Biotechnology Journal
www.biotechnology-journal.com

The SSV-Seq 2.0 PCR-Free Method Improves the Sequencing of Adeno-Associated Viral Vector Genomes Containing GC-Rich Regions and Homopolymers

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

Adeno-associated viral vectors (AAV) are efficient engineered tools for delivering genetic material into host cells. The commercialization of AAV-based drugs must be accompanied by the development of appropriate quality control (QC) assays. Given the potential risk of co-transfer of oncogenic or immunogenic sequences with therapeutic vectors, accurate methods to assess the level of residual DNA in AAV vector stocks are particularly important. An assay based on high-throughput sequencing (HTS) to identify and quantify DNA species in recombinant AAV batches is developed. Here, it is shown that PCR amplification of regions that have a local GC content >90% and include successive mononucleotide stretches, such as the CAG promoter, can introduce bias during DNA library preparation, leading to drops in sequencing coverage. To circumvent this problem, SSV-Seq 2.0, a PCR-free protocol for sequencing AAV vector genomes containing such sequences, is developed. The PCR-free protocol improves the evenness of the rAAV genome coverage and consequently leads to a more accurate relative quantification of residual DNA. HTS-based assays provide a more comprehensive assessment of DNA impurities and AAV vector genome integrity than conventional QC tests based on real-time PCR and are useful methods to improve the safety and efficacy of these viral vectors.

1. Introduction

AAVs vectors are widely used as viral vectors to deliver therapeutic DNA. With the success of clinical trials using recombinant AAVs (rAAVs), regulatory bodies have increased requirements for the quality control (QC) of these new drugs. In particular, the presence of residual DNA in the final product is of significant concern, given the potential risk of oncogenicity, immunogenicity, and decreased gene transfer efficiency.[1] The consequences of co-injecting DNA contaminants along with AAVs depends on multiple criteria, including the type, nature (i.e., free or encapsidated, fragmented, unmethylated), and quantity of DNA impurities. To limit these risks, the Food and Drug Administration recommends that residual host cell DNA (HCD) levels not exceed 10 ng per parental dose (http://www.nvic.org/cmstemplates/nvic/pdf/fda/fda-briefing-09192012.pdf). This can be difficult to achieve in some cases, especially when high doses of AAV vectors are required to lead to a therapeutic effect (e.g., in the treatment of Duchenne muscular dystrophy[2] or spinal muscular atrophy)[3]. HCD is usually quantified using real-time PCR, a targeted technique that analyzes few DNA species and suffers from high inter-laboratory variability due to a lack of standardized protocols and instruments used.[4] We previously developed the Single-Stranded Virus Sequencing (SSV-Seq) method for the analysis of residual DNA in AAV vector stocks.[5] The SSV-Seq protocol, based on Illumina high-throughput sequencing (HTS), has been adapted for the analysis of AAV vectors generated either by plasmid transfection of HEK293 mammalian cells[6] or baculovirus infection of Sf9 insect cells.[7] Using this method, we showed that DNA impurities mainly originate from the vector plasmid or the baculovirus genome for HEK293- and Sf9-based manufacturing platforms, respectively, and that residual sequences proximal to the inverted terminal repeats (ITR) predominate.[7,8] SSV-Seq determines the relative percentage of each DNA species and provides information on vector genome identity through computational analysis of single nucleotide variants (SNV) and the sequencing coverage over the rAAV genome. The growing interest in the use of HTS-based methods for rAAV QC has also led to the development of other sequencing protocols to analyze the identity[9,10] or integrity[11–13] of AAV vector genomes.

Here, we show that a high local GC content and the presence of G/C-homopolymers in the AAV vector genome impair PCR amplification efficiency during library preparation, decreasing the sequencing coverage of these regions. To address this issue, we have optimized library preparation using a PCR-free

E. Lecomte, S. Saleun, M. Bolteau, A. Guy-Duché, Dr. O. Adjali, Dr. V. Blouin, Dr. M. Penaud-Budloo, Dr. E. Ayuso
INSERM UMR1089
Translational Gene Therapy Laboratory
University of Nantes
Centre Hospitalier Universitaire of Nantes, Nantes 44200, France
E-mail: magalie.penaud-budloo@univ-nantes.fr

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/biot.202000016

DOI: 10.1002/biot.202000016

protocol. We used this novel method, SSV-Seq 2.0, to analyze a vector genome harboring a cytomegalovirus (CMV) early enhancer/chicken beta-actin (CAG) promoter, which is well known as a difficult template for PCR and sequencing. The PCR-free protocol improved the coverage evenness of the CAG-containing rAAV genome and consequently, led to more accurate quantification of the residual DNA relative to the reads aligned to the rAAV reference. HTS-based assays offer the most exhaustive means of controlling AAV vector quality and purity, and to assess risks associated with residual nucleic acids.

2. Experimental Section

2.1. rAAV Vector Production and Purification

The rAAV2/8-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 rAAV genome, which consists of the cytomegalovirus enhancer fused to the 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 ITR from the plasmid pSub201.[14] The co-transfected helper plasmid pDP8 contains the helper genes E2a, E4, and VA RNA from adenovirus 5 (Ad5), and allows expression of AAV2 Rep proteins under the control of the mouse mammary tumor virus LTR promoter and a shortened p5 promoter, and of AAV serotype eight viral proteins from the natural p40 promoter.[15] The AAV vector was produced and purified by ultracentrifugation on a double CsCl gradient as previously described.[16] The vector genome titer was determined by real-time PCR targeting AAV2 ITRs, as previously described.[17]

2.2. Identification of GC-Rich Regions and Homopolymers in the rAAV Vector Sequence

GC-rich regions in the rAAV2/8-CAG-GFP vector sequence were identified using NTContent (http://github.com/emlec/NTContent) from the SSV-Conta package, applying the following parameters: window and step sizes of 200 and 20, respectively, or of 50 and 25, respectively. Mononucleotide repeats of ≥ 6 nucleotides and simple sequence repeats (SSR) were localized along the AAV vector genome using the MISA-web server (https://webblast.ipk-gatersleben.de/misa/) applying 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, and 10/2, maximum length of sequence between two SSRs to register as compound SSR—100, and output file parameter—GFF.[18]

2.3. Preparation of Fragmented PhiX174 DNA

Sequencing libraries were prepared using PCR-free kits from 200 ng of fragmented PhiX174 DNA. For fragmentation, 1.5 µg of PhiX174 RF II DNA (NEB, Ipswich, MA) was diluted in a final volume of 100 µL of 10 mM Tris, 1 mM EDTA, pH 8.0 in 0.5 mL Bioruptor microtubes and sonicated using Bioruptor UCD-200 (Diagenode, Seraing, Belgium) at low power (160 W) for 12 pulses (30 s ON/90 s OFF). Buffer exchange was performed with 10 mL Tris-HCl pH 8.0 using the kit NucleoSpin Gel and PCR Clean-up (Macherey-Nagel, Düren, Germany). The profile of the fragmented PhiX DNA was analyzed with the Agilent Bioanalyzer 2100 using the High Sensitivity DNA kit (Agilent Technologies, Santa Clara, CA). The average fragment size was 289 bp. DNA was quantified using the Qubit 1× dsDNA High Sensitivity Assay kit (ThermoFisher Scientific, Waltham, MA) before library preparation.

