Molecular Therapy
Methods & Clinical Development
Original Article

Recombinant AAV batch profiling by nanopore sequencing elucidates product-related DNA impurities and vector genome length distribution

Florian Dunker-Seidler,[1,2] Kathrin Breunig,[1,2] Magdalena Haubner,[1] Florian Sonntag,[1] Markus Hörer,[1] and Rebecca C. Feiner[1]

[1]Ascend Advanced Therapies GmbH, Fraunhoferstraße 9b, 82152 Planegg-Martinsried, Germany

During production, recombinant adeno-associated virus (rAAV) capsids are equipped with heterogeneous genetic payloads including undesired DNA impurities as well as truncated vector genomes. Comprehensive analysis of encapsidated DNA by long-read next-generation sequencing is destined to guide platform optimization and provide crucial insights into safety of gene therapies. We used nanopore sequencing for in-depth profiling of an rAAV9 batch produced using our proprietary split two-plasmid system in a 50-L bioreactor. We compared three methods for single-strand to double-strand DNA conversion and their impact on the sequencing data. We observed a distinct library size profile but comparable impurity distribution. We contrasted recent nanopore sequencing advancements such as the V14 chemistry and dorado basecalling software with the widespread V9 chemistry and detected a markedly increased read quality. Our data highlight a high vector batch quality with low plasmid-derived and host cell DNA impurities of random origin, critical for mitigating associated safety risks. Finally, we compared nanopore data with orthogonal SMRT sequencing data and observed a higher base quality, but largely similar length and impurity profiles. Taken together, nanopore sequencing is a state-of-the-art method for comprehensive, in-depth rAAV vector batch analysis during all stages of gene therapy development.

INTRODUCTION

Recombinant adeno-associated virus (rAAV) is the most common vector for *in vivo* gene therapy and seven currently available AAV-based therapeutics have been approved by the Food and Drug Administration, the European Medicines Agency, and additional regulatory bodies worldwide.[1] Although major advances and maturation have been made in the gene therapy field in recent years, many challenges remain to ensure the safety and efficacy of AAV gene therapeutics. Despite its generally favorable safety profile, recently multiple severe adverse events were recorded during clinical trials and post market authorization using high vector dose, even resulting in the death of some patients.[2–4] This highlights the urgent need to develop therapeutics with increased efficacy enabling lower doses. While even wild-type (WT) AAVs are traditionally deemed non-pathogenic,[5] it is noteworthy that in recent reports WT AAV2 infections under specific circumstances were associated with severe hepatitis in children.[6–8] However, unlike WT AAV2, rAAVs used in gene therapies are devoid of replication competency even in the presence of helper virus and require additional WT-AAV infection for efficient mobilization.[9]

During AAV manufacturing, alongside the desired AAV vector genome with the transgene, distinct DNA sequences derived from the production plasmids, the host cell genome or recombinants thereof are packaged in small, but not insignificant amounts.[10] These represent a crucial part of the so-called product-related impurities and can have important implications on drug product safety and efficacy, demanding a detailed and comprehensive characterization. Whereas nowadays quantitative or digital polymerase chain reaction (PCR)-based methods remain the standard for impurity quantification especially in the regulatory environment,[11] they suffer from limited insights as any PCR-based method relies on pre-designed primers. Thus, all impurities that do not arise from the short sequence between the primers remain unexplored.

In contrast, next-generation sequencing (NGS) allows for a comprehensive analysis of all encapsidated DNA sequences and does not require any *a priori* information. NGS applications can be classified according to their read length into short-read (35–350 bases) and long-read (up to several thousand bases) sequencing. The most prevalent sequencing technologies are Illumina (short read), SMRT sequencing, and nanopore sequencing (both long read).[12] Whereas Illumina has been used to characterize rAAV vector sequence identity and to investigate product-related impurity profiles, the technology cannot resolve the highly structured inverted terminal repeats (ITRs) that are especially prone to mutation.[13–15] Because of its short reads, Illumina also fails to elucidate native vector genome length distribution. Recent protocols using landmarks to enable the re-construction of long reads from Illumina

Received 16 August 2024; accepted 20 January 2025;
https://doi.org/10.1016/j.omtm.2025.101417.

[2]These authors contributed equally

Correspondence: Rebecca Feiner, Ascend Advanced Therapies GmbH, Fraunhoferstraße 9b, 82152 Planegg-Martinsried, Germany.
E-mail: rebecca.feiner@ascend-adv.com

data[16] have to our knowledge not been applied for rAAV characterization. Long-read NGS methods overcome these limitations acquiring sequence reads of several thousand nucleotides from single molecules albeit with reduced sequencing depth and an originally increased error rate.[12] SMRT sequencing, commercialized by Pacific Biosciences, has been used to characterize entire vector genome sequences with their size distribution as well as the distribution of product-related DNA impurities.[17–20] However, SMRT sequencing requires relatively large input amounts (e.g., 1E12-1E13 vg for the commercial AAV sequencing services) but can provide up to 4 million reads with the latest device versions (Sequel IIe). A high accuracy comparable with short-read methods is achieved by sequencing each circularized SMRTbell repeatedly to build a consensus read from the results.

Nanopore sequencing, commercialized by Oxford Nanopore Technologies, relies on base-specific current alterations as a DNA strand is passed through a nanopore. The raw electric signal is basecalled by a neuronal network-based software.[21] Nanopore sequencing can provide reads from several thousand to even million bases in length; however, per base read quality is lower than Illumina and SMRT sequencing.[21] Nanopore sequencing has been initially used for rAAV sequencing with a transposase-based library preparation protocol leading to random vector genome shearing and thereby losing the native length distribution.[22] Another study described a ligation-based protocol yielding ITR to ITR full-length reads, but the quality of reads was reduced compared with SMRT sequencing leading to reporting of false-positive mutations in the rAAV vectors and even the production plasmids.[23] However, these studies used outdated sequencing technologies and basecalling algorithms prior to the introduction of the V14 sequencing chemistry and the dorado basecalling software by Oxford Nanopore Technologies. Thus, we wanted to test if these innovations can improve the quality of sequencing data making nanopore sequencing suitable for vector quality control and to guide vector cassette design or manufacturing process development.

Here, we present the nanopore sequencing-based characterization of a recombinant AAV9 batch with a reporter transgene produced in a 50-L bioreactor using up-to-date nanopore sequencing technology with a comparison with commercially available SMRT sequencing.

RESULTS

Enzymatic DNA conversion resulted in ITR-primed hairpin formation

The single-stranded genome of AAV must be converted into double-stranded DNA (dsDNA) to be amenable for ligation-based sequencing library preparation. Several approaches for this crucial step in the library preparation have been published ranging from annealing of (+) and (−) genomes,[19] second-strand synthesis by DNA polymerase I,[13] to omitting a dedicated second-strand synthesis step altogether.[22] To get a deeper insight into the DNA conversion step, we tested three approaches in parallel: annealing, second-strand synthesis by DNA polymerase I, and second-strand synthesis with reverse transcriptase, which also displays DNA-dependent DNA polymerase activity but is active at higher temperatures.[24] To our knowledge, reverse transcriptase has not been employed for this purpose previously. We hypothesized that this enzyme might be well suited for second-strand DNA synthesis especially at challenging ITR structures due to the higher reaction temperatures. As a model rAAV batch for this study we selected an rAAV9 vector batch encoding a secreted embryonic alkaline phosphatase (SEAP) reporter transgene produced in HEK293 suspension cells in a 50-L bioreactor. Non-encapsidated DNA was removed by Denarase treatment and the vectors were purified from the crude lysate by affinity chromatography. After isolation of the encapsidated DNA and conversion with the three methods, we inspected the input DNA and the converted DNA by agarose gel electrophoresis. We observed a dominant band at the expected size of our transgene for all three methods. As a comparison we denatured the input DNA to obtain single-stranded DNA (ssDNA), which ran at a lower size and no band was detected at the ssDNA size in either converted sample (Figure 1A). This suggested that all three methods are suitable for DNA conversion. To further inspect the conversion products, we analyzed them via TapeStation automated gel electrophoresis. We detected a very similar profile of the samples converted by annealing, reverse transcriptase, and DNA polymerase I with a single dominant peak at the expected full genome size (Figure 1B). Finally, we subjected the untreated viral vector, the extracted viral DNA, and the converted DNA to alkaline gel electrophoresis (Figure 1C). The high pH of the buffer not only ruptures the viral capsid, but also denatures the dsDNA to ssDNA.[25] We detected the expected ssDNA size (∼3,000 nt) for the viral vector, the DNA input, and the annealed sample, but surprisingly a much higher ssDNA size of around 6,000 nt for the sample converted by DNA polymerase I and reverse transcriptase. This larger molecule is likely to arise from hairpin formation through 3′ ITR extension during dsDNA conversion instead of random hexamer priming (Figure S1). Second-strand synthesis methods have been first described for the use with Illumina sequencing. This short-read method does not preserve the native DNA length, which provides an explanation why the extension during enzymatic conversion has never been intensively studied in the past. To our knowledge the first mention of a 3′ ITR extension which generates a template for SMRT sequencing was published just recently.[26]

