Characterization of genomic heterogeneity in rAAV preparations using short- and long-read next generation sequencing

Tina Strauss, Alpha B. Diallo, Irene Zolotukhin, Elizabeth Tseng, Kristina Weber, Matthew Burg, Nilay Patel, Brett Palaschak, Stewart Coleman, Jennifer L. Marlowe and Adam S. Cockrell

Solid Biosciences, Charlestown, MA, USA, Form Bio, Dallas/Austin, TX, Pacific Biosciences, Menlo Park, CA

Introduction

Recombinant adeno-associated virus (rAAV) is a popular delivery vehicle for therapeutic transgenes to correct monogenic disorders. As the demand for large-scale rAAV vector manufacturing rises, the safety, purity and efficacy of rAAV productions should be reproducible. Triple transfection (TT) is a common production method to produce rAAV that requires the use of three plasmids: (i) Transgene encodes the therapeutic genetic cargo flanked by ITRs and packaged into rAAV particles; (ii) Rep/Cap encodes the rAAV enzymatic and structural proteins that comprise rAAV vector particles; (iii) Ad Helper (Helper) encodes adenovirus helper genes to facilitate rAAV production (Fig. 1). Following AAV-vector production via TT the genetic cargo may be a heterogenous mixture of molecular DNA species that include full-length transgene (the desired genetic cargo), truncated transgene, fragments of host genomic DNA, DNA fragments from Rep/Cap and Helper plasmid, DNA fragments from Transgene plasmid backbone (e.g. bacterial origin of replication or bacterial antibiotic resistance encoding gene) and DNA chimeras (i.e. molecular recombinants of the aforementioned species) (Fig. 1). To characterize the heterogeneity of rAAV productions we performed short (Illumina) and long (PacBio) read sequencing analysis. Long read sequencing was previously demonstrated to reveal novel insights into the heterogeneity of molecular species in rAAV preparations (Tran et al. 2022, Tai et al. 2018, Tran et al. 2020). Combining long read sequencing with a well-developed bioinformatics pipeline might facilitate the application of learnings from rAAV vector characterizations to sophisticated artificial intelligence (AI) platforms, providing foresight into Transgene vector design that ultimately reduce costs and shorten timelines in early R&D pipelines.

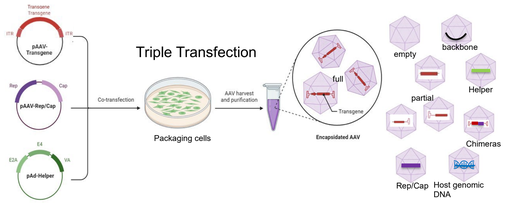

Figure 1. Triple Transfection in packaging cells leads to heterogeneity of genetic cargo encapsidated into AAV-vector particles. Heterogenous mixture of genetic cargo after TT can include full-length and truncated (partial) transgene, fragments of host genomic DNA, fragments of plasmid DNA originating from the Rep/Cap and Helper-plasmid, fragments derived from the backbone of the transgene plasmid and DNA chimeras (combinations of different DNA fragments involved in TT).

Materials and Methods

Illumina (short-read sequencing) and PacBio (long-read sequencing) were used as orthogonal methods for the analysis of several rAAV productions with Azenta's AAV sequencing workflows. Viral DNA was extracted from rAAV and used in library preparation including adapter ligation. The samples were sequenced on NOVAseq/HIseq (Illumina) and Sequel or Sequel IIe instruments (PacBio). Data analysis was performed with Azenta's, FormBio's, or PacBio's bioinformatic pipelines* specific to the sequencing method focusing on the alignment of the sequence reads to the different reference sequences that include sequences from plasmids defined here as: Transgene (ITR-Transgene-ITR), Backbone (Transgene plasmid outside of ITRs), Rep/Cap, Helper, Chimeras (Transgene with other reference) and Host (human genomic DNA sequences contributed by HEK 293 cells) used during TT.

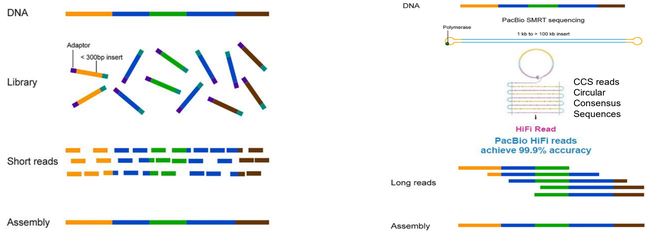

Figure 2. Workflow for Illumina and PacBio based NGS sequencing shows that PacBio CCS reads produce long reads (up to 100kb) with high fidelity compared to Illumina's short 300bp reads. Long read sequencing supports resolution of full-length molecules, affording novel insights into the presence of multiple molecular species.