2.4. Library Preparation for Illumina Sequencing

For the original SSV-Seq protocol which included a PCR amplification step, DNA sequencing libraries were prepared from 2×10^{11} vg (vector genome; two replicates of 1×10^{11} vg) based on the free ITR qPCR titer,[17] and 200 ng of double-stranded DNA (dsDNA) quantified by spectrophotometry (Nanodrop OneC; ThermoFisher Scientific).[5]

Three kits were tested for 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). PCR-free sequencing libraries were prepared following the suppliers' instructions, except for the purification and the ligation steps. For each kit, the adapters were replaced with home-made Illumina-compatible P5/P7 adapters and DNA purification was performed using SPRIselect reagent (Beckman Coulter, Brea, CA).[5]

The SSV-Seq 2.0 protocol was performed using 8×10^{11} vg (four replicates of 2×10^{11} vg) of an rAAV vector batch. After DNA extraction and second-strand synthesis, dsDNA concentration was determined by fluorimetry using the Qubit 1× dsDNA High Sensitivity Assay kit (ThermoFisher Scientific). Two tubes per sample were prepared with 150 ng of DNA in a final volume of 100 µL of 10 mM Tris, 1 mM EDTA, pH 8.0. DNA was sonicated using the Bioruptor UCD-200 (Diagenode, Liege, Belgium) as described in Lecomte et al. to reach an average target size of ≈300 bp.[5] Fragmented DNA from the two tubes was pooled (300 ng) and purified using 1.6× SPRIselect reagent (Beckman Coulter). The magnetic beads with bound DNA were then washed twice with 360 µL of freshly prepared ethanol 80% and DNA was eluted in 20 µL of ultrapure DNase/RNase-free distilled water (dH_2O). Libraries were then prepared using the NxSeq AmpFREE Low DNA kit (Lucigen, which combined the end-repair and A-tailing steps. The following mix was prepared in a 0.2-mL PCR tube: 17 µL of previously sheared and purified DNA, 25 µL 2× buffer, and 8 µL enzyme mix. The one-step reaction was performed using the Applied Biosystems Veriti Thermal Cycler (ThermoFisher Scientific) for 20 min at 25 °C with a heated (72 °C) lid, followed by a 20-min cycle at 72 °C and holding at 4 °C. Illumina-compatible P5/P7 adapters were prepared as previously described and diluted at 15 µM in dH_2O.[5] After DNA repair and A-tailing, 3 µL of diluted adapters and 4 µL of ligase were added to the 50-µL reaction volume. Adapter ligation in a thermocycler (30 min at 25 °C) was immediately followed by double-1× SPRI purification. Each SPRI purification step includes two washes with 180 µL ethanol 80%. The first elution

was performed in 50 µL dH_2O and the second in 16 µL ultrapure dH_2O.

2.5. QC of DNA Sequencing Libraries

The quality of the DNA libraries was verified by microchip electrophoresis using the High Sensitivity DNA kit (Agilent Technologies). Electropherograms were obtained after migration in the Agilent Bioanalyzer 2100 instrument and after analysis using Agilent 2100 Expert software. The total DNA concentration of the libraries was determined on the Qubit 4 fluorometer using the Qubit 1× dsDNA High Sensitivity Assay kit (ThermoFisher Scientific). Adapter-ligated DNA fragments were quantified by real-time PCR using the Universal qPCR Kapa SYBR Fast kit (Roche) before Illumina sequencing.

2.6. Illumina Sequencing

1% PhiX Control v3 DNA (Illumina, San Diego, CA) was added to DNA libraries before sequencing, providing QC for cluster generation and Illumina sequencing. The libraries were denatured and diluted following instructions for the HiSeq protocol (Part # 15050107 v03). Cluster generation was performed using the cBot system and the HiSeq Rapid PE Cluster Kit v2 (Illumina). HTS was performed using the HiSeq Rapid SBS Kit v2 (Illumina) on the Illumina HiSeq 2500 system (Illumina) with the following parameters: rapid run paired-end mode and read size, 94 bp.

2.7. Bioinformatics Analysis

Base call files were converted into FASTQ files using Illumina bcl2fastq2 Conversion Software (Illumina). The following programs, included in the SSV-Conta package (https://github.com/emlec/SSV-Conta), were then used to quantify and characterize all DNA species present in an rAAV vector batch: Quade, a FASTQ files demultiplexer; Sekator, an adapter trimmer; RefMasker, to mask sequence homologies; ContaVect, to analyze residual DNAs.[5] Briefly, FASTQ files were demultiplexed with Quade according to the barcodes. 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 when each base of the barcodes had a PHRED quality score ≥ 25. Passed paired-end reads were trimmed using Sekator according to sequence quality and the adapter removed, as previously described.[5] 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: phage φX174 genome (GenBank accession number J02482.1), phage λ genome (J02459.1), rAAV genome, plasmid backbone sequence, plasmid helper sequence, Ad5 sequence (nucleotides 1–4344 of human Ad5, complete genome, AC_000008), and the human genome (GRCh38 primary assembly). Regarding homologies between two reference sequences, RefMasker masked the homologous region on the second reference sequence following the above list order, replacing the nucleotides with an N-base symbol before aligning the reads. 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. Reads were aligned to each reference sequence using BWA-MEM with the option -M (bwa version 0.7.15).[19] Finally, the relative percentage of each DNA species was calculated by dividing the number of reads (properly and improperly paired) that aligned to the reference by the total number of reads mapped. Unmapped and Lowmapq reads (Figure S1, Supporting Information), as well as reads aligned to the PhiX and phage lambda genomes, were excluded from the calculation. Sequencing coverage along each base of the vector plasmid was generated using SSV-Coverage, a program included in the SSV-Conta package. The FASTQ data for this study have been deposited in the European Nucleotide Archive at EMBL-EBI under accession number PRJEB38306 (https://www.ebi.ac.uk/ena/data/view/PRJEB38306).

2.8. Graphical Representations

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

2.9. Statistics

Statistical analyses were applied to samples with at least three replicates. Data were expressed as mean ± SD. A one-tailed non-parametric Mann–Whitney *U*-test was performed to compare two independent groups. Differences were considered statistically significant at $p \leq 0.05$. Analyses were performed using GraphPad Prism v5.01.