Successful nanopore sequencing of the ITR-primed hairpin required the latest technology

To clarify the molecular identity of this DNA species, we processed the converted DNA into nanopore sequencing libraries using the recently introduced V14 chemistry. Raw sequencing data were processed using the new basecaller dorado. The size distribution of trimmed and quality filtered raw reads confirmed our observations from alkaline gel electrophoresis. Whereas the annealed library displayed the expected peak at ∼3,000 nt, the samples converted by DNA polymerase I and reverse transcriptase displayed a peak around 5,800 nt (Figure 1D). We annotated the features of representative reads of this population and revealed a hairpin structure, with a single ITR in the center of the read (Figure 1E). Notably, this structure is indistinguishable from a genuine dsDNA in standard agarose gel electrophoresis.

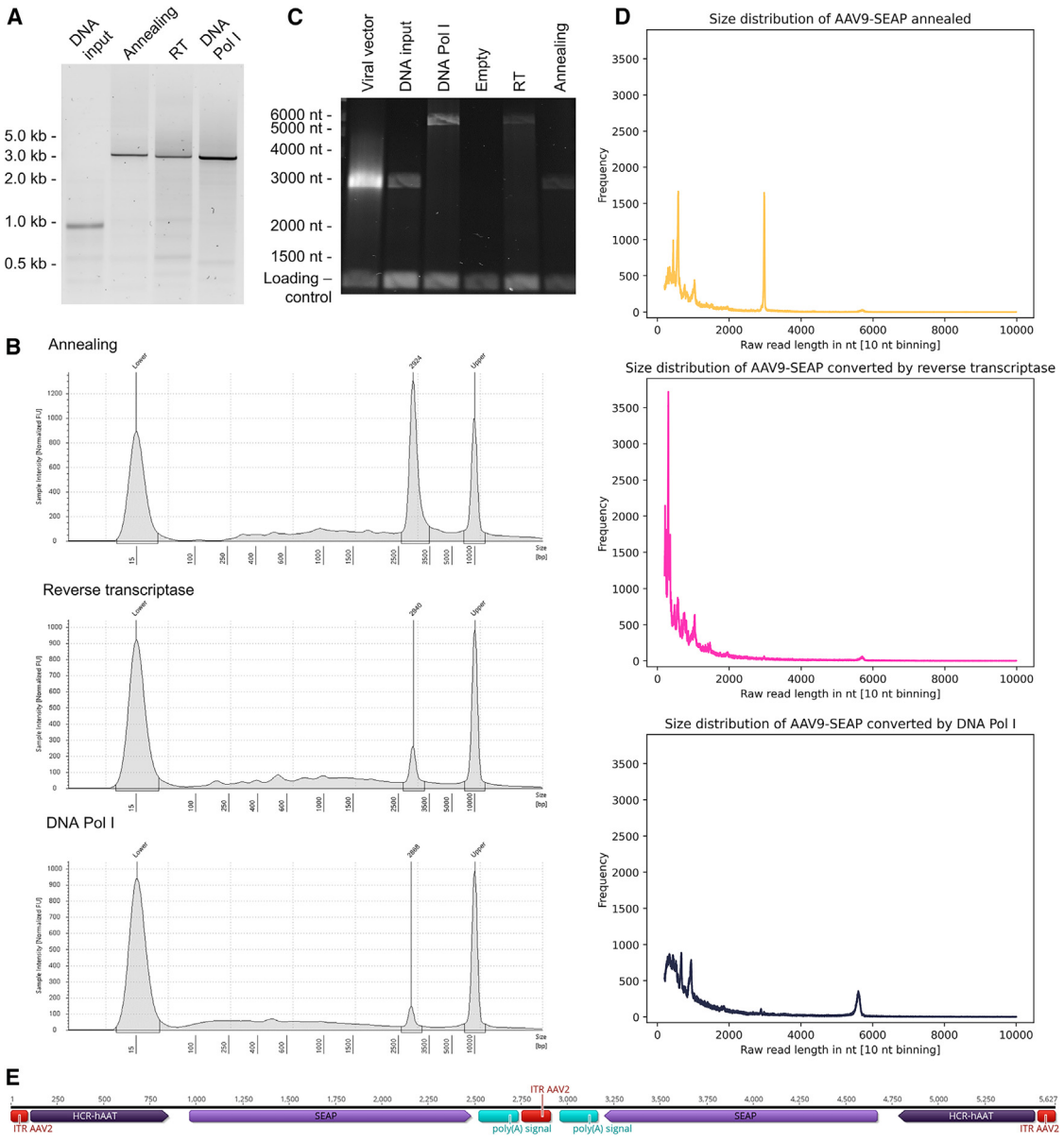

Figure 1. Enzymatic second-strand synthesis of viral DNA resulted in hairpin formation

(A) Agarose gel electrophoresis of ssDNA input and converted dsDNA revealed a virtually full conversion. The size of ssDNA of extracted rAAV does not match the double-stranded marker. (B) Tape Station profiles of dsDNA converted by annealing and enzymatically with reverse transcriptase and DNA polymerase I displayed a similar size profile with a single peak at the expected full genome size (2,985 nt). (C) Alkaline gel electrophoresis revealed a double-sized product after enzymatic conversion. In contrast, viral vector, input DNA and annealed samples ran at the expected ssDNA size of 3 kb. (D) Size distribution of raw nanopore sequencing reads. The annealed sample revealed a peak at 3 kb, but the enzymatically converted samples revealed a peak at 5.6–5.8 kb. (E) A representative read after dsDNA conversion by DNA polymerase I was annotated in Geneious Prime and displayed a hairpin structure with ITRs at both read ends and a single ITR in the middle.

We wondered why the molecular mechanism for enzymatic second-strand synthesis remained undiscovered in the past.[13] Thus, we used additional dsDNA converted by annealing and DNA polymerase I to generate sequencing libraries using the older and by now outdated V9 chemistry to have better comparability with previous data. After reverting to the former library preparation kit V9, the R9.4.1 flow cell and the guppy basecalling algorithm, we only detected marginal amounts of the double-sized species in the raw reads (Figure S2A). Instead, we observed a ~3,000 nt main population regardless of the DNA conversion step. This confirmed that the latest technology is required to successfully sequence the entire hairpin structure and explained why the molecular mechanism behind second-strand

synthesis was not discovered in the past for nanopore sequencing. As the V9 chemistry has been discontinued by Oxford Nanopore Technologies, the detailed reason for the inability to efficiently sequence the hairpin structure by previous chemistries remains unresolved.

Nanopore sequencing captured full-length reads but with a size bias for shorter reads

Besides the main peaks in the size distribution of the raw read length profiles, the individual samples comprising the three different conversion methods (annealing, reverse transcriptase, DNA polymerase I) also displayed a high proportion of shorter reads. We detected only a marginal main peak and many short reads <1,000 nt using reverse transcriptase, which is thus not the recommended default enzyme for DNA conversion. In contrast, a clear main peak with some short reads was recorded for the conversion by annealing and DNA polymerase I (Figure 1D). The short reads were not derived from premature sequencing termination, as for the length distribution analysis we required reads to have barcodes on both ends. We hypothesized that the share of truncated genomes may be higher than anticipated from TapeStation and agarose gel electrophoresis or that nanopore sequencing may favor short DNA fragments.

To clarify a potential size bias of nanopores sequencing, we digested λ phage DNA with EcoRI and HindIII to obtain equimolar fragments of a wide size range (Figure S2B). When we processed the fragmented λ phage DNA into a library and sequenced it by nanopore sequencing, we observed a clear size-dependent bias with the shorter reads being sequenced in far higher numbers (Figure S2C). This suggested that for a quantitative rAAV vector genome length analysis it is vital to co-sequence a lambda digest for post-sequencing size balancing as described previously for SMRT sequencing.[18] To verify that the secondary structure, especially the ITRs of the transgene itself, did not create any truncations, we digested the production plasmid with NotI to obtain an ITR-to-ITR transgene fragment that is identical to the sequence of a full-length AAV vector payload and a backbone fragment. The restriction fragments were used for library preparation and nanopore sequencing and two clear peaks of the expected sizes with marginal numbers of shorter reads were observed (Figure S2D). This experiment confirmed that nanopore sequencing is agnostic to sequence composition but has a size bias for shorter reads.

