Accurate Identification and Quantification of DNA Species by Next-Generation Sequencing in Adeno-Associated Viral Vectors Produced in Insect Cells

Magalie Penaud-Budloo, Emilie Lecomte, Aurélien Guy-Duché, Sylvie Saleun, Alain Roulet, Celine Lopez-Roques, Benoît Tournaire, Benjamin Cogné, Adrien Léger, Véronique Blouin, et al.

Accurate Identification and Quantification of DNA Species by Next-Generation Sequencing in Adeno-Associated Viral Vectors Produced in Insect Cells

Magalie Penaud-Budloo,[1] Emilie Lecomte,[1] Aurélien Guy-Duché,[1] Sylvie Saleun,[1] Alain Roulet,[2,3] Céline Lopez-Roques,[2,3] Benoît Tournaire,[1] Benjamin Cogné,[1] Adrien Léger,[1,†] Véronique Blouin,[1] Pierre Lindenbaum,[4] Philippe Moullier,[1,5] and Eduard Ayuso[1,*]

[1]*INSERM UMR1089, University of Nantes, Centre Hospitalier Universitaire, Nantes, France;* [2]*INRA, GeT-PlaGe, Genotoul, Castanet-Tolosan, France;* [3]*INRA, UAR1209, Castanet-Tolosan, France;* [4]*INSERM UMR1087, Nantes, France; and* [5]*Department of Molecular Genetics and Microbiology, University of Florida, College of Medicine, Gainesville, Florida.*

[†]*Present address: EMBL, European Bioinformatics Institute, Cambridge, United Kingdom.*

Recombinant adeno-associated viral (rAAV) vectors have proven excellent tools for the treatment of many genetic diseases and other complex diseases. However, the illegitimate encapsidation of DNA contaminants within viral particles constitutes a major safety concern for rAAV-based therapies. Moreover, the development of rAAV vectors for early-phase clinical trials has revealed the limited accuracy of the analytical tools used to characterize these new and complex drugs. Although most published data concerning residual DNA in rAAV preparations have been generated by quantitative PCR, we have developed a novel single-strand virus sequencing (SSV-Seq) method for quantification of DNA contaminants in AAV vectors produced in mammalian cells by next-generation sequencing (NGS). Here, we describe the adaptation of SSV-Seq for the accurate identification and quantification of DNA species in rAAV stocks produced in insect cells. We found that baculoviral DNA was the most abundant contaminant, representing less than 2.1% of NGS reads regardless of serotype (2, 8, or rh10). *Sf*9 producer cell DNA was detected at low frequency (≤0.03%) in rAAV lots. Advanced computational analyses revealed that (1) baculoviral sequences close to the inverted terminal repeats preferentially underwent illegitimate encapsidation, and (2) single-nucleotide variants were absent from the rAAV genome. The high-throughput sequencing protocol described here enables effective DNA quality control of rAAV vectors produced in insect cells, and is adapted to conform with regulatory agency safety requirements.

Keywords: AAV vectors, residual DNA, next-generation sequencing, baculovirus, insect cells, gene therapy

INTRODUCTION

Recombinant adeno-associated viral (rAAV) vectors are among the most promising currently available tools for gene delivery *in vivo*. Encouraging preclinical data have been corroborated by successful results in clinical trials of gene therapy for genetic diseases including Leber congenital amaurosis and hemophilia B,[1] and convincing outcomes of phase 1/2 clinical trials have led to an increase in the number of biotechnology and pharmaceuticals companies in this field. However, the large-scale commercial production of high-quality rAAV vectors will also require the development of new methodologies to fully characterize these drugs. Moreover, given the potential presence of a variety of process impurities (chemicals, proteins, and nucleic acid contaminants) in the final AAV vector products, effective methods for the accurate characterization of genetically engineered vectors will be essential.[2,3] During rAAV production in mammalian cells, capsids can illegitimately internalize DNA in addition to recombinant genomes.

*Correspondence: Dr. Eduard Ayuso, INSERM UMR1089, IRS2-Nantes Biotech, 22 Boulevard Benoni Goullin, 44200 Nantes, France. E-mail: eduard.ayuso@univ-nantes.fr

DOI: 10.1089/hgtb.2016.185

Quantitative PCR (qPCR), Sanger sequencing, and functional assays have revealed the presence of several undesirable DNA sequences, including prokaryotic sequences from plasmids and cell genome fragments, in purified rAAV particles.[4–7] We previously demonstrated that these sequences can be transferred *in vivo*[8] and may constitute a potential hazard to patients.[9,10] The level of DNA contamination in final rAAV preparations is thus an important safety concern, particularly if undesirable sequences are packaged in virions and if high-vector doses are required to achieve a therapeutic level (e.g., for the systemic treatment of genetic diseases such as hemophilia or muscular dystrophies). We have developed a novel protocol, known as *single-strand virus sequencing* (*SSV-Seq*), for the analysis of DNA impurities in rAAV stocks. Using SSV-Seq combined with Illumina (San Diego, CA) high-throughput sequencing, we can exhaustively quantify and characterize DNA species copurified and encapsidated in rAAV particles produced by plasmid transfection of mammalian cells.[11]

As an alternative to HEK293 cells, the baculovirus–*Sf*9 platform has been established as a scalable, Good Manufacturing Practice-compatible system for the production of rAAV vectors.[12] The production of rAAV vectors by infection of the *Spodoptera frugiperda* insect cell line (*Sf*9) with three baculovirus expression vectors (BEVs) was first described in 2002 by Kotin's group.[13] Subsequent improvements were introduced, reducing the number of BEVs to two: a BEV-rep-cap allowing the expression of Rep78/52 and VP1/VP2/VP3 proteins and a BEV-AAV carrying the gene of interest flanked by inverted terminal repeats (ITRs).[14] The baculovirus–*Sf*9 system has been used for the production of Glybera, an AAV-based product for the treatment of lipoprotein lipase deficiency and the first gene therapy drug approved by the European Medicines Agency (EMA).[15] One of the major concerns noted in the EMA assessment report for Glybera pertained to the characterization and elimination of baculovirus-derived impurities in the final product.[16] The encapsidation of DNA impurities in rAAV particles is usually partially assessed by qPCR using primers and probes designed for the detection of relevant sequences in the baculovirus backbone or the producer cell line genome. By contrast, next-generation sequencing (NGS) allows exhaustive characterization and distribution of all contaminant DNA sequences. In the present study, we adapted our previously described SSV-Seq protocol[11] to quantify DNA species in rAAV stocks produced in the *Sf*9 insect cell line. We developed a specific bioinformatics pipeline for the baculovirus–*Sf*9 production system, and single-molecule real-time (SMRT) sequencing was used to generate an accurate reference sequence for the recombinant baculovirus genome. Using NGS, we show that DNA derived from the BEVs represents less than 2.1% of the DNA content in rAAV stocks, irrespective of serotype. Furthermore, we successfully quantified DNA derived from the *Sf*9 producer cell line, which accounted for a small percentage of the total reads (≤0.03%). In summary, we have developed a novel, NGS-based analytic technique that conforms to regulatory agency requirements and could serve as an effective tool for the characterization of DNA impurities in clinical rAAV lots.

MATERIALS AND METHODS

Plasmid construction

The donor plasmid pFB-GFP contains the human cytomegalovirus (CMV) promoter, the enhanced green fluorescent protein (eGFP) reporter gene, and the 3′-untranslated region (3′-UTR) of the human hemoglobin β (HBB) gene. The recombinant genome in the pFB-GFP plasmid is flanked by two flop-oriented ITRs of AAV serotype 2 derived from the pSub-201 plasmid.[17] The ITR upstream of the CMV promoter lacks 15 bp in the external A region, and the ITR downstream of the HBB 3′-UTR is truncated by 17 bp with respect to the wild-type ITR2. The donor pMB-GFP-Puro plasmid contains the CMV promoter, the eGFP transgene followed by an encephalomyocarditis virus internal ribosome entry site (IRES), a puromycin resistance sequence, and the HBB 3′-UTR. The pMB-GFP-Puro rAAV genome is delimited by the wild-type flip and flop ITRs from AAV serotype 2. The recombinant AAV cassettes (including ITRs) were validated by Sanger sequencing and subsequently used for the generation of the recombinant baculoviruses.

Generation of recombinant baculovirus

The BEV-GFP and BEV-GFP-Puro recombinant baculoviruses were generated using the donor plasmids pFB-GFP and pMB-GFP-Puro, respectively. BEV-rep2-cap8, BEV-rep2-cap2, and BEV-rep2-caprh10 were generated using the donor plasmids pFB-rep2-cap8 (kindly provided by R.M. Kotin, University of Massachusetts Medical School, Worcester, MA), pFB-rep2-cap2 (kindly provided by O.W. Merten, Généthon, Évry, France), and pFB-rep2-caprh10 (cloned in our laboratory), respectively. These BEVs encoding the Rep and Cap proteins follow the design described in Smith and colleagues,[14] allowing production of AAV vectors using a dual baculovirus system.

