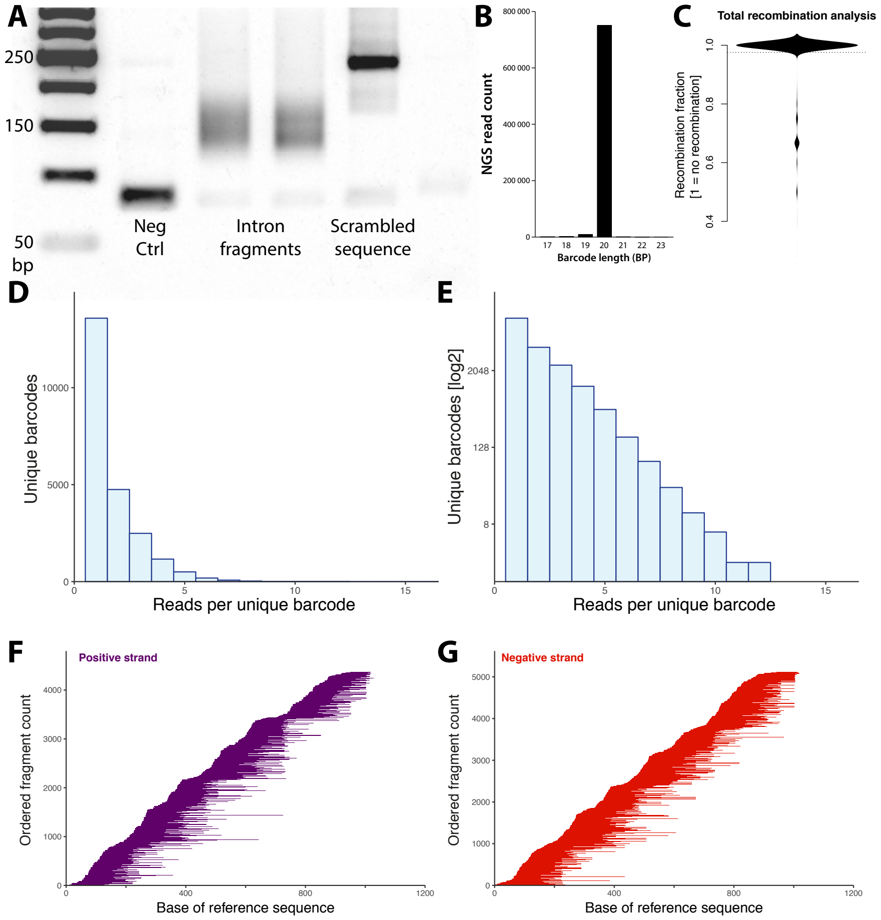
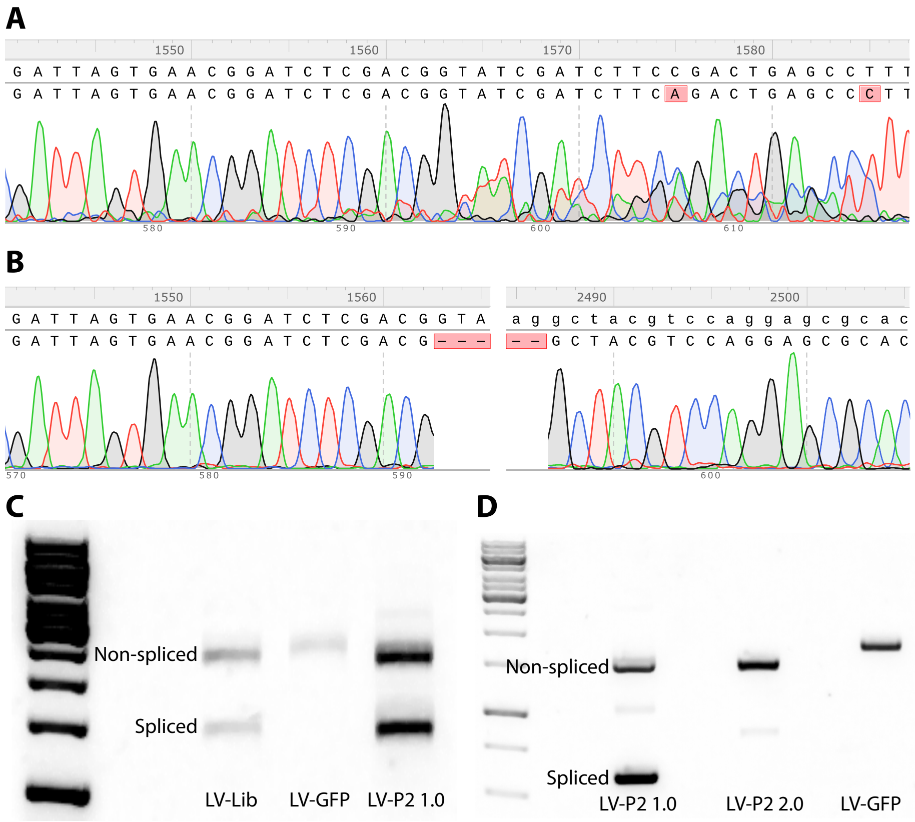
Supplemental figure S1