The V14 chemistry and dorado basecaller enhanced read and base quality

For a long time nanopore sequencing was marked by a reduced sequence quality compared with other sequencing technologies.[23] Recently, however, Oxford Nanopore Technologies released major performance improvements: the SQK-NBD114.24 library preparation kit, the R10.4.1 flow cell, and the dorado basecalling algorithm.[27] The upgrade from V9 to V14 chemistry featured a new, improved motor protein (E8.2) for the sequenced DNA and reformulated buffers, but also a new type of flow cell R10.4.1 using a nanopore with a larger reader head. On the bioinformatic side, Oxford Nanopore Technologies released the completely new basecaller dorado that is based on PyTorch, a high-performance deep learning algorithm and very large training datasets for enhanced performance. We tested whether these advancements qualify nanopore sequencing as a distinguished method for rAAV batch profiling. We compared the most recent nanopore technology V14 (SQK-NBD114.24/R10.4.1/basecalling dorado sup) with the widely used but outdated technology V9 (SQK-LSK109/R9.4.1/basecalling guppy hac). After demultiplexing and barcode trimming, we quality filtered the reads with nanoq using identical settings. We inspected the raw read quality with fastqc and observed a clear increase in average base quality from around 18.0 to 24.6 with a considerable portion of V14 reads even reaching qualities >30 (Figure 2A). We aligned the reads to the production plasmid references and used Alfred to determine alignment accuracy. The higher raw read quality enhanced the average alignment match rate from 96.8% to 98.0% (Figure 2B) and reduced the average rate of indels from 1.42% to 0.55% (Figure 2C). False indels are especially frequent sequencing artifacts of nanopore sequencing[28,29] and thus their reduction is probably a direct result of technological advances increasing raw read quality. To better dissect the influence of chemistry and basecalling improvements, we re-basecalled the raw data using dorado for the V9 data and guppy for the V14 data. Notably, whereas there was no difference for V9 data, we detected a marked improvement of base quality using V14 and dorado compared with V14 and guppy (Figure S3). We thus conclude that the better raw signal quality of the V14 chemistry provides the foundation for the dorado basecaller to realize its full potential. Taken together our data provide compelling evidence that recent advances of nanopore sequencing yield high quality datasets suitable for rAAV batch analysis.

Characterization of the AAV9-SEAP vector batch with a focus on encapsidated impurities

During manufacturing, most filled capsids are equipped with the correct transgene sequence; however, a small but not insignificant proportion exhibits undesired DNA fragments from the manufacturing plasmids. In this study, we used a split two-plasmid system with an AAV vector cassette/*cap* plasmid and a *rep*/AdV helper plasmid.[30] We characterized encapsidated DNA by mapping the raw reads to three references: the AAV vector genome from ITR to ITR, the vector plasmid backbone (comprising the bacterial backbone and the *cap* expression cassette) and the *rep*/AdV helper plasmid (Figure 3A). The helper plasmid shares a part of the backbone, comprising the origin of replication and the antibiotic resistance cassette, with the AAV vector cassette/*cap* plasmid. We removed this double region *in silico* from the helper plasmid reference to avoid mis-mapping as in contrast to the vector plasmid backbone it does not bear any neighboring packaging signals such as ITRs or a p5 promoter sequence.[31,32] A similar strategy masking identical regions in the references before mapping was applied to solve the mis-mapping problem previously.[14,15]

First, we inspected the length distribution of reads that were mapped to the AAV vector genome. In contrast to the raw reads which show the actual length of the DNA molecule (Figure 1D), the mapped proportion revealed as expected a peak at the transgene length of 2,985 nt

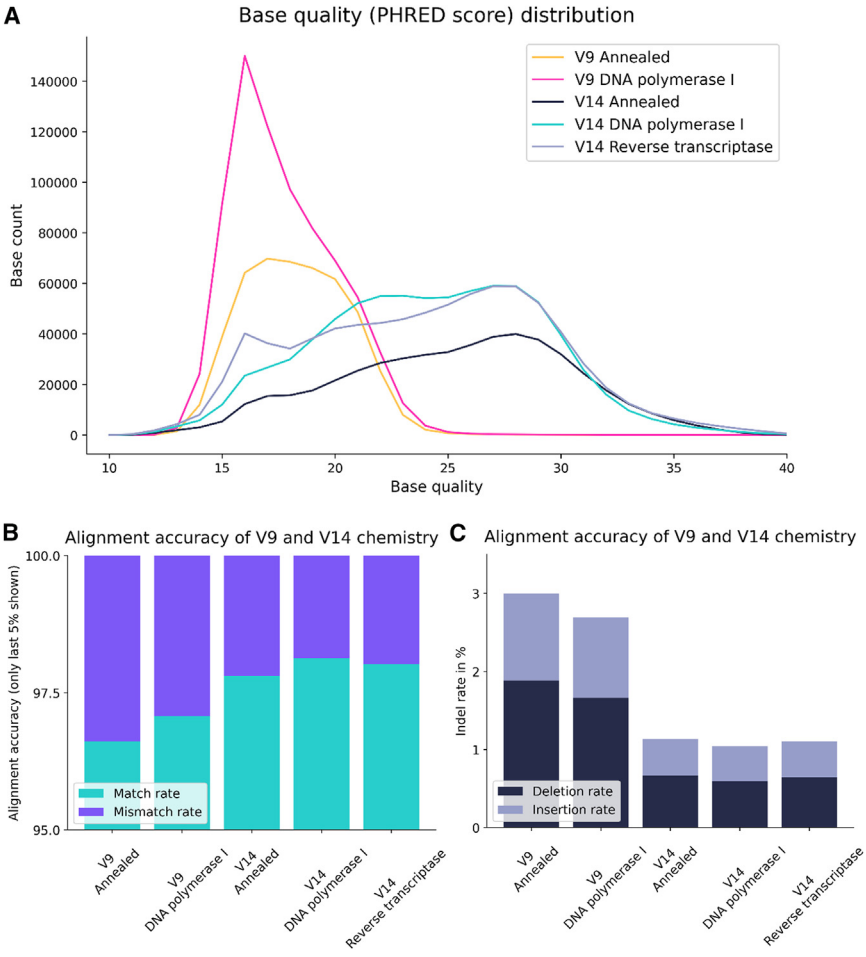

Figure 2. The V14 chemistry led to substantially improved read quality

(A) Comparison of base quality (PHRED score) between V9 and V14 chemistry of converted samples showed higher PHRED scores for V14 chemistry-treated samples. (B) The alignment accuracy of V9 and V14 chemistry was determined by Alfred and exhibited an increased match rate in V14 samples. (C) Indel rates determined by Alfred revealed a clear reduction in the V14 chemistry. Note that false indels are the dominant sequencing error in nanopore sequencing.

regardless of the library preparation method (Figure 3B). Once a read is mapped to the reference sequence the non-aligning ends are soft-clipped to focus on the well-aligned portions of the read. In this specific case, the hairpin structure which is generated during the 3′ ITR extension during second-strand synthesis would not match the reference and is soft-clipped. This can be visualized with the Integrative Genomics Viewer (IGV)[33] (Figure S4). As expected from the raw read profiles, the full-length peak was more prominent in the annealed sample and in the sample converted by DNA polymerase I. Notably, most shorter reads mapped to the AAV vector genome as well, especially the ITR sequences. After filtering of all sequences mapping to the AAV vector genome and plasmid derived impurities, we mapped the remaining reads to a human genome reference to explore encapsidated host cell DNA (HCD) impurities. This strategy reduces mis-mapping of reads to the host cell genome that in fact originate from the vector genome containing human promoter sequences and a transgene with high homology to human alkaline phosphatases. Overall, we observed comparable DNA profiles across all samples, with over 95% of the reads aligning to the AAV vector genome. The largest impurity source was the vector plasmid backbone with 1.65%–3.69% dependent on the second-strand synthesis protocol. In comparison, helper plasmid and HCD impurities were less abundant with around or less than 1% and less than 0.1%, respectively (Figure 3C; Table 1).

