Impact of Genome Size and Sequence Composition on AAV Vector Genome Integrity

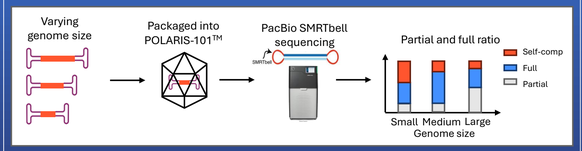
Jie Tan[1], Irene Vicente Larena[1], Hong Ma[2], Jennifer Gifford[1], Adam Cockrell[1], Nilay Patel[1], Ian Cook[1], Ryan Irving[1], Phillip W.L. Tai[1], Glen B. Banks[1]

[1] Solid Biosciences Inc., Charlestown, USA, [2] UMass Chan Medical School, Horae Gene Therapy Center, Worcester, United States

INTRODUCTION

Adeno-associated virus (AAV) vectors are widely used in gene therapy, yet maintaining high genome purity and integrity during vector production remains a critical challenge. In this study, we systematically investigated how construct design, genome size, and sequence features affect the manufacturing yield and packaging quality of single-stranded (ssAAV) and self-complementary (scAAV) vectors. We utilized POLARIS-101™ (formerly known as AAV-SLB101) as the packaging capsid, Solid Biosciences' next-generation capsid engineered to enhance gene delivery to striated muscles. It is based on AAV9 serotype with an RGD-containing peptide inserted into variable region VIII. Clinical studies have demonstrated that it offers robust transduction capabilities across multiple skeletal and cardiac muscle tissues at a clinically safe dose.

Schematic workflow of the study



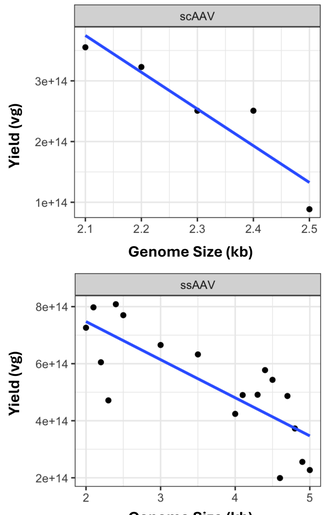
A panel of constructs spanning a dense sampling of genome sizes was packaged into the POLARIS-101™ capsid and evaluated by PacBio long-read sequencing to quantify the fraction of full-length, partial, and dimeric packaged genomes.

Construct design

Construct design for ssAAVs and scAAVs with varying genome sizes. Constructs share a common expression cassette (eGFP flanked by human TNNT2 5' and 3' UTRs) and were length-tuned across 2.0 kb to 5.0 kb for ssAAV and 2.1 kb to 2.5 kb for scAAV via stepwise UTR shortening.

Reference:
1. Talevich, E. et al. Standardized Nomenclature and Reporting for PacBio HiFi Sequencing and Analysis of rAAV Gene Therapy Vectors. Preprint at https://doi.org/10.1101/2024.05.07.592296 (2024).
2. Cer, R. Z. et al. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Research 41, D94–D100 (2012).
3. Lorenz, R. et al. ViennaRNA Package 2.0. Algorithms Mol Biol 6, 26 (2011).

AAV GENOME SIZE AND PRODUCTION YIELD

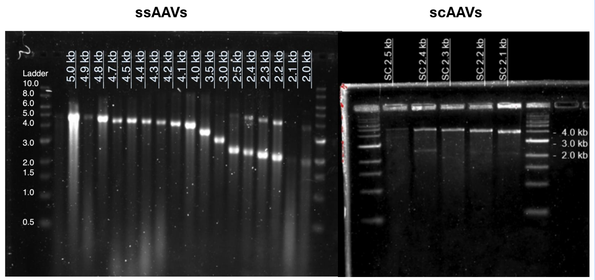


scAAV	corr	p.value
Production yield	-0.93	<0.05

ssAAV	corr	p.value
Production yield	-0.79	<0.001

Scatter plots and correlation statistics indicate a strong negative relationship between genome size and production yield for both vector types.

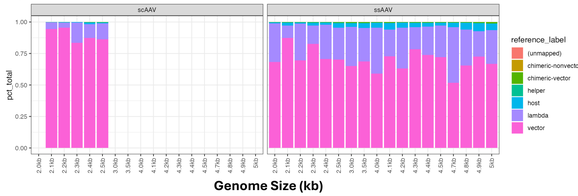
GENOME INTEGRITY PROFILING BY ELECTROPHORETIC GEL



Gel electrophoresis of ssAAVs and scAAVs packaging genomes with different sizes. The gel of ssAAVs shows a strong band at the expected genome size. For ssAAV with size between 2.0kb and 2.5kb, it shows bands 2x the expected size, suggesting self-complementary genome formation. Gel of scAAVs shows a band at the expected size for each construct.

