OMTM, Volume *9*

Supplemental Information

Adeno-associated Virus Genome Population Sequencing Achieves Full Vector Genome Resolution and Reveals Human-Vector Chimeras

Phillip W.L. Tai, Jun Xie, Kaiyuen Fong, Matthew Seetin, Cheryl Heiner, Qin Su, Michael Weiand, Daniella Wilmot, Maria L. Zapp, and Guangping Gao

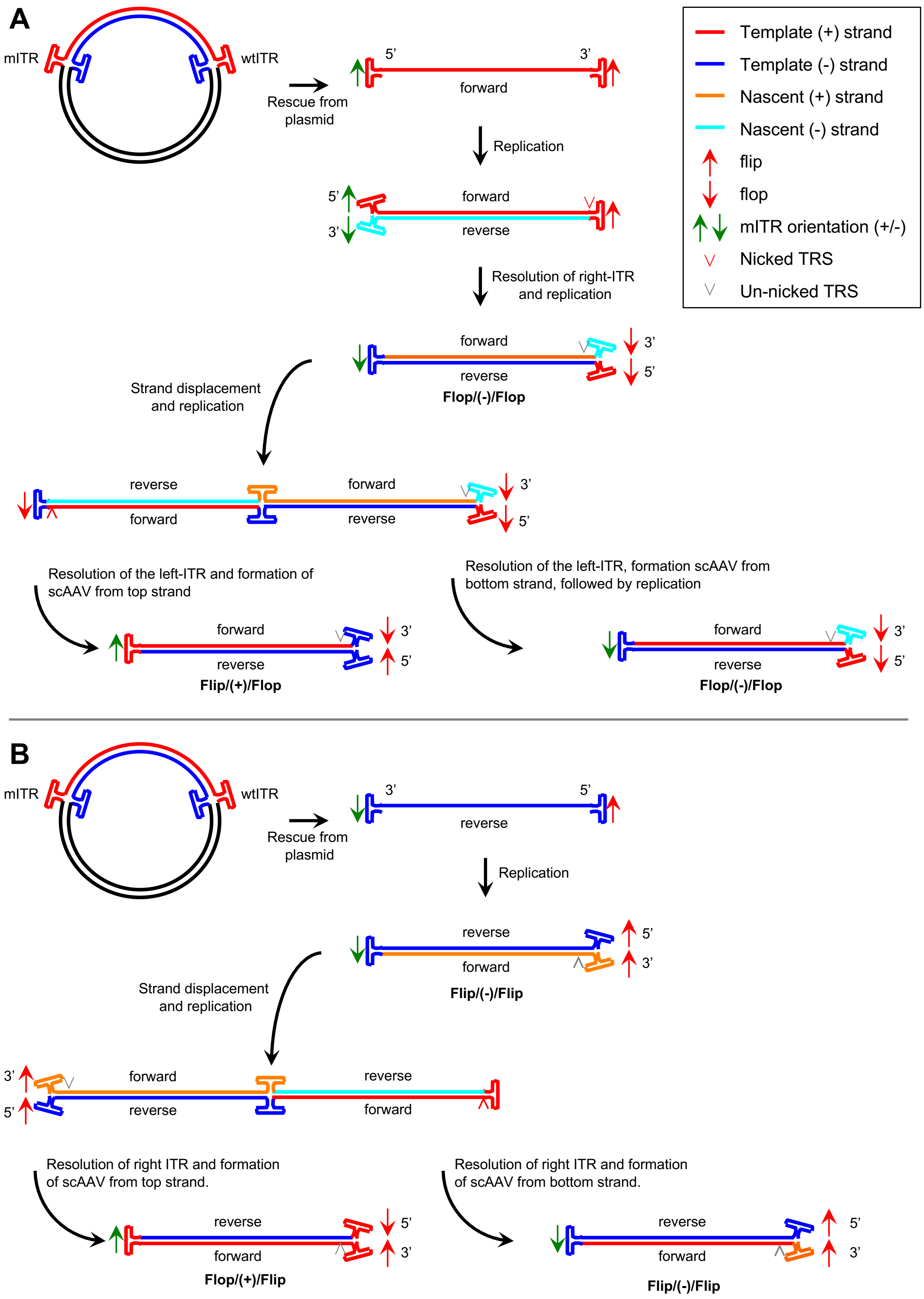
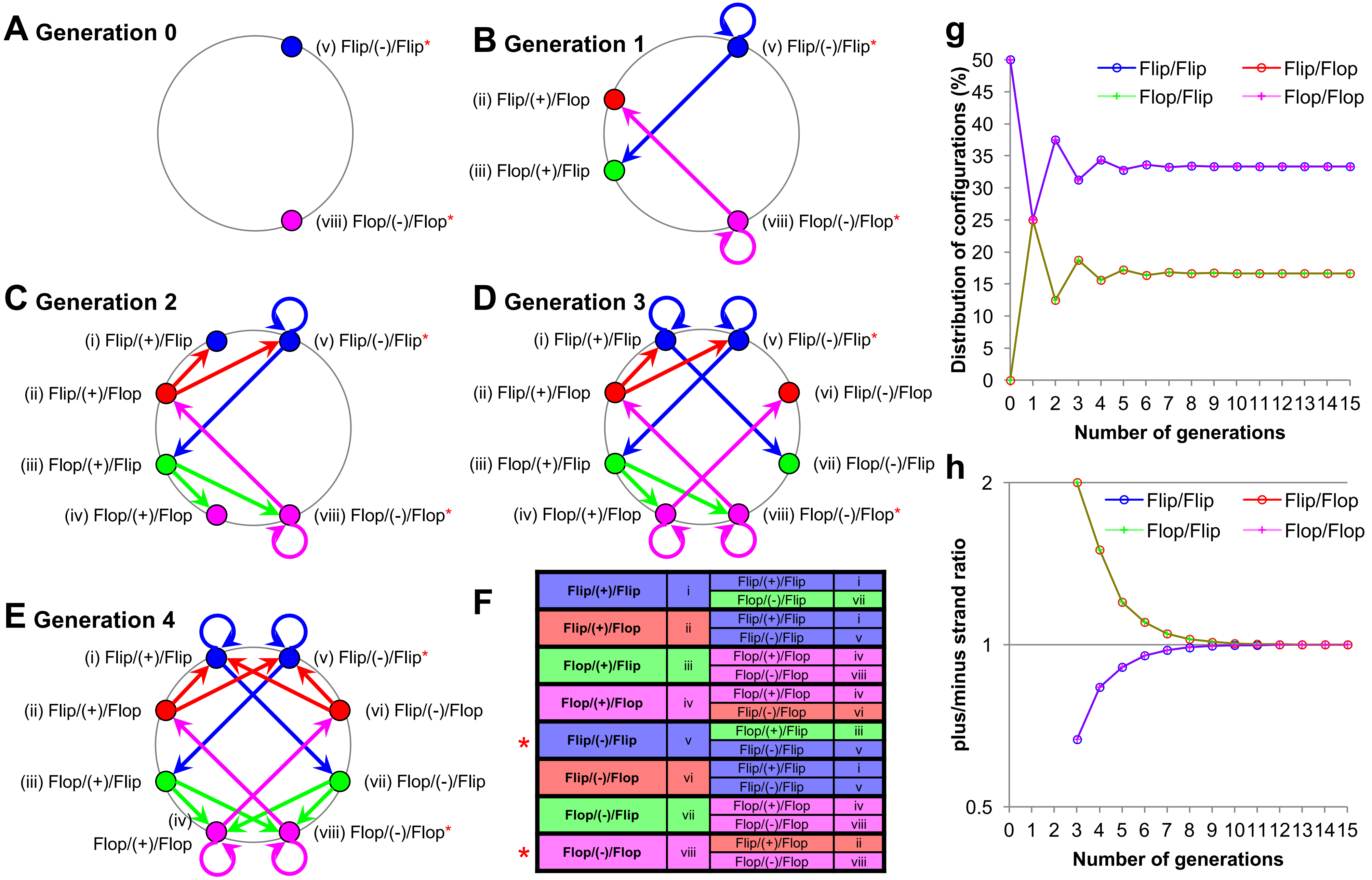
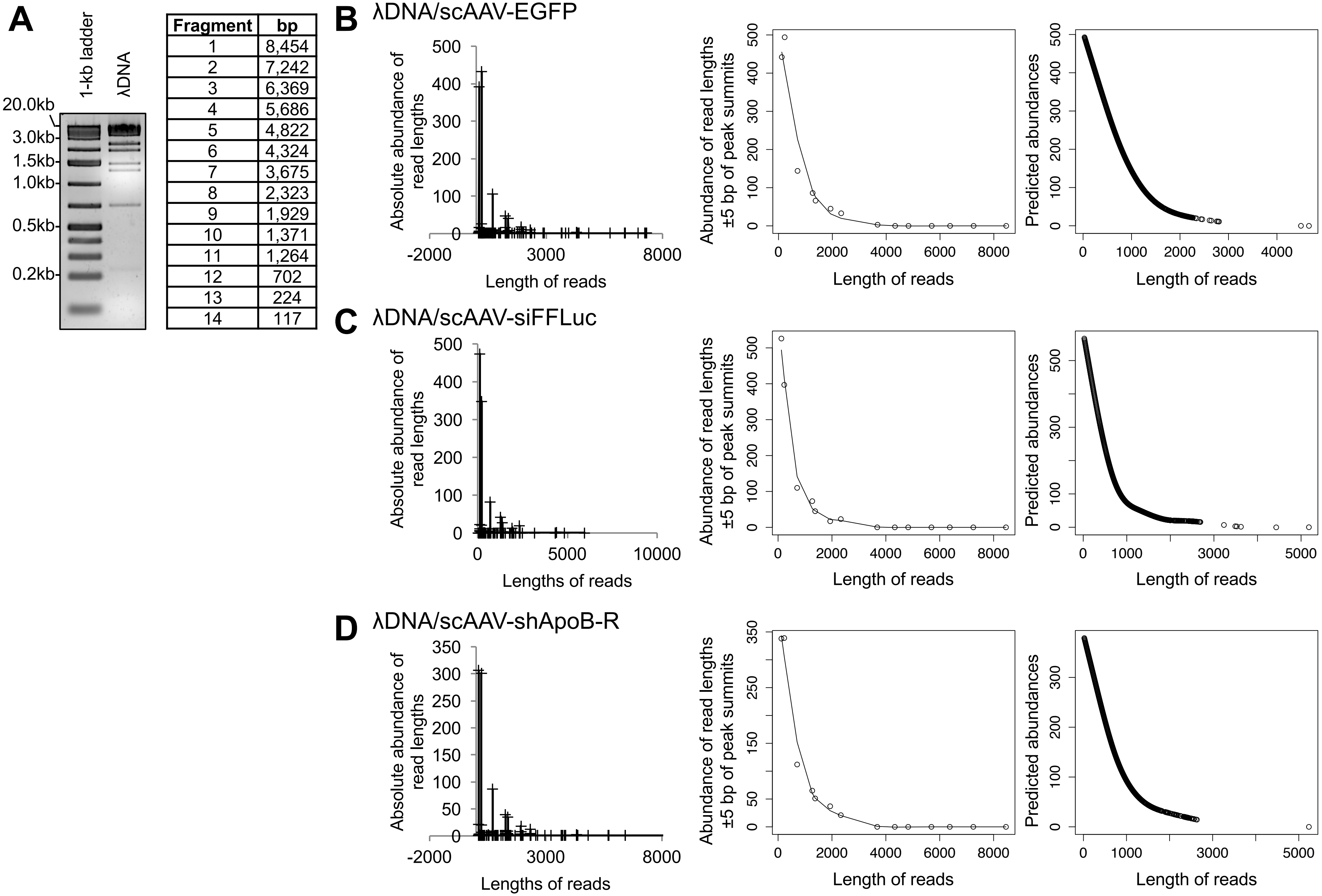