We inspected the mapping of the V14 sample converted by annealing, reverse transcriptase, and DNA polymerase I more closely using IGV.[33] We observed an overall similar mapping coverage pattern between the three conversion methods (Figures S5–S7). To evaluate both plasmid backbone impurity coverage and AAV vector genome coverage, mapping was conducted for the linearized plasmid (Figures S5A, S6A, and S7A). IGV analysis revealed a dominant coverage within the ITR-to-ITR vector cassette completely overshadowing the spurious impurity packaging. When splitting the coverage maps into AAV vector genome mapping and vector plasmid backbone, we spotted a relatively smooth coverage across the AAV vector genome with slightly higher coverage at both ITRs. The high read quality was highlighted by the absence of frequent mutations outside of the expected flip/flop polymorphisms in the ITR regions (Figures S5B, S6B, and S7B). When we inspected the read coverage on the AAV vector plasmid backbone, we detected an increased coverage at the ends of the reference adjacent to the AAV vector genome ITRs and upstream of the p5 promoter element. Both ITR and p5 promoter reverse packaging have been postulated previously as main drivers of impurity packaging.[31,32] Our mapping coverage data also indicated that most impurity reads are relatively short and did not span the entire cap gene (Figures S5C, S6C, and S7C). The helper plasmid mapping revealed very low amounts of impurity with a single significant hotspot upstream of the p5 promoter (Figures S5D, S6D, and S7D). The p5 promoter is required to drive *rep* gene expression from the plasmid. This region encodes the non-coding VA RNA required for AAV replication and inhibition of the innate immunity factor protein kinase R (PKR).[34,35]

Secondary structures can lead to AAV vector genome truncations reducing vector efficacy[18]; thus we used pysam to extract individual

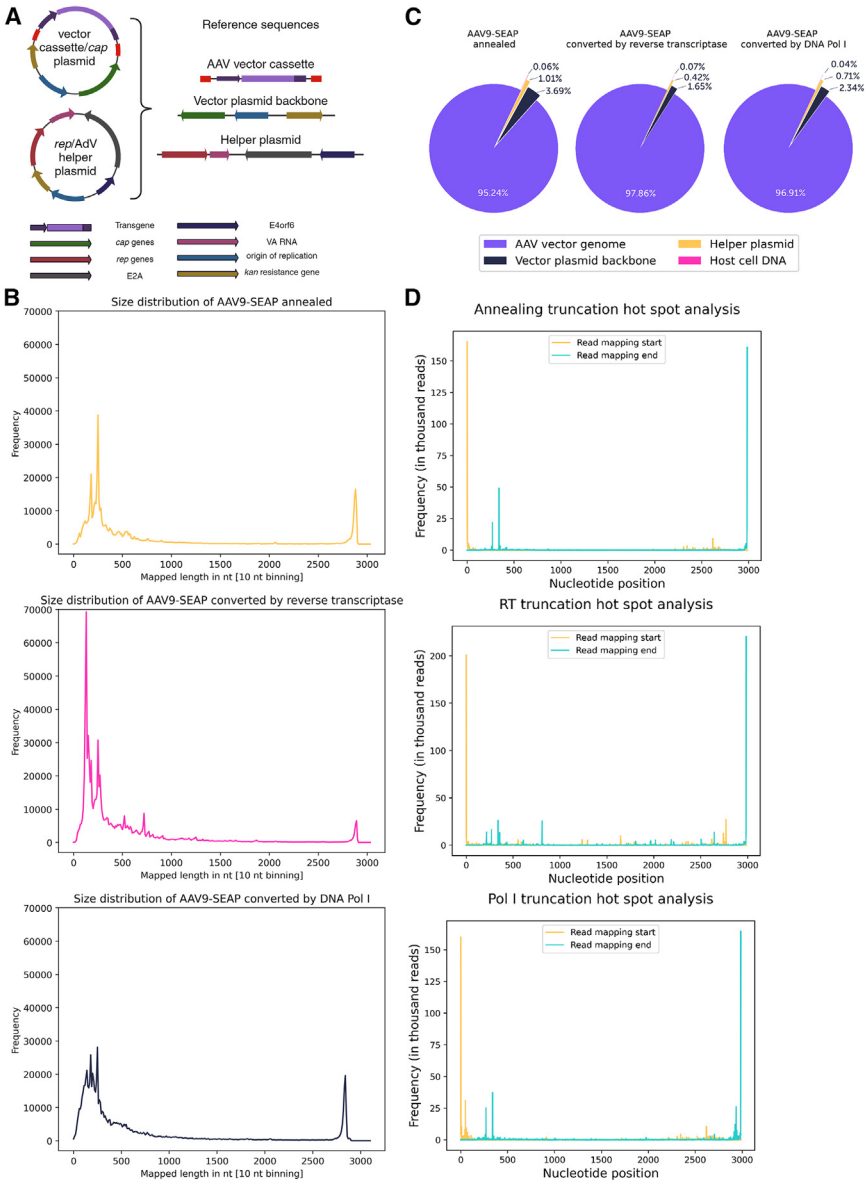

Figure 3. Nanopore sequencing is a method of choice for AAV batch analysis

(A) The split two-plasmid system and reference sequences used in this study. The separation of AAV vector genome and vector plasmid backbone (cap expression cassette and plasmid backbone) enabled detailed impurity analysis and characterization. (B) Mapped length of converted samples showed a peak for transgene size (2,985 nt) with high frequency in annealed and DNA polymerase I-converted sample. (C) Mapping distribution of libraries to the reference sequences revealed a high percentage of AAV vector genome mapping throughout all samples with the vector cassette plasmid backbone as the main impurity source. Host cell DNA-derived reads played only a marginal role. (D) Truncation hotspot analysis revealed very few intermediate read start and end events throughout the AAV vector genome for annealed and Pol I-converted sample. These were located in the highly structured promoter region (nucleotide positions 178–921). In comparison, the sample converted by reverse transcriptase revealed an increased number of truncation events.

read mapping start and endpoints for plotting. When we applied this truncation hotspot analysis to our V14 samples converted by annealing and DNA polymerase I, we observed a low frequency of premature vector termination with only two small peaks inside the promoter region. In contrast, the sample converted with reverse transcriptase revealed more truncation events probably due to the lower processivity of reverse transcriptase (Figure 3D).

In-depth analysis of HCD impurities

Our data revealed a very low content of HCD impurities of less than 0.1%. Still, this impurity is of special interest for regulatory authorities because of its potential oncogenicity and immunogenicity of aberrant polypeptides.[10] Therefore, although no detailed guidance exists for gene therapeutics, some researchers orient themselves using the recommendation of the World Health Organisation (WHO) for vaccines and biologics such as antibodies to reduce the amount of residual HCD to <10 ng per dose.[36] This limit, however, is very hard to meet with rAAV therapeutics[37] and encapsidated DNA might pose additional risks because of its efficient uptake into cells. Thus, pharmaceutical companies must rely on customized de-risking approaches, e.g., transcriptional profiling and NGS, which are primed to play a vital role in this task.

In addition to the molecular status (single- versus double-strand packaging), one crucial feature of HCD impurities is their length, as the WHO considers DNA fragments <200 bp as less critical with reduced safety implications.[36] We analyzed the mapped read length of HCD of the three libraries produced with the V14 chemistry and detected overall similar size profiles with the majority of reads <500 nt and only a few reads reaching a length of >1,000 nt (Figure 4A). Next, we investigated the reads classified according to their chromosome mapping origin to detect potential hotspots. To account for the distinct chromosome lengths, we normalized the read number by the chromosome size and observed an overrepresentation of chromosome 19-derived sequences in all three samples (Figure 4B). We closely inspected the mapping region of individual reads on chromosome 19% and ~80% of reads were derived from a single locus around chr19:44,924,250 (Figure S8A). This locus contains a regulatory sequence for the ApoE gene which is part of the HCR-hAAT promoter used to drive transgene expression in our vector construct.

Table 1. Detailed analysis of mapping locations by nanopore sequencing revealed >95% transgene mapping and <0.1% HCD mapping

	V14 annealed		V14 reverse transcriptase		V14 DNA polymerase I	
Reference sequence	Absolute no.	Mapping (%)	Absolute no.	Mapping (%)	Absolute no.	Mapping (%)
AAV vector genome	478,561	95.24	765,928	97.86	766,784	96.91
Vector plasmid backbone	18,536	3.69	13,442	1.65	18,510	2.34
Helper plasmid	5,056	1.01	3,436	0.42	5,645	0.71
Host cell DNA	326	0.06	555	0.07	356	0.04
Total	502,479	100	813,361	100	791,295	100

Thus, we suggest that those reads are derived from a rare mis-assignment of reads originating primarily from the transgene to the human genome. To test this hypothesis, we extracted the soft-clipped part of reads mapping to chromosome 19 with samextractclip from the jvarkit package and remapped the clipped parts to the production plasmid references. We confirmed the initial mis-assignment of a large proportion as 61.7% (sample converted by annealing), 90.9% (sample converted by DNA polymerase I), and 83.8% (sample converted by reverse transcriptase) clipped sequences mapped to the production plasmid references. We speculate that, in very rare cases, sequencing errors or vector genome rearrangements can lead to algorithm mistakes during the heuristic steps of minimap2 leading to false-negative alignment rejection.