Tn*7* site-specific transposition of the cassette of interest in the bacmid backbone bMON14272 was performed by transformation of 10 ng of the donor plasmids in *Escherichia coli* DH10Bac competent bacteria in accordance with the instructions in the Bac-to-Bac expression system manual (Thermo Fisher Scientific, Waltham, MA). The recombinant bacmids were validated for the presence of the insert DNA by PCR using the primers M13-pUC-F (5′-CCAGTCACGACGTTGTAAAACG) and M13-pUC-R (5′-AGCGGATAACAATTTCACACAGG) from either side of the insert and the set of primers M13-pUC-F and BAC-G (5′-AGCCACCTACTCCCAACA TC) targeting the gentamicin resistance sequence in the insertion cassette, and by Sanger sequencing. One microgram of each bacmid DNA was then transfected into 10^6 insect cells cultivated in 6-well plates, using 9 μl of Cellfectin II reagent (Thermo Fisher Scientific). The supernatants (P1 stocks) were recovered 96 hr posttransfection. P1p clones were then isolated from the P1 stocks by plaque assay. One clone per recombinant baculovirus was selected, based on the infectious titer expressed as plaque-forming units (PFU) and genetic stability of the insert after five passages. The BEV P2 stocks were then generated after amplification of P1p stocks in *Sf*9 cells seeded in spinner flasks. The identity of the baculoviral genomes was verified by Sanger sequencing from PCR products of P2 stock DNA extracts.

rAAV production and purification

For rAAV2/8 production in mammalian cells, HEK293 cells were cotransfected with the pDP8 helper plasmid and the aforementioned pFB-GFP or pMB-GFP-Puro vector plasmid and then purified by double cesium chloride (CsCl) gradient ultracentrifugation as described previously.[18]

For rAAV production in insect cells, *Spodoptera frugiperda Sf*9 cells (Invitrogen/Thermo Fisher Scientific) were grown at 27°C in Sf-900 III SFM in a spinner flask or 2-L bioreactor cultures (Thermo Fisher Scientific). *Sf*9 cells were infected at a density of 10^6 cells/ml with a BEV-rep-cap and a BEV-AAV at a multiplicity of infection (MOI) of 0.05 PFU/baculovirus. Four days after infection, cells were lysed by the addition of 0.5% Triton X-100 (Sigma-Aldrich, St. Louis, MO) for 2.5 hr at 27°C with agitation, and treated for an additional 2 hr at 37°C with Benzonase (5 U/ml; Merck, Darmstadt, Germany). The crude bulk was clarified by centrifugation for 5 min at $1000\times g$ or by depth filtration using Opticap disposable capsule filters (Merck Millipore, Billerica, MA). Vectors were then purified by double CsCl gradient ultracentrifugation[19] or by affinity chromatography on an AVB Sepharose high-performance column (GE Healthcare Life Sciences, Little Chalfont, UK).[14] AVB-purified rAAV vectors were concentrated by tangential flow filtration, using a MicroKros 100-kDa mPES fiber filter (Spectrum Labs, Rancho Dominguez, CA). CsCl- or AVB-purified AAV-GFP-Puro vectors and the AVB-purified AAV-GFP vectors were formulated in Dulbecco's phosphate-buffered saline (DPBS) (Lonza, Verviers, Belgium) containing 0.001% Pluronic F-68 (Sigma-Aldrich). CsCl-purified AAV-GFP vectors were formulated in DPBS without Pluronic F-68. For qPCR analysis, 3 μl of each purified rAAV stock was pretreated or not with 20 U of DNase I (Roche, Basel, Switzerland) before DNA extraction in a total volume of 200 μl of DNase reaction buffer (13 m*M* Tris [pH 7.5], 0.12 m*M* $CaCl_2$, 5 m*M* $MgCl_2$) for 45 min at 37°C. The vector genome (VG) copy number was determined after DNA extraction, using a High Pure viral nucleic acid kit (Roche), by free ITR,[20] HBB2pA, or eGFP qPCR assays (Supplementary Table S1; Supplementary Data are available online at www.liebertpub.com/hgtb). Baculoviral DNA contamination was quantified by Bac (*Autographa californica* nucleopolyhedrovirus [AcMNPV] DNA polymerase) and Genta (gentamicin resistance) qPCR. The copy numbers of the rep-cap sequences and the *Sf*9 genome were determined by Rep52 (rep2) and *Sf*9 (*Sf* ribosomal protein L37A) qPCR, respectively (Supplementary Table S2). Vector purity was evaluated by SDS-PAGE followed by silver staining (PlusOne silver stain kit; GE Healthcare Life Sciences) of 2×10^{10} vector genomes of each rAAV stock.

Baculovirus filtration experiments

One milliliter of the BEV-GFP stock was pretreated or not with 0.5% Triton X-100 (Sigma-Aldrich) for 2.5 hr at 27°C (similarly to the cell lysis step performed at harvest for rAAV production in *Sf*9 cells). In a total volume of 2.5 ml of DPBS (Lonza), 10^7 copies of untreated or Triton-pretreated BEV were spiked into 10^{11} vector genomes of purified rAAV2/8-CAG-hGFP vector produced by transfection in HEK293 cells. The rAAV-to-BEV copy ratio was on the same order of magnitude as the ratio observed by qPCR in purified rAAV stocks produced in *Sf*9 cells. The mixture was subjected to three successive filtrations through 0.45-μm Millipore PVDF Millex-HV, 0.22-μm Millipore PES Millex-GP, and 0.1-μm Millipore PVDF Millex-VV syringe filters (Thermo Fisher Scientific). Vector genomes and BEV copies were quantified after each filtration step by free ITR and Bac qPCR, respectively.

Library preparation and Illumina sequencing according to the SSV-Seq protocol

For each rAAV stock, DNA extraction and second-strand synthesis were performed in triplicate as described previously by Lecomte and colleagues[11] using 10^{11} vector genomes, as determined by free ITR qPCR.[20] The Illumina libraries were prepared according to an improved protocol including dual-indexing as follows: adaptor ligation was performed with preannealed adapters compatible with Illumina TruSeq Universal P5 and P7 adapters containing a 6-base index. Thus, each sample was identified by a unique combination of two barcodes. One percent of PhiX Control v3 DNA (Illumina) was spiked into the libraries before sequencing with a HiSeq 2500 system, using the rapid run paired-end mode (2×100 bp).

For the analysis of rAAV vectors produced in insect cells, the internal normalizer was prepared by admixing DNA species found in rAAV preparations in proportions comparable to the quantities detected by qPCR, that is, 10^{11} copies of rAAV genome as determined by free ITR qPCR; 2×10^7 copies of bacmid AAV; 2×10^7 copies of bacmid rep-cap; and 10^3 copies of *Sf*9 genomic DNA. For every rAAV genome and rep-cap combination, the internal normalizer was analyzed in the same Illumina sequencing run as the corresponding vector. The donor plasmids were digested with restriction endonucleases to release the rAAV genome from ITR to ITR. DNA fragments were purified by gel electrophoresis and subsequently extracted with NucleoSpin gel and a PCR clean-up kit (Macherey-Nagel, Düren, Germany). Two micrograms of each bacmid (AAV and rep-cap) was fragmented at an average size of 1.5 kb by sonication for 15 sec at 160 W, using a Bioruptor (Diagenode, Seraing, Belgium). *Sf*9 genomic DNA was extracted with a Gentra Puregene blood kit (Qiagen, Hilden, Germany). In accordance with the manufacturer's recommendations, 6 μg of genomic DNA was sheared per g-TUBE (Covaris, Woburn, MA) to achieve a mean fragment size of 6 kb. The environmental control consisted of a quantity of phage λ DNA corresponding to 10^{11} copies of the vector genome (i.e., 484 ng). The internal normalizer and the negative control were processed in accordance with the SSV-Seq protocol without DNase pretreatment.

Bioinformatic processing of SSV-Seq data

SSV-Seq quantification and characterization of DNA species in rAAVs were performed with the following four scripts: Quade (Fastq files demultiplexer; https://github.com/a-slide/quade), Sekator (adapter trimmer; https://github.com/a-slide/Sekator), RefMasker (sequence homologies masker; https://github.com/a-slide/RefMasker), and ContaVect (DNA contaminant analyzer; https://github.com/a-slide/contavect) bundled in SSV-Seq (https://github.com/emlec/SSV-Seq/releases, v1.0). Briefly, Fastq files generated with CASAVA (Illumina) were first demultiplexed with Quade according to their barcodes. A minimum barcode base quality of 25 (PHRED quality score) was required to retain the read. Next, the adapters were trimmed, using Sekator. After trimming, the minimum read size was 30 bases. The distribution of DNA contaminants was determined with RefMasker and ContaVect. The reference sequences were indicated in the ContaVect configuration file in the following order: the phage ϕX174 genome (GenBank accession number J02482.1), the phage λ genome, the rAAV genome, the bacmid backbone sequence, the rep-cap sequence, and the *Sf*21 genome (GenBank accession number JQCY00000000.2).[21] The reference sequence for the bacmid backbone bMON14272 was generated by *de novo* assembly from SMRT Pacific Biosciences (Menlo Park, CA) sequencing data. Using RefMasker, homologies between two reference sequences were masked on the second reference sequence, 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 reads and mapped reads that did not fulfill these criteria were excluded. An *in silico* control was generated to determine the prediction accuracy of the mapping process, using Fastq Control Sampler software (https://github.com/a-slide/fastq_control_sampler). Paired Fastq files were generated from the rAAV genome (650,000 reads), rep-cap sequences (5,000 reads), and bacmid backbone (105,000 reads) with the following parameters: size of the randomly generated reference sequence, 100,000 bp; size of the reads, 100 bp; sonicated fragment size, 250–1000 bp. Sequencing coverage along each base of a reference sequence was determined with Samtools v1.3 (http://www.htslib.org/) depth. Graphic representations of the sequencing coverage were generated with Gnuplot v5.0.3.