3. Results

3.1. A High GC Content and Homopolymers in the AAV Vector Genome Lead to a Poor SSV-Seq Sequencing Coverage

The SSV-Seq protocol consists of the following successive experimental steps (**Figure 1**): 1) facultative DNase pretreatment, 2) DNA extraction from rAAV stocks, 3) second-strand DNA synthesis using random hexamers, 4) library preparation, and 5) Illumina sequencing.[5] Illumina-compatible sequencing libraries are prepared using a custom protocol (Figure 1, step 4). 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 via 15 PCR cycles. Finally, the library is paired-end sequenced using the Illumina HiSeq platform, and the data processed using our dedicated bioinformatics pipeline (https://github.com/emlec/SSV-Conta).[5]

The PCR amplification step has been described as the principal source of bias during sequencing library preparation.[20] Indeed, AT-[21] and GC-rich[22] fragments are less efficiently amplified than other regions in the genome, potentially leading to bias and consequent lower sequencing coverage. Sequencing library preparation protocols that include a PCR step

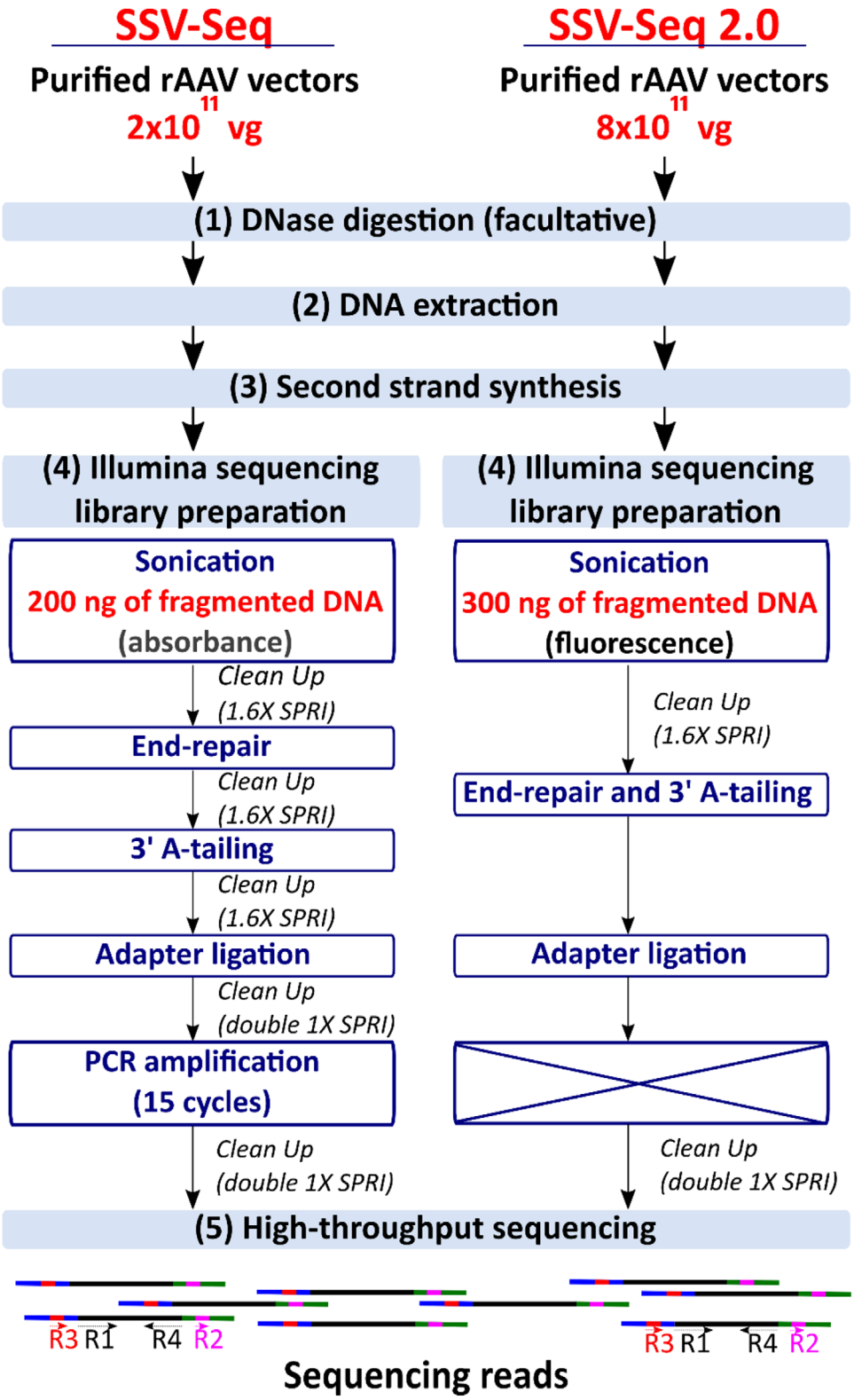


Figure 1. SSV-Seq 2.0 workflow. The SSV-Seq protocol was described in Lecomte et al. (left panel).[5] The optimized SSV-Seq 2.0 protocol is represented on the right side. A total quantity of 8×10^{11} vg of purified rAAV vector sample is required as input. Pretreatment with an endonuclease (Baseline-ZERO) and an exonuclease (Plasmid-Safe DNase) can be performed before DNA extraction to specifically identify and quantify DNA encapsidated in rAAV capsids. A second strand synthesis step is performed, followed by PCR-free DNA library preparation. Finally, HTS is performed using the Illumina HiSeq platform (rapid run mode 2×94 pb).

can thus result in uneven distribution of read coverage across the DNA. In the context of AAV vector analysis by HTS, this bias could lead to underestimation of AT- and/or GC-rich sequences.

To more precisely determine the impact of base composition on Illumina sequencing coverage, we first developed a new program, NTContent (https://github.com/emlec/NTContent), which is based on a sliding-window analysis and requires a DNA sequence in FASTA format as input. NTContent generates a tab-delimited text file composed of two columns, indicating for a given position the percentage of the requested nucleotide combination (Figure S2, Supporting Information). NTContent was applied to a 3.2-kb rAAV vector genome sequence containing the CAG promoter followed by the GFP reporter gene and the SV40 polyadenylation signal sequence. The CAG promoter was chosen as it is GC-rich and known as a difficult template for PCR amplification. **Figure 2**a shows the percentage of GC along the rAAV genome and the sequencing coverage obtained by SSV-Seq from a rAAV8 vector batch and its corresponding plasmid vector. For both the plasmid and the rAAV sample, two major drops in sequencing coverage appeared in the CAG promoter around positions 666 (asterisk, region 1) and 1421 (asterisk, region 2) of the rAAV genome. The superimposed graphs revealed that both drops in coverage are related to high GC content. Next, a more in-depth analysis of the percentage of GC was performed using NTContent, taking into account nature (A, T, G, or C) of 50 successive bases (step size, 25 bases). This analysis showed that the two sharp drops coincided with regions composed of > 90% GC (Table S1, Supporting Information).

