SSV-Seq 2.0, a more accurate PCR-free method for high-throughput sequencing of adeno-associated viral vector genomes

Emilie Lecomte[1], Sylvie Saleun[1], Mathieu Bolteau[1], Aurélien Guy-Duché[1], Oumeya Adjali[1], Véronique Blouin[1], Magalie Penaud-Budloo[1], and Eduard Ayuso[1]

[1]INSERM UMR 1089

May 18, 2020

Abstract

Adeno-associated viral vectors (AAV) are one of the most efficient engineered tools for delivering genetic material into host cells. The commercialization of AAV-based drugs goes hand in hand with the need to increase manufacturing capacities and to develop appropriate quality controls. In particular, accurate methods to assess the level of residual DNA in AAV vector stocks are needed, considering the potential risk of co-transferring oncogenic or immunogenic sequences with the therapeutic vectors. Our laboratory has developed an assay based on high-throughput sequencing (HTS) to exhaustively identify and quantify DNA species in recombinant AAV batches. Compiled with a computational analysis of the single nucleotide variants (SNV), Single-Stranded Virus Sequencing (SSV-Seq) also provides information regarding rAAV genome identity. In this study, we show that the PCR amplification of regions with high GC content and including mononucleotide stretches can be biased during the DNA library preparation, leading to drops in the sequencing coverage along the AAV vector genome. To circumvent this problem, we have developed a PCR-free protocol, named SSV-Seq 2.0, that is optimized for the sequencing of rAAV genomes that contain sequences with a high percentage of GC and homopolymers, such as the CAG promoter. HTS-based assays are indispensable to provide accurate data to the regulatory agencies regarding nucleic acids content in AAV vector batches and to improve the safety and efficacy of these viral vectors.

INTRODUCTION

The adeno-associated virus (AAV) is widely used as a viral vector to deliver therapeutic DNA. With the success of clinical trials using recombinant AAV (rAAV), the regulatory bodies have increased the level of requirements regarding the quality control (QC) of these new drugs. In particular, the presence of residual DNA in the final product is of significant concern due to the potential risk of oncogenicity, immunogenicity and decrease in gene transfer efficiency (Wright, 2014). The consequences of co-injecting DNA contaminants in patients with vectors depends on multiple criteria, such as; the type, the nature (i.e. free or encapsidated, fragmented, unmethylated) and the quantity of DNA impurities. To limit these risks, the Food and Drug Administration (FDA) recommends a level of residual host cell DNA (HCD) below 10 ng per parental dose (Food and Drug Administration, 2012), which might be difficult to not exceed in some cases, for example when high dose of AAV vectors is required to reach the therapeutic effect, such as for the treatment of Duchenne muscular dystrophy (Crudele and Chamberlain, 2019) or Spinal muscular atrophy (Al-Zaidy and Mendell, 2019). Quantification of HCD is most often based on real-time PCR, a targeted technique that only analyze a few numbers of DNA species. In addition, it is subjected to high variability due to a lack of harmonization (Ayuso et al., 2014; Dorange, F and Le Bec, C, 2018). To provide the community with a more exhaustive QC assay, our laboratory reported the Single-Stranded Virus Sequencing (SSV-Seq) method for the analysis of residual DNA in AAV vector stocks (Lecomte et al., 2019). SSV-Seq is based on Illumina high-throughput sequencing (HTS) and allows to identify and quantify all DNA impurities that are co-purified (encapsidated or not) with rAAV particles. The protocol has been adapted for the analysis of

AAV vectors generated either by plasmid transfection of HEK293 mammalian cells (Lecomte et al., 2015) or baculovirus infection of Sf9 insect cells (Penaud-Budloo et al., 2017). Using this method, we showed that DNA impurities mainly originate from the vector plasmid or the baculovirus genome for the HEK293- or Sf9-based manufacturing platform, respectively, with a predominance of residual sequences proximal to the inverted terminal repeats (ITR) (Penaud-Budloo et al., 2018a). In addition to the relative percentage of each DNA species, SSV-Seq can provide information regarding vector genome identity with a computational analysis of the single nucleotide variants (SNV) and the sequencing coverage over the recombinant AAV genome. Since then, and as sign of interest for HTS-based methods applied to rAAV quality control, other protocols have been developed to analyze the identity (Guerin et al., 2020; Maynard et al., 2019) or integrity (Radukic et al., 2019; Tai et al., 2018; Xie et al., 2017) of AAV vector genomes by sequencing.

In this study, we show that a high GC content and the presence of homopolymers in the AAV vector genome impaired the efficiency of PCR amplification during the preparation of sequencing library, leading to a decrease in coverage in the SSV-seq protocol. To solve this issue, we have optimized the library preparation using a PCR-free protocol. The novel method, SSV-Seq 2.0, has been used to analyze a vector genome harboring a CMV early enhancer/chicken beta-actin (CAG) promoter, that is well known to be a difficult template for PCR and sequencing. HTS-based assays represent the most exhaustive way to control the AAV vector quality and purity, and to fulfill the regulatory agencies requirements in term of residual nucleic acids.

MATERIALS AND METHODS

rAAV vector production and purification

The recombinant AAV2/8-CAG-GFP vector was produced in adherent HEK293 cells by double-plasmid transfection. The 6.6 kbp vector plasmid pAAV-CAG-GFP-SV40pA-ISceI harbors the 3179 bp recombinant AAV genome which is composed of the cytomegalovirus enhancer fused to the CMV early enhancer/chicken beta-actin (CAG) promoter, followed by the enhanced green fluorescent protein (eGFP) reporter gene and a simian virus 40 (SV40) polyadenylation signal. The rAAV genome is flanked by AAV2 inverted terminal repeats (ITR) from the plasmid pSub201 (Samulski et al., 1987). The co-transfected helper plasmid pDP8 (Grimm et al., 1998) contains the helper genes E2a, E4 and VA RNA of Adenovirus 5 and allows the expression of AAV2 Rep proteins under the mouse mammary tumor virus (MMTV) LTR promoter and a shortened p5 promoter and of the viral proteins VP of AAV serotype 8 from the natural p40 promoter. The AAV vector was produced and purified by ultracentrifugation on double cesium chloride gradient as described in Ayuso et al. (Ayuso et al., 2010). The titer in vector genomes (vg) was determined by real-time PCR targeting the AAV2 ITRs and following the conditions described by D'Costa et al. (D'Costa et al., 2016).

Identification of GC-rich regions and homopolymers in rAAV vector sequence

GC-rich regions were identified in the recombinant AAV2/8-CAG-GFP vector sequence using the program NTContent (*http://github.com/emlec/NTContent*), included in the SSV-Conta package. The following parameters were used: window size, 200; step size, 20 or window size, 50; step size, 25. Mononucleotide repeats composed of at least six nucleotides and simple sequence repeats (SSR) were localized along the AAV vector genome using the MISA-web server (*https://webblast.ipk-gatersleben.de/misa/*) (Beier et al., 2017) and the following parameters: SSR motif length/min. no. of repetitions, 1/6, 2/2, 3/2, 4/2, 5/2, 6/2, 7/2, 8/2, 9/2, 10/2 and max. length of sequence between two SSRs to register as compound SSR, 100, output file parameter, GFF.