Supplemental figure S1 | Analysis of splice acceptor plasmid library, related to Figure 1

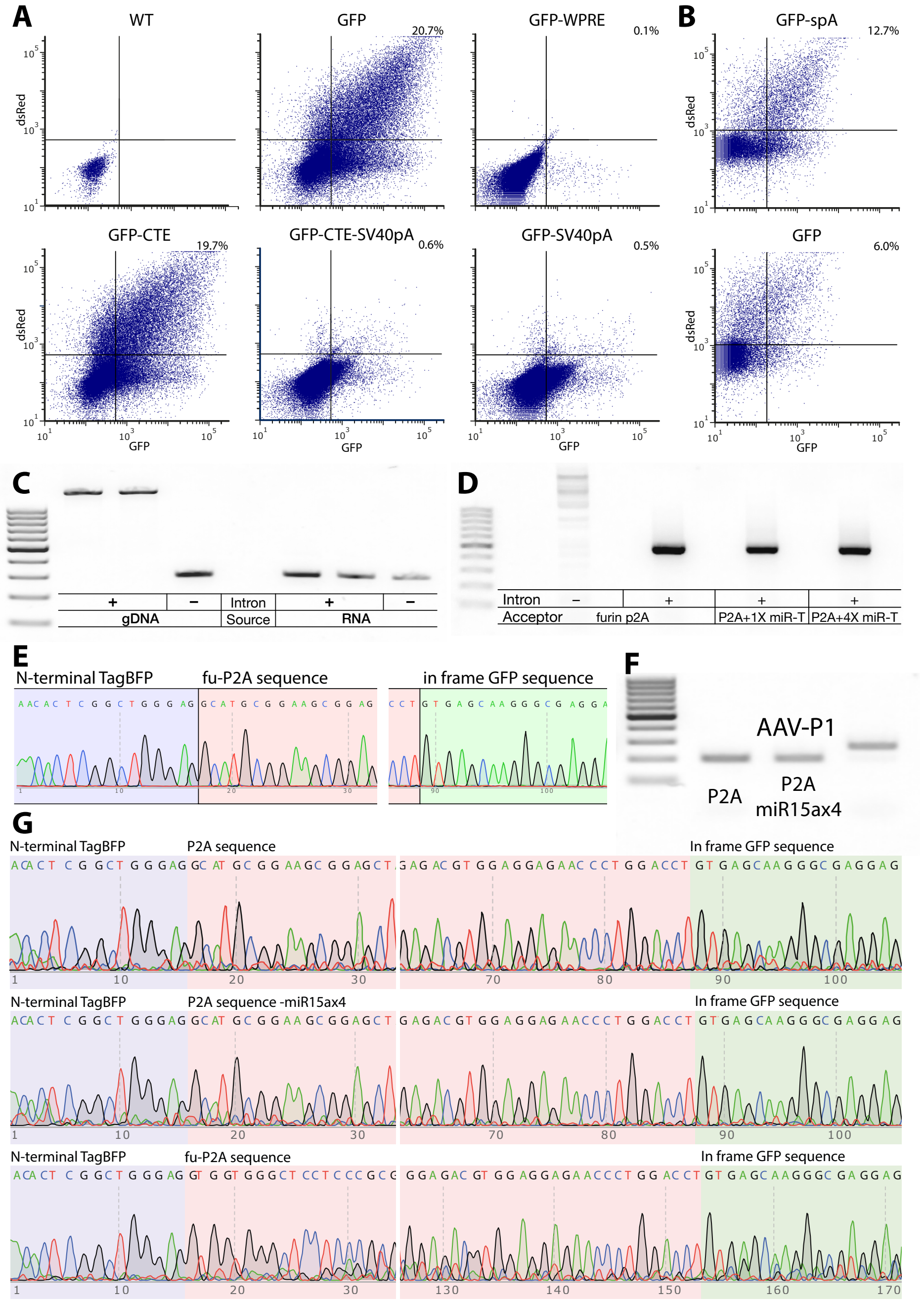
(A) PCR showing insertion and size distribution of empty vector (Neg ctrl), intron fragment and scrambled sequenced in cloning vector. (B) Length distribution of molecular barcodes in final LV-vector as assessed by Ion torrent sequencing. The barcode size was confirmed to be 20 nt for the majority of clones. (C) Bean plot showing the fraction of sequences exposed through recombination between plasmids i.e., where the same barcode is linked to multiple intron fragments. One equals to all fragments being the same for each specific barcode. As the input sequences were generated using a PCR-free protocol, this library displayed perfect coherence between barcode identity and Synapsin I fragment inserted as a binding domain, meaning that one barcode only linked to one fragment (see (Davidsson et al. 2016) for additional information). This is a key requirement as the re-use of a barcode would result in an ambiguous readout from the mRNA samples. (D-E) Evaluation of the diversity and distribution of fragments in the final library represented in a linear (D) and log2 (E) scale showing that the distribution follows closely a Poisson distribution with a slight inflation at the singlet reads. (F-G) Distribution of all unique intron fragments recovered from the plasmid library aligned to the original complete intron sequence and ordered by the 5' start base divided up by insertion orientation into the LV plasmid either in the positive strand orientation (F) or the negative strand orientation (G).