To further investigate the origin of the drops in sequencing coverage, we analyzed the presence of A, T, C, and G nucleotide stretches in the rAAV genome (Figure 2b,c). Long G/C homopolymers have been reported as a source of bias during sequencing on Illumina systems, including HiSeq.[23] We observed a succession of C, T, and G homopolymers in region 1 of the CAG promoter between positions 588 and 676 (Figure 2b). C and G stretches were also detected in region 2, upstream from the drop in coverage between positions 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.[24] For SSV-Seq, PCR amplification is performed using PfuUltra II Fusion HotStart DNA Polymerase. The active site of this polymerase contains 6 nucleotides. We thus used MISA software to search for nucleotide repeats ≥6. Regions 1 and 2 of the CAG promoter contain 6 out of 14 homopolymers of ≥ 6 bases detected in the AAV vector genome (Table S2, Supporting Information). In particular, the first region in which a drastic coverage drop was observed included two stretches of 8 and 16 mononucleotides (C and G, respectively).

Overall, we can conclude that drastic drops in sequencing coverage correlate with the presence of long stretches of G and C nucleotides in the rAAV vector genome, consistent with the very high GC percentage illustrated in Figure 2a. To determine whether this bias occurs at the PCR stage and/or during HiSeq Illumina sequencing, we developed a PCR-free protocol, which we then compared with the PCR-enriched SSV-Seq method.

3.2. Development of a PCR-free Protocol for Sequencing Library Preparation

Two parameters are critical to adapt the SSV-Seq library preparation protocol to a PCR-free method: i) 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 because they contain all elements required for bridge amplification on Illumina flowcells (i.e., sequences complementary to the flowcell oligonucleotides, sequence targets of the P5/P7 sequencing primers, and 6-base indexes).[25] We selected six commercially available PCR-free kits based on their compatibility with Illumina technology (Table S3, Supporting Information). From these we tested three kits (Kapa HyperPrep, NxSeq AmpFREE Low DNA, and NEBNext Ultra II) that fulfilled the following criteria: i) a small amount of fragmented DNA (≤200 ng) is required as input, ii) end repair and A-tailing steps are combined into a single step, iii) home-made adapters can be used, and iv) compatibility with Illumina paired-end sequencing. To test the PCR-free kits and optimize the library preparation step, PhiX174 RF II DNA was chosen as its genome length (5.4 kbp) is close to that of the wild-type AAV genome (4.7 kb). Libraries were prepared from 200 ng of fragmented PhiX174 DNA following manufacturers' instructions, except two steps: home-made instead of commercial adapters were used for ligation; and the post-ligation cleanup and size selection steps were replaced with double purification with 1× SPRI beads (Figure 1, SSV-Seq 2.0). Libraries were prepared in triplicate to determine the robustness of each protocol. The efficiency of the three kits for generating sequencing libraries was compared qualitatively and quantitatively. DNA quality was controlled by high-sensitivity capillary electrophoresis (**Figure 3**a). For all three kits, electropherograms obtained using an Agilent chip revealed a negligible amount of free adapters in the final DNA libraries (Figure 3a, black arrow). Next, the number of adapter-ligated molecules in the libraries was quantified by Kapa qPCR using primers targeting the P5 and P7 sequences of the adapters, corresponding to the Illumina flowcell binding sequences (Figure 3b). The ligation step of the NxSeq kit was the most efficient, resulting in a library DNA concentration of 8.2 nM. Therefore, the NxSeq AmpFREE Low DNA kit was selected over the two other kits and included in the novel SSV-Seq 2.0 protocol for PCR-free library preparation.

Next, the optimized PCR-free protocol was tested by preparing a DNA library from an rAAV sample. The total number of vector genomes required as input was increased from 2×10^{11} vg for the original SSV-Seq method to 8×10^{11} vg for the SSV-Seq 2.0 protocol. After DNA extraction and second strand synthesis, 300 ng of fragmented DNA, as determined by fluorometric quantification, was used as input for PCR-free library preparation. The PCR-free library DNA, devoid of free adapters, was quantified by Kapa qPCR. The mean concentration of adapter-ligated fragments was 4.5 ± 0.2 nM in a final volume of 16 µL, compared with 49.8 ± 5.1 nM in a final volume of 30 µL for the PCR-based protocol, which is above the minimum concentration required for Illumina sequencing. In conclusion, the quality and quantity of library DNA obtained using the optimized SSV-Seq 2.0 protocol are sufficient for the analysis of AAV vectors by HiSeq Illumina sequencing.

