Leveraging HTS technologies to develop methods for in-depth characterization of rAAV products.

E Lecomte[1], S Maestro[1], S Stinus[1], DL Bowie[1], A Pérez San Vicente[1], A Navarro Nieto[1], O Aldasoro Zabala[1], L Agundez[2], F Bastida[2], N Pawlowski[3], A Lecanda Sanchez[3], R Palacios[4], S Cooper[4], R Fraser[4], S Brahma[5], J Smith[5], S Devi Velivela[5], C Trigueros Fernandez[1], A François[1]

[1]Viralgen Vector Core S.L., San Sebastián, Gipuzkoa, 20009, Spain. [2]TAAV Biomanufacturing Solutions, S.L.U, San Sebastián, Gipuzkoa, 20009, Spain.
[3]Bayer AG, Analytical Development & Clinical QC, Wuppertal, 42117, Germany. [4]Asklepios BioPharmaceutical, Inc. (AskBio), Edinburgh, EH25 9RG, United Kingdom.
[5]Asklepios BioPharmaceutical, Inc. (AskBio), Research Triangle Park, NC 27709, USA.

Introduction

With the success of clinical trials, gene therapy today may represent hope for patients suffering from genetic diseases. Recombinant adeno-associated virus (rAAV) products are extensively characterised prior to release. Among the **Critical Quality Attributes (CQAs)** used for release testing, those related to the identity and purity of rAAV products are scrutinized to ensure patient safety and treatment efficacy. Although quantitative polymerase chain reaction (qPCR) PCR and Sanger sequencing are currently among the gold standard techniques, the rapid development of high-throughput sequencing (HTS) technologies has paved the way for a much more exhaustive characterisation of rAAV products DNA content.

Studies carried out so far show that both short-read and long-read sequencing technologies are of importance for the characterization of DNA, each with its own strengths[1,2,3,4]. Short-read sequencing offers high-sequencing depth and a low sequencing error rate, enabling exhaustive identification and quantification of contaminating DNA, and detection of mutations at unprecedented levels. On the other hand, long-read sequencing technologies offer the opportunity to sequence DNA at single-molecule level, allowing the identification and quantification of truncation events in rAAV genomes as well as determining the size of contaminating DNA.

We have developed and established all the laboratory protocols, from sample and library preparation to sequencing, as well as the bioinformatics pipelines for providing information on CQAs, including:

- **Non-payload sequences characterization – PURITY** (identity, and relative quantification). **Illumina**
- **Non-payload sequences characterization – PURITY** (identity, and size distributions). **Oxford Nanopore Technologies**
- **Payload sequence INTEGRITY** (DNA lengths and truncation hotspots). **Oxford Nanopore Technologies**
- **Payload sequence IDENTITY** (identification of Single Nucleotide Polymorphisms, indels). **Illumina**

Materials and Methods

To evaluate the sensitivity and accuracy of each method, DNA controls were generated, used in spiking experiments, and then sequenced either using the NextSeq2000 sequencer (Illumina technology) or the MinION/GridION sequencers (Oxford Nanopore Technologies).

Briefly,
- to assess the limit of quantification (LoQ) of non-payload sequences, we used 3 different DNA mimicking the contamination of an rAAV product by plasmid DNA and host cell DNA, which were mixed in different proportions.
- to assess the DNA molecules size range that can be identified using ONT sequencing, we used a mixture of 6 DNA fragments of different lengths, from 186 to 7,988 bp, obtained by restriction enzyme digestion of a large plasmid.
- to assess the limit of detection (LoD) of AAV genome truncations, a deleted version of a 6,861 bp plasmid was generated by restriction enzyme digestion at position 4,436 bp, which was spiked in the full-length plasmid in different proportions.
- to assess the LoQ of SNP and InDel, we generated a plasmid with different types of mutations, which was spiked in the corresponding unmutated plasmid in different proportions.