SNVs in the rAAV genome were analyzed from the BAM files containing the reads aligned on the rAAV sequence and using Samtools mpileup and bcftools call v1.3.

Calculations of coverage divergence

To compare the read coverage on the HBB 3′-UTR region of the rAAV vectors with that of the internal normalizer, the following mathematical formula was applied: Fold change = [median (HBB 3′-UTR reads/total AAV cassette reads)$_{\text{AAV sample}}$]/

[median (HBB 3′-UTR reads/total AAV cassette reads)$_{\text{internal normalizer}}$].

The difference in read frequency between the rAAV genome and the BEV region close to the 5′-ITR was calculated as follows: log(median number of total AAV cassette reads/median number of reads aligned to the 5 kb region before 5′-ITR). The difference in read frequency between the rAAV genome and the BEV region close to the 3′-ITR was calculated according to the same formula, with the median number of reads aligned to the 1.2-kb region downstream of the 3′-ITR.

To estimate the variation between the rAAV genome coverage and the bacmid backbone coverage at the furthest position from the rAAV insertion site, the BEV-AAV sequence was linearized at the site opposite the vector genome insertion. The median number of reads aligned to the rAAV cassette was then divided by the median number of reads aligned to the 0.5- to 1.5-kb region of the linearized BEV-AAV.

Statistics

Because of the limited number of samples, statistical analyses were performed using nonparametric tests with PRISM 5 software (GraphPad Software, San Diego, CA). A two-tailed Spearman's rank correlation test was used to compare the percentages of DNA contaminants calculated by qPCR and SSV-Seq.

RESULTS

Optimization of SSV-Seq and bioinformatics tools for the analysis of rAAV vectors produced in insect cells

SSV-Seq coupled with Illumina sequencing is a powerful technique for the exhaustive quantification of individual DNA species in rAAV stocks.[11] To facilitate sample assignment from mixed clusters, we improved on our previously published methodology[11] by (1) developing a new bioinformatics program for sample demultiplexing and (2) preparing dual-indexed Illumina libraries instead of single-indexed libraries. First, we developed a new demultiplexing program, named Quade (http://a-slide.github.io/Quade/), to remove low index quality reads. The program's performance was evaluated by spiking 1% of PhiX DNA Control v3 DNA in each sample before the HiSeq sequencing run. PhiX DNA was used not only for monitoring sequencing quality and to create cluster diversity, but also as a negative control for the demultiplexing step. Indeed, no read should be attributed to PhiX genome as its DNA is devoid of adaptors (barcode). In an analysis of a batch of rAAV2/8-GFP vector produced in HEK293 cells, the number of reads attributed to PhiX, using the previous CASAVA demultiplexing program (included in the Illumina sequencing analysis software), was 59,713, as compared with 5839 reads using the Quade demultiplexing program. Thus, our new program reduces read misattribution during demultiplexing by a mean of 11-fold (coefficient of variation [CV], 17.8%; $n=5$). Because double indexing has been shown to improve sample identification using the Illumina platform,[22] we added a second barcode to the P5 adapter. After preparation of a dual-indexed library from the same rAAV batch described previously, only 37 reads were attributed to the PhiX control. Thus, dual-indexing overcame the majority of demultiplexing errors, resulting in a mean reduction in the number of misattributed reads of 1514-fold (CV, 21%; $n=5$). Using the combined strategy of dual-indexing and index quality filtering, less than 20 reads were attributed to PhiX for each sample analyzed, indicating a low level of read misattribution.

To ensure exact attribution of each Illumina read in bioinformatics analyses, accurate reference sequences are also required. In the case of rAAV vectors produced in insect cells by dual baculovirus infection,[14] the following DNA species can potentially be found in the preparation: (1) the rAAV vector genome, (2) *Spodoptera frugiperda* (*Sf*9) cellular DNA, (3) recombinant BEV-AAV DNA, and (4) recombinant BEV-rep-cap DNA. Given that the *Sf*9 genome sequence is not available, we assigned all reads to the *Sf*21 genome. Because *Sf*9 is a subclone of *Sf*21 cells, the genome identity between these two cell lines should be extremely high. For the BEV reference, we first inferred a theoretical sequence based on the E2 strain sequence of AcMNPV published by Maghodia and colleagues[23] (GenBank accession number KM 667940) and the bacmid bMON14272 cloning strategy described by Luckow and colleagues.[24] An initial coverage graph for BEV-AAV was drawn from the Illumina reads aligned to this theoretical reference, but revealed an uneven profile (Supplementary Fig. S1). To obtain a more accurate BEV reference sequence, the "empty" bacmid bMON14272 was extracted from DH10Bac bacteria and sequenced, using Pacific Biosciences technology. Taking advantage of the long reads generated by SMRT sequencing, we assembled a new reference sequence for the bacmid bMON14272 and compared it with the theoretical sequence. This comparison revealed only 4 single-nucleotide polymorphisms (SNPs) and 40 insertions/deletions (INDELs) that differed between the two reference sequences (Supplementary Fig. S2),

which do not justify the observed drops in the BEV coverage (Supplementary Fig. S1). We thus hypothesized that the Illumina sequencing technology used for SSV-Seq generates a bias over GC-rich and AT-rich regions, as previously reported.[25] Indeed, the coverage drops along the BEV sequence correlated with AT-rich regions (Supplementary Fig. S3). By contrast, this bias was not observed for the bacmid sequence obtained by SMRT sequencing (Supplementary Fig. S2). Consequently, we used the new bacmid reference assembled from SMRT sequencing for all subsequent computational analyses.

Quantification of DNA species in rAAV preparations produced in insect cells

Using the updated SSV-Seq protocol, a massive number of reads was obtained from rAAV-GFP preparations (serotypes 2, 8, and rh10) produced in *Sf*9 cells by dual-BEV infection. For each sample, we mapped more than 11 million total reads to the reference sequences (Supplementary Table S3). As expected, read alignment revealed that the vast majority corresponded to the rAAV genome (more than 97.8%) regardless of serotype or purification process (double CsCl gradient ultracentrifugation or AVB affinity chromatography) (Table 1). Baculoviral DNA was the most common contaminant, but accounted for less than 2.07% of the total reads. Low levels of the rep-cap sequence were observed (≤0.10%). SSV-Seq quantification of insect cell DNA attributed ≤0.03% of total reads to this genome, indicating greater sensitivity of SSV-Seq versus qPCR-based analysis (Tables 1 and 2). Because residual DNA contamination is usually measured by qPCR,[26] we compared our SSV-Seq results with those obtained by qPCR determination of the residual baculovirus, rep-cap, and *Sf*9 copy numbers. The percentages of DNA contaminants as determined by qPCR (Supplementary Table S2), normalized by DNA reference sequence size, correlated with the SSV-Seq data (Spearman's rho, $r = 0.9912$) (Fig. 1). Together, our data demonstrate that baculoviral DNA accounts for the majority of illegitimately encapsidated DNA.

Copurified versus encapsidated DNA contaminants

Before extraction of rAAV DNA content, we performed a drastic nuclease treatment, which

Table 1. *Percentages of DNA species in rAAV-GFP preparations obtained by next-generation sequencing and inferred from qPCR assay*

		AAV2/2	*AAV2/8*	*AAV2/rh10*			
		CsCl	*CsCl*	*CsCl*		*AVB*	
Method	*Reference sequence*	*+DNase*	*+DNase*	*+DNase*	*–DNase*	*+DNase*	*–DNase*
SSV-Seq	rAAV genome	98.54	98.90	98.28	98.91	97.80	98.35 98.58
	Baculovirus backbone	1.38	1.03	1.61	1.03	2.07	1.56 1.34
	Rep-Cap	0.05	0.05	0.08	0.05	0.10	0.07 0.06
	Sf genome	0.03	0.02	0.03	0.01	0.03	0.02 0.02
qPCR	rAAV genome	98.92	99.30	98.99	98.97	98.94	98.20 99.02
	Baculovirus backbone (Bac)	0.97	0.60	0.88	0.88	0.93	1.60 0.86
	Baculovirus backbone (*Genta*)	*11.15*	*5.95*	*5.66*	*5.69*	*5.30*	*7.10* *4.52*
	Rep-Cap	0.11	0.10	0.13	0.16	0.13	0.20 0.12
	Sf genome	<LOQ	<LOQ	<LOQ	<LOQ	<LOQ	<LOQ

LOQ, limit of quantification.