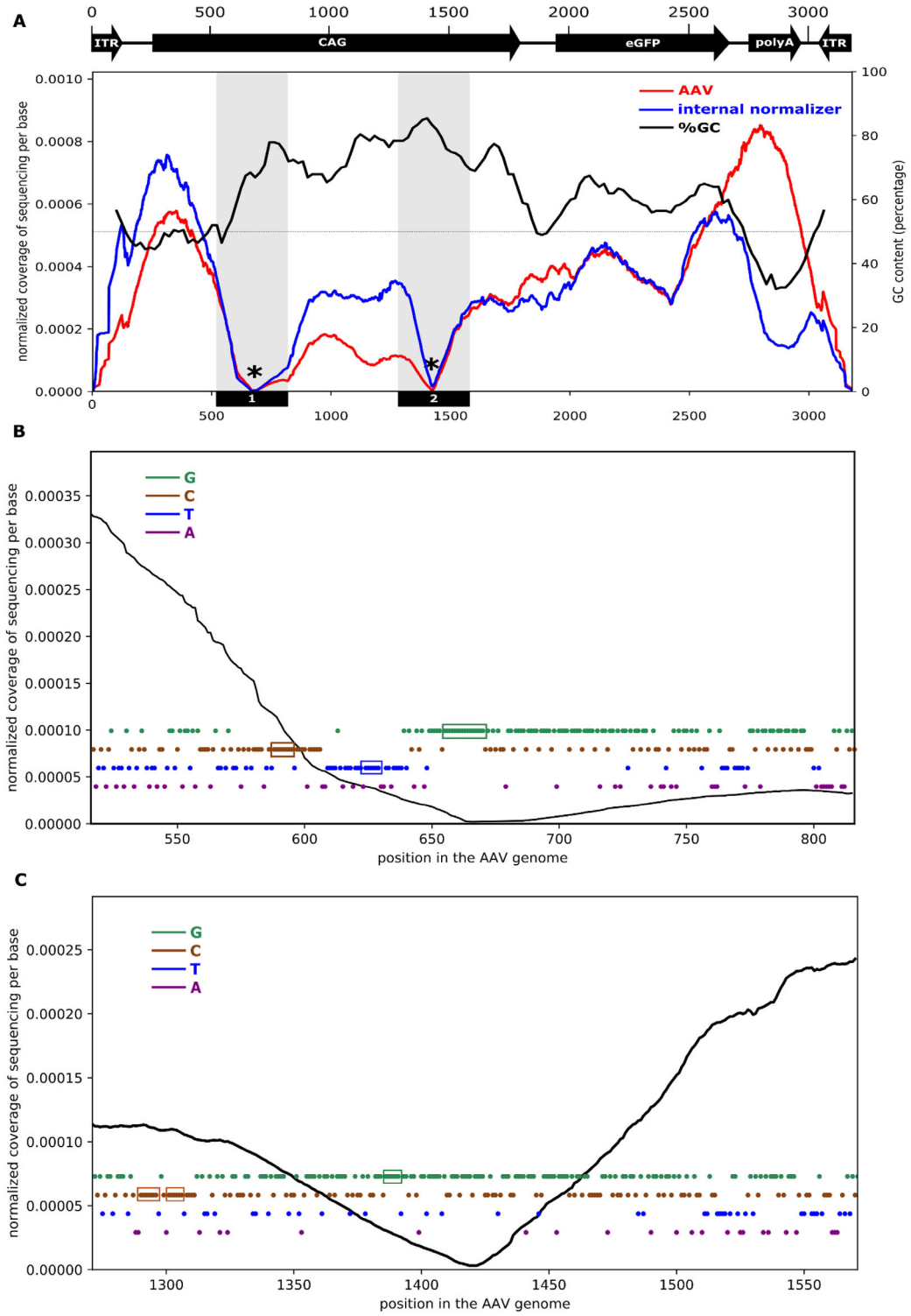
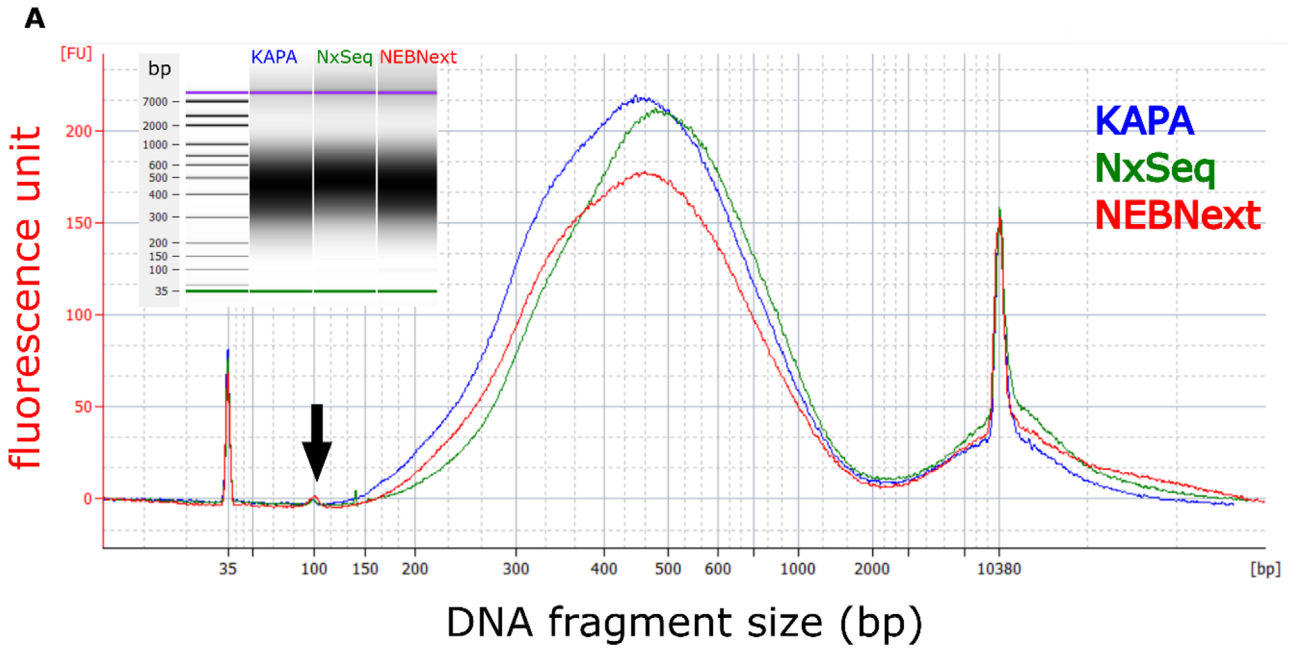


Figure 2. Impact of GC and homopolymer content on sequencing coverage in the rAAV genome. a) Sequencing coverage and GC percentage along the AAV vector genome. The sequencing coverages obtained for the 3.2-kb AAV8-CAG-GFP vector (red) and the internal normalizer (vector plasmid) (blue) were normalized by dividing the read coverage at each base by the sum of the coverage for all bases mapped to the rAAV genome. The internal normalizer consists of a mix of all DNA molecules that are expected to be found in rAAV stocks. Grey boxes indicate two 300-bp regions showing a drastic drop in sequencing coverage. Regions 1 and 2 were centered around the minimal number of reads at positions 666 and 1421 of the rAAV genome, respectively. The GC percentage (black) was determined using the NTContent program, applying the following parameters: window size, 200 bases; step size, 20 bases. The rAAV genome map is represented above the graph. b,c) Nucleotide content of b) region 1 and c) region 2. Each base is 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 coverage is represented as black lines.



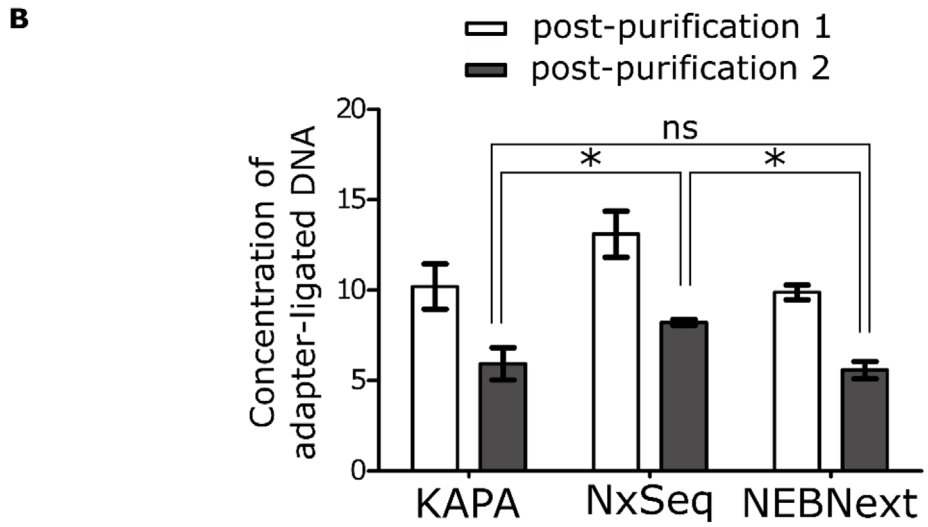


Figure 3. Comparison of PCR-free kits for the preparation of Illumina-compatible sequencing libraries. Libraries were prepared in triplicate from 200 ng of fragmented PhiX174 DNA (average fragment size 289 bp) using three different PCR-free kits: Kapa HyperPrep (Kapa), NxSeq AmpFREE Low DNA (NxSeq), and NEBNext Ultra II (NEBNext). a) Size distribution profiles of DNA libraries. DNA library quality was determined using an Agilent Bioanalyzer 2100. One representative electropherogram per kit is shown (Kapa, blue; NxSeq, green; NEBNext, red). The black arrow indicates the localization of free adapter dimers. b) The concentration of adapter-ligated fragments determined by Kapa qPCR. The concentration of DNA fragments ligated with two adapters was determined by Kapa qPCR, after the first (post-purification 1) and second (post-purification 2) SPRI bead purification steps. The concentration was 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 ± SD of library concentration from three replicates. ns, $p > 0.05$; *, $p \leq 0.05$ (one-tailed Mann–Whitney *U*-test).