References:

Tran et al. 2022, Human and Insect Cell-Produced Recombinant Adeno-Associated Viruses Show Differences in Genome Heterogeneity

Tai et al. 2018, Adeno-associated Virus Genome Population Sequencing Achieves Full Vector Genome Resolution and Reveals Human-Vector Chimeras

Tran et al. 2020, AAV-Genome Population Sequencing of Vectors Packaging CRISPR Components Reveals Design-Influenced Heterogeneity

Figure 1 created with Biorender.com and Figure 2 modified from: Chen, Zhiao & He, Xianghuo (2021), Application of third-generation sequencing in cancer research and PacBio

*Elizabeth Tseng Github tutorial: https://github.com/Magdoll/AAV/

Results

Recombinant AAV Productions Can be Reproducibly Characterized Across NGS Platforms

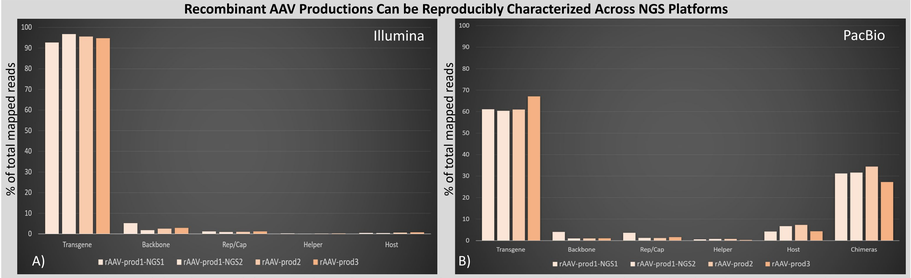

Figure 3. Illumina (A) and PacBio (B) NGS analysis of rAAV vectors from identical or different production batches show similar results.

- A) Illumina (short read) and B) PacBio (long read) NGS analyses of three rAAV-productions using the same transgene and same capsid.
- rAAV-prod1-NGS2 is an NGS-duplication of rAAV-prod1-NGS1 (same rAAV-production undergoing NGS analysis).
- rAAV-prod2 and 3 are rAAV productions that use identical plasmid ratios or slightly different ratios, respectively.
- Sequence reads were aligned to corresponding reference sequences and show percentage of total mapped reads for each reference.
- All four NGS analyses show similar percentages for the referenced categories for A) Illumina and B) PacBio analysis.

Purification of rAAV Full Capsids Modifies the Molecular Heterogeneity Profile

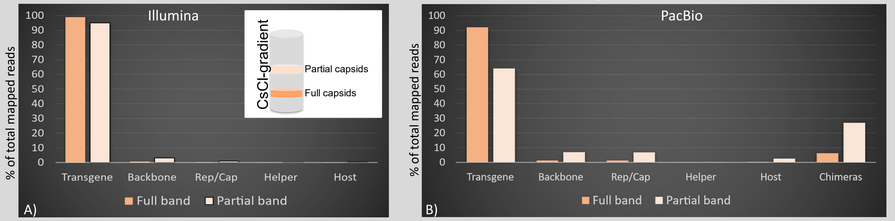

Figure 4. Additional enrichment using a CsCl-density gradient shows increased Transgene-specific reads with PacBio analysis in full capsids.

- Full and partial bands from a rAAV production (one capsid, one transgene) have been separated on a CsCl-gradient, pulled and independently analyzed using A) Illumina and B) PacBio NGS analysis.
- PacBio analysis shows increased transgene reads (~ +30%) and less chimeras (~ -30%) in full band compared to partial band.
- Differences in the Illumina analysis are minor between full and partial band, most likely due to short read alignments that cannot cover chimeric events between molecular species.