Figure S1. Model for rolling-hairpin replication of scAAVs. (A) The mITR and wtITR are presumed to be rescued by a combination of a Holliday junction resolvase and AAV-Rep. The plus (+) strand is replicated from the 3'-ITR. This forms an intramolecular double-stranded genome with an open mITR region. Resolution of the wtITR and replication from the self-primed 3'-mITR by either host-DNA repair or by Rep generates a Flop/(-)/Flop molecule. The 3'-ITR initiates strand-displacement replication to form an intermediate molecule containing a duplicated intramolecular, double-stranded, genome. The first generation of scAAVs is depicted as resolution of the left TRSs and the synthesis of two daughter scAAV genomic forms. **(B)** Plasmid rescue also generates a minus (-) strand template. Replication of the (-) strand produces a Flip/(-)/Flip molecule. Subsequent strand displacement and replication occurs in a similar fashion to create two additional scAAV forms. Nomenclatures for flip/flop configurations are: 5'-ITR / ± strand-ness / 3'-ITR.



	Flip/(+)/Flip	i	Flip/(+)/Flip	i
			Flop/(-)/Flip	vii
	Flip/(+)/Flop	ii	Flip/(+)/Flip	i
			Flip/(-)/Flip	v
	Flop/(+)/Flip	iii	Flop/(+)/Flop	iv
			Flop/(-)/Flop	viii
	Flop/(+)/Flop	iv	Flop/(+)/Flop	iv
			Flip/(-)/Flop	vi
*	Flip/(-)/Flip	v	Flop/(+)/Flip	iii
			Flip/(-)/Flip	v
	Flip/(-)/Flop	vi	Flip/(+)/Flip	i
			Flip/(-)/Flip	v
	Flop/(-)/Flip	vii	Flop/(+)/Flop	iv
			Flop/(-)/Flop	viii
*	Flop/(-)/Flop	viii	Flip/(+)/Flop	ii
			Flop/(-)/Flop	viii

Figure S2. Flip/flop configuration outcomes predicted by rolling-hairpin replication of scAAV genomes. (A-E) Diagrams of flip/flop configurations originating from the two source molecular forms: Flip/(-)/Flip and Flop/(-)/Flop (red asterisks) (see Figure S1). Each node represents a possible flip/flop configuration: Flip/Flip (blue), Flip/Flop (red), Flop/Flip (green), and Flop/Flop (magenta). By the 3rd generation, all possible configurations are represented. **(F)** Table representation of each flip/flop configuration yielding two daughter forms. **(G)** Distributions of flip/flop configurations based on prediction model for each replication round, extended to 15 generations. By the fifth generation, a steady-state ratio of 2:1:1:2 is reached. **(H)** Plot of plus-to-minus ratios for each flip/flop configuration for every replication generation.



Fragment	bp
1	8,454
2	7,242
3	6,369
4	5,686
5	4,822
6	4,324
7	3,675
8	2,323
9	1,929
10	1,371
11	1,264
12	702
13	224
14	117

Figure S3. Read length distributions of BstEII-digested λDNA. **(A)** Agarose-gel of EtBr-stained BstEII-digested λDNA and fragment-lengths generated by digestion according to manufacturer's description. **(B-D)** Analysis of λDNA spike-ins for SMRT sequencing reads of **(B)** scAAV-EGFP, **(C)** scAAV-siFFLuc, and **(D)** scAAV-shApoB-R vector libraries distributed by length. Left plots displays the absolute read abundances distributed by length. Center plots displays the polynomial-splines fit to the data points. Right plots display the predicted abundances of all observed SMRT sequencing read lengths mapping to the vector genome.