Supplemental figure S2



Supplemental figure S2 | Removal of aberrant cis-splicing in lentiviral vectors, related to Figure 4
Aberrant cis-splicing in the original lentiviral construct (LV 1.0) was observed and confirmed by PCR reaction and Sanger sequencing (A-C). In the LV sequence between the 5' LTR and the splice acceptor there are 17 putative 5' SS as this sequence is only conclusively defined by a 4bp consensus sequence (RG|AT). Through Sanger sequencing of the shorter band from the integrated P2 splice acceptor, we found that this is indeed the case and that the vast majority originates from splicing at the 5'SS positioned 1564bp from the 5'LTR. The exact location of the deletion was identified using Sanger sequencing of the amplicons (A-B) for the correctly sized LV-GFP (A) and the short band from LV-P2 1.0 (B) in (C). The LV-P2 short band displayed a deletion from base 1565 of the LV genome (counting from the left LTR) to the 3' splice site of the splice acceptor. (C) Plasmids containing splice acceptor (LV-Lib and LV-P2 1.0) had parts spliced out during LV production. LV-GFP, without a splice acceptor, did not show cis-splicing during LV production. (D) To circumvent this aberrant splicing, we generated a novel LV backbone with a single-nucleotide mutation in this 5'SS consensus sequence, i.e., G1564A. After a mutation of the *de novo* 5' splice site in the lentiviral plasmid (named LV-P2 2.0), LV-P2 2.0 shows no sign of cis-splicing compared to LV-P1 1.0 that has the splice site intact. LV-GFP, without splice acceptor was used as control in PCR.

Supplemental figure S3



Supplemental figure S3 | Generation and validation of a bi-directional lentiviral construct and splice acceptors expressing full length GFP, related to Figure 5

(A-B) For termination of transcription and stabilization of mRNA of the gene expressed in trans, we assessed the WPRE, sv40pA, CTE and a synthetic pA (spA) sequence. While all constructs express both transgenes very well when delivered to cells using transient transfection, the WPRE and sv40pA sequences efficiently disrupted the production of the LV vectors (A). Both the CTE and the spA sequences however resulted in functional LV vectors which expressed both GFP and dsRed2 at well correlated levels (A). However, only the spA sequence significantly enhanced expression levels of GFP compared to a vector without 3'UTR/transcript termination sequence (B) and thus this was the vector design utilized going forward. (C) The TagBFP (advantageous over iRFP in a cell culture setting in that it can be viewed with a DAPI filter) was chosen for the insertion of the Synapsin I intron to generate a novel splice donor vector. Similar to eGFP, we inserted the intron at base 423 of TagBFP at an AG|GC sequence. This position is also in reading frame meaning that a novel coding sequence inserted through trans-splicing can be expressed without added sequences. To assess that this intron placement is still highly functional we produced LV vectors CMV-TagBFP|PGK-iRFP and CMV-tagBF[intron]P|PGK-iRFP and stably transduced HEK293T cells. (C) PCR of correct cis-splicing of