Preparation of fragmented PhiX174 DNA

The sequencing libraries were prepared using PCR-free kits from 200 ng of fragmented PhiX174 DNA. For the fragmentation, 1.5 µg of PhiX174 RF II DNA (NEB, Ipswich, MA) was diluted in a final volume of 100 µL TE 10:1 in 0.5 mL Bioruptor microtubes and sonicated using Bioruptor UCD-200 (Diagenode, Seraing, Belgium) at LOW power (160 W) during 12 pulses of 30s ON/90s OFF. A buffer exchange was performed with 10 mL Tris-HCl pH8.0 using the kit NucleoSpin Gel and PCR Clean-up (Macherey-Nagel, Düren, Germany). The profile of the fragmented PhiX DNA was controlled on the Agilent Bioanalyzer 2100 instrument using the High Sensitivity DNA kit (Agilent Technologies, Santa Clara, CA), and the average fragment size was determined to be 289 bp. DNA was quantified using the Qubit 1X dsDNA High Sensitivity Assay kit (ThermoFisher Scientific, Waltham, MA) before library preparation.

Library preparation for Illumina sequencing

For the original SSV-Seq protocol including a PCR amplification step, DNA sequencing libraries were prepared from 2x10^{11}vector genomes (2 replicates of 1x10^{11} vg) as determined by free ITR qPCR (D'Costa et al., 2016) and 200 ng of double-stranded DNA (dsDNA) as determined by spectrophotometry using the Nanodrop OneC (ThermoFisher Scientific, Waltham, MA). The complete SSV-Seq method is described in Lecomte et al. (Lecomte et al., 2019).

Three kits were tested for the PCR-free library preparation: KAPA HyperPrep kit (Roche, Basel, Switzerland), NxSeq AmpFREE Low DNA kit (Lucigen, Middleton, WI) and NEBNext Ultra II (New England Biolabs, Ipswich, MA). The PCR-free sequencing libraries were prepared following the instructions of the suppliers, at the exception of the purification steps and the adapter ligation. The adapters of the three kits were replaced by home-made Illumina-compatible P5/P7 adapters (Lecomte et al., 2019) and DNA purifications were carried out using the SPRIselect reagent (Beckman Coulter, Brea, CA).

The SSV-Seq 2.0 protocol was realized from 8x10^{11}vector genomes (4 replicates of 2x10^{11} vg) of a rAAV vector batch. After DNA extraction and second-strand synthesis (Lecomte et al., 2019), the concentration of dsDNA was determined by fluorimetry using the Qubit 1X dsDNA High Sensitivity Assay kit (ThermoFisher Scientific, Waltham, MA). Two tubes per sample were prepared with 150 ng of DNA in a final volume of 100 µL TE10:1 pH8. DNA was sonicated using the Bioruptor UCD-200 (Diagenode, Liege, Belgium) and in accordance with the conditions described in Lecomte et al. (Lecomte et al., 2019) to reach an average target size centered on approximatively 300 bp. The fragmented DNA of the two tubes was pooled (300 ng) and purified using 1.6X of SPRIselect reagent (Beckman Coulter, Brea, CA). The magnetic beads with bound DNA were then washed two times with 360 µL of freshly prepared ethanol 80% and DNA was eluted in 20 µL of ultrapure DNase/RNase free distilled water (dH2O). Libraries were then prepared using the NxSeq AmpFREE Low DNA kit (Lucigen, Middleton, WI) that combined the end-repair and the A-tailing steps. The following mix was prepared in 0.2 mL PCR tube: 17 µL of previously sheared and purified DNA, 25 µL of 2X buffer and 8 µL of enzyme mix. The one-step reaction was realized using the Applied Biosystems Veriti Thermal Cycler (ThermoFisher Scientific, Waltham, MA, USA) for 20 min at 25°C with a 72°C heated lid, followed by a 20-min cycle at 72°C and hold at +4°C. Illumina-compatible P5/P7 adapters were prepared as described in Lecomte et al. (Lecomte et al., 2019) and diluted at 15 µM in dH2O. After DNA repair and A-tailing, 3 µL of diluted adapters and 4 µL of ligase were added in the 50-µL reaction volume. Adapter ligation took place in a thermocycler for 30 min at 25°C and was immediately followed by a double-1X SPRI purification. Each SPRI purification step includes two washes with 180 µL ethanol 80%. The first elution was performed in 50 µL of dH2O and the final elution was done in 16 µL of ultrapure distilled water.

Quality control of the DNA sequencing libraries

The quality of the DNA libraries was controlled by electrophoresis on a microchip using the High Sensitivity DNA kit (Agilent Technologies, Santa Clara, CA). Electrophoregrams were obtained after migration in the Agilent Bioanalyzer 2100 instrument and after analysis using the Agilent 2100 Expert Software. The total DNA concentration of the libraries was determined on the Qubit 4 fluorometer using the Qubit 1X dsDNA High Sensitivity Assay kit (ThermoFisher Scientific, Waltham, MA). Adapter-ligated DNA fragments were quantified by real-time PCR using the Universal qPCR KAPA SYBR Fast kit (Roche, Basel, Switzerland) before Illumina sequencing.

Illumina sequencing

One percent of PhiX Control v3 DNA (Illumina, San Diego, CA) was spiked into DNA libraries before sequencing. The libraries were denatured and diluted according to the Denaturing and Diluting Libraries for the HiSeq protocol (Part # 15050107 v03). The cluster generation was realized using the cBot system and the HiSeq Rapid PE Cluster Kit v2 (Illumina, San Diego, CA). The high-throughput sequencing was performed using the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA) on the Illumina HiSeq 2500 system (Illumina, San Diego, CA) with the following parameters: rapid run paired-end mode and read size of 94 bp.

Bioinformatics Analysis

Base call (BCL) files were converted into FASTQ files with the Illumina bcl2fastq2 Conversion Software (Illumina, San Diego, CA). Programs that are included in the SSV-Conta package (*https://github.com/emlec/SSV-Conta*) were then used to quantify and characterize all DNA species that are present in a rAAV vector lot: Quade, a FASTQ files demultiplexer, Sekator, an adapter trimmer, RefMasker to mask sequence homologies and ContaVect to analyze residual DNAs (Lecomte et al., 2019). Briefly, FASTQ files were demultiplexed with Quade according to their barcodes. The paired-end reads were assigned to a sample when the combination of the two barcodes (index read 1 and index read 2) was correct and if each base of the barcodes had a PHRED quality score of at least 25. Passed paired-end reads were trimmed using Sekator, according to the sequence quality and removing the adapter, as described in Lecomte et al (Lecomte et al., 2019). The distribution of residual DNA was determined using RefMasker and ContaVect programs. The reference sequences were indicated in the ContaVect configuration files in the following order: the phage φX174 genome (GenBank accession number J02482.1), the phage λ genome (J02459.1), the rAAV genome, the plasmid backbone sequence, the plasmid helper sequence, the adenovirus 5 (Ad5) sequence (nucleotides 1 to 4344 of the Human adenovirus 5 complete genome, AC_000008) and the human genome (GRCh38 primary assembly). Using RefMasker, homologies between two reference sequences were masked on the second reference sequence in the list order, replacing homologous nucleotides with an N base symbol. ContaVect was run, applying the following main parameters: minimum mean read quality, 30; minimum quality mapping for read validation, 20; minimum mapping size, 25 bases. Unmapped and mapped reads that did not fulfill these criteria were excluded. Sequencing coverage along each base of the vector plasmid was generated using the program SSV-Coverage, a program included in the SSV-Conta package. Sequencing data have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under the accession number PRJEB38306 (*https://www.ebi.ac.uk/ena/data/view/PRJEB38306*).

Graphical representations

Graphs were generated using the Python plotting library Matplotlib. Figures were post-processed using Inkscape v0.92.3 software to add captions.