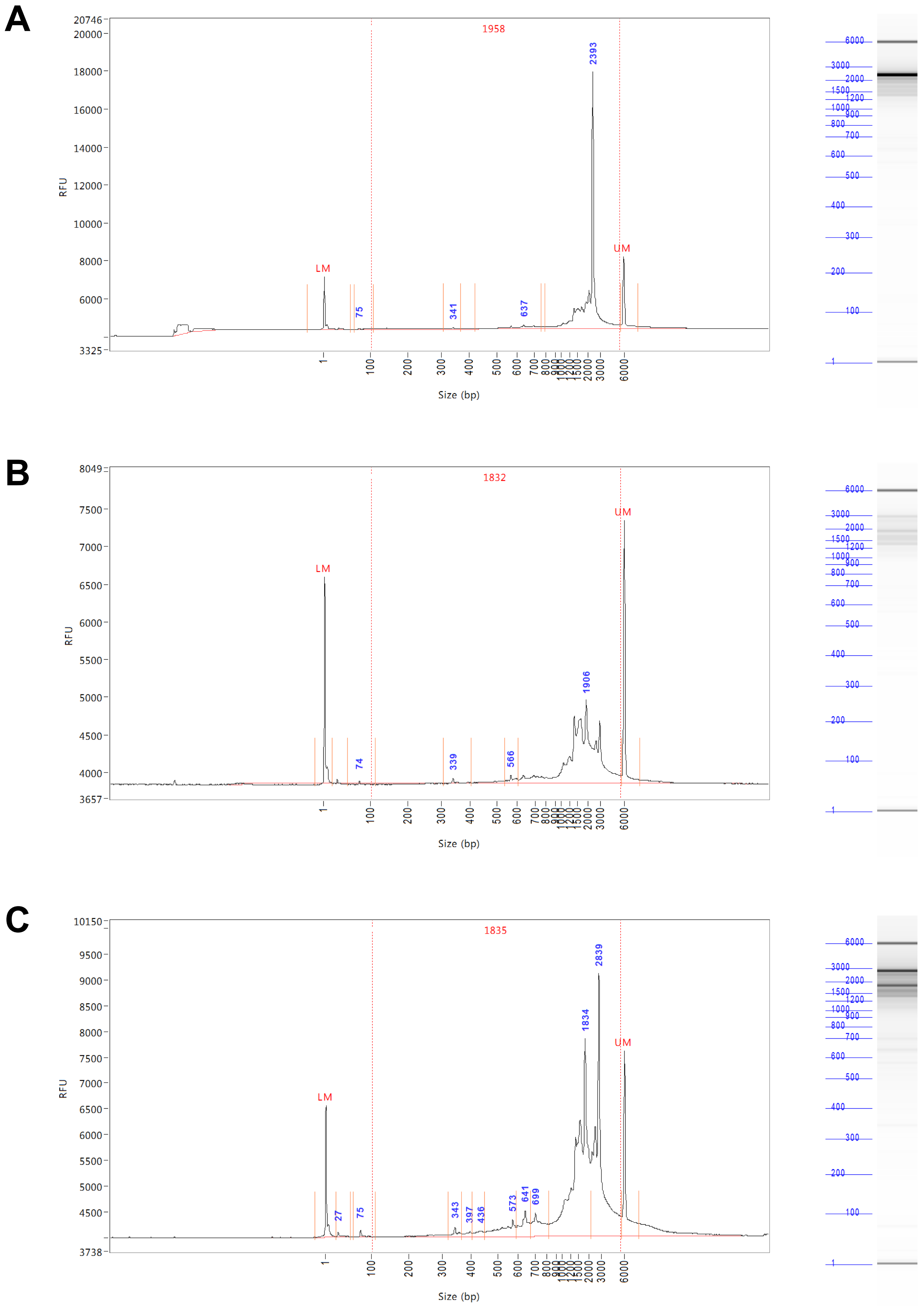


Figure S4. Fragment analyses of purified vector genomes by capillary electrophoresis of **(A)** scAAV-EGFP, **(B)** scAAV-siFFLuc, and **(C)** scAAV-shApoB-R rAAV preparations. Note that fragment sizes indicated at peak summits are approximate to actual fragment sizes.