Each sample was analyzed using the methods developed internally, following the 4 main steps indicated as follows:

Sample preparation → Library preparation → Sequencing → Analysis

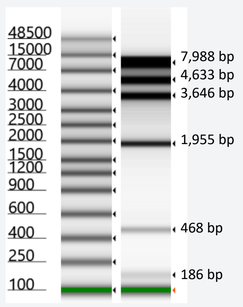
The same methods were then applied to a research-grade single-stranded AAV vector produced at Viralgen to characterize its DNA content. Both second-strand synthesis and thermal annealing approaches were used for sample preparation.

Method performance, focusing on sensitivity and accuracy

A

Sample condition	Reference sequence	Number of reads	EXPECTED proportion	EXPERIMENTAL proportion
DNA mixture mimicking **contaminant-free** rAAV production	rAAV sequence	38,707,476	100%	99.989%
	Helper plasmid	1,048	0%	0.003%
	Human genome + integrated Ad5 sequence	3,043	0%	0.008%
DNA mixture mimicking rAAV production containing **0.01% contaminants**	rAAV sequence	62,045,294	100%	99.972%
	Helper plasmid	7,742	0.010%	**0.012%**
	Human genome + integrated Ad5 sequence	9,588	0.010%	**0.015%**

B



C

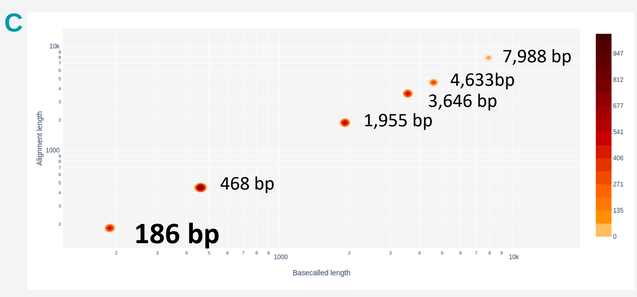
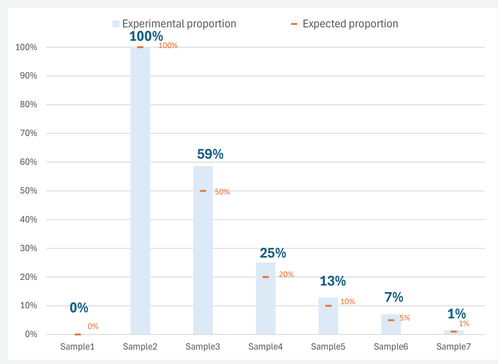


Fig 1. Non-payload sequences characterization - PURITY

(A) Three different DNAs were mixed in different proportions (100%, 0%, 0.01%), mimicking contamination of a rAAV production by a plasmid DNA and gDNA from the producer cell line, and sequenced using Illumina NextSeq2000. **The method allows to detect and quantify DNA down to a level of 0.01%.** (B) A mixture of 6 DNA fragments of different lengths, from 186 to 7,988 bp, was prepared and analysed by TapeStation, prior to ONT sequencing. (C) After alignment to the plasmid sequence, the 6 DNA fragments were retrieved, including the smallest fragment, showing that the **method allows to detect and identify DNA fragments with a size of at least 186 bp**. Read length (x-axis) and alignment length (y-axis) obtained with pycoQC[5].

A



B

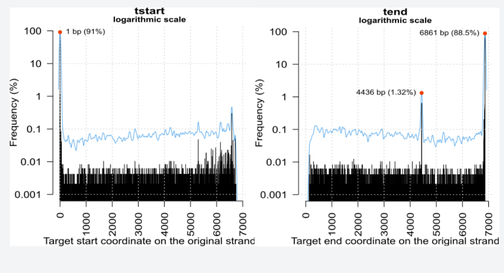


Fig 2. Payload sequence INTEGRITY

(A) After ONT sequencing, the percentage of truncation was assessed for 7 samples, containing only the full-length plasmid, only the truncated plasmid, or a mixture of both composed of 50%, 20%, 10%, 5%, and 1% of the truncated plasmid. (B) The start and end positions of the reads aligned on the full-length plasmid sequence were used to identify and quantify the truncation, following a normalization step. **The method allows to detect and quantify sequence truncation hotspots down to a level of 1%.**