Statistics

Statistical analysis was applied to samples with at least 3 replicates. Data were expressed as mean ± standard deviation. One-tailed nonparametric Mann-Whitney test was performed to compare two independent groups. Differences were considered statistically significant at *p[?]0.05. Analyses were performed using GraphPad Prism v5.01.

RESULTS

Analysis of the sequencing coverage of rAAV vector genomes containing GC-rich regions and mononucleotide stretches

SSV-Seq is a powerful technique intended, primarily, for the analysis of residual DNA in AAV vector lots. However, based on high-throughput sequencing, this method also provides information about the vector genome identity, including the identification of single nucleotide variants (SNV) and indels. The SSV-Seq protocol consists of the following successive experimental steps (**Figure 1**) (Lecomte et al., 2019): (1) a facultative DNase pretreatment, (2) the DNA extraction from rAAV stocks, (3) the second-strand DNA synthesis using random hexamers, (4) the library preparation and (5) the Illumina sequencing. Briefly, before particle disruption, residual DNA that is not protected by the AAV capsid can be removed by a facultative nuclease treatment. After DNA extraction, the single-stranded rAAV genome is converted into double-stranded DNA (dsDNA) using random hexamers to generate a template compatible with Illumina sequencing library preparation workflow. Illumina-compatible sequencing libraries are then prepared using a custom protocol (**Figure 1, protocol**). DNA is sheared by sonication, end-repaired and A-tailed and adapters are ligated*via* a 3-prime T-overhang. DNA fragments that are flanked by adapters are amplified by 15 cycles of PCR. Finally, the library is paired-end sequenced using the Illumina HiSeq platform and the data are processed using our dedicated bioinformatics pipeline (*https://github.com/emlec/SSV-Conta*) (Lecomte et al., 2019).

The PCR amplification step has already been described as being the principal source of bias during sequencing library preparation (Aird et al., 2011). Indeed, AT- (Oyola et al., 2012) and GC-rich (Benjamini and Speed, 2012) fragments are less efficiently amplified than other regions in the genome which can lead to a bias resulting of a lower sequencing coverage. Consequently, sequencing library preparation protocols including a PCR step can cause an uneven distribution of reads coverage across the DNA. In the context of AAV vector analysis by HTS, this bias could cause an underestimation of AT- and/or GC-rich sequences and the impossibility to identify SNV and indels in these regions.

To determine more precisely the impact of the base composition on the Illumina sequencing coverage, we first developed a new program, named NTContent (*https://github.com/emlec/NTContent*). This program is based on a sliding window analysis and requires a DNA sequence in FASTA format as input. It generates a tab-delimited text file composed of two columns indicating for each given position the percentage of the requested nucleotide combination (**Supplementary Figure S1**). The NTContent program was applied to a rAAV genome sequence containing the CAG promoter followed by the GFP reporter gene and the SV40 polyadenylation signal sequence. The CAG promoter was chosen because it has high GC content and it is known to be difficult template for PCR amplification. **Figure 2a** shows the percentage of GC along the rAAV genome, the sequencing coverage obtained by SSV-Seq from a rAAV8 vector batch and from its corresponding plasmid vector. Both for the plasmid and the rAAV sample, two major drops in the sequencing coverage appeared in the CAG promoter around position 666 (asterisk, region 1) and position 1421 (asterisk, region 2) of the rAAV genome. The superimposed graphs revealed that the two sequencing drops in the CAG promoter are clearly related to a high GC content. A more precise analysis of the percentage of GC was performed using the same program NTContent but taking into account the nature (A, T, G, or C) of 50 successive bases with a step of 25 bases. This analysis showed that the two steep drops coincide with regions composed of more than 90% of GC **(Supplementary Table S1).**

In order to further investigate the origin of the sequencing drops, the presence of A, T, C, and G nucleotide stretches along the rAAV genome was analyzed (**Figure 2b-c**). Of note, long G/C homopolymers have already been reported as a source of bias during sequencing on Illumina systems such as HiSeq instrument (Shin and Park, 2016). A succession of C, T and G homopolymers was observed in region 1 of the CAG promoter between position 588 and 676 (**Figure 2b**). C and G stretches were also detected in region 2, upstream the steep drop, between position 1290 and 1391 (**Figure 2c**). A previous study suggested that the presence of repetitive mononucleotides at the active site of a polymerase can lead to its dissociation from the DNA (Fazekas et al., 2010). For SSV-Seq, the PCR amplification is realized using the PfuUltra II Fusion HotStart DNA Polymerase. The number of nucleotides that fills the active site of this polymerase is 6. Thus, the software MISA was used to seek nucleotide repeats equal or greater than 6. Regions 1 and 2 of the CAG promoter contain 6 out of 14 homopolymers of [?] 6 bases detected in the AAV vector genome (**Supplementary Table S2**). In particular, the first region showing a drastic coverage drop includes two stretches of 8 and 16 mononucleotides (C and G, respectively).

Overall, we can conclude that drastic drops in the sequencing coverage correlate with the presence of long stretches of G and C nucleotides in the rAAV vector genome which is consistent with the very high GC percentage illustrated on **Figure 2a** . It remains to be determined if this bias occurs at the PCR step and/or during the HiSeq Illumina sequencing. To address this question, a PCR-free protocol has been developed and compared to the PCR-enriched SSV-Seq method.

Development of a PCR-free protocol for the sequencing library preparation

Two parameters are critical for adapting the SSV-Seq library preparation protocol to a PCR-free method: (i) the reduction of the number of steps to avoid DNA loss during beads-based cleanup and (ii) the use of appropriate adapters (**Figure 1**). SSV-Seq adapters are suitable for use in a PCR-free protocol (Kozarewa et al., 2009) because they contain all elements required for the bridge amplification on Illumina flow cells, i.e. the sequences complementary to the flow cell oligonucleotides, the sequence targets of the P5/P7 sequencing primers and 6 base-indexes. Among PCR-free kits that are commercially available, 6 kits were first selected because of their compatibility with Illumina technology (**Supplementary Table S3)** . Three out of 6 kits were tested, i.e. KAPA HyperPrep (Roche), NxSeq AmpFREE Low DNA (Lucigen) and NEBNext Ultra II (New England Biolabs) based on the following criteria: (i) a low amount of fragmented DNA ([?] 200 ng) is required as input, (ii) the end repair and A-tailing steps are combined into a unique step, (iii) home-made adapters can be used and (iv) kits are compatible with Illumina paired-end sequencing. Libraries were prepared from 200 ng of fragmented PhiX174 DNA following the instructions of each kit, at the exception of two steps: the SSV-Seq adapters were used for the ligation instead of the commercial ones, and the post-ligation cleanup and size selection steps were replaced by a double purification with 1X SPRI beads (**Figure 1, SSV-Seq 2.0**). Libraries were prepared in triplicate to determine the robustness of each protocol. The efficiency of the 3 kits for generating sequencing libraries was compared qualitatively and quantitatively. DNA quality was controlled by high-sensitivity capillary electrophoresis **(Figure 3a**). The electrophoregrams obtained on Agilent chip showed negligible amounts of free adapters in the final DNA libraries after the two SPRI purification steps whatever the kit used. Then, the number of adapter-ligated molecules in the libraries was quantified by qPCR Kapa assay using primers targeting the P5 and P7 sequences of the adapters, that correspond to the binding sequences to the Illumina flow cell (**Figure 3b**). The ligation step of the kit NxSeq was the most efficient, leading to a library DNA concentration of 8.2 nM. For this reason, the NxSeq AmpFREE Low DNA kit was preferred over the two other kits and was included in the novel SSV-Seq 2.0 method for the PCR-free sequencing library preparation.