Transgene Cassette Design Influences the Profile of the Molecular Species Packaged Into the Same Capsid

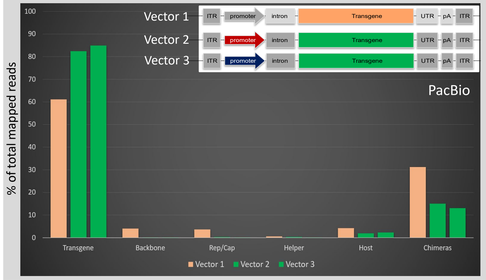

Figure 5. Different expression cassettes packaged with the same serotype show different packaging efficiencies based on insert reads.

- Three rAAV-vector samples packaged into the same capsid.
- Vector 1 contains a different transgene cassette compared to vectors 2 and 3.
- Vectors 2 and 3 only differ in the promoter but are otherwise completely identical.
- Vector 1 shows lowest transgene percent while vector 2 and vector 3 display similar percentages across the transgene and other references.
- Chimera percentage is decreased by more than 50% with vectors 2 and 3.

Results

Backbone Configurations Can Influence the Profile of Molecular Species Packaged Into the Same Capsid

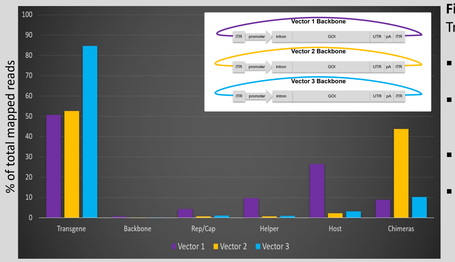

Figure 6. Backbone configuration can lead to higher Transgene packaging and less heterogeneity.

- PacBio NGS analysis of three rAAV-vector samples packaged in the same capsid.
- Vectors are identical in their expression cassette between the ITRs but differ in their backbone sequences.
- Vector 3 displays highest transgene reads, vectors 1 and 2 show similar percentages.
- Vectors 1 and 2 show high heterogeneity, vector 2 exhibits highest percentage of Chimeras, while vector 1 shows the highest amount of packaged host genomic DNA.

Thorough Bioinformatics Analysis Reveals Detailed Profiles of Packaged Molecular Species

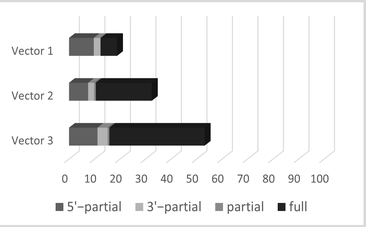

Figure 7. Sophisticated Bioinformatics Pipelines Support Profound Molecular Characterization

- Proprietary bioinformatics pipelines developed by FormBio and PacBio reveal novel insights into the molecular composition of the three vectors assessed in Fig. 6.
- Specific molecular species (5'-partial, 3'-partial, partial, and full length) were quantified based on read counts.
- Vector 3 exhibited the highest percent of full-length ITR-to-ITR reads.
- Importantly, the >80% of transgene reads shown for vector 3 in Fig. 6 is comprised of multiple different molecular species.

Artificial Intelligence-Based Machine Learning Can Inform Vector Modifications That Can Limit Heterogeneity

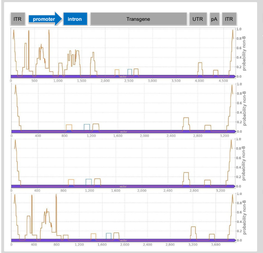

Figure 8. AI-derived output for non-B secondary structure probability plotted as a function of sequence position may predict truncation hotspots.

- FormBio's AI platform provides secondary structure predictions and codon optimizations that may inform optimal vector design, with the goal of increasing full-length transgene species while limiting packaging of all other molecular species.
- These four different histograms represent four different vectors that differ only in the promoter (i.e. constitutive, ubiquitous and cardiac-specific promoters) and intron (presence or absence) sequence.
- The propensity for secondary structure formation in 2 of the 4 different vectors suggests that modifications within the promoter region may limit truncations due to secondary structures.

Conclusions

- Short and long read NGS were used as orthogonal methods to develop a high-level molecular profile of rAAV vectors.
- Long read sequencing exhibited reproducibility across different vector productions.
- Molecular heterogeneity within rAAV productions was determined by several factors including purification processes, vector expression cassette, vector backbone, and choice of capsid.
- Sophisticated bioinformatics analyses provided thorough profiling of the molecular species packaged into a rAAV capsid.
- Building AI platforms with novel characterization data may inform vector designs that can lead to reduction in cost and shorter timelines to a therapeutic product.