A

Mutation	ALLELE FREQUENCY Sample1 100%	Sample2 50%	Sample3 20%	Sample4 10%	Sample5 5%
InsGAATTC	90%	42%	19%	8%	5%
T > A	99%	47%	18%	10%	5%
DelT	95%	45%	18%	9%	4%
InsG	95%	44%	16%	9%	4%
TG > GT	83%	40%	21%	9%	5%
InsTG	94%	45%	18%	9%	5%
GGTGTG > GGGG	92%	44%	17%	9%	4%

B

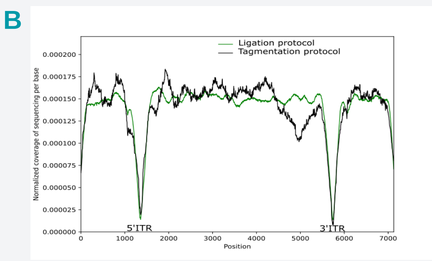
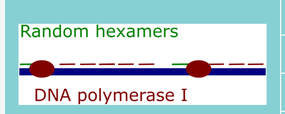
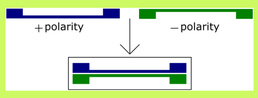


Fig 3. Payload sequence IDENTITY / FIDELITY

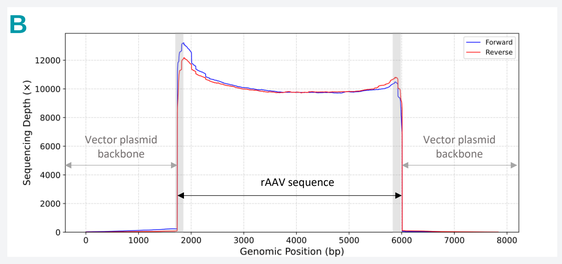
(A) Seven mutations were introduced into a plasmid, corresponding to 3 insertions, 2 deletions and 2 substitutions. The mutated plasmid was combined with the original plasmid at levels of 100%, 50%, 20%, 10%, and 5%. Samples were sequenced by NGS. **The method allows to detect and quantify with accuracy mutations down to a level of 5 %.** (B) A plasmid containing AAV ITRs was sequenced after PCR-free library preparation based on tagmentation or ligation to assess the sequencing coverage. Sequencing coverage is 6 to 7-fold lower in the ITRs compared to the rest of the plasmid but still considered as sufficient for ITR mutation detection thanks to the high-sequencing depth.

In-depth characterization of rAAV product

A

Sample preparation	Reference sequence	Number of reads	Relative quantification
Second-strand synthesis 	rAAV sequence	157,956	**97.15%**
	Vector plasmid backbone	3,055	1.88%
	Rep-Cap plasmid	1,166	0.72%
	Helper plasmid	182	0.11%
	Human genome + integrated Ad5 sequence	231	0.14%
Thermal-annealing 	rAAV sequence	164,169	**98.86%**
	Vector plasmid backbone	1,581	0.95%
	Rep-Cap plasmid	286	0.17%
	Helper plasmid	21	0.01%
	Human genome + integrated Ad5 sequence	11	0.01%

B



C

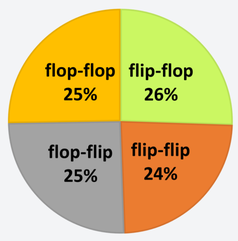


Fig 4. Characterization of a rAAV harboring CMV promoter and eGFP sequences, following either second-strand synthesis or thermal annealing.

(A) Percentages of DNA populations in the rAAV preparation. The main contamination comes from the vector plasmid backbone, as expected[6,7]. More DNA contaminants are detected after second-strand synthesis compared to thermal annealing. (B) Sequencing coverage obtained with long-reads along the vector plasmid sequence after thermal annealing. Positions of the 5' and 3' ITRs are highlighted in grey. The whole rAAV genome sequence is covered, including ITRs. (C) Pie chart summarizing percentages of flip and flop ITR conformations among all full-length genomes. The vector plasmid used to produce the rAAV contains a flop 5'-ITR and a flip 3'-ITR, which means that flop-flop, flip-flip, and flip-flop conformations appeared during rAAV genome replication in the transfected cells.