The optimized PCR-free protocol was then tested for the analysis of a rAAV vector batch, in parallel to the original SSV-Seq technique (Lecomte et al., 2019). The new protocol was applied to a purified rAAV8-CAG-GFP vector batch that was produced in HEK293 cells by plasmid transfection (**Figure 1, PCR-free protocol**). The initial amount of rAAV vectors needed for the analysis and determined by free ITR qPCR (D'Costa et al., 2016) had to be increased from 2x10^{11}vector genomes for the original SSV-Seq method to 8x10^{11} vg for the SSV-Seq 2.0. For each protocol, an additional replicate was used to control the second strand synthesis step. After DNA extraction and second strand synthesis, 300 ng of fragmented DNA, as determined by fluorometric quantification, was used as input for the PCR-free library preparation. The PCR-free library DNA, devoid of free adapters, was quantified by qPCR Kapa. The concentration of adapter-ligated fragments was on average 4.5 +- 0.2 nM in a final volume of 16 µL compared to a mean of 49.8 ± 5.1 nM in a final volume of 30 µL for the PCR-mediated protocol, which is above the minimal concentration required for Illumina sequencing. In conclusion, our optimized SSV-Seq 2.0 protocol allows to reach the quality and quantity of library DNA necessary for the analysis of AAV vectors by HiSeq Illumina sequencing.

The SSV-Seq 2.0 protocol improves the sequencing coverage along regions of the rAAV genome that are both rich in GC and homopolymers

After sequencing of the DNA library prepared from the AAV8-CAG-GFP vector following the PCR-free novel method, the sequencing reads were passed through our dedicated bioinformatics pipeline, named SSV-Conta (Lecomte et al., 2019). SSV-Conta is intended for the determination of the proportion of residual DNA species in a rAAV batch and of the analysis of the coverage along the vector genome. The percentage of reads that passed the quality and adapter trimming steps were higher than 94% for both protocols, although a slightly lower percentage was observed for the PCR-free libraries (**Supplementary Table S4).**The filtered reads were then aligned to the vector plasmid to visualize the sequencing coverage along the two GC-rich regions in the CAG promoter described above. Using the PCR-free protocol, the coverage along these regions was significantly restored compared to the PCR-enriched protocol (**Figure 4**), indicating that the PCR amplification step is one of the major causes of artefactual drop in the sequencing coverage. In addition to the rAAV vector genome, read alignment was realized for other DNA species, i.e. the vector plasmid backbone, the helper plasmid and the HEK293 cell genome. The number of reads aligned to each reference is shown in **Supplementary Table S5** . Overall, a minimum of 15.2 M and 23.8 M reads per sample was mapped to the known references for the PCR-free and PCR protocol, respectively. Finally, the percentage of each DNA species was calculated as described in Lecomte et al. (Lecomte et al., 2019) (**Table 1**). Similar to the SSV-Seq method (Lecomte et al., 2015), the novel SSV-Seq 2.0 protocol is highly reproducible, as indicated by the coverage graph of each replicate (**Figure 4**). Consistently with a better coverage along the rAAV genome and less-biased sequencing of the GC-rich regions, the optimized PCR-free method leads to a higher percentage of reads aligned to the rAAV-CAG-GFP genome (93.9 ± 0.4% and 91.9 ± 0.3% of the total mapped reads for PCR-free and PCR protocols, respectively). As described in our previous study (Lecomte et al., 2015), the predominant DNA contaminant originates from the vector plasmid backbone. The relative percentage of this contaminant was reduced using the PCR-free protocol, since more reads were attributed to the rAAV genome (5.7±0.4% of the total mapped reads for SSV-Seq 2.0 compared to 7.6±0.3% for SSV-Seq). Consequently, the SSV-Seq protocol slightly overestimates the percentage of DNA contaminants when the rAAV vector genome is composed of sequences that are difficult to amplify by PCR. In conclusion, the SSV-Seq 2.0 method is the most accurate approach for the high-throughput sequencing analysis of AAV vector genomes that contain regions with a high level of GC and homopolymers.

DISCUSSION

The goal of this manuscript was to develop a more accurate method to characterize DNA species contained in rAAV batches. Several technological platforms exist for the manufacturing of rAAV vectors for their use in gene therapy, either using mammalian or insect cells (Penaud-Budloo et al., 2018b). It is known that both upstream and downstream processes may impact the purity of the final product, including the amount and type of residual DNA. In order to assess the risk for the patient to co-transfer undesired DNA sequences with AAV vectors, an exhaustive identification and quantification of these DNAs is of utmost importance and can be achieved thanks to methods based on high-throughput sequencing technologies. We have previously described a protocol based on Illumina sequencing, called Single-Stranded DNA Virus Sequencing (SSV-Seq), to control rAAV purity in term of DNA contaminants (Lecomte et al., 2019, 2015). This protocol includes a PCR step during the library preparation, which could be affected by some type of bias inherent to the PCR technique, such as the presence of AT- (Oyola et al., 2012) and GC-rich regions (Aird et al., 2011). Several solutions have been proposed to reduce these artifacts, either by optimizing PCR conditions (Quail et al., 2011) or by developing alternative methods for library amplification (van Dijk et al., 2014). Here, we decided to be more drastic in order to improve our SSV-Seq protocol shifting towards a PCR-free library preparation kit. Our study clearly shows a correlation between a high GC and homopolymers content and a poor sequencing coverage. In order to avoid data misinterpretation, for example as a large deletion or a biological under-representation of a particular sequence in the rAAV particles population, it is of great importance to screen the rAAV genome for GC-rich regions and homopolymers prior to sequencing-based analysis. To this purpose, the software MISA and the new bioinformatics tool NTContent developed here (available at*https://github.com/emlec/NTContent)* can be extremely useful as prediction tools. In order to monitor any potential bias in SSV-Seq, an internal normalizer is also processed in parallel to the rAAV samples. Composed of a mix of the plasmid vector and other potential residual DNA species (producer cell DNA, helper plasmids), this control enables to visualize and compare the sequencing coverages obtained from the rAAV sample and from the plasmid vector, as shown on **Figure 2** .

A coverage drop in the CAG promoter, which has a local GC percentage higher than 90%, was previously observed using SSV-Seq (Kondratov et al., 2017). The same observation has been reported by another group using Fast-Seq, a technique based on Tn5 tagmentation (Maynard et al., 2019). Kondratov et al. have shown that a PCR-free protocol could outperformed a PCR-enriched method (8 amplification cycles) in term of sequencing coverage along GC-rich regions of the AAV vector genome (Kondratov et al., 2017). The authors used the Accel-NGS 2S PCR-Free DNA Library Kit from Swift Biosciences to prepare libraries. An initial amount of 4x10^{11} vg of a rAAV5-CAG-GFP vector and an input of 220 ng of dsDNA was used in their protocol. The Accel-NGS workflow includes two DNA repair steps and two adapter ligation steps and requires the use of specific adapters that are not compatible with low-throughput applications. Similar to our data, the authors were able to reduce the sequencing bias due to a high GC content by evicting PCR amplification, although coverage drops were still detected in the CAG promoter. Therefore, biases due to the sequencing technology need to be further assessed. Indeed, all sequencing technologies exhibit error-rate biases in low- ([?]10%) and high-GC ([?]75%) regions, and at long homopolymers (Ross et al., 2013). G-rich sequences can also be at the origin of sequence-specific errors (SSE) using Illumina technology (Dohm et al., 2008) and may cause false SNV (Shin and Park, 2016). However, in our study, no SNV has been observed in the CAG promoter.