3.3. The SSV-Seq 2.0 Protocol Improves Sequencing Coverage in GC- and Homopolymer-Rich Regions of the rAAV Genome

To assess the advantages of the SSV-Seq 2.0 protocol over the original protocol, DNA libraries were prepared using both PCR and PCR-free protocols from an rAAV8-CAG-GFP batch produced in HEK293 cells and purified by ultracentrifugation on CsCl gradients. After Illumina sequencing of the DNA libraries, reads were passed through SSV-Conta, our dedicated bioinformatics pipeline.[5] SSV-Conta is designed to determine the proportion of residual DNA species in an rAAV batch and to analyze coverage along the vector genome. For both protocols, over 94% of reads passed the quality and adapter trimming steps, although this percentage was slightly lower for the PCR-free libraries (Table S4,

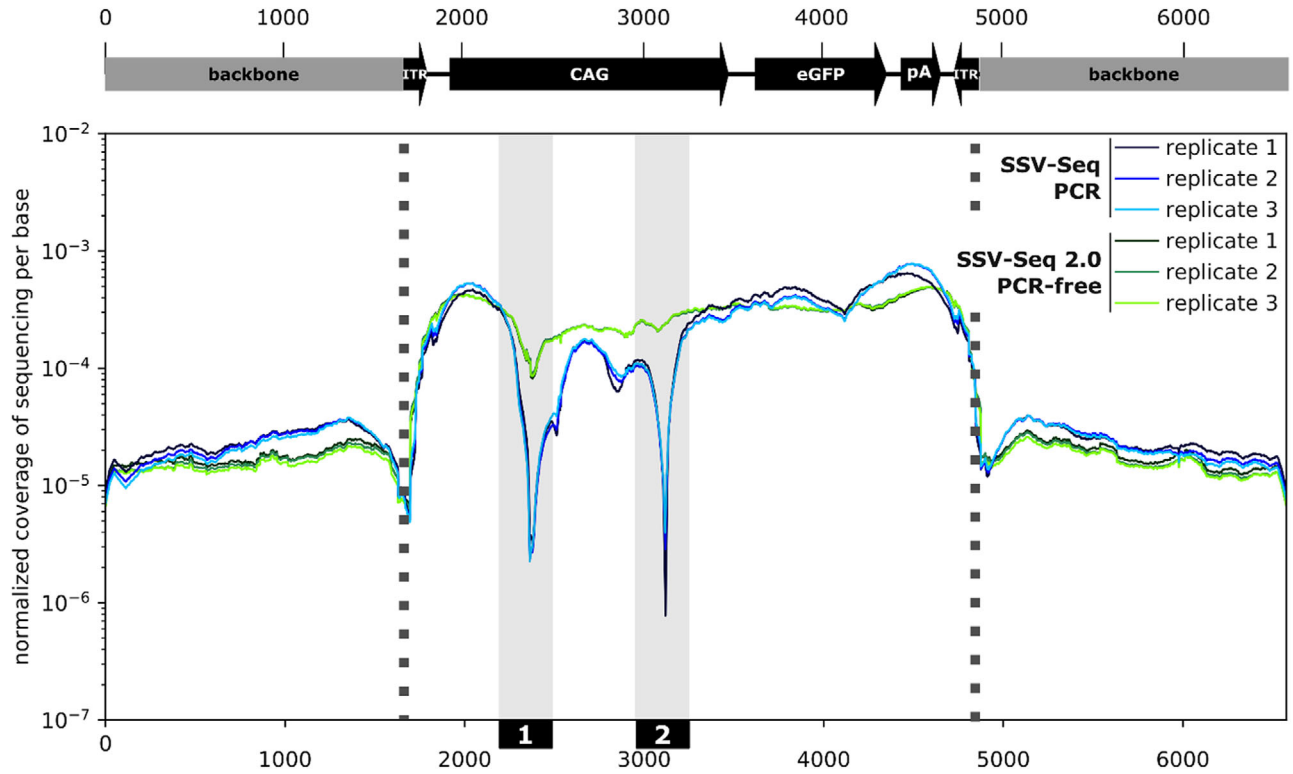


Figure 4. Sequencing coverage along the rAAV vector genome and the plasmid backbone. Sequencing libraries were prepared from a purified rAAV-CAG-GFP vector batch following the original SSV-Seq protocol (blue) or the SSV-Seq 2.0 PCR-free protocol (green). Each library was prepared in triplicate. Sequencing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

Supporting Information). The filtered reads were then aligned to the vector plasmid to visualize sequencing coverage in the two aforementioned GC-rich regions of the CAG promoter (**Figure 4**). Coverage of these regions was significantly improved using the PCR-free versus the PCR-enriched protocol, indicating that the PCR amplification step is one of the main causes of an artefactual drop in sequencing coverage. Specifically, the PCR-free protocol increased the mean total bases in sequencing coverage of the CAG promoter from 2.4% to 5.7% for region 1 and from 3.2% to 8.3% for region 2, approaching the theoretical value of 9.3% for even coverage of a 300-bp sequence in the rAAV genome. In addition to the rAAV vector genome, read alignment was performed 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 sequence is shown in Table S5, Supporting Information. Overall, a minimum of 15.2 and 23.8 M reads per sample were mapped to the known references for the PCR-free and PCR protocols, respectively. Independent of the method used, 97% of the Unmapped and Lowmapq reads obtained for the rAAV sample corresponded to Lowmapq reads. Of the Lowmapq reads, >95% aligned to the rAAV genome.

Finally, the proportion of each DNA species was calculated by dividing the filtered reads that aligned to each reference sequence by the total mapped reads and expressing as relative percentages (**Table 1**). For comparison with qPCR-based quantification methods of residual DNA, Table S6, Supporting Information, shows the relative percentages of each contaminant calculated from the copy number and the length of each DNA species. Consistent with better coverage of the CAG promoter, the optimized PCR-free method resulted in a higher percentage of reads aligned to the rAAV-CAG-GFP genome (93.9 ± 0.4% and 91.9 ± 0.3% of total mapped reads for the PCR-free and PCR protocols, respectively). We previously reported that the predominant DNA contaminant originates from the vector plasmid backbone.[6] 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% and 7.6 ± 0.3% of total mapped reads for SSV-Seq 2.0 and SSV-Seq, respectively). We also investigated the presence of SNV in the vector genome, but no SNV was observed in the CAG promoter using Samtools mpileup and bcftools call. Thus, we could not assess if the PCR-free protocol provides any advantage for the SNV analysis.