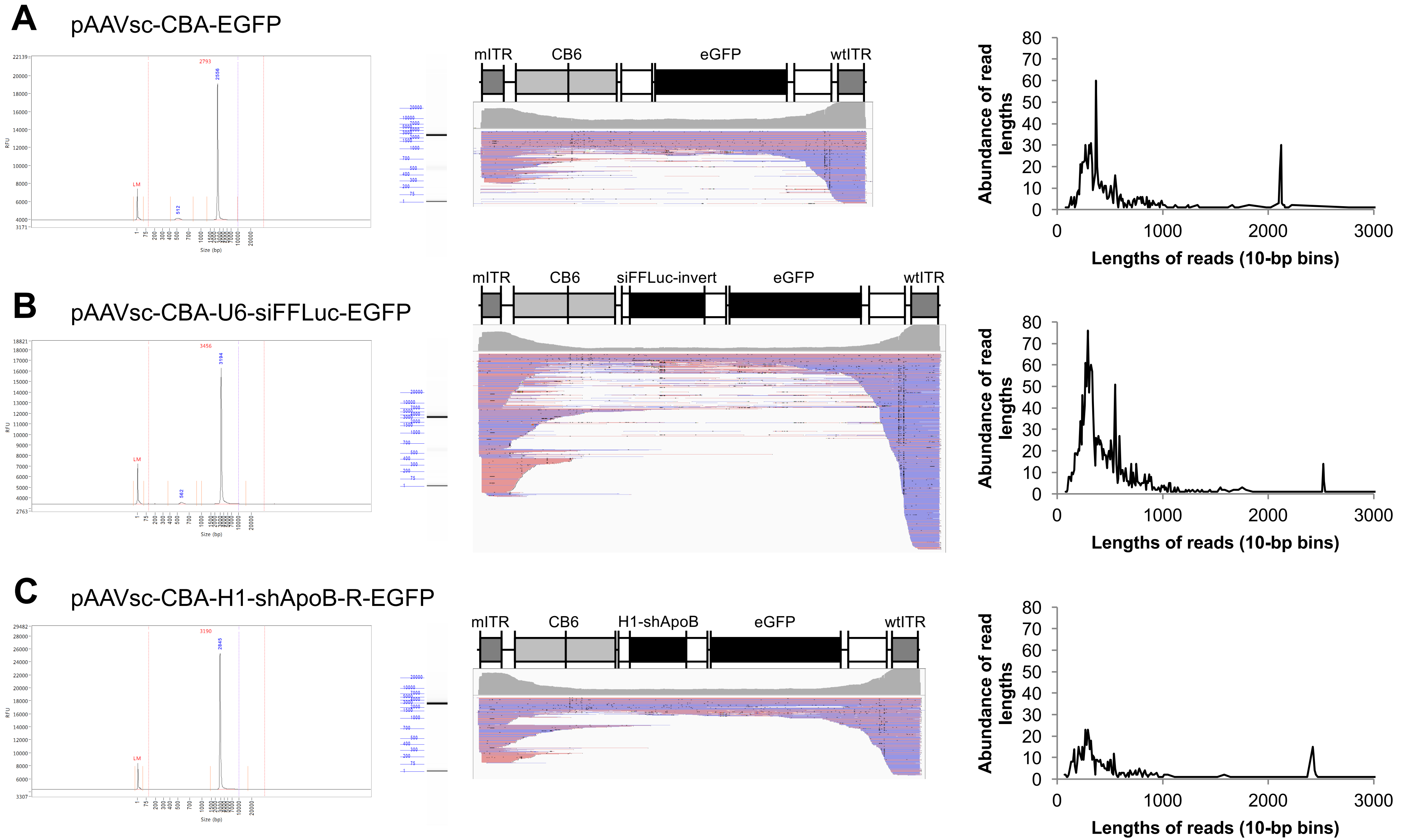


Figure S5. Preparation of SMRT libraries or sequencing error results in truncated reads. Plasmid DNA constructs of **(A)** scAAV-EGFP, **(B)** scAAV-siFFLuc, and **(C)** scAAV-shApoB-R vectors were cut with PacI and the digestion fragment was subjected to SMRT sequencing analyses to determine whether AAV-GPseq can reliably sequence through the ITRs. Left panels show by fragment analyses that the isolated PacI-digestion fragments used as DNA input for sequencing have uniform sizes. Analyses of SMRT sequence reads resulted in an overrepresentation of truncated reads that span the mITRs and the wtITRs (center panels). Right panels summarize the abundance of read counts distributed by read length.