Using our SSV-Seq 2.0 PCR-free method, we still detected minor coverage drops, as for example within the eGFP transgene (**Figure 4**). Independent of a PCR amplification bias, this could be related to the sequencing technology itself. Indeed, MiSeq sequencing that used the same four-channel sequencing chemistry than HiSeq has been shown to disfavor the “CCNGCC” motif in the GFP coding sequence (Van den Hoecke et al., 2016). On the other hand, sequencing technologies such as single molecule real-time (SMRT) sequencing (Pacific Biosciences) is described as giving a less biased coverage across GC-rich regions (Ross et al., 2013). Offering long read lengths, single molecule sequencing technologies also allow to study rAAV vector genome integrity (Radukic et al., 2019; Tai et al., 2018; Xie et al., 2017). Interestingly, rAAV genome truncations have been detected at hairpin-like structures using the AAV-GPseq SMRT-based assay, creating self-complementary viral genomes (Xie et al., 2017). Improving rAAV genome sequencing, and particularly through ITR and ITR-plasmid junctions would also be of great interest in the field. Recently, an HTS-based assay has been developed to identify off-target nuclease activity after AAV-mediated genome edition *in vivo* (Breton et al., 2020). PCR and adapter optimizations have been realized in this protocol, named ITR-Seq, to specifically amplify ITR-genomic DNA junctions. Combining multiple sequencing technologies could provide complementary information and reduce the risks associated to inherent technical errors of each platform. For instance, SSV-Seq based on Illumina technology that gives a high sequencing depth is likely the preferred method to identify and characterize residual DNA in rAAV stocks and perform SNV analysis, while AAV-GPseq based on SMRT sequencing is more adapted to AAV vector genome integrity (truncated rAAV genomes). The novel SSV-Seq 2.0 protocol allows to circumvent PCR biases and improves the HTS analysis of rAAV genomes harboring regions with high percentage of GC content and long mononucleotide stretches, such as those often found in promoters.

FIGURE LEGENDS

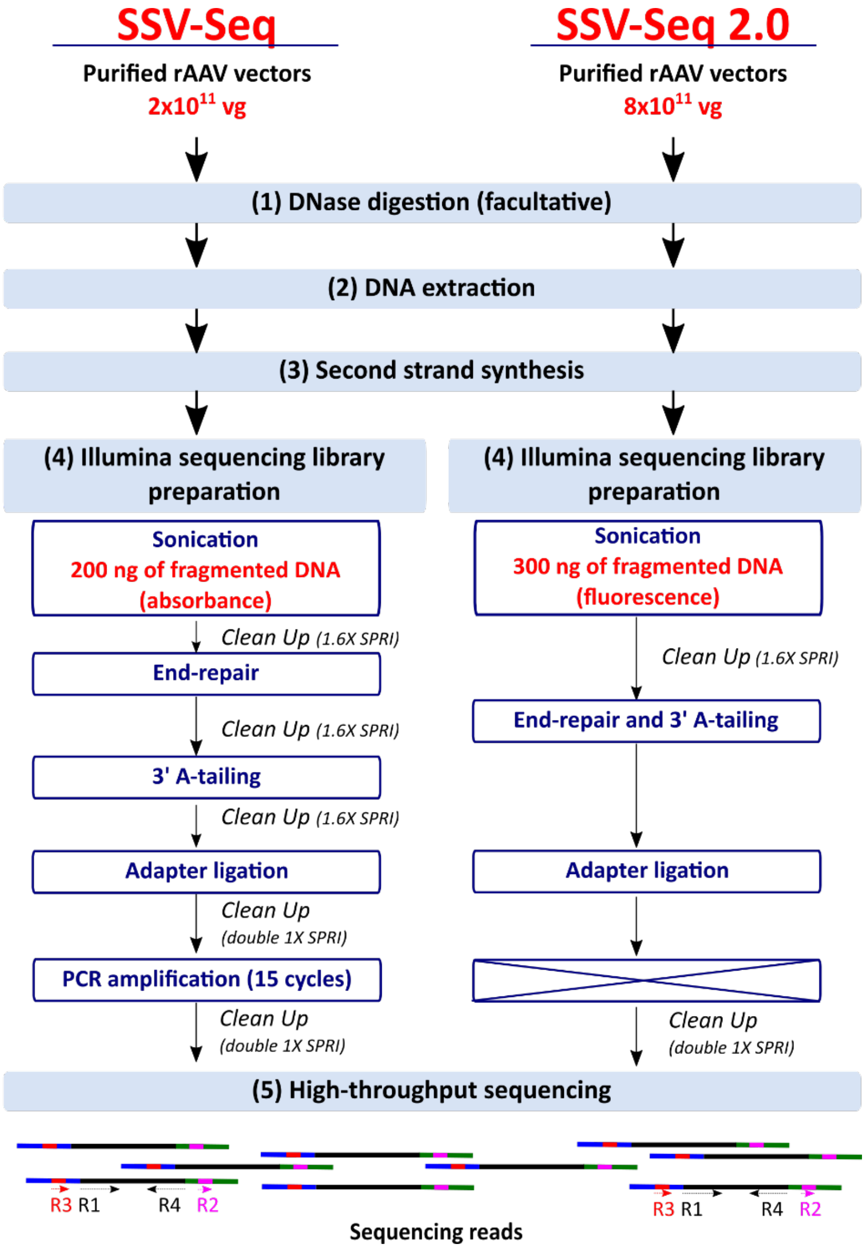
SSV-Seq
Purified rAAV vectors
2x10^11 vg
SSV-Seq 2.0
Purified rAAV vectors
8x10^11 vg
(1) DNase digestion (facultative)
(2) DNA extraction
(3) Second strand synthesis
(4) Illumina sequencing library preparation
Sonication
200 ng of fragmented DNA (absorbance)
Clean Up (1.6X SPRI)
End-repair
Clean Up (1.6X SPRI)
3' A-tailing
Clean Up (1.6X SPRI)
Adapter ligation
Clean Up (double 1X SPRI)
PCR amplification (15 cycles)
Clean Up (double 1X SPRI)
(4) Illumina sequencing library preparation
Sonication
300 ng of fragmented DNA (fluorescence)
Clean Up (1.6X SPRI)
End-repair and 3' A-tailing
Adapter ligation
Clean Up (double 1X SPRI)
(5) High-throughput sequencing
R3 R1 R4 R2
R3 R1 R4 R2
Sequencing reads

Figure 1. SSV-Seq 2.0 workflow. The SSV-Seq protocol was described in Lecomte et al. (Lecomte et al., 2019) (left panel). The optimized SSV-Seq 2.0 protocol is represented on the right side. A total quantity of 8x10^{11} vector genomes of purified rAAV vector sample is required as input. A pretreatment with a DNases cocktail can be performed prior to DNA extraction to specifically identify and quantify DNA that are encapsidated in rAAV capsids. A second strand synthesis step is carried out, followed by the PCR-free DNA library preparation. Finally, high-throughput sequencing is performed using Illumina HiSeq platform (rapid run 2x94 pb).