GENOME INTEGRITY PROFILING BY LONG-READ SEQUENCING

A) Percentage of reads mapping to each reference



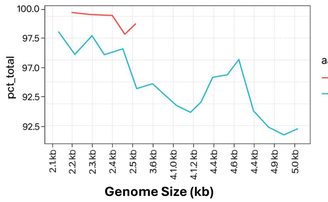
A. Percentage of reads mapping to each reference, including vector, helper plasmid, host, lambda spike-in, and chimeric. Sequencing data were analyzed by the LAAVA pipeline[1].

B. Percentage of reads mapping to the vector genome vs. non-vector reference after excluding the lambda spike-in reads. Overall low percentage (1%-10%) of non-vector contamination reads in all vectors. Small genomes trend with lower levels of packaged contaminants. scAAVs tend to have lower contamination read ratio than ssAAVs.

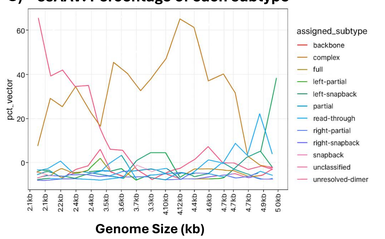
C. Distribution of vector genome read subtypes across genome sizes for ssAAVs. As genome size increases from 2kb, the full ratio increases first (except 2.5kb) and reaches the maximum at 4.3-4.4kb. Then it decreases and drops sharply beyond 4.8kb. The unresolved-dimer reads contain full ITR-to-ITR in both strands and thus considered as full. Unresolved-dimer read ratio is high for 2.0 kb to 2.5 kb and decreased to a baseline ratio of <15% for larger genome sizes. This is consistent with the gel result showing band at 2x expected size for genome sizes <= 2.5kb.

D. Distribution of vector genome read subtypes across genome sizes for scAAVs. Full genome percentage starts to decrease when genome size is equal or larger than 2.4kb.

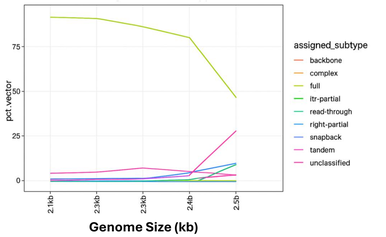
B) Percentage of reads mapping to vector (lambda excluded)



C) ssAAV: Percentage of each subtype

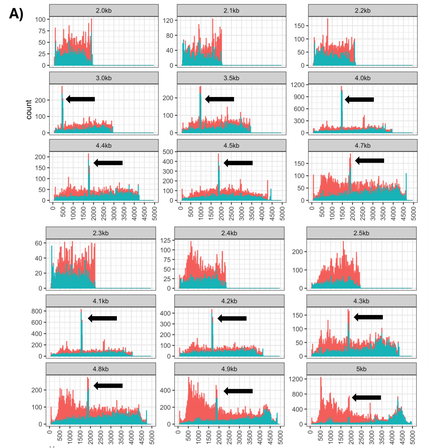


D) scAAV: Percentage of each subtype



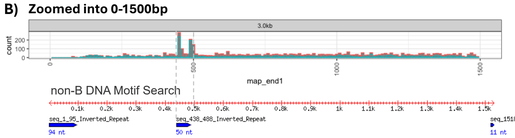
AN INVERTED REPEAT LEADS TO GENOME FRAGMENTATION

A)



Genome size	Partial perc from hotspot
3.0kb	7.18
3.5kb	6.46
4.0kb	7.94
4.1kb	6.86
4.2kb	5.34
4.3kb	1.75
4.4kb	1.43
4.5kb	4.31
4.7kb	3.22
4.8kb	3.21
4.9kb	2.70
5.0kb	2.57

B) Zoomed into 0-1500bp



C) DNA secondary structure (seq_438_488 highlighted)

A. Distribution of read-end positions for left and right partial reads. A fragmentation hotspot (indicated by arrows) in the 5' UTR of the human TNNT2 gene emerges in constructs larger than 3.0 kb. Percentage of partial reads from the hotspot is shown in the table **B.** Zoomed view of the 0–1500 bp region of the 3.0 kb construct overlaid with non-B DNA motif annotations2. The fragmentation hotspot co-localized with a predicted inverted repeat. **C.** Illustration of DNA secondary structure (fold prediction by RNAfold3 via Geneious) consistent with the inverted-repeat motif prediction. DNA sequence from 439 to 488 is highlighted.

CONCLUSIONS

- ssAAV payloads achieve the highest integrity near ~4.3–4.4 kb and exhibit sharp truncation beyond ~4.8 kb, whereas scAAV integrity declines above ~2.4 kb. ssAAV at 2.5 kb also shows high genome truncation.
- Small ssAAV genomes do not increase packaged non-vector DNA, and scAAVs show lower impurity than ssAAVs at matched sizes
- Local sequence features such as inverted repeats contribute to fragmentation, though having a smaller impact than oversizing, highlighting the combined importance of length control and sequence-level refinement.

Presented at American Society of Gene & Cell Therapy, May 11-15, 2026, Boston, MA