The relative quantities of rAAV genome and DNA contaminants are indicated as the percentage of total sequences. The samples correspond to rAAV2/2-GFP, AAV2/8-GFP, and AAV2/rh10-GFP preparations purified by cesium chloride gradient ultracentrifugation (CsCl) or affinity chromatography (AVB). For SSV-Seq, Fastq files containing reads were processed by ContaVect, using the following references: AAV-GFP genome; bacmid backbone; rep2-cap2, rep2-cap8, or rep2-caprh10 sequences; and the *Sf*21 genome (used as an approximation of the *Sf*9 cell genome). Quantification was performed by counting the number of reads mapped to each reference sequence, and expressing them as a percentage of the total number of mapped reads (discarding unmapped reads). For qPCR, the percentage of DNA species was calculated from the copy number obtained by free ITR qPCR for the rAAV genome, Rep52 qPCR for the rep-cap sequence, and *Sf*9 qPCR for the *Sf* genome, and normalized to each reference size. The percentage of the baculovirus backbone DNA was determined by extrapolation of the copy number obtained by Bac qPCR or Genta qPCR (in italics). The percentage calculated from Genta qPCR data was not considered when determining the distribution of DNA contaminants. The LOQ of *Sf*9 qPCR is 2×10^4 copies/ml. The two values in each cell indicate two technical replicates.

Table 2. *Percentages of DNA species in rAAV-GFP-Puro preparations depending on the purification process*

Method	*Reference sequence*	*CsCl*	*AVB*	*AVB + CsCl*
SSV-Seq	rAAV genome	99.60	99.25	99.41
	Baculovirus backbone	0.38	0.71	0.56
	Rep-Cap	0.01	0.03	0.03
	Sf genome	0.005	0.01	0.01
qPCR	rAAV genome	99.72	99.53	99.63
	Baculovirus backbone (Bac)	0.24	0.41	0.33
	Baculovirus backbone (*Genta*)	*1.16*	*1.38*	*1.21*
	Rep-Cap	0.03	0.06	0.05
	Sf genome	<LOQ	<LOQ	<LOQ

LOQ, limit of quantification.

The relative quantities of rAAV genome and DNA contaminants are indicated as the percentage of total sequences. The samples correspond to rAAV2/8-GFP-Puro preparations purified by cesium chloride gradient ultracentrifugation (CsCl), affinity chromatography (AVB), or affinity chromatography followed by cesium chloride gradient ultracentrifugation (AVB + CsCl). Each sample was treated with DNase treatment before SSV-Seq or qPCR. For SSV-Seq, Fastq files containing reads were processed by ContaVect, using the following references: AAV-GFP-Puro genome, bacmid backbone, rep2-cap8 sequence, and the *Sf*21 genome (used as an approximation of the *Sf*9 cell genome). Quantification was performed by counting the number of reads mapped to each reference, and expressing them as a percentage of the total mapped reads (discarding unmapped reads). For qPCR, the percentage of DNA species was calculated from the copy number obtained by free ITR qPCR for the rAAV genome, Rep52 qPCR for the rep-cap sequence, and *Sf*9 qPCR for the *Sf* genome, and normalized to each reference size. The percentage of the baculovirus backbone DNA was determined by extrapolation of the copy number obtained by Bac qPCR or Genta qPCR (in italics). The percentage calculated from Genta qPCR data was not considered when determining the distribution of DNA contaminants. The LOQ of *Sf*9 qPCR is 2×10^4 copies/ml.

allowed us to distinguish nuclease-resistant DNA impurities packaged in viral particles from nuclease-sensitive DNA species, which are not protected by rAAV or baculovirus capsids. After DNase treatment, the percentages of baculoviral DNA, rep-cap, and *Sf*9 DNA contaminants increased slightly for CsCl- and AVB-purified AAV2/rh10-GFP vectors (Table 1) and for CsCl-purified rAAV2/8-GFP vectors (data not shown). Conversely, qPCR revealed a modest decrease in the percentage of DNA contaminants (Table 1). This discrepancy may be due to the more drastic DNase enzymatic treatment used for the SSV-Seq assay as compared with the DNase I treatment used for the qPCR assay. Indeed, if unpackaged rAAV genomes constituted a significant proportion of the DNA species in vector preparations, DNase hydrolysis of unprotected vector genomes would be expected to increase the percentage of other DNA contaminants via a pendulum effect. The results obtained after DNase treatment suggest that baculoviral DNA is encapsidated in rAAV virions or protected in baculovirus nucleocapsids.

To better determine whether baculovirus DNA is packaged in rAAV vectors, we evaluated the sensitivity of naked baculoviral DNA to SSV-Seq

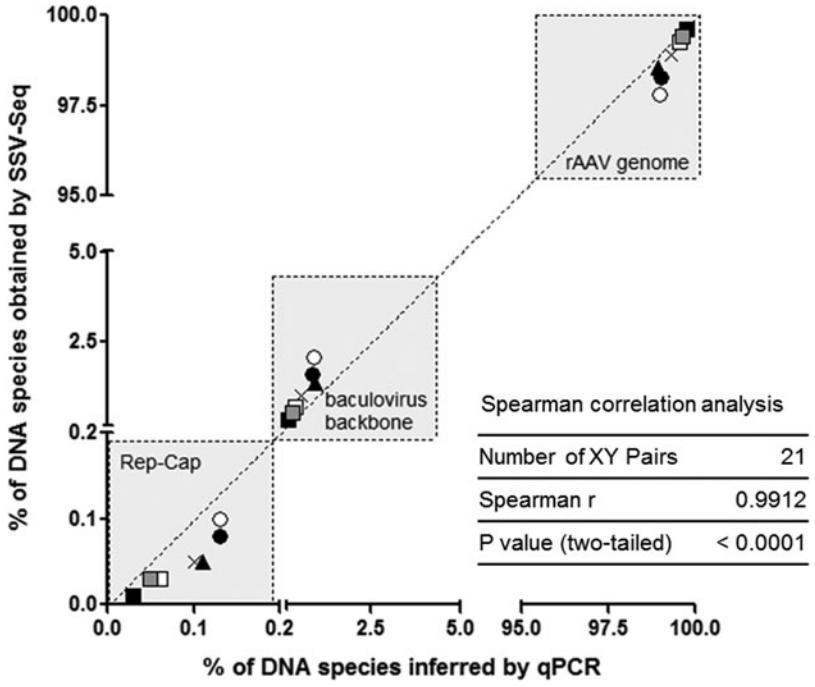


Figure 1. Correlation between the percentages of DNA species determined by SSV-Seq and those inferred from qPCR data. The percentages of DNA contaminants obtained by qPCR and SSV-Seq from seven rAAV batches are clustered in three gray squares corresponding to the rAAV genome, the baculovirus backbone, and the rep-cap reference sequences. Recombinant AAV2/2-GFP (*triangle*), AAV2/8-GFP (*cross*), AAV2/rh10-GFP (*circle*), and AAV2/8-GFP-Puro (*square*) vectors were purified by CsCl (*black*), AVB (*white*), or AVB + CsCl (*gray*) and pretreated with DNase before analysis. For real-time PCR, the percentage of baculovirus backbone was extrapolated from Bac qPCR data. A *diagonal dotted line* illustrates the theoretical perfect correlation between the two methods. The *x* and *y* axes are symmetrical, represented on linear scales and truncated between 0.2–0.2% and 5–95% to highlight intersample variability. The experimental correlation between the methods was evaluated using a two-tailed nonparametric Spearman's test with a 95% confidence interval.

DNase pretreatment. To simulate free baculoviral DNA contamination of an rAAV stock, 10^{11} vector genomes of purified rAAV, produced by plasmid transfection in mammalian cells, was spiked with 10^8 copies of bacmid. Samples were treated or not with the PS-DNase and Baseline Zero cocktail, as described in Lecomte and colleagues.[11] No baculoviral DNA was detected by qPCR after DNase treatment (Bac qPCR limit of quantification [LOQ], 3.3×10^3 copies), indicating that the percentage contamination determined by SSV-Seq did not correspond to free baculoviral DNA. At this stage, contamination of the final stocks by baculoviral particles copurified with rAAV particles could not be ruled out. Indeed, baculoviral DNA could be protected by the AcMNPV nucleocapsid and thus remain inaccessible to DNase treatment. Given that Triton X-100 is known to disrupt the envelope of multicapsid AcMNPV virions, but not the nucleocapsid,[27] a BEV stock was pretreated or not with 0.5% Triton X-100 for 2.5 hr. The DNase cocktail used for SSV-Seq had no effect on 10^8 copies of baculoviral DNA protected either in BEV nucleocapsids (Triton pretreatment) or in intact BEV particles (no Triton pretreatment). To confirm the origin of the contaminant baculoviral DNA, we used a series of filters that discriminate between these two virus types (rAAV and BEV) in final vectors stocks (Fig. 2). While no rAAV loss was observed after filtration, passage through 0.45-, 0.22-, and 0.1-μm filters led to overall decreases of 1 and 2 log units in the copy number of intact bacu-