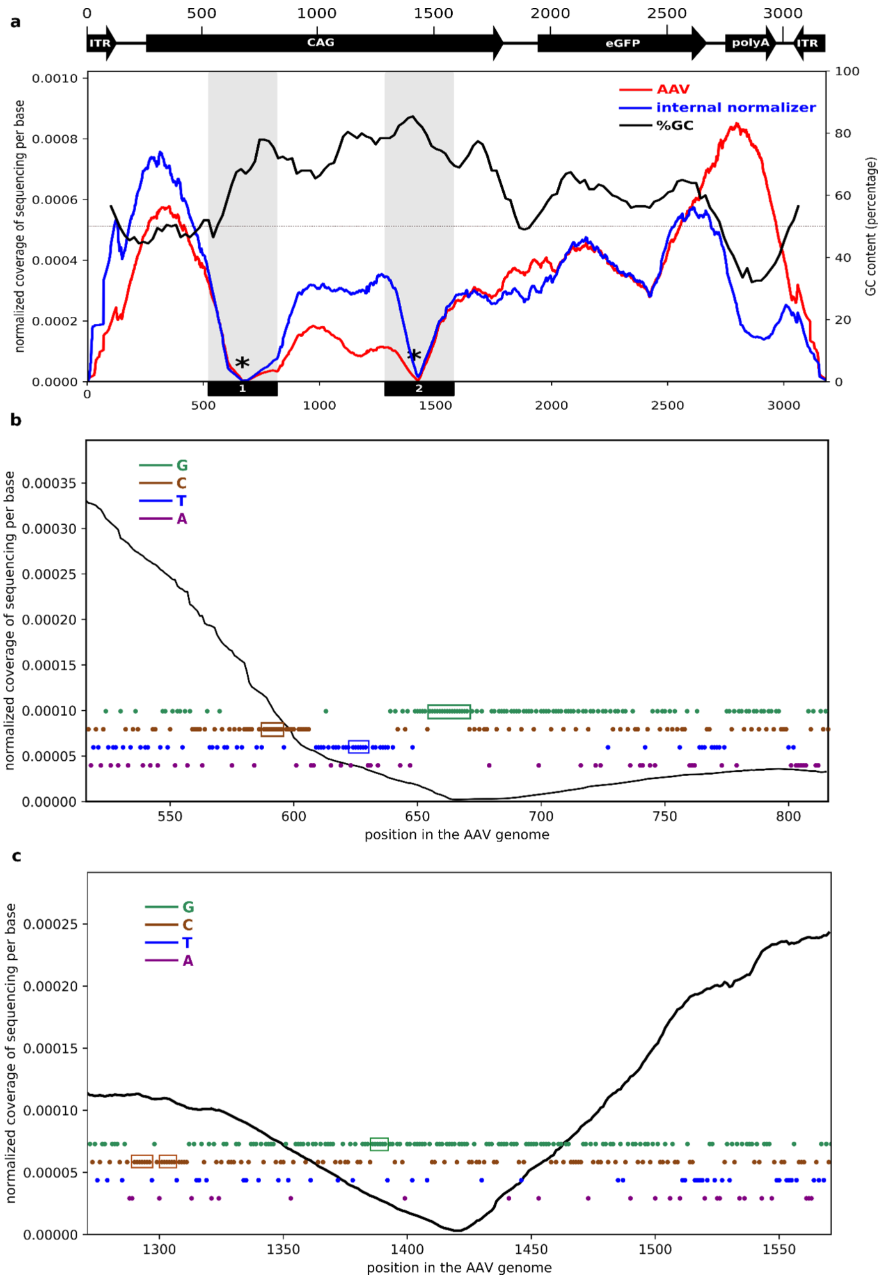

Figure 2. Impact of GC and homopolymers content on the sequencing coverage of the recombinant AAV genome. (a) Sequencing coverage and percentage of GC along the AAV vector genome. The sequencing coverages obtained from the 3.2-kb AAV8-CAG-GFP vector (red) and the internal normalizer (vector plasmid) (blue) were normalized by dividing the read coverage at each base by the sum of the coverage for all bases mapped along the rAAV genome. The grey boxes indicate two 300 bp-regions showing a drastic sequencing coverage drop. Region 1 and 2 were centered around the minimal number of reads at position 666 and 1421 of the rAAV genome, respectively. The percentage of GC (black) was determined using the program NTContent with the following parameters: window size, 200 bases and step, 20 bases. The rAAV genome map is represented above the graph. **(b, c)**Nucleotide content along regions 1 (b) and 2 (c). Each base was represented at each position by a colored dot: G (green), C (brown), T (blue) and A

(purple). Colored boxes represent homopolymers of [?] 6 nucleotides. Magnified sequencing coverages are represented as black lines.

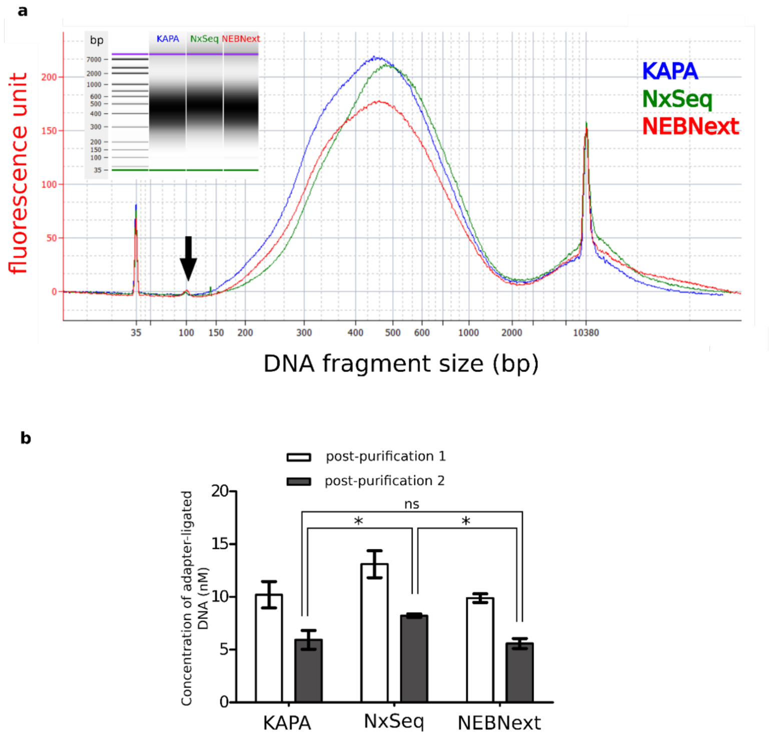

Figure 3. Comparison of PCR-free kits for the preparation of Illumina-compatible sequencing libraries. Libraries were prepared in triplicate from 200 ng of fragmented PhiX174 DNA (average fragment size 289 bp) using three different PCR-free kits: KAPA HyperPrep (KAPA), NxSeq AmpFREE Low DNA (NxSeq) and NEBNext Ultra II (NEBNext).**(a)** Size distribution profiles of DNA libraries. The quality of DNA libraries was determined on Agilent Bioanalyzer 2100 instrument. One representative electrophoregram per kit is shown (KAPA in blue, NxSeq in green and NEBNext in red). The black arrow indicates the localization of free adapter dimers. **(b)** Concentration of adapter-ligated fragments determined by qPCR KAPA. The concentration of DNA fragments that are ligated with two adapters was determined by qPCR

KAPA, after the first (post-purification 1) and the second (post-purification 2) SPRI beads purification step. The concentration obtained in a total volume of 50 µL after the first purification step was normalized to the final volume of the libraries (20 µL). Bars represent the mean concentration ± standard deviation of libraries from 3 replicates. Concentrations obtained were compared by a one-tailed Mann-Whitney test. ns, p > 0.05; *, p [?] 0.05.