TagBFP[+intron]. Left-PCR on genomic DNA where TagBFP[+intron] shows larger bands corresponding to retention of the Synapsin I intron, compared to TagBFP[-intron]. Right- PCR on cDNA from extracted RNA. TagBFP[+intron] and TagBFP[-intron] show the same size after intron is spliced away in TagBFP[+intron]. To confirm this, we Sanger sequenced the cDNA derived amplicon from the intron containing TagBFP and found this to be splicing at the original AG|GC junction resulting in a fully functional TagBFP (data not shown).

The correct trans-splicing between TagBFP and GFP would result in a significant fraction of the TagBFP amino poly-peptide being fused to the N-terminal of the GFP. Therefore, we explored three alternative approaches to ensure a functional GFP protein after trans-spicing while ensuring lack of protein translation from the splice acceptor in the absence of the splice donor. We used either a previously published flexible linker (GS 15) (Argos 1990), or a ribosome skipping sequence (P2A) to separate the poly-peptides of TagBFP and GFP at the ribosome (Ryan et al. 1991) (Figure 5A). In a third construct, we added a furin cleavage site upstream of the P2A sequence (fu-P2A) to cleave any TagBFP-GFP fusion protein escaping the ribosome skipping at the Golgi apparatus (Thomas 2002). In the first assessment of the GS15 sequence it was found that this sequence induced a second aberrant 3' splice-site and thus resulted in a mRNA with the GFP out of reading frame. Therefore, this linker was excluded from further analysis. RT-PCR of double transfection with three different splice acceptors (fu-P2A and P2A with 1 or 4 miR targets) on stable cells lines expressing TagBFP[+intron] or TagBFP[-intron]. Only transfection on TagBFP[+intron] gave rise to trans-splicing between splice donor and splice acceptor. Using the TagBFP splice donor, and the full length GFP splice acceptor under control of the positively selected P1 binding domain we found that both linking designs P2A and fu-P2A worked well with no major difference in efficacy, and with no indications (assessed using gel electrophoresis and Sanger peak analysis) of multiple splice variants (D-E).

To remove any splice acceptor mRNA escaping trans-splicing or being expressed in cells lacking the target intron, we evaluated if a microRNA target (miR-T) site can be inserted into the splice acceptor intronic sequence, without affecting the trans-splicing efficacy. The rationale behind this is that the spliceosome is active at mRNA transport out of the nucleus. The miR induced digestion of mRNAs on the other hand depends on the Dicer protein, which has been shown to be an exclusively cytoplasmic protein (Much et al. 2016). In the on-target setting, the endogenous miR (if localized in the nucleus) may bind to the splice acceptor, but would not be able to digest the RNA before the trans-splicing occurs. If there is no trans-splicing, the splice acceptors intronic sequence would be included in the mRNA exported from the nucleus, and then digested by Dicer. The miR-T chosen for this was designed towards the miR15a, as this has been shown to be very broadly expressed (Wang et al. 2014). Using the double-transient transfection in HeLa cells, we found that inserting one or four copies of the miR15a target did not affect the efficacy of the trans-splicing on the mRNA level (D) (F-G) Sanger sequencing (G) and PCR (F) on trans-spliced mRNA showing correctly spliced mRNA originating from TagBFP[+intron], and LV-derived splice acceptor LV-P1 2.0 expressing full length GFP, and either P2A, P2A+mir15a targets or fu-P2A.

Supplemental references

Argos P. 1990. An investigation of oligopeptides linking domains in protein tertiary structures and possible candidates for general gene fusion. J Mol Biol 211: 943-958.

Davidsson M, Diaz-Fernandez P, Schwich OD, Torroba M, Wang G, Bjorklund T. 2016. A novel process of viral vector barcoding and library preparation enables high-diversity library generation and recombination-free paired-end sequencing. Sci Rep 6: 37563.