Surprisingly, the sample converted by reverse transcriptase also revealed a strong overrepresentation of chromosome 11. However, a close inspection revealed that >95% of the assigned reads map to the telomere region and revealed a low-complex sequence typical of library preparation or sequencing artifacts (Figure S8B). This suggests again that using reverse transcriptase for conversion is not ideal and should be avoided. Overall, the detailed inspection of HCD mapping pattern clearly indicated that the number of true HCD reads is probably even further reduced by 10%–50% compared with the numbers reported by automatic quantification of mapped reads (Figures 3C; Table 1).

In summary, we provide evidence that genuine HCD reads originate randomly without specific enrichment, which is a crucial finding to further de-risk HCD packaging in HCD risk assessments. Our data point to a low level of spurious packaging of HCD and support the high vector quality obtained using this manufacturing platform. In addition, we argue that nanopore sequencing can provide comprehensive data to support chemicals, manufacturing and controls (CMC) risk assessments, and regulatory filings.

Taken together, our data highlight the high quality of an rAAV developmental batch and the suitability of nanopore sequencing to comprehensively characterize encapsidated DNA of rAAV vector batches.

Comparison with SMRT sequencing data from the same vector batch

SMRT sequencing is especially widespread for AAV sequencing, with its outstanding base quality and long-read length-spanning entire rAAV vector payloads. Therefore, we compared our nanopore sequencing data with the data from a commercially available AAV SMRT sequencing service from the same rAAV vector batch. We filtered the provided raw reads the same way as our nanopore sequencing reads and inspected the quality by fastqc. The sequencing data were of outstanding quality with the vast majority of bases achieving PHRED scores >85 (Figure 5A). We inspected the raw read length distribution and observed a dominant peak at the expected size of 2,985 nt (Figure 5B), comparable with our AAV9-annealed nanopore data (Figure 1D). However, we encountered a slight reduction of shorter reads, potentially suggesting a somewhat reduced size bias. In addition, SMRT sequencing revealed defined peaks for the shorter reads (Figure 5B), in contrast to a broad size distribution observed by nanopore sequencing (Figure 1D). After mapping the data to the same reference sequence described above (Figure 3A), we observed an overall comparable mapping origin distribution. Whereas the percentage of reads mapping to HCD and the helper plasmid was nearly identical to the corresponding nanopore sequencing, we observed a clear, 2-fold increase in reads mapping to the vector plasmid backbone and a parallel reduction of reads mapping to the AAV vector genome (Figure 5C; Table 2).

Chimeric reads, i.e., reads that map both to the AAV vector genome payload and to a distinct reference, are challenging to assign for mapping algorithms and data can be strongly influenced by the distinct numbers for SMRT and nanopore sequencing, which might explain the differences. Following a reviewer's suggestion, we investigated chimeric reads in our nanopore data and detected 1.8% of chimeric reads for samples converted with reverse transcriptase, 2.8% for DNA polymerase I-converted samples and 3.4% for annealed samples. As a marked contrast to these relatively similar numbers, we detected 8.9% of chimeras in the SMRT sequencing, which is more than 2-fold increase compared with nanopore data (Table S1).

We finally inspected the alignment file with the IGV browser and verified that, despite the distinct read distribution, the mapping coverage pattern of both the AAV vector genome and the plasmid-related impurities was comparable between SMRT and nanopore sequencing (Figure S9).

SMRT sequencing is used to determine vector truncation hotspots and vector genome instabilities.[19,38] When we compared the SMRT results with our nanopore sequencing data, we observed a very high agreement between both sequencing methods with two minor

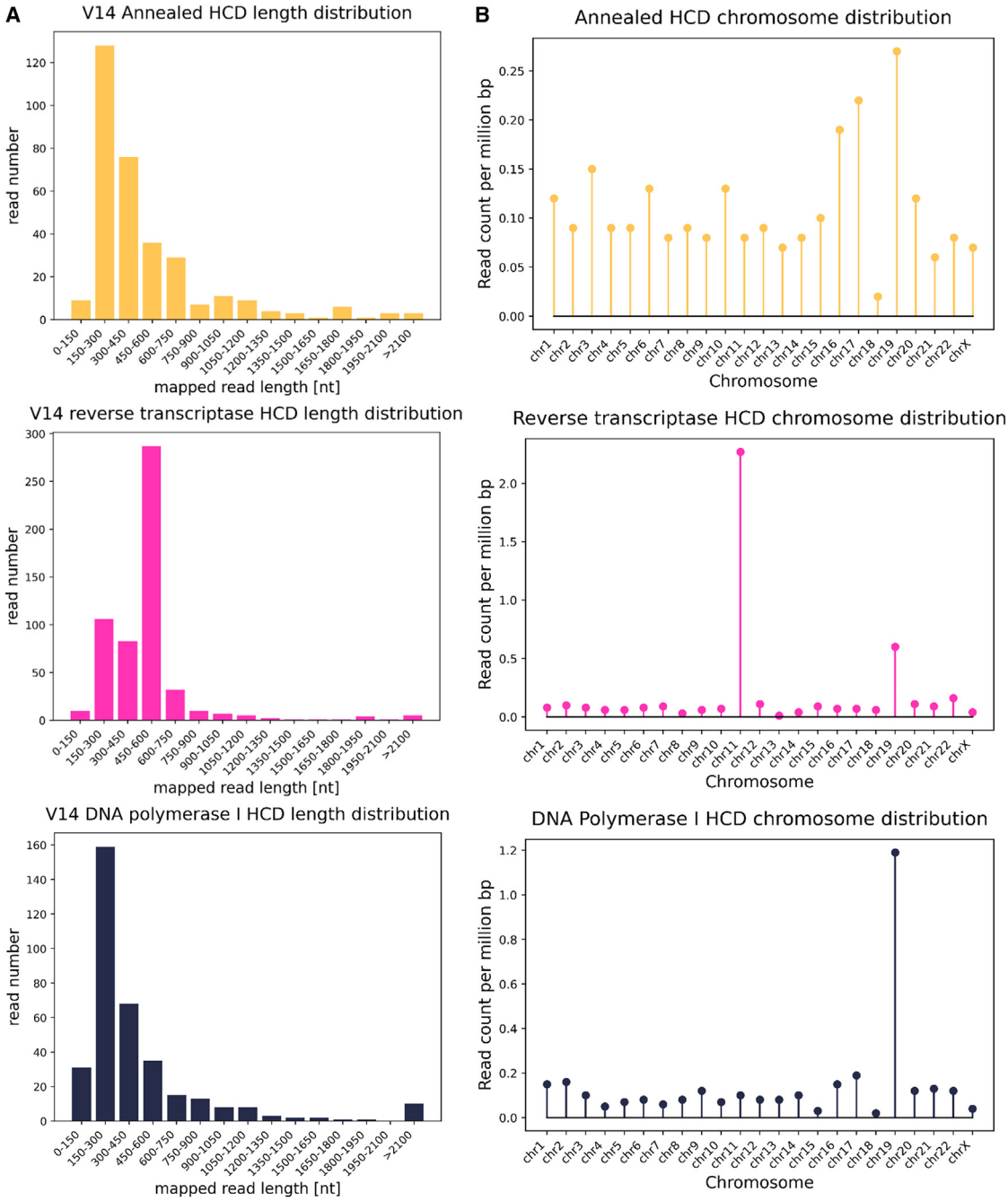

Figure 4. In-depth analysis of encapsidated host cell DNA (HCD)-derived impurities

(A) Size histograms revealed a comparable size distribution of HCD impurities dominated by shorter sequences and only few reads exceeding 1,000 nt. (B) The chromosome distribution of reads revealed an overrepresentation of chromosome 19 in all reads and a strong overrepresentation for chromosome 11 in the RT-converted library.

truncation hotspots in the transgene promoter region (Figures 5D and 3D). Overall, our results clearly indicate that, despite the lower raw read quality, nanopore sequencing can provide insights into rAAV vector batches largely comparable with SMRT sequencing, especially elucidating encapsidated DNA impurities.