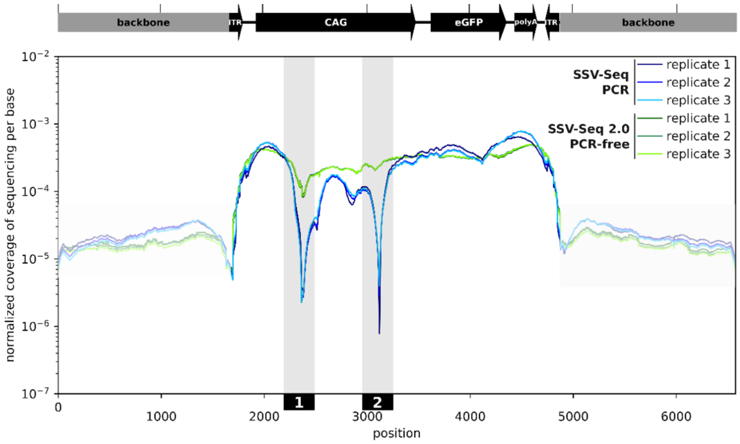

Figure 4. Sequencing coverage along the rAAV vector genome and the plasmid backbone. Sequencing libraries were prepared from a purified rAAV-CAG-GFP vector batch following the original SSV-Seq protocol (blue shades) or SSV-Seq 2.0 PCR-free protocol (green shades). Each library was prepared in triplicate. The sequencing coverage is normalized by dividing the read coverage at each base by the sum of the coverage for all bases mapped along the vector plasmid. Grey boxes represent the two 300 bp-region where the number of reads drops. The vector plasmid map is represented above the graph.

Reference sequence	Replicate	SSV-Seq (PCR) (%)	SSV-Seq 2.0 (PCR-free) (%)
rAAV genome	1	91.58	93.62
	2	91.87	93.70
	3	92.15	94.36
Vector plasmid backbone	1	7.92	6.02
	2	7.64	5.61
	3	7.33	5.32
Helper plasmid	1	0.37	0.27
	2	0.37	0.59
	3	0.36	0.24
Human genome (*)	1	0.13	0.10
	2	0.13	0.09
	3	0.15	0.09

Table 1. Percentage of DNA species in a rAAV vector batch after high-throughput sequencing:

comparison of the original SSV-Seq protocol with the SSV-Seq2.0 PCR-free method. The relative quantity of each DNA species is indicated in percentages. The rAAV8-CAG-GFP stock was produced by plasmid transfection in HEK293 cells and purified by ultracentrifugation on cesium chloride gradients. For SSV-Seq computational analyses, Fastq files were processed by ContaVect. (*) This reference corresponds to the human genome (GRCh38) and the Ad5 genome fragment integrated into the HEK293 cell line genome. Libraries were prepared in triplicates, without DNase pre-treatment (total residual DNA).

Acknowledgments

We thank all the staff of the Vector Core (CPV) at the University Hospital of Nantes (*https://umr1089.univ-nantes.fr/plateaux-technologiques/*) for production of the AAV reference vectors and for technical assistance. Illumina sequencing was conducted at the Genomics core facility of Nantes, France. We are most grateful to the Genomics and Bioinformatics Core Facility of Nantes (GenoBiRD, Biogenouest) for its technical support. We particularly thank Adrien Léger for his advices regarding the development of bioinformatics tools.

Financial support

This research was supported by the Fondation d'Entreprise Thérapie Génique en Pays de Loire, the Commissariat Général à l'Investissement (ANR Program, Investissements d'Avenir, Preindustrial Gene Therapy vector consortium), the Centre Hospitalier Universitaire (CHU) of Nantes and the Institut National de la Santé et de la Recherche Médicale (INSERM).

Authors contribution

E.L. designed the study, performed experiments and bioinformatics analyses, compiled data in figures and tables and wrote the manuscript. S.S. performed experiments, optimizations of the PCR-free protocol and interpreted data. M.B. performed experiments and created the SSV-Coverage program. A.G-D developed the NTContent program. V.B. supervised the production of AAV viral vectors and reviewed the manuscript. O.A. provided funding and supervised research. M.P-B and E.A. coordinated the work, supervised research, interpreted data, provide funding and wrote the manuscript.

Author Disclosure Statement

E.A. is inventor in several patents related to AAV technology and consultant for companies in the field of AAV gene therapy.

References

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., Gnirke, A., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12, R18. https://doi.org/10.1186/gb-2011-12-2-r18

Al-Zaidy, S.A., Mendell, J.R., 2019. From Clinical Trials to Clinical Practice: Practical Considerations for Gene Replacement Therapy in SMA Type 1. Pediatr. Neurol. 100, 3–11. https://doi.org/10.1016/j.pediatrneurol.2019.06.007

Ayuso, E., Blouin, V., Lock, M., McGorray, S., Leon, X., Alvira, M.R., Auricchio, A., Bucher, S., Chtarto, A., Clark, K.R., Darmon, C., Doria, M., Fountain, W., Gao, G., Gao, K., Giacca, M., Kleinschmidt, J., Leuchs, B., Melas, C., Mizukami, H., Müller, M., Noordman, Y., Bockstael, O., Ozawa, K., Pythoud, C., Sumaroka, M., Surosky, R., Tenenbaum, L., van der Linden, I., Weins, B., Wright, J.F., Zhang, X., Zentilin, L., Bosch, F., Snyder, R.O., Moullier, P., 2014. Manufacturing and characterization of a recombinant adeno-associated virus type 8 reference standard material. Hum. Gene Ther. 25, 977–987. https://doi.org/10.1089/hum.2014.057

Ayuso, E., Mingozzi, F., Montane, J., Leon, X., Anguela, X.M., Haurigot, V., Edmonson, S.A., Africa, L., Zhou, S., High, K.A., Bosch, F., Wright, J.F., 2010. High AAV vector purity results in serotype- and tissue-independent enhancement of transduction efficiency. Gene Ther. 17, 503–510. https://doi.org/10.1038/gt.2009.157

Beier, S., Thiel, T., Münch, T., Scholz, U., Mascher, M., 2017. MISA-web: a web server for microsatellite prediction. Bioinforma. Oxf. Engl. 33, 2583–2585. https://doi.org/10.1093/bioinformatics/btx198

Benjamini, Y., Speed, T.P., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 40, e72. https://doi.org/10.1093/nar/gks001

Breton, C., Clark, P.M., Wang, L., Greig, J.A., Wilson, J.M., 2020. ITR-Seq, a next-generation sequencing assay, identifies genome-wide DNA editing sites in vivo following adeno-associated viral vector-mediated genome editing. BMC Genomics 21, 239. https://doi.org/10.1186/s12864-020-6655-4

Crudele, J.M., Chamberlain, J.S., 2019. AAV-based gene therapies for the muscular dystrophies. Hum. Mol. Genet. 28, R102–R107. https://doi.org/10.1093/hmg/ddz128

D'Costa, S., Blouin, V., Broucque, F., Penaud-Budloo, M., François, A., Perez, I.C., Le Bec, C., Moullier, P., Snyder, R.O., Ayuso, E., 2016. Practical utilization of recombinant AAV vector reference standards: focus on vector genomes titration by free ITR qPCR. Mol. Ther. Methods Clin. Dev. 5, 16019. https://doi.org/10.1038/mtm.2016.19

Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36, e105. https://doi.org/10.1093/nar/gkn425

Dorange, F, Le Bec, C, 2018. Analytical approaches to characterize AAV vector production & purification: Advances and challenges. Cell Gene Ther. Insights 4, 119–129. https://doi.org/10.18609/cgti.2018.015

Fazekas, A., Steeves, R., Newmaster, S., 2010. Improving sequencing quality from PCR products containing long mononucleotide repeats. BioTechniques 48, 277–285. https://doi.org/10.2144/000113369

Food and Drug Administration, 2012. Cell lines derived from human tumors for vaccine manufacture. Vaccines and related biological product advisory committee meeting. Available online: http://www.nvic.org/cmstemplates/nvic/pdf/fda/fda-briefing-09192012.pdf.