To determine whether PCR amplification of the residual DNA was also subjected to bias, the sequences of the two main contaminants (i.e., the vector plasmid backbone and the helper plasmid) were analyzed using NTContent. Neither sequence contained a local percentage of GC > 90%, as observed for the CAG promoter in the rAAV vector genome. No difference in sequencing coverage was observed between both protocols for the vector plasmid backbone (Figure 4, grey boxes). A more in-depth analysis was performed on the helper plasmid (see Figure S3, Supporting Information, for superimposed graphs of helper plasmid sequencing coverage and GC content). No sharp drop was observed in the coverage. Nucleotide stretches of ≥ 6 nucleotides were

searched in the helper plasmid sequence using the program MISA (Table S7, Supporting Information). Only 9 of 36 homopolymers corresponded to C- or G-stretches. Consistently, the PCR and PCR-free protocols resulted in comparable sequencing coverage along with the helper plasmid (Figure S3, Supporting Information), suggesting that the presence of multiple successive G- and C-stretches, in correlation with a high local GC percentage, is the main cause of the PCR amplification defect. Finally, the sequencing depth of HEK293 DNA was too low (Table S5, Supporting Information) to analyze the PCR-related artifact.

In conclusion, the PCR-free protocol improves the coverage evenness of the rAAV genome, leading to a reduced relative percentage of residual DNA. The SSV-Seq 2.0 method is more adapted than the PCR-enriched original method to analyze AAV vector genomes containing GC- and homopolymer-rich regions.

4. Discussion

The goal of this study was to develop a more accurate method to characterize DNA species present in rAAV batches. Several technological platforms are used to manufacture rAAV vectors for use in gene therapy, using either mammalian or insect cells.[8] Both upstream and downstream processes are known to potentially impact the purity of the final product, including the amount and type of residual DNA. Exhaustive identification and quantification of these DNA species is essential to assess the risk of co-transfer of undesired DNA sequences along with AAV vectors and can be achieved using HTS-based methods. We previously described the Illumina sequencing-based protocol SSV-Seq, to control rAAV purity by quantifying DNA contaminants.[5,6] The library preparation stage of SSV-Seq includes a PCR step, which may introduce some type of bias inherent to PCR at AT-[21] and GC-rich regions.[20] Indeed, all sequencing technologies exhibit error-rate biases in GC-poor (≤10%) and GC-rich (≥75%) regions, and those containing long homopolymers.[26] Illumina sequencing technology can also produce sequence-specific errors in G-rich sequences[27] potentially giving rise to false SNV discovery.[23] Several solutions have been proposed to reduce these artifacts, either through optimizing PCR conditions[28] or developing alternative library amplification methods.[29] To improve our SSV-Seq protocol we opted for a more drastic approach, switching to a PCR-free library preparation kit. Our findings show a clear correlation between a high GC and homopolymer content and poor sequencing coverage. To avoid data misinterpretation (e.g., large deletions or biological under-representation of a particular sequence in the rAAV particle population), it is essential to screen the rAAV genome for GC-rich regions and homopolymers before using sequencing-based analysis. To this end, MISA software and the new bioinformatics tool NTContent presented here (available at https://github.com/emlec/NTContent) can be extremely useful prediction tools. To monitor for potential bias in SSV-Seq, an internal normalizer is processed in parallel with 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 visualization and comparison of the sequencing coverage obtained for the rAAV sample and the plasmid vector (Figure 2).

A coverage drop in the CAG promoter, in which the local GC percentage exceeds 90%, has been reported by authors using SSV-Seq,[30] and others using Fast-Seq, a technique based on Tn5 tagmentation.[9] Kondratov et al. reported that a PCR-free protocol outperformed a PCR-enriched method (eight amplification cycles) in terms of sequencing coverage in GC-rich regions of the AAV vector genome.[30] Those authors used the Accel-NGS 2S PCR-Free DNA Library Kit (Swift Biosciences) for library preparation, with an initial amount of 4×10^{11} vg of a rAAV5-CAG-GFP vector and an input of 220 ng dsDNA. The Accel-NGS workflow includes two DNA repair steps and two adapter ligation steps and requires the use of specific adapters that are incompatible with low-throughput applications. In line with our findings, the authors reduced sequencing bias due to high GC content by removing the PCR step, although coverage drops were still detected in the CAG promoter and the eGFP transgene (Figure 4). Amplification biases may be introduced during the clonal bridge amplification PCR used for cluster generation on the Illumina flowcell, even though the conditions used for bridge PCR differ from those used for the SSV-Seq library preparation. Indeed, the use of Bst DNA polymerase and the inclusion of formamide in buffers may improve PCR efficiency in GC-rich sequences. Independent of PCR amplification bias, the coverage drops may be related to the sequencing technology itself. Indeed, MiSeq sequencing using the same four-channel sequencing system as HiSeq has been shown to disfavor the CCNGCC motif in the GFP coding sequence.[31] Conversely, sequencing technologies such as single-molecule real-time (SMRT) sequencing (Pacific Biosciences) appear to provide less biased coverage across GC-rich regions.[26] Offering long read lengths, single-molecule sequencing technologies also enable the study of rAAV vector genome integrity.[11–13] Interestingly, AAV-GPseq SMRT-based assays have detected rAAV genome truncations at hairpin-like structures that form self-complementary viral genomes.[11] Improving rAAV genome sequencing, particularly sequencing of ITR and ITR-plasmid junctions, is also of great interest. An HTS-based assay was recently developed to identify off-target nuclease

Table 1. Percentage of DNA species in a rAAV8-GFP vector batch after HTS: comparison of the original SSV-Seq protocol with the SSV-Seq2.0 PCR-free method.

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[a)]	1	0.13	0.10
	2	0.13	0.09
	3	0.15	0.09

[a)] This reference corresponds to the human genome (GRCh38) and the Ad5 genome fragment integrated into the HEK293 cell line genome.

activity after the AAV-mediated genome edition in vivo.[32] In that protocol, named ITR-Seq, PCR, and adapter optimizations are performed to specifically amplify ITR-genomic DNA junctions. Combining multiple sequencing technologies could provide complementary information and reduce the risks associated with the 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, while SMRT sequencing-based AAV-GPseq is better suited to the analysis of AAV vector genome integrity (truncated rAAV genomes). The novel SSV-Seq 2.0 protocol presented here circumvents PCR-related bias and improves the HTS analysis of rAAV genomes with GC-rich regions and long mononucleotide stretches, as often found in promoters.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors thank the staff of the Vector Core (CPV) at the University Hospital of Nantes (https://umr1089.univ-nantes.fr/plateaux-technologiques/) for producing the AAV vectors and for technical assistance. Illumina sequencing was conducted at the Genomics Core Facility of Nantes, France. The authors are most grateful to the Genomics and Bioinformatics Core Facility of Nantes (GenoBiRD, Biogenouest) for technical support. Special thanks to Adrien Léger for his advice on the development of bioinformatics tools. 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 ANR-10-DBPS-01), the Centre Hospitalier Universitaire (CHU) of Nantes, and the Institut National de la Santé et de la Recherche Médicale (INSERM).