DISCUSSION

NGS has attracted high interest for rAAV vector batch analytics, because unlike PCR-based methods, it provides comprehensive insights into encapsidated DNA heterogenicities and impurities.[39] In recent years, several groups have published valuable data of

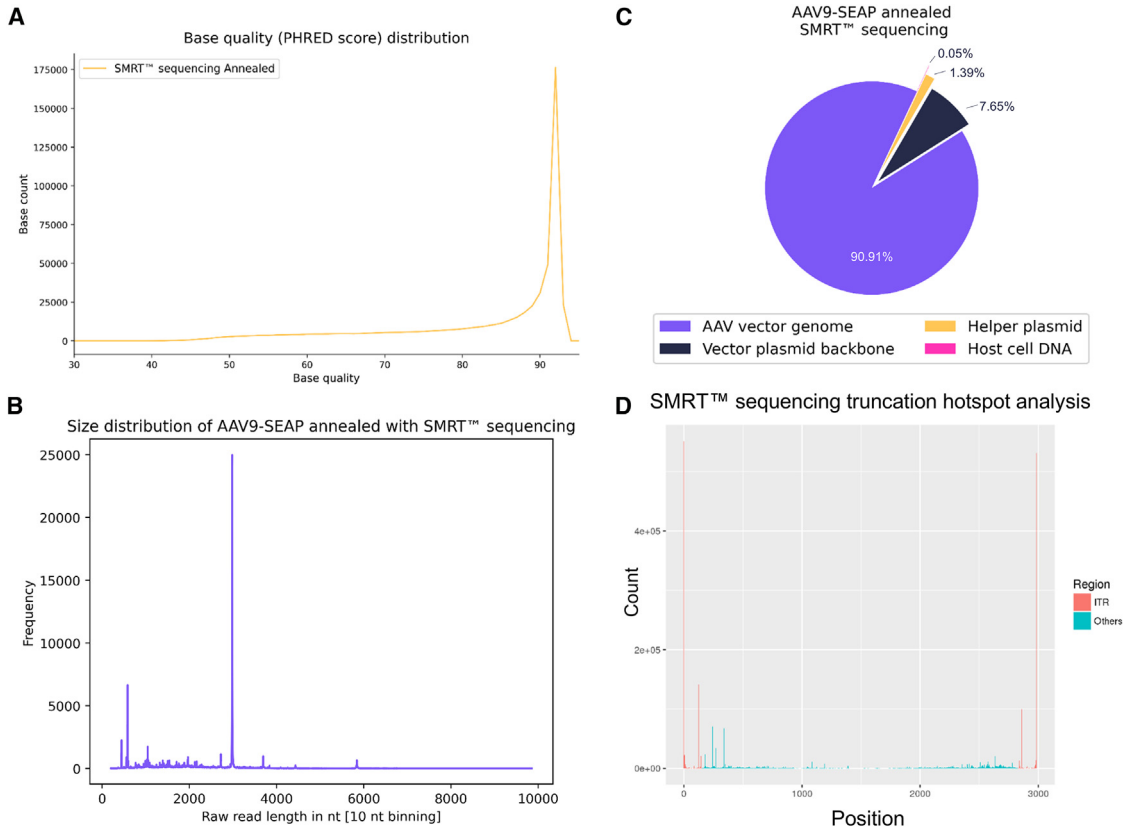

Figure 5. A commercial SMRT sequencing service based on annealing revealed comparable results with nanopore sequencing

(A) Raw read (PHRED score) quality distribution of SMRT sequencing confirmed outstanding read quality. (B) The raw read distribution revealed fewer short reads, but overall comparable results with nanopore sequencing, with a dominant peak at 2,985 nt. (C) Read distribution revealed >90% of reads mapping to the AAV vector genome. Helper plasmid and HCD impurity read rates were almost identical to nanopore sequencing. Vector cassette plasmid backbone impurities were 2-fold higher than nanopore sequencing, potentially because of different handling of chimeric reads containing parts of the AAV vector genome but extending into the vector cassette plasmid backbone. (D) SMRT truncation hot spot analysis revealed comparable results (Figure 3D).

sequenced rAAVs using various second-strand synthesis protocols as well as a broad range of sequencing methods.[38] However, typically for a fast-paced research field, data on the compatibility between distinct wet lab processes and sequencing methods have been lacking, limiting informed choices of protocols and analysis tools. In this study, we evaluated the applicability of nanopore sequencing for rAAV batch sequencing using different dsDNA conversion protocols and compared the data both with the outdated, but widely used V9 chemistry, and with SMRT sequencing.

Table 2. Detailed analysis of mapping locations by SMRT sequencing revealed >90% AAV vector genome mapping and <0.1% HCD mapping with ~12% chimeric read content

	SMRT sequencing annealed	
Reference sequence	Absolute no.	Mapping (%)
AAV vector genome	496,852	90.91
Vector plasmid backbone	41,825	7.65
Helper plasmid	7,606	1.39
Host cell DNA	268	0.05
Total	546,551	100

We started our evaluation by performing three different protocols for conversion of the extracted ssDNA into dsDNA amenable for library preparation: annealing of (+) and (−) genomes and second-strand synthesis by DNA polymerase I were used previously,[13,18] whereas second-strand synthesis with reverse transcriptase was developed during this study.

Enzyme-mediated second-strand synthesis by both DNA polymerase I and reverse transcriptase led to sequence reads of around twice the expected length, confirming our alkaline gel electrophoresis results. This finding was undetectable during native agarose gel electrophoresis and size fragment analysis by TapeStation, as these methods cannot distinguish between a genuine dsDNA and a pseudo double-stranded hairpin DNA. Therefore, it has likely

been overlooked so far by other researchers performing second-strand synthesis.[13–15] Interestingly, the widely used V9 nanopore sequencing chemistry did not efficiently resolve the entire hairpin sequence, also masking the molecular identity of the converted DNA. We suggest that the hairpin is formed by priming of the free hydroxyl group of the 3′ ITR. This intramolecular priming mechanism is faster than intermolecular binding of random hexamer primers, represents the natural biology enabling AAV replication (Figure S1), and has been exploited for second-strand synthesis prior to SMRT sequencing.[26]

Profiting from technological innovations in nanopore sequencing, we provide a comprehensive dataset investigating the quality of a vector batch produced using our proprietary split two-plasmid system in a 50-L bioreactor. We detected overall low DNA impurity levels confirming the high quality of the plasmid design and the upstream manufacturing platform. The detection of DNA impurities was only weakly impacted by the DNA conversion protocol and, in the case of the helper plasmid and HCD impurities, we also obtained very good data agreement between nanopore and SMRT sequencing. In contrast, nanopore sequencing revealed a considerably higher number of reads mapping to the AAV vector genome and a ∼50% lower level of vector cassette plasmid backbone impurities. IGV confirmed a similar mapping pattern with a considerable amount of backbone impurities starting in the AAV vector genome, namely the ITR region, and originating from ITR readthrough and ITR reverse packaging.[40] The handling of these chimeric reads is difficult for mapping algorithms, strongly dependent on the read length, and the position of read clipping determines its mapping location. We suggest that differences in chimeric read handling are the main reason for the disagreement of plasmid backbone-derived impurities and SMRT sequencing revealed a notably higher number of chimeric reads. We speculate that this is rooted in the distinct library preparation protocols as the typically longer incubation time during SMRTbell adapter ligation might slightly foster DNA fragment fusion. A detailed molecular characterization of this finding, however, extends beyond this study. Taken together, we propose to monitor and report the read mapping coverage pattern in addition to quantitative impurity distributions.

Besides product-related impurities, the rAAV transgene integrity is an important feature of vector safety and potency. Our nanopore sequencing-based truncation analysis revealed data aligning with data acquired by a commercial SMRT sequencing, confirming that nanopore sequencing is also suitable for this analysis. Whereas the transgene in this study was a model SEAP gene, our truncation analysis can also be applied for therapeutic transgenes and promoters, exposing vector instabilities earlier in development thus saving time and resources.

So far, nanopore sequencing has rarely been used for rAAV batch sequencing, even though, with its long-read length, low input amount, and competitive sequencing depth, it is ideally suited for batch characterization during process development and research. We speculate that its low usage is likely rooted in its previous lower read accuracy.[21] Even though the average base quality of consensus-based methods such as SMRT sequencing remains out of reach for nanopore sequencing, using the latest V14 chemistry and basecalling software we obtained a high number of more than 2 million high-quality sequencing reads per flow cell. This is 4 times higher than a typical, successful SMRT sequencing run and 10 to 100 times higher than reported previously with older chemistry and flow cell combinations.[23] In contrast, the required vector input is lower compared with SMRT sequencing (5E10-5E11 vg instead of 1E12-1E13 vg for SMRT sequencing), allowing nanopore sequencing assays during early development when vector amount is usually strongly limited. We did not observe any unexpected mutations that were frequent enough (>20%) to appear in the IGV mapping consensus, and alignment analysis revealed an overall accuracy of around 98%.