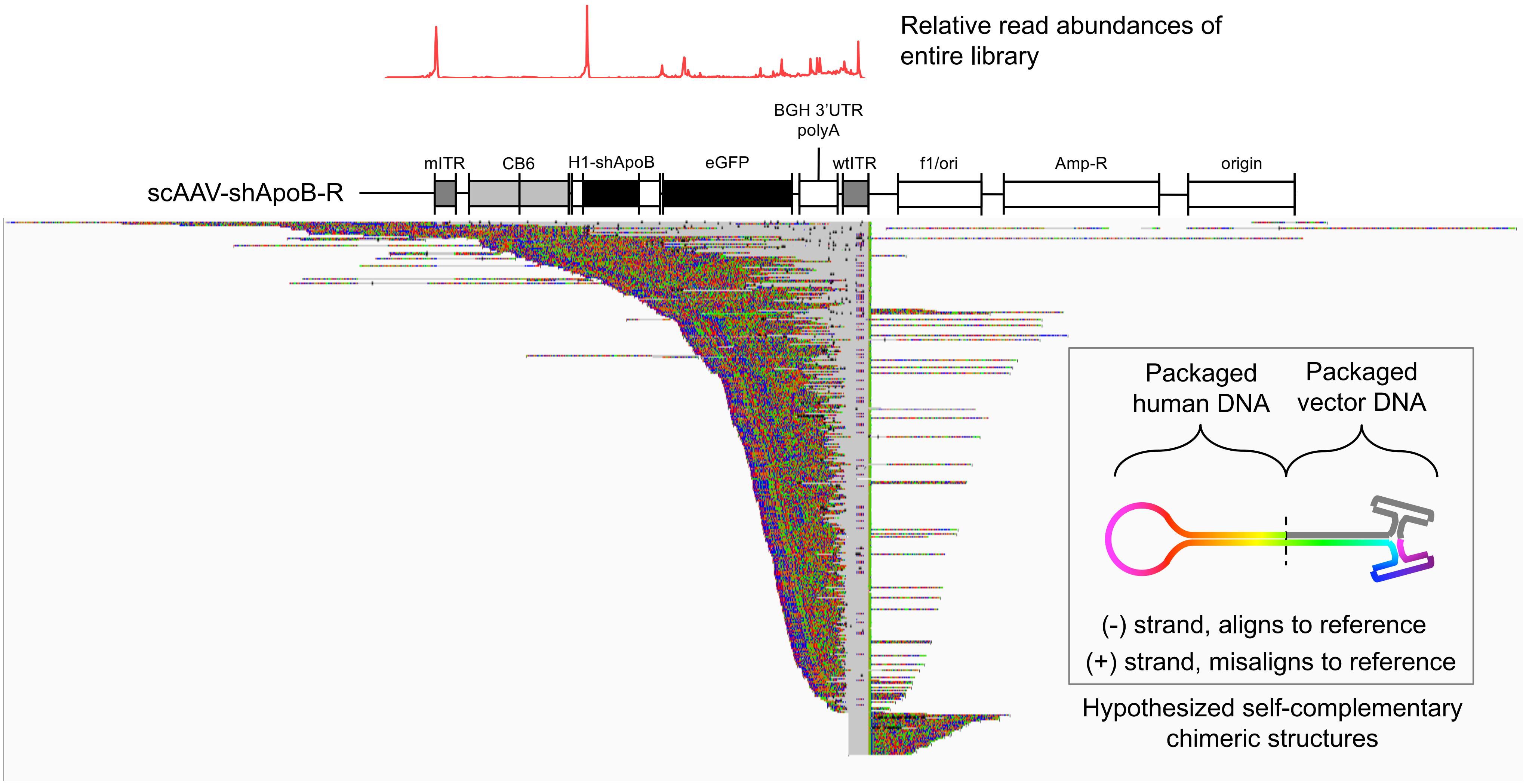
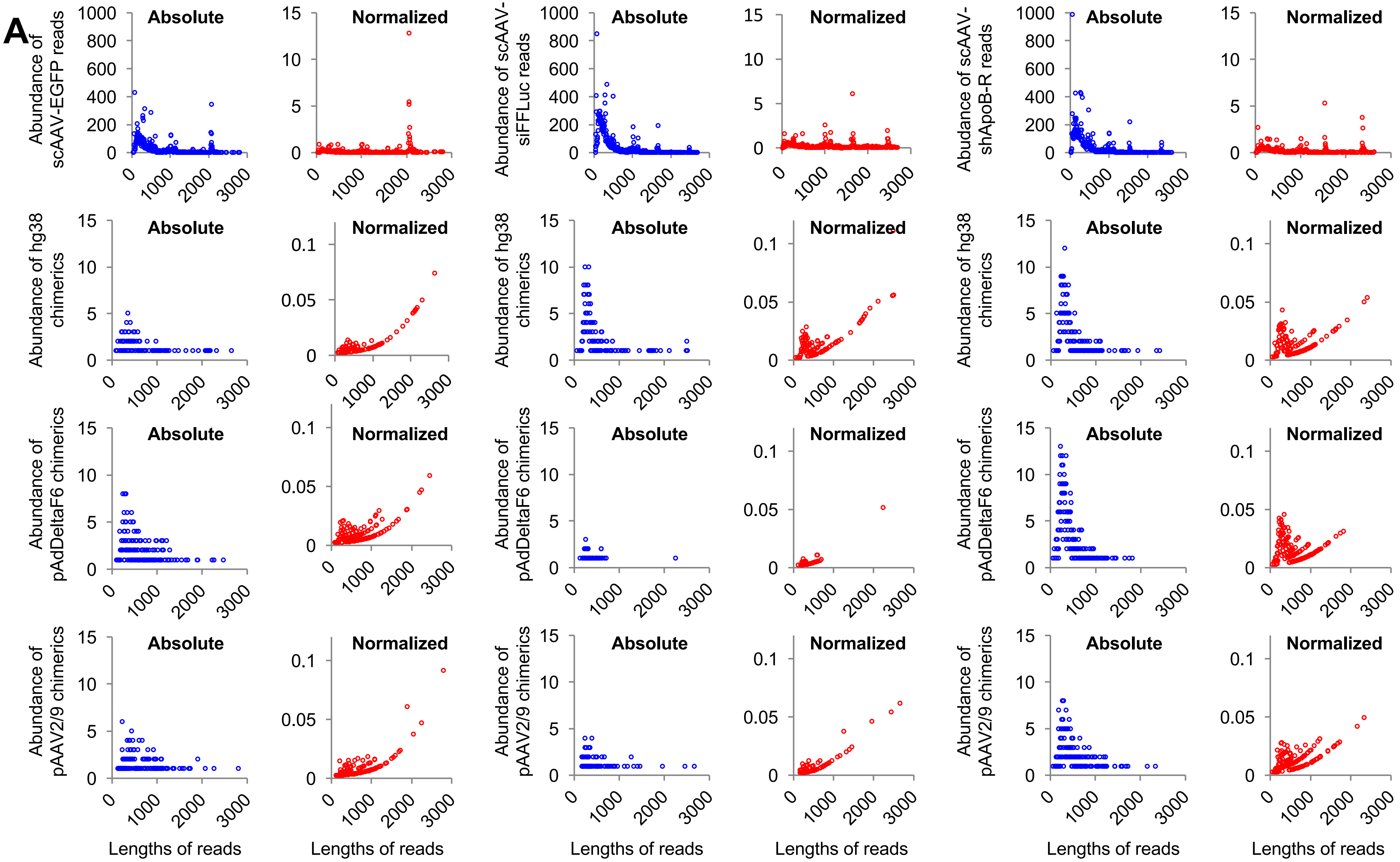