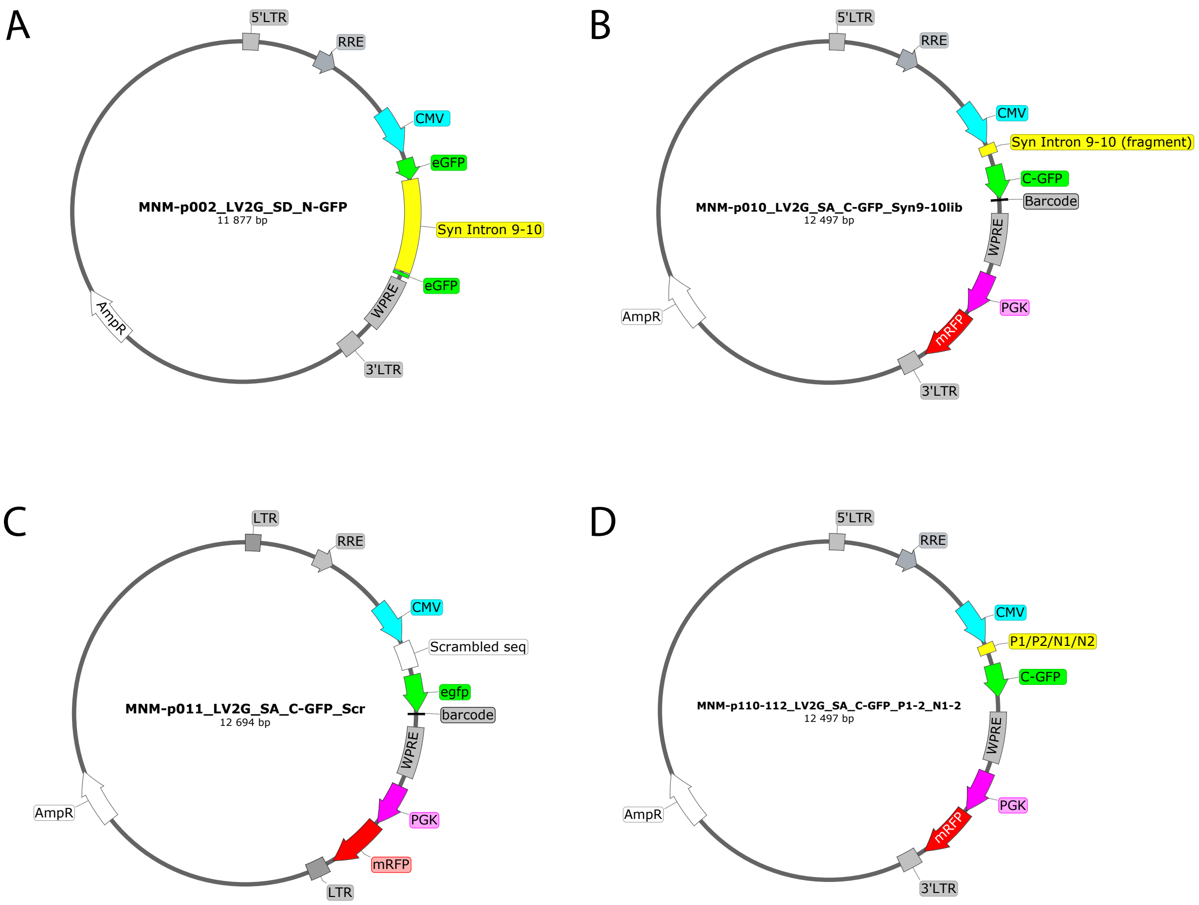
Much C, Auchynnikava T, Pavlinic D, Buness A, Rappsilber J, Benes V, Allshire R, O'Carroll D. 2016. Endogenous Mouse Dicer Is an Exclusively Cytoplasmic Protein. PLoS Genet 12: e1006095.

Ryan MD, King AM, Thomas GP. 1991. Cleavage of foot-and-mouth disease virus polyprotein is mediated by residues located within a 19 amino acid sequence. J Gen Virol 72 (Pt 11): 2727-2732.

Thomas G. 2002. Furin at the cutting edge: from protein traffic to embryogenesis and disease. Nat Rev Mol Cell Biol 3: 753-766.

