagaaa  | 1140 |
| gagcgcaaga agaagaaaaa gaaggggccc agcgaggaac ctgaagagga agaacctgac  | 1200 |
| gagagcatgc tggactggtg gtccaagtac ttcgcctcca tcgacacaat gaaggaacag  | 1260 |
| ctgagacagc acgagacaag cggcaccgac ctcgaagaga aagaagagat ggaatccgcc  | 1320 |
| gaaggactga agggccctat gaagtccaaa gagaagtcta gggccgcca agaagagaaa   | 1380 |
| aaaaagaaga accagtctcc tggaccaggc cagggatctg aggctcccga aaagaaaaag  | 1440 |
| gccaagatcg acgagctgaa ggtgtacccc aaagagctgg aaagcgagtt cgacagcttc  | 1500 |

|                                                                    |      |
|--------------------------------------------------------------------|------|
| gaggactggc tgcacacctt caatctgctg agaggaaaga caggcgacga cgaggatggc  | 1560 |
| agcactgaag aagagagaat cgtcggcaga ttcaagggca gcctgtgctg gtacaaggtg  | 1620 |
| ccactgcctg aggacgtgtc cagagaggct ggctacgata ctacctacgg catgttccaa  | 1680 |
| ggcatcccta gcaacgaccc catcaatgtg ctctgtcgga tctatgtcgt gcggggccact | 1740 |
| gatctgcatc ccgccgatat caacggcaag gcagacccct atatcgctat caagctgggg  | 1800 |
| aaaaccgaca tcagggacaa agagaactac atcagcaagc agctgaaccc cgtgttcggc  | 1860 |
| aagagcttcg acatcgaggc tagcttcccc atggaatcca tgctgaccgt ggccgtgtac  | 1920 |
| gactgggatc tcgtgggaac agacgacctg atcggagaga caaagattga cctggaaaac  | 1980 |
| cggttctact ccaagcaccc ggccacctgt ggaatcgccc agacctactc tatccacggc  | 2040 |
| tacaacatct ggcgggaccc catgaagcct agccagatcc tgaccaggct gtgcaaagaa  | 2100 |
| ggcaaggtcg acggccctca ctttggacct cacggccggg tcagagtggc caacagagt   | 2160 |
| ttcacaggcc cctccgagat cgaggatgag aacggccaga gaaagcccac cgatgagcat  | 2220 |
| gtggctctga gcgctctgag acactgggaa gatatcccta gagtgggctg cagactggtg  | 2280 |
| cccgagcacg tggaacaag acccctgctg aaccagaca agcccggaat cgaacagggc    | 2340 |
| agactcgaac tgtgggtcga catgttcctt atggacatgc ccgcacctgg cacaccactg  | 2400 |
| gacatcagcc ctaggaagcc caagaaatac gagctgcgcg tgatcgtgtg gaacaccgac  | 2460 |
| gaagtgggtg tggaagatga cgacttcttc accggcgaaa agtccagcga catcttcgtc  | 2520 |
| agaggatggc tgaagggaca gcaagaggat aagcaggaca ccgacgtgca ctaccacagc  | 2580 |
| cttacaggcg aaggcaactt taactggcgc tacctgtttc ctttcgacta cctggccgcc  | 2640 |
| gaagagaaga tcgtgatgtc caagaaagaa tctatgttca gctgggacga gacagagtac  | 2700 |
| aagatccccg ccagactgac cctgcagatc tgggatgccg atcacttcag cgccgacgac  | 2760 |
| tttctgggag ccatcgagct ggacctgaat agattcccca gaggcgcca gaccgccaag   | 2820 |
| cagtgcacaa tggaatggc cactggcgag gtcgacgtgc cactggtgtc tatcttcaag   | 2880 |
| cagaagcgcg tcaaaggctg gtggcccctg ctggctagaa acgagaacga cgagttcgag  | 2940 |
| ctgaccggaa aggtggaagc cgagctgcat ctgctgacag ctgaagaggc cgagaagaat  | 3000 |
| cctgtgggcc tcgctaggaa tgagcccgat cctctggaaa agcccaacag acccgatacc  | 3060 |
| gccttcgtgt ggtttctgaa cccactgaag tccatcaagt acctgatctg taccgggtac  | 3120 |

|                                                                   |      |
|-------------------------------------------------------------------|------|
| aagtggctga ttatcaagat cgtgctggcc ctgctggggc tgctgatgct tgctctgttc | 3180 |
| ctgtactccc tgcctggcta tatggtcaag aagctgctgg gcgccggcgc tcgggctgac | 3240 |
| tacaaagacc atgacggtga ttataaagat catgacatcg actataagga tgacgatgac | 3300 |
| aaatgaggta ccaattcctc acctgcatc tcgatgcttt atttgtgaaa tttgtgatgc  | 3360 |
| tattgcttta tttgtaacca ttataagctg caataaaca gttaacaaca acaattgcat  | 3420 |
| tcattttatg tttcaggttc agggggaggt gtgggaggtt ttttaaa               | 3467 |

<210> 173  
 <211> 33  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                      |    |
|--------------------------------------|----|
| <400> 173                            |    |
| gtaagtattg ctttcatttt tgtctttttt taa | 33 |

<210> 174  
 <211> 30  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                                  |    |
|----------------------------------|----|
| <400> 174                        |    |
| gtaagttctt gctttgttca aactgtctat | 30 |