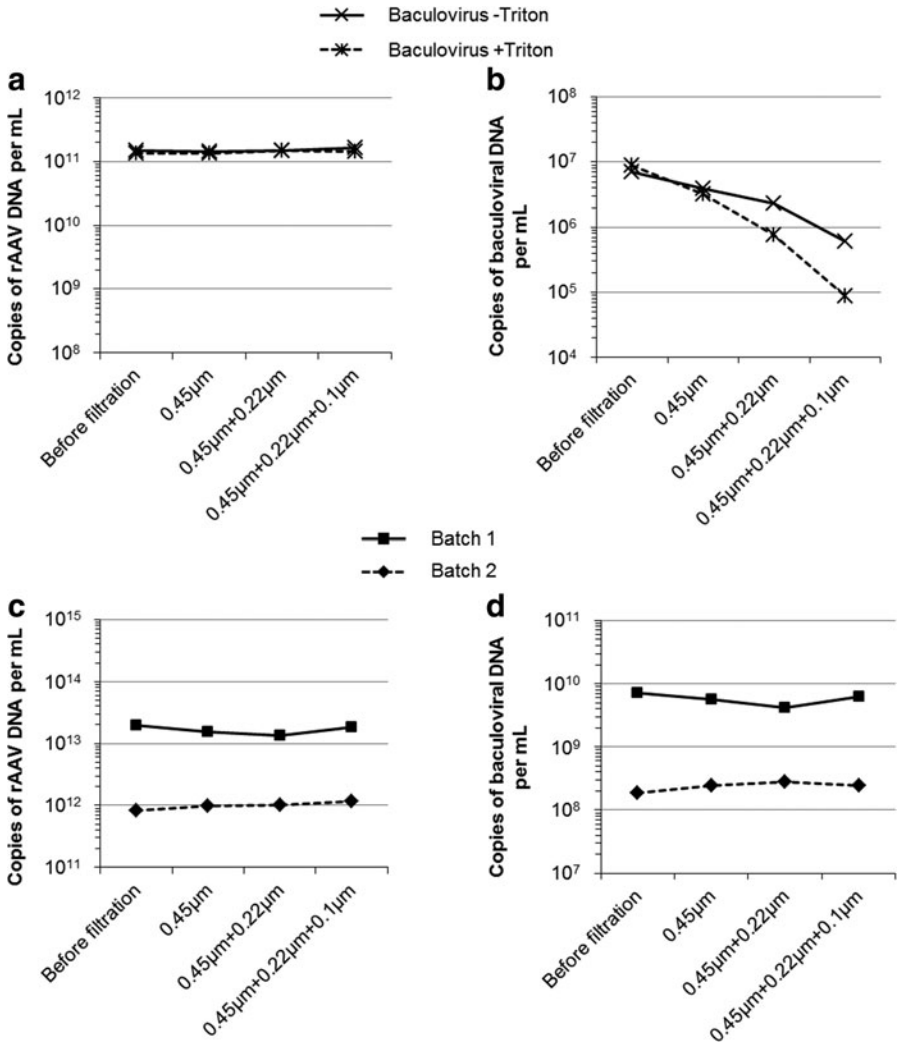


Figure 2. Baculovirus filtration assay. **(a** and **b)** Recombinant baculovirus (10^7 copies) was spiked into 10^{11} vector genomes of CsCl-purified rAAV2/8-GFP produced by plasmid transfection in HEK293 cells. Baculovirus stock was pretreated or not with 0.5% Triton X-100 for 2.5 hr at 27°C. The BEV and rAAV mixture diluted in 2.5 ml of DPBS was subjected to successive filtrations through 0.45-, 0.22-, and 0.1-μm filters. **(c** and **d)** Two batches of rAAV2/8-GFP produced in *Sf*9 cells and purified by CsCl gradient ultracentrifugation were filtered, using the same procedure. rAAV vector genomes were quantified by free ITR2 qPCR **(a** and **c)** and baculoviral DNA by Bac qPCR **(b** and **d)**.

lovirus virions and Triton X-100-pretreated baculovirus, respectively. Notably, the same successive filtrations applied to two batches of CsCl-purified rAAV-GFP had no effect on the level of contaminant baculoviral DNA. Moreover, AVB-purified vectors were concentrated via a tangential flow filtration step, using a 100-kDa cutoff hollow fiber that efficiently removes baculovirus virions from rAAV preparations.[28] Taken together, our data suggest that baculoviral DNA detected by SSV-Seq after DNase treatment is packaged in rAAV particles (or strongly associated with the vector capsid and thus inaccessible to nucleases).

Given that residual DNA packaged in rAAV particles cannot be eliminated by purification methods, modification of upstream processes is the best strategy to reduce levels of these contaminants. For instance, in rAAVs produced in HEK293 cells, increasing the rAAV genome length to the natural packaging capacity has been shown to reduce the illegitimate encapsidation of vector plasmid DNA.[29] With these findings in mind, we attempted to reduce the illegitimate packaging of the baculoviral backbone by increasing the transgene length to 4.6 kb and by replacing the truncated ITRs (rAAV-GFP) with integral wild-type ITRs (rAAV-GFP-Puro). The percentage of baculoviral DNA in CsCl-purified rAAV2/8-GFP-Puro (4572 bp) preparations was reduced 2.7-fold compared with rAAV2/8-GFP (3298 bp) stocks (Table 2). While this difference needs to be confirmed using additional samples, this fold reduction is well beyond the coefficient of variation calculated for the rAAV2/rh10-GFP technical duplicate in our study (Table 1). Moreover, rep-cap and *Sf*9 read percentages in rAAV2/8-GFP-Puro were reduced by 5- and 4-fold, respectively, compared with rAAV2/8-GFP vector. Complementary experiments will be required to confirm the beneficial effects, as regards DNA contamination, of using wild-type instead of truncated ITRs and increasing transgene size.

In parallel, we investigated whether the removal of empty capsids by CsCl ultracentrifugation modified the distribution of impurities in rAAV preparations. We compared the percentages of contaminants between rAAV2/8-GFP-Puro stocks purified by (1) CsCl gradient ultracentrifugation only (CsCl), (2) affinity chromatography only (AVB), and (3) affinity chromatography followed by CsCl gradient ultracentrifugation (AVB + CsCl) (Table 2). The purity of the preparations was verified by SDS-PAGE and silver staining, and full capsid enrichment by CsCl ultracentrifugation was confirmed by Coomassie blue staining (Supplementary Fig. S4). Interestingly, the percentage of DNA contaminants appeared to correlate with the proportion of particles that did not contain the transgene. Indeed, the highest percentage of baculoviral DNA was detected in the AVB-purified preparation, which had the highest p/VG ratio as calculated from both eGFP qPCR and ELISA data.

These results demonstrate that our novel SSV-Seq-based methodology is a powerful technique capable of identifying and quantifying DNA impurities in rAAV stocks produced in insect cells.

Evaluation of recombinant AAV genome identity

Using Illumina sequencing data and the ContaVect bioinformatics pipeline, we examined the identity of the recombinant vector genome packaged in rAAV particles. In particular, we focused our analysis on the read coverage along the vector cassette and the single-nucleotide variants (SNVs) in the vector genome sequence. Figure 3a shows the number of reads at each nucleotide position of the rAAV-GFP genome normalized by the total number of reads aligned to the rAAV genome. To identify possible enrichment of specific DNA motifs, the depth of coverage along the recombinant genome was compared with that of an internal normalizer consisting of a mix of all DNA species potentially present in rAAV stocks (i.e., rAAV cassette, bacmid AAV, bacmid rep-cap, and *Sf*9 genomic DNA). As shown in Fig. 3a, rAAV-GFP genomes in capsids of serotypes 2, 8, and rh10 all showed coverage profiles similar to that of the internal normalizer, except at the HBB 3′-UTR-pA region. The coverage profile differed from that of the plasmid control in this region, which was overrepresented by an average of 1.84-fold in rAAV-GFP particles produced in *Sf*9 cells. Similarly, the HBB region appeared to be slightly overrepresented in CsCl-purified rAAV2/8-GFP-Puro particles produced in insect cells (1.48-fold with respect to the internal normalizer) (Fig. 3b) but not in their counterparts produced in mammalian cells (Fig. 3c). Moreover, the overall shape of the coverage graph for the 4.6-kb rAAV2/8-GFP-Puro vector mimicked the 3.3-kb rAAV2/8-GFP cassette profile of analogous regions (CMV, eGFP, and HBB 3′-UTR). Interestingly, read distribution along the GFP-Puro cassette was independent of both the purification method (Fig. 3b) and the production system (Fig. 3c), except for the aforementioned HBB sequence. In conclusion, read distribution along the rAAV genome is homogeneous and identical regardless of serotype, purification procedure, and transgene length.