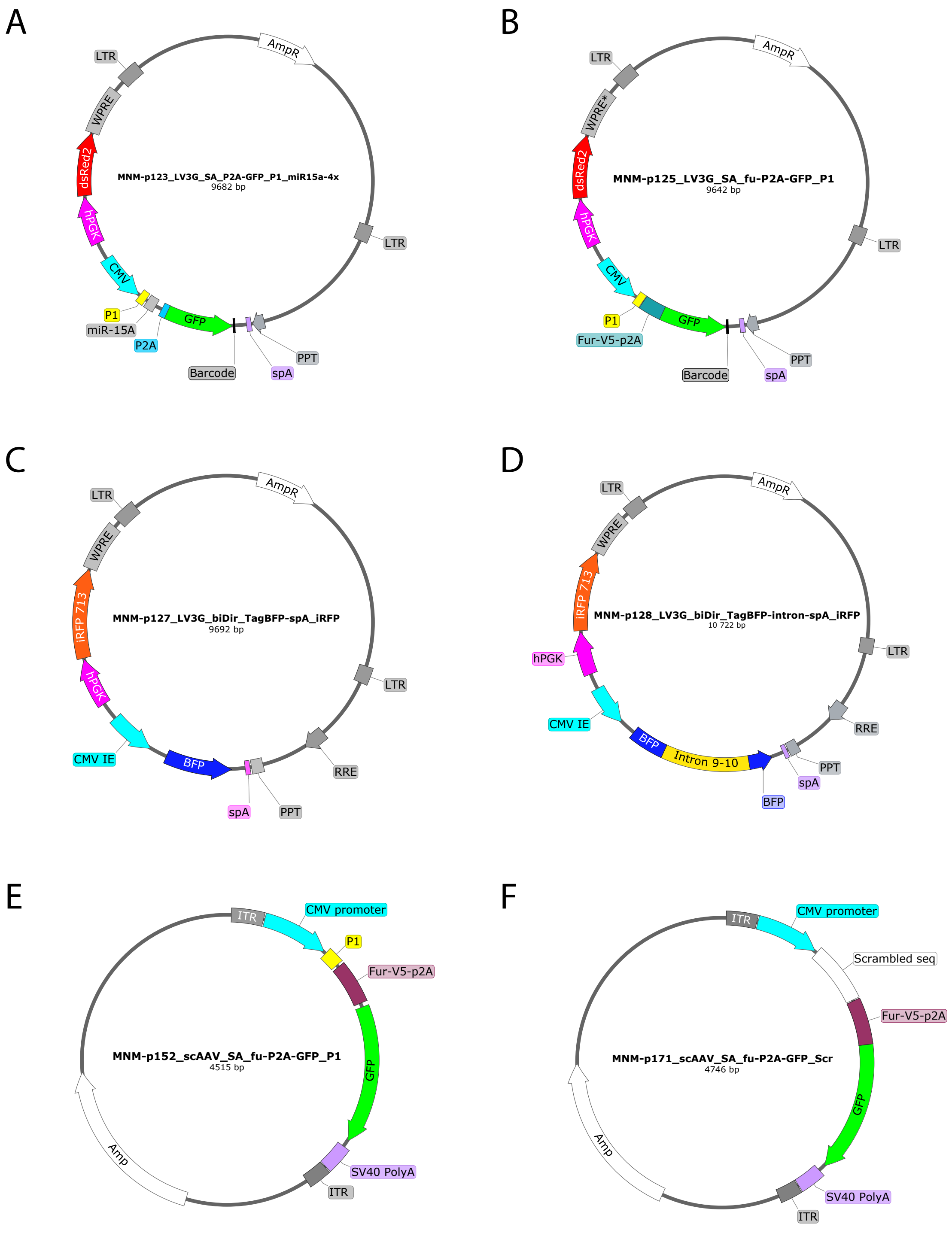
Wang WX, Danaher RJ, Miller CS, Berger JR, Nubia VG, Wilfred BS, Neltner JH, Norris CM, Nelson PT. 2014. Expression of miR-15/107 family microRNAs in human tissues and cultured rat brain cells. Genomics Proteomics Bioinformatics 12: 19-30.

Supplemental figure S4



Supplemental figure S4 | Plasmid maps of 2nd generation LV vectors
(A)-(C) utilized in Fig 2 & 3.(D) utilized in Fig2 H.
Full sequences can be obtained at: http://RNA2018.neuromodulation.se

Supplemental figure S5



Supplemental figure S5 | Plasmid maps of 3rd generation LV vectors and scAAV (A)-(D) utilized in Fig5 B-D. (E) & (F) utilized in Fig5 D-G.
Full sequences can be obtained at: http://RNA2018.neuromodulation.se