<210> 175  
 <211> 27  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

|                               |    |
|-------------------------------|----|
| <400> 175                     |    |
| gtaagtattc ttttgttctt cactcat | 27 |

<210> 176  
 <211> 32  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 176  
 gtaagtattt ttttactcct catttttact cc 32

<210> 177  
 <211> 36  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 177  
 gtaagtattt ttttacgggtt atattctcct ttcccc 36

<210> 178  
 <211> 28  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 178  
 gtaagtattt tctgttggtt attttcag 28

<210> 179  
 <211> 39  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 179  
 gtaagtattg gggttgatta tgtgtgggac ggtgtaagg 39

<210> 180  
 <211> 35  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 180  
 gtaagtattt cctctttctt tccatgggtt ggcct 35

<210> 181  
 <211> 35  
 <212> DNA  
 <213> Artificial sequence  
  
 <220>  
 <223> Synthetic nucleic acid sequence  
  
 <400> 181  
 gtaagtatta ccagagattc gtagacctgc ttgac 35

<210> 182  
 <211> 39  
 <212> DNA  
 <213> Artificial sequence  
  
 <220>  
 <223> Synthetic nucleic acid sequence  
  
 <400> 182  
 tggggctggg cagagggttg aggggagagg gtcctgggg 39

<210> 183  
 <211> 28  
 <212> DNA  
 <213> Artificial sequence  
  
 <220>  
 <223> Synthetic nucleic acid sequence  
  
 <400> 183  
 tcatgggtgg gttcattggg tgggttca 28

<210> 184  
 <211> 23  
 <212> DNA  
 <213> Artificial sequence  
  
 <220>  
 <223> Synthetic nucleic acid sequence  
  
 <400> 184  
 tagggcgcag tagtccaggg ttt 23

<210> 185  
 <211> 30  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 185  
 ttctctgtgg ggtggcattc tctgctctct 30

<210> 186  
 <211> 29  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 186  
 gggttatggg acctcaggga taagggacc 29

<210> 187  
 <211> 15  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 187  
 cggggatggg ggtca 15

<210> 188  
 <211> 23  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 188  
 tggggggagg tcatgggggg agg 23

<210> 189  
 <211> 24  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 189  
 gttggtgggt tcatgttggt gggt 24

<210> 190  
<211> 29  
<212> DNA  
<213> Artificial sequence  
  
<220>  
<223> Synthetic nucleic acid sequence  
  
<400> 190  
gggtttcggg ttttcaggtg gtcgttggt 29

<210> 191  
<211> 29  
<212> DNA  
<213> Artificial sequence  
  
<220>  
<223> Synthetic nucleic acid sequence  
  
<400> 191  
ggtggtcgtt gggtcatttg ggctattgg 29

<210> 192  
<211> 29  
<212> DNA  
<213> Artificial sequence  
  
<220>  
<223> Synthetic nucleic acid sequence  
  
<400> 192  
tttgggctat tggtaagg ggcgagggg 29

<210> 193  
<211> 29  
<212> DNA  
<213> Artificial sequence  
  
<220>  
<223> Synthetic nucleic acid sequence  
  
<400> 193  
agggggcgag gggtcaggta ttcggtatt 29

<210> 194  
<211> 29  
<212> DNA  
<213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 194  
 ggtattcggg atttcaagggt aacaggtaa 29

<210> 195  
 <211> 29  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 195  
 aggtaacagg taatcagggt ttcggggttt 29

<210> 196  
 <211> 29  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 196  
 tcttactttt gtaaacttta tggtttgtg 29

<210> 197  
 <211> 28  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 197  
 cacgtattct cgggtacggac gttacaga 28

<210> 198  
 <211> 13  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 198  
 taagctggta tcc 13



<210> 199  
<211> 34  
<212> DNA  
<213> Artificial sequence  
  
<220>  
<223> Synthetic nucleic acid sequence  
  
<400> 199  
cactaactct ttttcccccc tttttttttt acag 34

<210> 200  
<211> 36  
<212> DNA  
<213> Artificial sequence  
  
<220>  
<223> Synthetic nucleic acid sequence  
  
<400> 200  
tactaactct ttcttttttc ctttccttct tcacag 36

<210> 201  
<211> 43  
<212> DNA  
<213> Artificial sequence  
  
<220>  
<223> Synthetic nucleic acid sequence  
  
<400> 201  
cactaactct gtcatactta tcctgtcct tttttttcca cag 43

<210> 202  
<211> 45  
<212> DNA  
<213> Artificial sequence  
  
<220>  
<223> Synthetic nucleic acid sequence  
  
<400> 202  
cactaactct ctttcttttt cttccctcct ctcccccaac tgcag 45

<210> 203  
<211> 38  
<212> DNA  
<213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 203  
 cactaactct tttttttttt tttttttttt tacagcag 38

<210> 204  
 <211> 13  
 <212> DNA  
 <213> Artificial sequence

<220>  
 <223> Synthetic nucleic acid sequence

<400> 204  
 taagctggta tcc 13

<210> 205  
 <211> 8  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> branch point sequence

<400> 205  
 tactaaca 8

<210> 206  
 <211> 49  
 <212> DNA  
 <213> Artificial Sequence

<220>  
 <223> polyadenylation signal

<400> 206  
 aataaaatat ctttattttc attacatctg tgtgttggtt ttttgtgtg 49