Thus, with the latest improvements, in contrast to Namkung et al.,[23] we regard nanopore sequencing basecall quality as being well-suited for in-depth rAAV batch analytics comparable with SMRT sequencing. We suggest that discrepancies between this study and others[22,23] are the result of advances in chemistry, basecalling, and bioinformatic analysis (Table S2). For example, for this study we profited from the long-read-optimized mapping algorithm minimap2 instead of the generic short read mapper BWA-MEM used by Namkung et al.[23] It is important to note that neither SMRT nor nanopore NGS are currently suited for GMP release testing but are powerful methods for additional characterization analysis. Efforts to translate both methods into GMP QC labs are under way and great progress has been made for nanopore sequencing in viral safety testing.

Compared with size distribution on agarose gels and TapeStation, nanopore sequencing displayed a surprisingly high number of short reads indicating a high degree of partial genomes. The sequencing of a linearized AAV vector genome confirmed that the native fragment sizes are not altered during sequencing library preparation or nanopore sequencing, irrespective of secondary structures such as ITRs. However, using an equimolar λ phage DNA input, we detected a strong bias for shorter reads, potentially because of their higher diffusion speed. We also detected shorter fragments, albeit a lower number, when using SMRT sequencing, so they were not a result of low read quality and following mis-assignment of reads. Notably, in contrast to nanopore sequencing the profile of the shorter reads revealed multiple sharp peaks. This pattern of distinct DNA lengths is neither reflected by Tape Station nor agarose gel electrophoresis. In comparison, the smooth shorter read profile of nanopore sequencing corresponded better to electrophoresis results, but the size bias appeared to be stronger. This size bias will require back-calculation with a size standard for fully quantitative vector size distribution as described for SMRT sequencing.[18] Importantly, even after balancing, the short sequences will make a substantial part of the sample and many of them contained an ITR sequence, which contributes to the payload of "partially filled" capsids.[20] It should be noted that the simplified downstream process used for manufacturing

of the AAV9-SEAP development batch did not include a full capsid enrichment step and thus the batch contains more partial and empty capsids than batches purified using a "full"-enrichment step. A thorough characterization of full, partial, and empty particles in this rAAV batch, however, extends beyond the scope of this study.

In summary, we provide conclusive evidence that nanopore sequencing enables an in-depth characterization of DNA packaged into AAV capsids. A major limitation of this study is that, because of limited resources and vector material, we analyzed only a single vector batch and did not perform technical replicates to comprehensively characterize the method robustness.

Using the new chemistry and bioinformatics tools, nanopore sequencing is a method of choice for in-depth rAAV characterization. It furthermore supports continuous rAAV platform and process development in an iterative manner. By comprehensive biochemical characterization and nanopore sequencing we demonstrated that current second-strand synthesis protocols lead to formation of a pseudo double-stranded hairpin. We verified a high vector batch quality with low HCD impurities, which were of random origin, for rAAV batches produced with our proprietary manufacturing platform. We revealed a size bias in nanopore sequencing, but no sequence-specific bias. These findings can help us to tailor future rAAV batch sequencing analyses and thereby improve AAV gene therapy's safety and efficacy.

MATERIAL AND METHODS

Plasmid design for vector production

Viral vectors were produced using a previously described proprietary split two-plasmid system.[30] It consists of a first plasmid containing the vector genome from ITR to ITR and the AAV9 capsid gene and a second plasmid that provides the *rep* gene as well as indispensable sequences from the adenoviral helper genes VA RNA, E2, and E4 required for AAV multiplication. The AAV vector genome encoded a SEAP reporter construct (Invivogen, Toulouse, France) under the control of a liver-specific HCR-hAAT promoter[41] flanked by AAV2 ITR sequences with a total vector genome length of 2,985 bases.

Virus batch

Viral vectors for nanopore sequencing analysis were produced using a proprietary HEK293 suspension cell-based process in a 50-L bioreactor (Biostat STR 50L, Sartorius, Göttingen, Germany). HEK293 cells were transfected with the AAV vector cassette/*cap* and the *rep*/AdV helper plasmid using PEIpro (Polyplus, Illkirch, France). Cells were harvested at 72 h post transfection, lysed mechanically (CF2, Constant Systems, Daventry, UK) and non-encapsidated DNA was removed by Denarase (c-LEcta, Leipzig, Germany) treatment. The viral vectors were purified using a proprietary process based on affinity chromatography with a POROS CaptureSelect AAVX affinity resin (Thermo Fisher Scientific, Waltham, MA).

Viral DNA extraction

DNA was extracted from 9E11 vector genomes with the High Pure Viral Nucleic Acid Kit (Roche Diagnostics, Mannheim, Germany) according to the manufacturer's instructions. The extracted DNA was quantified, and quality checked using a nanophotometer (NP80, Implen, München, Germany). The extracted DNA was used to prepare all five nanopore libraries in this study.

DNA conversion by annealing

Single-stranded DNA (150–200 ng) was diluted in the appropriate amount of 10× NEBuffer 2 (New England Biolabs, Frankfurt, Germany) and denatured at 95°C for 5 min in a PCR thermocycler (C1000 touch, Bio-Rad, München, Germany). The samples were slowly cooled to 25°C by setting the ramp speed to 0.1°C/s and inserting additional hold steps for 2 min every 10°C. Annealed dsDNA was purified using the Monarch PCR & DNA Cleanup Kit (New England Biolabs) according to the manufacturer's instructions.

Second-strand synthesis using DNA polymerase or reverse transcriptase

ssDNA (150–200 ng) was supplemented with 2 µL random primers (60 µM, New England Biolabs), 1 µL dNTPs (10 mM each, New England Biolabs), and filled up with nuclease-free water to 10 µL. Viral DNA was denatured for 5 min at 95°C and snap cooled on ice. Afterward a DNA polymerase (1 µL DNA polymerase I [New England Biolabs], 2 µL 10× NEBuffer 2, and 7 µL of water) or reverse transcriptase (1 µL Protoscript II [New England Biolabs], 4 µL 5× buffer, 2 µL 0.1 M DTT, and 3 µL water) master mix was added per sample, respectively. The samples were incubated at 25°C for 5 min, 37°C for 60 min, and 70°C for 10 min in a PCR cycler (C1000 touch, Bio-Rad, München, Germany) for DNA polymerase samples or at 25°C for 5 min, 48°C for 60 min, and 85°C for 5 min in a PCR cycler (C1000 touch, Bio-Rad). After second-strand synthesis, dsDNA was purified using the Monarch PCR & DNA Cleanup Kit after the manufacturer's instructions (New England Biolabs).

Native agarose gel electrophoresis

Agarose gels were prepared by dissolving the appropriate amount of agarose (GeneOn, Ludwigshafen, Germany) in 1× TAE buffer (Omega Bio-tek, Norcross, GA) by boiling in a microwave. GelRed (Biotium, Freemont, CA) was added to a 0.5× concentration from a 10,000× stock. The samples were supplemented with the appropriate amount of 6× purple loading dye (New England Biolabs), and electrophoresis was performed at 5 V/cm. DNA was visualized under UV light using a ChemiDoc station (Bio-Rad).

Alkaline gel electrophoresis

Alkaline gel electrophoresis of AAV and DNA was performed as described previously[25] with minor modifications. In brief, an alkaline agarose gel was prepared by dissolving 1% (w/v) agarose (GeneOn) in ultrapure water by boiling. After cooling to 50°C, the appropriate amount of 50× alkaline buffer (2.5 M NaOH, 50 mM EDTA) was added and the gel was cast. The samples were supplemented with the appropriate amount of 6× alkaline loading dye (18% Ficoll 400 [Carl Roth, Karlsruhe, Germany], 12%, v/v, 50× alkaline buffer, 24%, v/v, 10% SDS [Sigma-Aldrich, Steinheim, Germany], 0.25% xylene cyanole FF [Sigma-Aldrich], 0.15% bromocresol green [Carl

Roth]) and loaded on the gel. The gel was run at 4°C and 2.7 V/cm for 6 h and neutralized in 0.5 M Tris-HCl (pH 7.9). The DNA was stained with 1× SYBR Gold (Thermo Fisher Scientific, Eugene, OR) in 0.1 M Tris-HCl (pH 7.9), and visualized using a Fusion imaging system (Vilber, Eberhardzell, Germany).

Fragment analysis by TapeStation

Fragment sizes were determined with a TapeStation (Agilent, Waldbronn, Germany) device and the high sensitivity D5000 screentape system (Agilent) according to the manufacturer's instructions.