No truncation hotspots (at a level >1%) or mutations (at a frequency >5%) were identified in the gene of interest using long-reads and short-reads sequencing, respectively.

Discussion

Methods from sample preparation to bioinformatics analyses were developed to characterize both DNA starting materials (e.g. plasmids) and rAAV gene therapy products.

The use of DNA controls was essential throughout the development of the methods, including the bioinformatics part, to (1) detect and minimize biases, and (2) determine the sensitivity and accuracy of the methods, to be confident when analyzing the results obtained with more complex samples, such as rAAV.

Our methods were able (1) to identify and quantify Single Nucleotide Polymorphisms and indels with accuracy at a level of 5% (Fig. 3), (2) truncation hotspots were identified and quantified with accuracy at a level of 1% (Fig. 2), and (3) the presence of low levels (0.01%) of DNA contaminants was detected and quantified with accuracy (Fig. 1A). The method based on ONT sequencing was also able to detect and identify DNA fragments of different sizes down to 186 bp (Fig. 1C).

Illumina and Oxford Nanopore Technologies were then used for the in-depth characterization of DNA in a rAAV product. Since AAV genome is single-stranded and high-throughput sequencing technologies require double-stranded DNA, both second-strand synthesis and thermal annealing approaches were used prior to library preparation. The relative quantification of all the sequences was similar for both approaches, with more than 97% of the reads attributed to the rAAV sequence (Fig. 4A). However, the proportion of DNA contaminants quantified using second-strand synthesis was slightly higher compared to thermal annealing.

A majority of full-length rAAV genomes were sequenced by ONT (Fig. 4B), with flip and flop 5'-ITR and 3'-ITR conformations present in similar proportion (Fig 4.C).

Finally, no truncation hotspots or mutations were identified in the gene of interest using long-reads and short-reads respectively.

Conclusion

High-throughput sequencing is of great interest for providing information on 3 Critical Quality Attributes for rAAV products: **(1) payload sequence identity, (2) payload sequence integrity, and (3) residual DNA sequence identity and size distribution.**

HTS provides information which, when combined with other analytical methods such as analytical ultra-centrifugation or mass photometry for full/empty capsids characterization, ddPCR or qPCR for residual DNA quantification, and potency assays, will be beneficial to guide process development toward rAAV products with improved potency and safety profiles.

We believe our methods can be applied to the exhaustive characterization of both DNA starting materials (e.g. plasmids) and rAAV gene therapy products. The key parameters assessed in this study, i.e. sensitivity and accuracy, provided promising results and, consequently, validation of these HTS-based methods can be considered in the near future.

References

1. Lecomte E, et al. (2021). The SSV-Seq 2.0 PCR-free method improves the sequencing of adeno-associated viral vector genomes containing GC-rich regions and homopolymers. *Biotechnol. J.* 16(1):e2000016.
2. Tai PWL, et al. (2018). Adeno-associated virus genome population sequencing achieves full vector genome resolution and reveals human-vector chimeras. *Mol. Ther. Methods Clin. Dev.* 9, 130–141.
3. Namkung S, et al. (2022). Direct ITR-to-ITR nanopore sequencing of AAV vector genomes. *Hum. Gene Ther.* 33(21–22), 1187–1196.
4. Radukic MT, et al. (2020). Nanopore sequencing of native adeno-associated virus (AAV) single-stranded DNA using a transposase-based rapid protocol. *NAR Genom. Bioinform*. 2(4):lqaa074.
5. Leger A, et al. (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software*, 4(34):1236.
6. Chadeuf G, et al. (2005). Evidence for encapsidation of prokaryotic sequences during recombinant adeno-associated virus production and their in vivo persistence after vector delivery. *Mol Ther.* 12(4):744-53.
7. Lecomte E, et al. (2015). Advanced Characterization of DNA Molecules in rAAV Vector Preparations by Single-stranded Virus Next-generation Sequencing. *Mol Ther Nucleic Acids.* 4(10):e260.

EUSKO JAURLARITZA
GOBIERNO VASCO

Europar Batasunak kofinantzatua
Cofinanciado por la Union Europea