Grimm, D., Kern, A., Rittner, K., Kleinschmidt, J.A., 1998. Novel tools for production and purification of recombinant adenoassociated virus vectors. Hum. Gene Ther. 9, 2745–2760. https://doi.org/10.1089/hum.1998.9.18-2745

Guerin, K., Rego, M., Bourges, D., Ersing, I., Haery, L., Harten DeMaio, K., Sanders, E., Tasissa, M., Kostman, M., Tillgren, M., Makana Hanley, L., Mueller, I., Mitsopoulos, A., Fan, M., 2020. A Novel Next-Generation Sequencing and Analysis Platform to Assess the Identity of Recombinant Adeno-Associated Viral Preparations from Viral DNA Extracts. Hum. Gene Ther. https://doi.org/10.1089/hum.2019.277

Kondratov, O., Marsic, D., Crosson, S.M., Mendez-Gomez, H.R., Moskalenko, O., Mietzsch, M., Heilbronn, R., Allison, J.R., Green, K.B., Agbandje-McKenna, M., Zolotukhin, S., 2017. Direct Head-to-Head Evaluation of Recombinant Adeno-associated Viral Vectors Manufactured in Human versus Insect Cells. Mol. Ther. J. Am. Soc. Gene Ther. https://doi.org/10.1016/j.ymthe.2017.08.003

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., Turner, D.J., 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat. Methods 6, 291–295. https://doi.org/10.1038/nmeth.1311

Lecomte, E., Leger, A., Penaud-Budloo, M., Ayuso, E., 2019. Single-Stranded DNA Virus Sequencing (SSV-Seq) for Characterization of Residual DNA and AAV Vector Genomes. Methods Mol. Biol. Clifton NJ 1950, 85–106. https://doi.org/10.1007/978-1-4939-9139-6_5

Lecomte, E., Tournaire, B., Cogné, B., Dupont, J.-B., Lindenbaum, P., Martin-Fontaine, M., Broucque, F., Robin, C., Hebben, M., Merten, O.-W., Blouin, V., François, A., Redon, R., Moullier, P., Léger, A., 2015. Advanced Characterization of DNA Molecules in rAAV Vector Preparations by Single-stranded Virus Next-generation Sequencing. Mol. Ther. Nucleic Acids 4, e260. https://doi.org/10.1038/mtna.2015.32

Maynard, L.H., Smith, O., Tilmans, N.P., Tham, E., Hosseinzadeh, S., Tan, W., Leenay, R., May, A.P., Paulk, N.K., 2019. Fast-Seq, a simple method for rapid and inexpensive validation of packaged ssAAV genomes in academic settings. Hum. Gene Ther. Methods. https://doi.org/10.1089/hum.2019.110

Oyola, S.O., Otto, T.D., Gu, Y., Maslen, G., Manske, M., Campino, S., Turner, D.J., Macinnis, B., Kwiatkowski, D.P., Swerdlow, H.P., Quail, M.A., 2012. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics 13, 1. https://doi.org/10.1186/1471-2164-13-1

Penaud-Budloo, M., François, A., Clément, N., Ayuso, E., 2018a. Pharmacology of Recombinant Adeno-associated Virus Production. Mol. Ther. - Methods Clin. Dev. 8, 166–180. https://doi.org/10.1016/j.omtm.2018.01.002

Penaud-Budloo, M., François, A., Clément, N., Ayuso, E., 2018b. Pharmacology of Recombinant Adeno-associated Virus Production. Mol. Ther. - Methods Clin. Dev. 8, 166–180. https://doi.org/10.1016/j.omtm.2018.01.002

Penaud-Budloo, M., Lecomte, E., Guy-Duché, A., Saleun, S., Roulet, A., Lopez-Roques, C., Tournaire, B., Cogné, B., Léger, A., Blouin, V., Lindenbaum, P., Moullier, P., Ayuso, E., 2017. Accurate Identification and Quantification of DNA Species by Next-Generation Sequencing in Adeno-Associated Viral Vectors Produced in Insect Cells. Hum. Gene Ther. Methods 28, 148–162. https://doi.org/10.1089/hgtb.2016.185

Quail, M.A., Otto, T.D., Gu, Y., Harris, S.R., Skelly, T.F., McQuillan, J.A., Swerdlow, H.P., Oyola, S.O., 2011. Optimal enzymes for amplifying sequencing libraries. Nat. Methods 9, 10–11. https://doi.org/10.1038/nmeth.1814

Radukic, M.T., Brandt, D., Haak, M., Müller, K.M., Kalinowski, J., 2019. Nanopore sequencing of native adeno-associated virus single-stranded DNA using a transposase-based rapid protocol. bioRxiv 2019.12.27.885319. https://doi.org/10.1101/2019.12.27.885319

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., Jaffe, D.B., 2013. Characterizing and measuring bias in sequence data. Genome Biol. 14, R51. https://doi.org/10.1186/gb-2013-14-5-r51

Samulski, R.J., Chang, L.S., Shenk, T., 1987. A recombinant plasmid from which an infectious adeno-associated virus genome can be excised in vitro and its use to study viral replication. J. Virol. 61, 3096–3101.

Shin, S., Park, J., 2016. Characterization of sequence-specific errors in various next-generation sequencing systems. Mol. Biosyst. 12, 914–922. https://doi.org/10.1039/c5mb00750j

Tai, P.W.L., Xie, J., Fong, K., Seetin, M., Heiner, C., Su, Q., Weiand, M., Wilmot, D., Zapp, M.L., Gao, G., 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. https://doi.org/10.1016/j.omtm.2018.02.002

Van den Hoecke, S., Verhelst, J., Saelens, X., 2016. Illumina MiSeq sequencing disfavours a sequence motif in the GFP reporter gene. Sci. Rep. 6, 26314. https://doi.org/10.1038/srep26314

van Dijk, E.L., Jaszczyszyn, Y., Thermes, C., 2014. Library preparation methods for next-generation sequencing: tone down the bias. Exp. Cell Res. 322, 12–20. https://doi.org/10.1016/j.yexcr.2014.01.008

Wright, J.F., 2014. Product-Related Impurities in Clinical-Grade Recombinant AAV Vectors: Characterization and Risk Assessment. Biomedicines 2, 80–97. https://doi.org/10.3390/biomedicines2010080

Xie, J., Mao, Q., Tai, P.W.L., He, R., Ai, J., Su, Q., Zhu, Y., Ma, H., Li, J., Gong, S., Wang, D., Gao, Z., Li, M., Zhong, L., Zhou, H., Gao, G., 2017. Short DNA Hairpins Compromise Recombinant Adeno-Associated Virus Genome Homogeneity. Mol. Ther. J. Am. Soc. Gene Ther. 25, 1363–1374. https://doi.org/10.1016/j.ymthe.2017.03.028