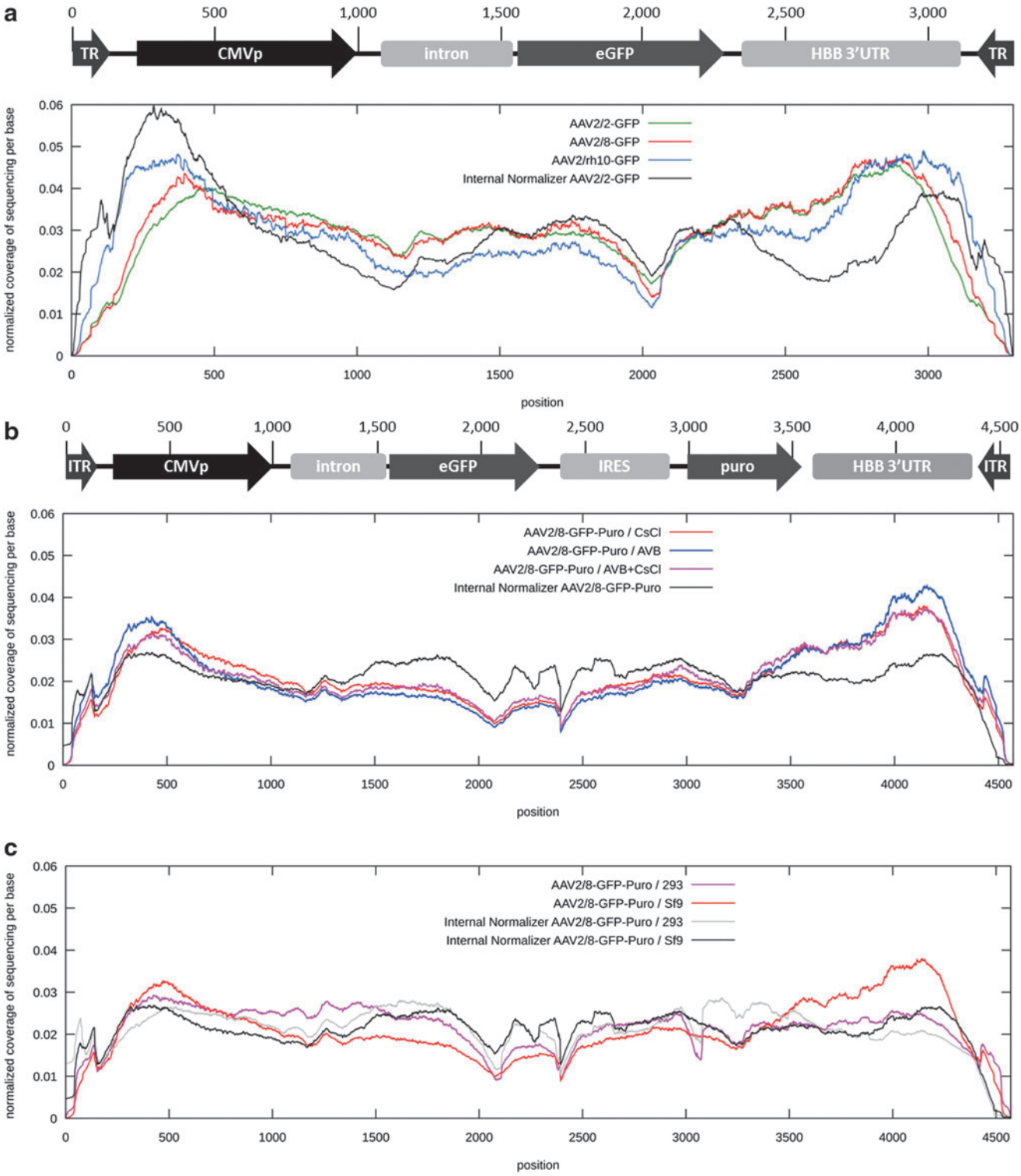


Figure 3. Sequencing coverage along the recombinant AAV genome. To compare samples independently of sequencing depth, sequencing coverage was normalized by dividing the read coverage at each base by the sum of the coverage for all bases mapped along the rAAV genome. The recombinant AAV genome map is represented to scale above the graph with coordinates in base pairs. All samples were pretreated with DNases before SSV-Seq analysis. **(a)** Sequencing coverage along the rAAV-GFP genome (3.3 kb). AAV vectors of serotype 2 (*green*), 8 (*red*), and rh10 (*blue*) were generated in the *Spodoptera frugiperda Sf*9 cell line by infection with a baculovirus expression vector (BEV) containing the rAAV genome and a second BEV mediating the expression of Rep and Cap proteins. Vectors were purified by CsCl gradient ultracentrifugation. The coverage profile of the rAAV2/2 internal normalizer was similar to that of the rAAV2/8 and rAAV2/rh10 normalizers and was chosen for representation purpose (*black*). **(b)** Sequencing coverage along the rAAV2/8-GFP-Puro genome (4.6 kb), depending on the purification process. rAAV-GFP-Puro vectors were produced using the *Sf*9–baculovirus platform and purified by CsCl gradient ultracentrifugation (CsCl, *red*), affinity chromatography (AVB, *blue*), or a combination of the two methods (AVB + CsCl, *magenta*). **(c)** Sequencing coverage along the rAAV2/8-GFP-Puro genome (4.6 kb), depending on the production system used. rAAV-GFP-Puro vectors were produced in *Sf*9 cells (*red*) or mammalian HEK293 cells (*magenta*) and purified by CsCl gradient ultracentrifugation. CMVp, cytomegalovirus promoter; eGFP, enhanced green fluorescent protein (CDS); intron, human β-globin intron; IRES, internal ribosome entry site; puro, puromycin resistance (CDS); HBB 3′-UTR, human β-globin 3′ untranslated region and polyadenylation signal; TR, inverted terminal repeats from serotype 2 originating from pSub-201 plasmid; ITR, flip (5′ end) and flop (3′ end) wild-type inverted terminal repeats from serotype 2.

Finally, we searched for single-nucleotide variants (SNVs) to verify the absence of mutations, in particular in regions of the genome essential for transgene expression (promoter and transgene). For all rAAV preparations tested, no nucleotide substitution, deletion, or insertion was detected in the whole GFP and GFP-Puro genomes (including ITRs) with respect to the internal normalizer (donor plasmid). In summary, the rAAV genome does not appear to undergo nucleotide variation during production in insect cells, although large deletions or insertions could not be assessed using SSV-Seq.

ITR proximity increases illegitimate packaging of baculoviral DNA in rAAV particles

The results of SSV-Seq analyses of two rAAV2/8 vectors purified by CsCl gradient ultracentrifugation are shown in Fig. 4. As expected, the sequencing coverage along the BEV-AAV shows that rAAV-GFP (Fig. 4a) and rAAV-GFP-Puro (Fig. 4b) cassettes are specifically encapsidated at a frequency 2.5 to 3 logs higher than backbone sequences in the vicinity of the ITRs. However, we found that the coverage along the baculoviral backbone was heterogeneous. In particular, the distribution of baculoviral sequences depended on ITR proximity; clear enrichment of backbone sequences was evident close to the ITRs, decreasing progressively with increasing distance from the rAAV genome. Identical coverage profiles were obtained for CsCl-purified rAAV-GFP vectors, serotypes 2 and rh10 (data not shown). Importantly, these findings were consistent with the results of qPCR titration analyses of the regions close to the ITR (Genta qPCR) and opposite the rAAV insertion site (Bac qPCR). In fact, the extrapolated percentage of baculoviral DNA contamination, based on the copy number of the gentamicin resistance gene (Genta), was on average 9.3-fold higher than the percentage extrapolated from the baculoviral DNA polymerase (Bac) copy number (Table 1). Furthermore, a close-up view of the 10-kb region adjacent to the 5′-ITR revealed a progressive decrease in reads up to approximately 5 kb downstream of the rAAV-GFP genome (Fig. 4c). Figure 4d shows a differential pattern close to the 3′-ITR compared with the 5′-ITR profile as the read distribution frequency levels off at around 1.2 kb upstream of the rAAV genome, and declines gradually after. These findings were independent of serotype, as similar coverage patterns along the BEV were observed for CsCl-purified rAAV2-GFP vectors of serotypes 8 and rh10 (data not shown). Interestingly, distribution patterns in the proximity of the 5′-ITR and 3′-ITR were identical for the 3.3- and 4.6-kb genomes (Fig. 4c and d). Moreover, the read distribution pattern close to the ITR was comparable for all three purification processes (CsCl, AVB, and AVB + CsCl) (Supplementary Fig. S5). Coverage profiles near ITRs, normalized to the total number of reads of the rAAV genome, did not differ between rAAV-GFP and rAAV-GFP-Puro vectors. Indeed, a global decrease in baculoviral sequence contamination was observed for the 4.6-kb vectors. In conclusion, distribution pattern analysis demonstrated that proximity to the ITR increases the probability of encapsidation of baculovirus sequences in rAAV particles.

In an attempt to investigate whether Rep-binding elements (RBEs) are related to specific DNA encapsidation, we searched for the following RBE motifs both in baculovirus and *Sf* genomes: (1) ITR RBE (CAGC GAGC GAGC GAGC), (2) AAVS1 RBE (GAGC GAGC GAGC GCGC), and (3) p5 RBE (GCCC GAGT GAGC ACGC).[30] No complete RBE has been found in *Sf* or baculovirus backbone. However, the BEV-GFP and BEV-GFP-Puro used in our study contained four and five short RBE motifs (GAGY GAGC), respectively (Fig. 4 and Supplementary Table S4). We found that short motif 2 (GAGT GAGC) was more represented in the rAAV particles than a random 8-bp motif (TCCA CTGG), due to their proximity to the ITRs (Fig. 4c and d). On the other hand, we have detected the presence of 8372 short RBE motifs in *Sf* genome (Supplementary Table S4). With a low number of *Sf* reads encapsidated in rAAV (<0.03%), the short motif 1 (GAGC GAGC) has a tendency to be more represented than the random *in silico* motif (Supplementary Table S4).