Conflict of Interest

E.A. is an inventor on several patents related to AAV technology and a consultant to companies working in the AAV gene therapy field.

Author Contributions

E.L. and S.S. contributed equally to this work. M.P.-B. and E.A. share senior authorship. E.L.: Conceptualization, investigation, project administration, software, visualization, writing-original draft, writing-review and editing; S.S.: Investigation, methodology, project administration, visualization; M.B.: Methodology, software; A.G.-D.: Software; O.A.: Resources; V.B.: Resources; E.A.: Funding acquisition, supervision, writing-original draft, writing-review and editing.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

AAV vectors, GC-content, high-throughput sequencing, homopolymers, PCR-free library

Received: May 14, 2020
Revised: September 29, 2020
Published online: November 8, 2020

[1] J. F. Wright, *Biomedicines* **2014**, *2*, 80.
[2] J. M. Crudele, J. S. Chamberlain, *Hum. Mol. Genet.* **2019**, *28*, R102.
[3] S. A. Al-Zaidy, J. R. Mendell, *Pediatr. Neurol.* **2019**, *100*, 3.
[4] F. Dorange, C. Le Bec, *Cell Gene Ther.Insights* **2018**, *4*, 119.
[5] E. Lecomte, A. Leger, M. Penaud-Budloo, E. Ayuso, in *Adeno-Associated Virus Vectors* (Ed: M. J. Castle), Humana Press, New York **2019**.
[6] E. Lecomte, B. Tournaire, B. Cogne, J. B. Dupont, P. Lindenbaum, M. Martin-Fontaine, F. Broucque, C. Robin, M. Hebben, O. W. Merten, V. Blouin, A. Francois, R. Redon, P. Moullier, A. Leger, *Mol. Ther.–Nucleic Acids* **2015**, *4*, e260.
[7] M. Penaud-Budloo, E. Lecomte, A. Guy-Duche, S. Saleun, A. Roulet, C. Lopez-Roques, B. Tournaire, B. Cogne, A. Leger, V. Blouin, P. Lindenbaum, P. Moullier, E. Ayuso, *Hum. Gene Ther.* **2017**, *28*, 148.
[8] M. Penaud-Budloo, A. Francois, N. Clement, E. Ayuso, *Mol. Ther.–Methods Clin. Dev.* **2018**, *8*, 166.
[9] L. H. Maynard, O. Smith, N. P. Tilmans, E. Tham, S. Hosseinzadeh, W. Tan, R. Leenay, A. P. May, N. K. Paulk, *Hum. Gene Ther.* **2019**, *30*, 195.
[10] K. Guerin, M. Rego, D. Bourges, I. Ersing, L. Haery, K. Harten De-Maio, E. Sanders, M. Tasissa, M. Kostman, M. Tillgren, L. Makana Hanley, I. Mueller, A. Mitsopoulos, M. Fan, *Hum. Gene Ther.* **2020**, *31*, 664.
[11] J. Xie, Q. Mao, P. W. L. Tai, R. He, J. Ai, Q. Su, Y. Zhu, H. Ma, J. Li, S. Gong, D. Wang, Z. Gao, M. Li, L. Zhong, H. Zhou, G. Gao, *Mol. Ther.* **2017**, *25*, 1363.
[12] P. W. L. Tai, J. Xie, K. Fong, M. Seetin, C. Heiner, Q. Su, M. Weiand, D. Wilmot, M. L. Zapp, G. Gao, *Mol. Ther.–Methods Clin. Dev.* **2018**, *9*, 130.
[13] M. T. Radukic, D. Brandt, M. Haak, K. M. Müller, J. Kalinowski, *NAR Genom Bioinform.* **2020**, *2*.
[14] R. J. Samulski, L. S. Chang, T. Shenk, *J. Virol.* **1987**, *61*, 3096.
[15] D. Grimm, A. Kern, K. Rittner, J. A. Kleinschmidt, *Hum. Gene Ther.* **1998**, *9*, 2745.
[16] E. Ayuso, F. Mingozzi, J. Montane, X. Leon, X. M. Anguela, V. Haurigot, S. A. Edmonson, L. Africa, S. Zhou, K. A. High, F. Bosch, J. F. Wright, *Gene Ther.* **2010**, *17*, 503.
[17] S. D'Costa, V. Blouin, F. Broucque, M. Penaud-Budloo, A. Francois, I. C. Perez, C. Le Bec, P. Moullier, R. O. Snyder, E. Ayuso, *Mol. Ther.–Methods Clin. Dev.* **2016**, *3*, 16019.
[18] S. Beier, T. Thiel, T. Munch, U. Scholz, M. Mascher, *Bioinformatics* **2017**, *33*, 2583.
[19] H. Li, R. Durbin, *Bioinformatics* **2009**, *25*, 1754.
[20] D. Aird, M. G. Ross, W. S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, A. Gnirke, *Genome Biol.* **2011**, *12*, R18.
[21] S. O. Oyola, T. D. Otto, Y. Gu, G. Maslen, M. Manske, S. Campino, D. J. Turner, B. Macinnis, D. P. Kwiatkowski, H. P. Swerdlow, M. A. Quail, *BMC Genomics* **2012**, *13*, 1.
[22] Y. Benjamini, T. P. Speed, *Nucleic Acids Res.* **2012**, *40*, e72.
[23] S. Shin, J. Park, *Mol. BioSyst.* **2016**, *12*, 914.
[24] A. Fazekas, R. Steeves, S. Newmaster, *BioTechniques* **2010**, *48*, 277.
[25] I. Kozarewa, Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman, D. J. Turner, *Nat. Methods* **2009**, *6*, 291.
[26] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, D. B. Jaffe, *Genome Biol.* **2013**, *14*, R51.
[27] J. C. Dohm, C. Lottaz, T. Borodina, H. Himmelbauer, *Nucleic Acids Res.* **2008**, *36*, e105.

[28] M. A. Quail, T. D. Otto, Y. Gu, S. R. Harris, T. F. Skelly, J. A. McQuillan, H. P. Swerdlow, S. O. Oyola, *Nat. Methods* **2012**, *9*, 10.
[29] E. L. van Dijk, Y. Jaszczyszyn, C. Thermes, *Exp. Cell Res.* **2014**, *322*, 12.
[30] O. Kondratov, D. Marsic, S. M. Crosson, H. R. Mendez-Gomez, O. Moskalenko, M. Mietzsch, R. Heilbronn, J. R. Allison, K. B. Green, M. Agbandje-McKenna, S. Zolotukhin, *Mol. Ther.* **2017**, *25*, 2661.
[31] S. Van den Hoecke, J. Verhelst, X. Saelens, *Sci. Rep.* **2016**, *6*, 26314.
[32] C. Breton, P. M. Clark, L. Wang, J. A. Greig, J. M. Wilson, *BMC Genomics* **2020**, *21*, 239.