Figure S6. Characterization of reads from scAAV-shAboB-R vector preparation that map to the human genome. Alignments are displayed with soft-clipped bases to demonstrate read segments that align to the vector genome reference (gray) and segments that do not (colored). The relative read abundances from Figure. 3C and the diagram of the construct reference is shown above the alignments to indicate plausible hotspots for strand switching. Majority of chimeric reads contain sequences that map to the wtITR, indicating an active mechanism for non-vector sequence packaging.



B Percentage of chimeric reads = $\dfrac{\text{adj.values}_{(\text{chimeras})}}{\text{adj.values}_{(\text{vector genomes})}}$ X 100

Figure S7. Calculation of chimeric-read abundances. **(A)** Absolute counts of reads mapping to vector genome sequence and chimeric-reads (blue plots) are normalized to the read-length distributions of λDNA spike-ins (Figure S3). **(B)** To obtain the percentage of chimeras in the vector genome populations, the totaled adjusted values of chimeric reads are simply divided by the totaled adjusted values of reads mapping to vector genomes (red plots). Calculated values are displayed in Table 1.

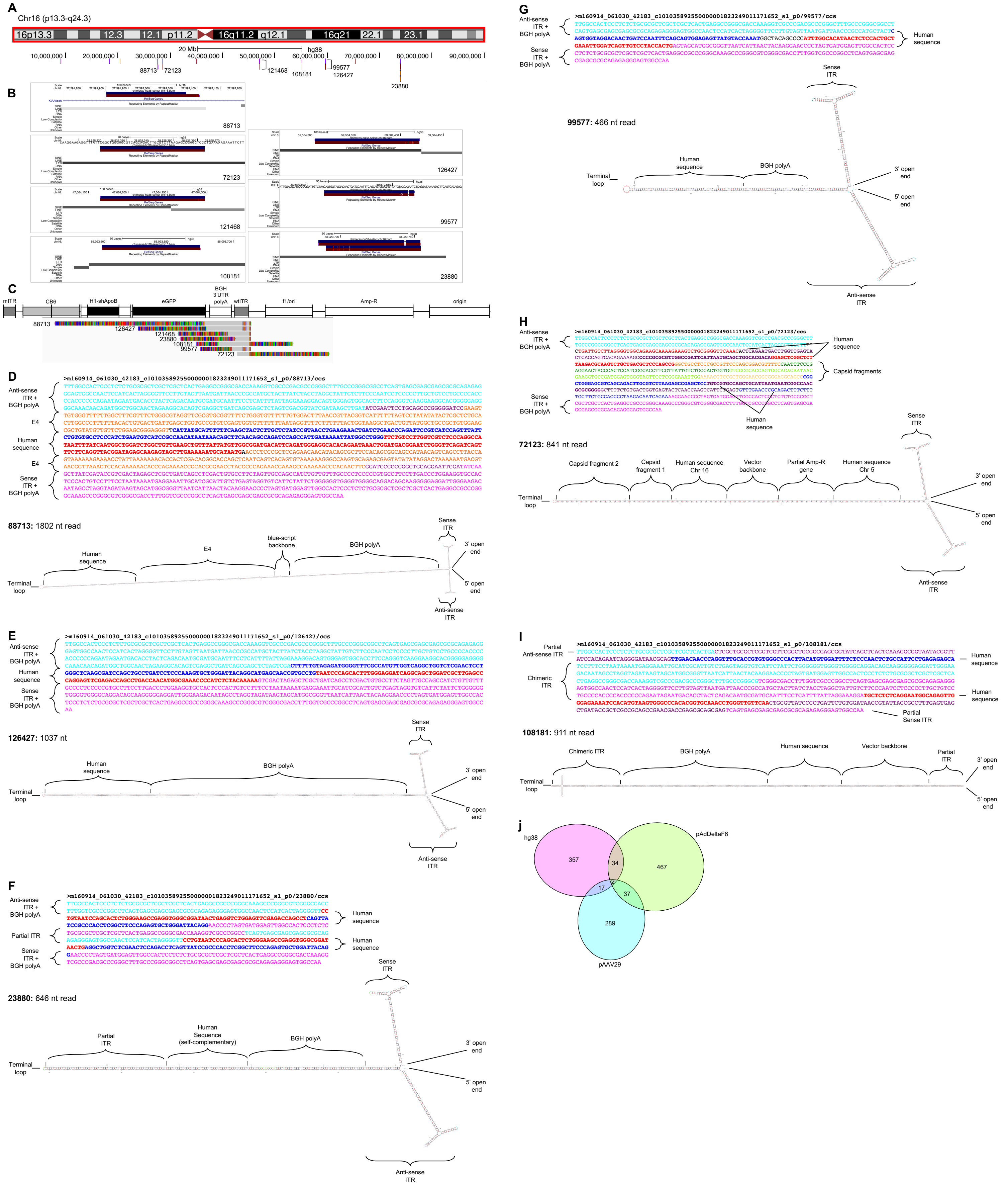


Figure S8. Characterization of chimeric reads that map to the human genome. (A) Seven chimeric reads that map to chr16 were interrogated. **(B)** These reads all share in common the feature of aligning twice (or more) to the same region of the human genome. **(C)** The chimeric reads also align to the wtITR. **(D-I)** Annotated sequences showing the diversity of chimeric forms among the seven selected reads. Each read is also accompanied by mfold structures. **(J)** Venn diagram showing the detection of chimeric reads that map to multiple genomic sources (human, pink; Ad-helper, green; and Rep-Cap packaging plasmid, blue).