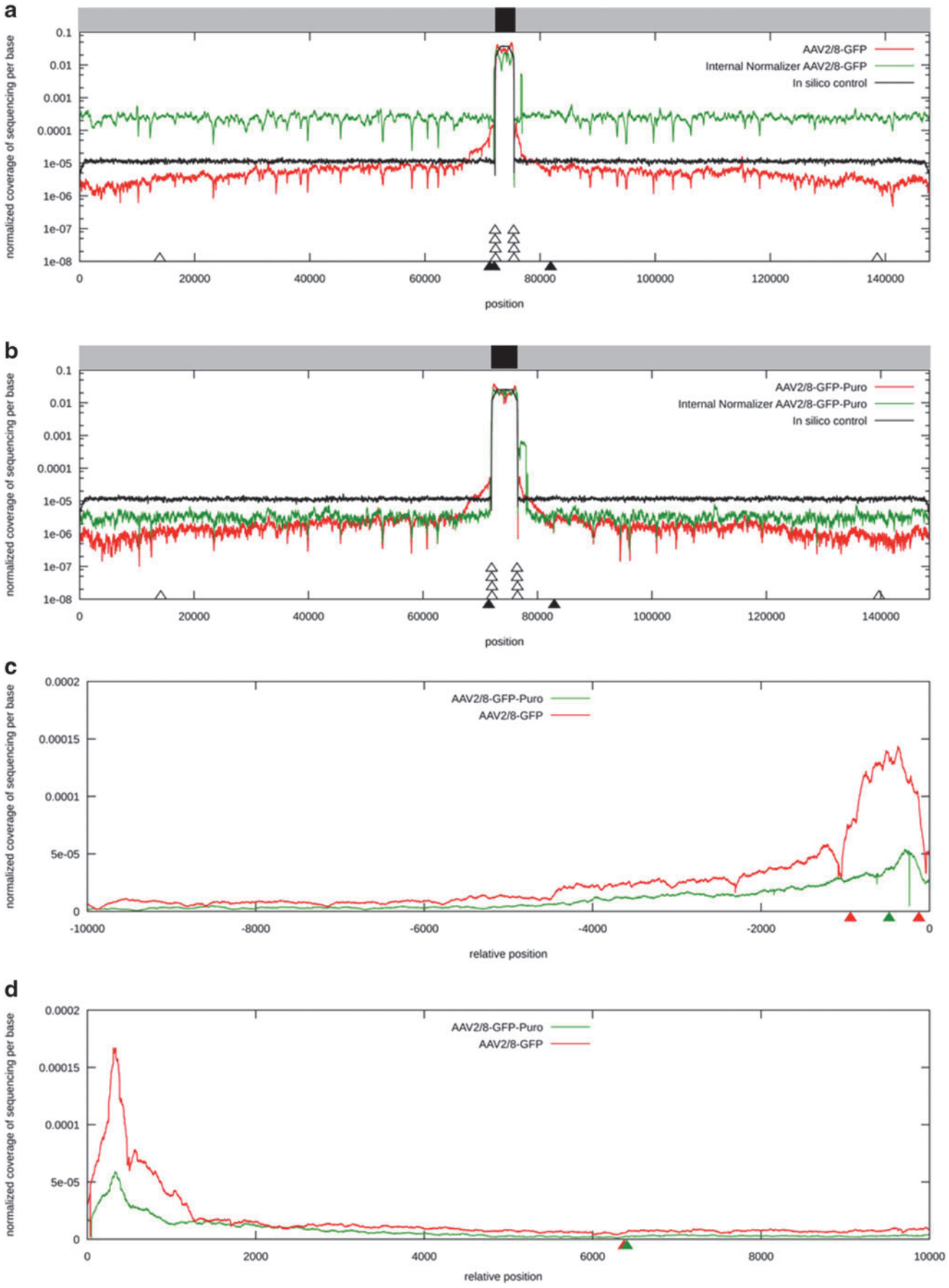
DISCUSSION

In 2012, for the market authorization of the first European gene therapy drug (Glybera), the EMA emphasized the need for assessment of residual baculovirus DNA. qPCR is the most common technique used to assess residual DNA contamination in rAAV stocks. However, qPCR analysis is partial, selective, and subject to intra- and inter-laboratory variability.[18] Furthermore, because of the large genome length, residual host cell DNA is usually undetectable (i.e., below the LOQ of the qPCR assay). This underscores the need for more sensitive assays to characterize product-related DNA impurities. Our laboratory has pioneered the development of an exhaustive NGS-based approach to quantify levels of DNA impurities relative to the rAAV genome in rAAV preparations produced by transfection of mammalian HEK293

cells.[11] The present study describes our development of a novel tool for the analysis of DNA impurities originating from the baculovirus–*Sf*9 production system. Using our SSV-Seq protocol, we determined that the predominant DNA contaminant is derived from the baculoviral genome, although levels of this contaminant did not exceed 2.1% in the seven batches tested. Interestingly, the proportion of residual baculoviral DNA in rAAV preparations was serotype-independent. However, the presence of empty/intermediate particles in purified stocks appears to increase the quantity of DNA contaminants. In agreement with this observation, the inclusion of a purification step to remove empty capsids was associated with a reduction in residual plasmid DNA in AAV2-hFIX vectors produced in HEK293 cells.[4] Importantly, our data strongly support the view that the vast majority of baculoviral DNA is packaged in rAAV virions. Therefore, DNA analysis-based techniques are not appropriate for the evaluation of baculovirus clearance in the context of rAAV manufacturing, and protein detection is preferable to qPCR analysis for the assessment of baculovirus contamination in rAAV batches. Importantly, risk assessment of residual baculoviral DNA should focus on sequences located within ±5 kb of the rAAV genome, as these are more prone to encapsidation. Our results are in line with those of a study that used sequential qPCR assays to assess levels of DNA contamination in rAAV vectors generated by plasmid transfection in mammalian cells.[31] Those authors showed that sequences located closer to the ITRs are detected at a 2 log-fold higher rate than distal sequences of the 18-kb plasmid backbone. Our findings are also consistent with qPCR and NGS analyses performed in advance of Glybera market authorization on the 3.6-kb rAAV1-LPL vector, in which baculovirus sequences adjacent to the LPL cassette were found to be preferentially packaged.[16] In the two cassettes used in the present study (3.3 and 4.6 kb), reverse packaging of baculoviral sequences appeared to occur up to ~5 kb downstream of the 5′-ITR but only up to approximately 1.2 kb upstream of the 3′-ITR. This discrepancy was not due to the nature of ITR sequences, as this effect was also observed for wild-type ITRs. Our results thus suggest a common mechanism for the illegitimate encapsidation of baculoviral and plasmid sequences by reverse priming initiated from each ITR. Further analyses are underway to determine whether the divergent distribution profiles around the 5′-ITR and 3′-ITR may be explained by DNA secondary structures or specific sequence motifs in the baculovirus, or by an ITR read-through mechanism. Crucially, in the Bac-to-Bac expression system, a kanamycin resistance gene is located in the region 5 kb upstream of the 5′-ITR and a gentamicin resistance sequence in the critical region 1.2 kb downstream of the 3′-ITR. Kanamycin and gentamicin are used for recombinant bacmid amplification and selection, using Bac-to-Bac technology. Considering the high doses of vectors necessary to treat systemic disease such as Duchenne muscular dystrophy (expected doses of $>10^{15}$ VG per patient)[32] and based on our analysis, the amount of gentamicin resistance sequence coinjected with rAAV vectors could reach 2.4×10^{12} copies. In mammalian cells, minicircles are used as a strategy to remove prokaryotic sequences from the vector and helper plasmids.[33] To ensure the absence of antibiotic sequences in rAAVs produced in insect cells, the appropriate strategy would be to use BEVs devoid of such sequences. For instance, some commercial systems, such as *Flash*BAC and BacMagic, enable the generation of BEVs by homologous recombination in insect cells and do not require antibiotic selection. Despite the absence of gentamicin, the risk of co-transferring baculoviral sequences together with AAV genome should be addressed. Although 12 baculoviral genes were identified by DNA microarray analysis as being transcribed in HeLa and BHK cells on AcMNPV infection,[34] further investigations should be carried out to establish whether intact baculoviral open reading frames can be encapsidated in rAAV and trigger aberrant expression of baculoviral polypeptides in patients. Nonetheless, baculoviruses are not pathogenic for mammals, and gene therapy studies using AcMNPV-derived vectors showed a good safety profile.[35]

The efficacy of AAV vectors in gene therapy has been consistently demonstrated in a growing number of clinical trials. The next challenge is to improve both the quality and efficiency of these new medical treatments. Illumina sequencing technology can provide insight into the purity of rAAV

Figure 4. Sequencing coverage along the baculovirus backbone. Global views of the read coverage along **(a)** BEV-AAV-GFP and **(b)** BEV-AAV-GFP-Puro. The rAAV genome (*black box*) and the baculovirus backbone (*gray boxes*) are annotated above the graphs. All rAAV2/8 vectors preparations were purified by CsCl gradient ultracentrifugation and pretreated with DNases before SSV-Seq analysis. Coverage profiles obtained for the vectors, the internal normalizer, and the *in silico*-generated control are depicted in *red*, *green*, and *black*, respectively. **(c)** Magnified coverage profiles of the 10-kb upstream regions of rAAV-GFP (*red*) and rAAV-GFP-Puro (*green*). **(d)** Magnified coverage profiles of the 10-kb downstream regions of rAAV-GFP (*red*) and rAAV-GFP-Puro (*green*). The depth of coverage is normalized to the number of reads aligned to the BEV-AAV reference sequence. *Triangles* represent short Rep-binding element (RBE) motifs 1 (GAGC GAGC, *open triangles*) and 2 (GAGT GAGC, *filled triangles*).



preparations and the integrity of rAAV genomes. Reassuringly, our SSV-Seq analysis of seven batches of rAAVs detected no SNVs in the recombinant genomes, a finding with positive implications regarding rAAV production in insect cells. Furthermore, no specific DNA motif was preferentially packaged in the two cassettes analyzed. However, as Illumina technology is not based on single-molecule sequencing (SMS), we cannot exclude the possibility that random truncations occur leading to rAAV genome heterogeneity. A 2012 study described a strategy to sequence single molecules of rAAV, using the Helicos Biosciences (Cambridge, MA) platform.[36] However, the average read length sequenced by Helicos technology is 32 nucleotides,[37] complicating read attribution. In the present study, we used Pacific Biosciences SMRT sequencing to generate the BEV reference sequence by *de novo* genome assembly. A new sequencing approach based on SMRT sequencing would constitute a valuable contribution to the field for the evaluation of genome heterogeneity in rAAV particles.

ACKNOWLEDGMENTS

Illumina sequencing was conducted at the UMR1087 Genomics core facility (Nantes, France). This work was performed in collaboration with the GeT core facility (Toulouse, France) (http://get.genotoul.fr), and was supported by the France Génomique National infrastructure, funded as part of the Investissements d'Avenir program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09). The authors thank Gérald Salin (INRA-UMR1388, Castanet-Tolosan, France) for help in developing the baculovirus genome assembly flowchart, and Christophe Klopp (GenoToul Bioinformatics core facility, Castanet-Tolosan, France) for expert assistance with the bioinformatics analysis of long reads. The authors are also grateful to Dr. Achille François for critical reading of the manuscript. This research was supported by the Fondation d'Entreprise Thérapie Génique en Pays de Loire, the Région Pays de Loire, the Commissariat Général à l'Investissement (ANR Program, Investissements d'Avenir, Preindustrial Gene Therapy Vector Consortium), the University of Nantes, and the Centre Hospitalier Universitaire (CHU) of Nantes and INSERM.

AUTHOR DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. Mingozzi F, High KA. Therapeutic *in vivo* gene transfer for genetic disease using AAV: progress and challenges. Nat Rev Genet 2011;12:341–355.
2. Wilson JM. A call to arms for improved vector analytics! Hum Gene Ther Methods 2015;26:1–2.
3. Wright JF. Manufacturing and characterizing AAV-based vectors for use in clinical studies. Gene Ther 2008;15:840–848.
4. Hauck B, Murphy SL, Smith PH, et al. Undetectable transcription of *cap* in a clinical AAV vector: implications for preformed capsid in immune responses. Mol Ther 2009;17:144–152.
5. Nony P, Chadeuf G, Tessier J, et al. Evidence for packaging of *rep-cap* sequences into adeno-associated virus (AAV) type 2 capsids in the absence of inverted terminal repeats: a model for generation of rep-positive AAV particles. J Virol 2003;77:776–781.
6. Ye G-J, Scotti MM, Liu J, et al. Clearance and characterization of residual HSV DNA in recombinant adeno-associated virus produced by an HSV complementation system. Gene Ther 2011;18:135–144.
7. Hüser D, Weger S, Heilbronn R. Packaging of human chromosome 19-specific adeno-associated virus (AAV) integration sites in AAV virions during AAV wild-type and recombinant AAV vector production. J Virol 2003;77:4881–4887.
8. Chadeuf G, Ciron C, Moullier P, et al. Evidence for encapsidation of prokaryotic sequences during recombinant adeno-associated virus production and their *in vivo* persistence after vector delivery. Mol Ther 2005;12:744–753.
9. Wright JF. Product-related impurities in clinical-grade recombinant AAV vectors: characterization and risk assessment. Biomedicines 2014;2:80–97.
10. Dolgin E. Gene therapies advance, but some see manufacturing challenges. Nat Med 2012;18:1718–1719.
11. Lecomte E, Tournaire B, Cogné B, et al. Advanced characterization of DNA molecules in rAAV vector preparations by single-stranded virus next-generation sequencing. Mol Ther Nucleic Acids 2015;4:e260.
12. Galibert L, Merten O-W. Latest developments in the large-scale production of adeno-associated virus vectors in insect cells toward the treatment of neuromuscular diseases. J Invertebr Pathol 2011;107(Suppl):S80–S93.
13. Urabe M, Ding C, Kotin RM. Insect cells as a factory to produce adeno-associated virus type 2 vectors. Hum Gene Ther 2002;13:1935–1943.
14. Smith RH, Levy JR, Kotin RM. A simplified baculovirus-AAV expression vector system coupled with one-step affinity purification yields high-titer rAAV stocks from insect cells. Mol Ther 2009;17:1888–1896.
15. Bryant LM, Christopher DM, Giles AR, et al. Lessons learned from the clinical development and market authorization of Glybera. Hum Gene Ther Clin Dev 2013;24:55–64.
16. European Medicines Agency. Assessment report: Glybera [cited 2016 Nov 22]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/002145/WC500135476.pdf
17. Samulski RJ, Chang LS, Shenk T. 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 1987;61:3096–3101.
18. Ayuso E, Blouin V, Lock M, et al. Manufacturing and characterization of a recombinant adeno-associated virus type 8 reference standard material. Hum Gene Ther 2014;25:977–987.
19. Ayuso E, Mingozzi F, Bosch F. Production, purification and characterization of adeno-associated vectors. Curr Gene Ther 2010;10:423–436.

20. D'Costa S, Blouin V, Broucque F, et al. Practical utilization of recombinant AAV vector reference standards: focus on vector genomes titration by free ITR qPCR. Mol Ther Methods Clin Dev 2016;5:16019.

21. Kakumani PK, Malhotra P, Mukherjee SK, et al. A draft genome assembly of the army worm, *Spodoptera frugiperda*. Genomics 2014;104:134–143.

22. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 2012;40:e3.

23. Maghodia AB, Jarvis DL, Geisler C. Complete genome sequence of the *Autographa californica* multiple nucleopolyhedrovirus strain E2. Genome Announc 2014;2:e01202–e01214.

24. Luckow VA, Lee SC, Barry GF, et al. Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in *Escherichia coli*. J Virol 1993;67:4566–4579.

25. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 2012;40:e72.

26. Wright JF, Zelenaia O. Vector characterization methods for quality control testing of recombinant adeno-associated viruses. Methods Mol Biol 2011;737:247–278.

27. Rueda P, Fominaya J, Langeveld JP, et al. Effect of different baculovirus inactivation procedures on the integrity and immunogenicity of porcine parvovirus-like particles. Vaccine 2000;19:726–734.

28. Hermens WTJMC, Smith JP. Removal of contaminating viruses from AAV preparations. Patent WO 2013/036118 A1.

29. Wright JF, Zelenaia O. Vectors comprising stuffer/filler polynucleotide sequences and methods of use. Patent WO 2014/144486 A2.

30. Feng D, Chen J, Yue Y, et al. A 16bp Rep binding element is sufficient for mediating Rep-dependent integration into AAVS1. J Mol Biol 2006;358:38–45.

31. Brimble MA, Zhou J, Morton C, et al. AAV preparations contain contamination from DNA sequences in production plasmids directly outside of the ITRs. Mol Ther 2016;24(Suppl 1):S218.

32. Le Guiner C, Montus M, Servais L, et al. Forelimb treatment in a large cohort of dystrophic dogs supports delivery of a recombinant AAV for exon skipping in Duchenne patients. Mol Ther 2014;22:1923–1935.

33. Schnödt M, Schmeer M, Kracher B, et al. DNA minicircle technology improves purity of adeno-associated viral vector preparations. Mol Ther Nucleic Acids 2016;5:e355.

34. Fujita R, Matsuyama T, Yamagishi J, et al. Expression of *Autographa californica* multiple nucleopolyhedrovirus genes in mammalian cells and upregulation of the host β-actin gene. J Virol 2006;80:2390–2395.

35. Airenne KJ, Hu YC, Kost TA, et al. Baculovirus: an insect-derived vector for diverse gene transfer applications. Mol Ther 2013;21:739–749.

36. Kapranov P, Chen L, Dederich D, et al. Native molecular state of adeno-associated viral vectors revealed by single-molecule sequencing. Hum Gene Ther 2012;23:46–55.

37. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet 2010;19:R227–R240.

Received for publication December 19, 2016;
accepted after revision April 13, 2017.

Published online: April 21, 2017.