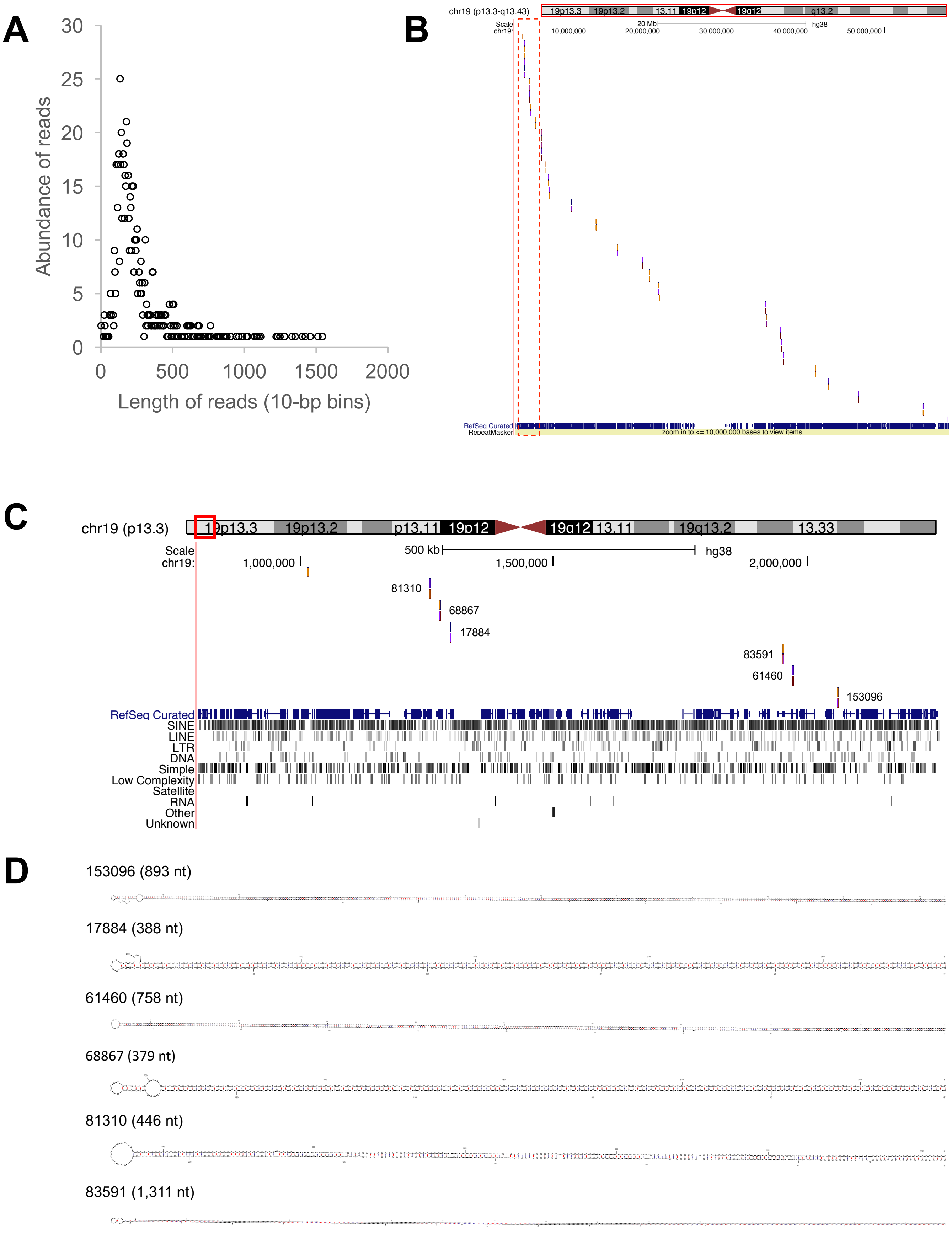


Figure S9. Foreign DNAs lacking ITRs (non-chimerics) can package into capsids as self-complementary strands. **(A)** Scatter plot showing the abundance of read lengths detected among vector genomes in the scAAV-EGFP preparation that map exclusively to the human genome. **(B)** UCSC genome browser alignment tracks of reads that exclusively map to hg38, chromosome 19. **(C)** Expanded view of 19p13.3 region on chromosome 19 (red box in panel B) indicate that many reads share in common the feature of aligning twice to the same region. Each aligned read is annotated with its unique read ID. Tracks for known RefSeq annotated transcripts and repetitive elements are shown below. **(D)** Mfold structures of six selected reads (from panel C) indicating self-complementation. Length of single-strand reads are displayed with each read ID.