Preparation of control samples for nanopore sequencing

DNA from λ phage (Oxford Nanopore Technologies, Oxford, UK) was double digested with EcoRI-HF and HindIII-HF for 4 h at 37°C and the enzymes were inactivated at 65°C for 20 min. The DNA was precipitated using 0.1 vol 3 M sodium acetate (pH 5.2) (Thermo Fisher Scientific, Vilnius, Lithuania) and 2.5 vol ethanol absolute. The pellet was collected by centrifugation at 4°C, washed with 70% ethanol, air dried, and resuspended in elution buffer (New England Biolabs). The vector genome/*cap* plasmid was digested with NotI-HF (New England Biolabs) for 4 h at 37°C and purified using the Monarch PCR & DNA Cleanup Kit (New England Biolabs) after the manufacturer's instructions.

Library preparation and nanopore sequencing

Nanopore sequencing libraries were prepared from 100 to 200 fmol of input dsDNA using a ligation-based library prep kit with barcodes (SQK-LSK109/EXP-NBD104 or SQK-NBD114.24; all Oxford Nanopore Technologies) according to the manufacturer's instructions. The library was sequenced using an R9.4.1 or R10.4.1 flow cell (Oxford Nanopore Technologies), respectively, on a MinION Mk1c (Oxford Nanopore Technologies) sequencer with live basecalling disabled. Sequencing was terminated when the flow cell was exhausted and no new sequencing information was gathered (usually between 24 and 48 h).

Basecalling with dorado and guppy software

Raw POD5 data were basecalled using dorado (https://github.com/nanoporetech/dorado, v.0.4.3) and the super accuracy (sup) basecalling model, without demultiplexing and barcode trimming. As comparison data from the V9 chemistry were basecalled with a high-accuracy (hac) model with guppy_basecaller (guppy v.6.5.7) without demultiplexing and barcode trimming. Basecalled fastq files were demultiplexed and barcodes were trimmed by dorado demux or guppy_barcoder, respectively, using default settings. For raw read length analysis, the settings were changed to require barcodes on both ends for demultiplexing. For comparison, V9 raw data were also basecalled, demultiplexed, and trimmed using dorado with the respective sup model (v.3.6) and V14 data were basecalled, demultiplexed, and trimmed with guppy (v.6.5.7)

Sequencing data analysis

Demultiplexed data were filtered with nanoq[42] using a quality cut-off of 9, a minimal read length of 200 nt and a maximum read length of 10,000 nt. No such size and quality filtering was performed for the analysis of the λ DNA digest and the digested vector cassette. Sequencing read quality and length analysis was performed using FastQC (https://github.com/s-andrews/FastQC) and Alfred.[43] Single fastq reads were inspected and auto-annotated using a custom feature library by Geneious Prime (https://www.geneious.com).

The raw reads were mapped to a reference sequence consisting of the ITR-to-ITR vector cassette, the vector plasmid backbone, and the unique helper plasmid region with minimap2[44] using the -x map-ont flag for nanopore data and secondary alignments were suppressed. Non-mapping reads were extracted as fastq reads from the BAM file using samtools[45] and re-mapped to the human genome (version GRCh38_no_alt_analysis_set_GCA_000001405.15) with minimap2 to investigate HCD impurities. Mapped data were quantified and analyzed with the samtools[45] and the pysam (https://github.com/pysam-developers/pysam) package and plotted using matplotlib.[46] Sorted and indexed mapping files were visualized using the IGV browser.[33] Additional HCD analysis was performed using samextractclip from the jvarkit package (https://github.com/lindenb/jvarkit).

For chimeric read analysis, reads were mapped with minimap2[44] to a linearized plasmid reference. Using samtools we quantified the number of reads mapping within the ITR-to-ITR vector cassette with a 10 nt overlap into the backbone region to balance barcode trimming errors as well as the number of reads upstream and downstream of the ITR-to-ITR vector cassette, respectively. Reads counted twice were treated as chimeras.

Library generation, SMRT sequencing, and data analysis

Extraction of vector DNA, annealing, SMRT sequencing, truncation hotspot analysis, and chimeric read quantification was performed by GENEWIZ/Azenta Life Science (Leipzig, Germany). For in-house analysis, raw fastq sequencing data were filtered with nanoq[42] using identical settings as for nanopore reads. Sequencing read quality and length analysis was performed using FastQC (https://github.com/s-andrews/FastQC).

The raw reads were mapped to a reference sequence consisting of the ITR-to-ITR vector cassette, the vector plasmid backbone, the unique helper plasmid region with minimap2[44] using the -x map-hifi flag for SMRT sequencing data and secondary alignments were suppressed. Non-mapping reads were extracted as fastq reads from the BAM file using samtools[45] and re-mapped to the human genome (version GRCh38_no_alt_analysis_set_GCA_000001405.15) with minimap2 to investigate HCD impurities. Mapped data were quantified and analyzed with the samtools[45] package and plotted using matplotlib.[46]

DATA AND CODE AVAILABILITY

The data that support the findings of this study are available in the main and supplemental figures of this publication. Source data are available from the corresponding author upon reasonable request, subject to the relevant legal agreements with Ascend

Advanced Therapies being in place. Sequencing data are publicly available in the European Nucleotide Archive repository (https://ebi.ac.uk/ena) using the project accession PRJEB78700.

ACKNOWLEDGMENTS

The authors wish to thank Andreas Schulze and Bettina Finkbeiner for providing the AAV9-SEAP plasmid as well as the helper plasmid. The rAAV9-SEAP vector batch for this study was kindly provided by the Ascend Process Development department. The authors acknowledge the help of Sonya Schermann, Dan Brittain, Elisabeth Schweigert, Sophie Beer, and Jason King in English language support and content review. Some figures and the graphical abstract were prepared using Biorender. The study was funded by Ascend Advanced Therapies.

AUTHOR CONTRIBUTIONS

M. Hörer conceived the initial idea for the study. F.D.-S., K.B., R.C.F., and M. Hörer designed the experiments. F.D.-S. and K.B. performed the experiments and analyzed the data. F.D.-S., M. Haubner, and R.C.F. set up the bioinformatic analysis pipeline. M. Haubner performed the chimeric read analysis. F.D.-S., K.B., M. Haubner, R.C.F., and F.S. interpreted the data. F.D.-S. and K.B. drafted the figures and the manuscript. R.C.F., F.S., and M. Hörer revised and edited the manuscript. R.C.F. and F.S. supervised the study.

DECLARATION OF INTERESTS

All authors are employees and hold shares or options of Ascend Advanced Therapies, a contract development and manufacturing organization that commercializes rAAV manufacturing, process development, and analytics. F.S. and M. Hörer have filed several patent applications on the split two-plasmid system used in this study.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtm.2025.101417.

REFERENCES

1. Reid, C.A., Hörer, M., and Mandegar, M.A. (2024). Advancing AAV production with high-throughput screening and transcriptomics. Cell Gene Ther. Insights *10*, 821–840. https://doi.org/10.18609/cgti.2024.095.

2. Philippidis, A. (2022). Novartis confirms deaths of two patients treated with gene therapy Zolgensma. Hum. Gene Ther. *33*, 842–844. https://doi.org/10.1089/hum.2022.29216.bfs.

3. Lek, A., Wong, B., Keeler, A., Blackwood, M., Ma, K., Huang, S., Sylvia, K., Batista, A.R., Artinian, R., Kokoski, D., et al. (2023). Death after high-dose rAAV9 gene therapy in a patient with Duchenne's muscular dystrophy. N. Engl. J. Med. *389*, 1203–1210. https://doi.org/10.1056/NEJMoa2307798.

4. Shieh, P.B., Kuntz, N.L., Dowling, J.J., Müller-Felber, W., Bönnemann, C.G., Seferian, A.M., Servais, L., Smith, B.K., Muntoni, F., Blaschek, A., et al. (2023). Safety and efficacy of gene replacement therapy for X-linked myotubular myopathy (ASPIRO): a multinational, open-label, dose-escalation trial. Lancet Neurol. *22*, 1125–1139. https://doi.org/10.1016/S1474-4422(23)00313-7.

5. Gonçalves, M.A.F.V. (2005). Adeno-associated virus: from defective virus to effective vector. Virol. J. *2*, 43. https://doi.org/10.1186/1743-422X-2-43.

6. Ho, A., Orton, R., Tayler, R., Asamaphan, P., Herder, V., Davis, C., Tong, L., Smollett, K., Manali, M., Allan, J., et al. (2023). Adeno-associated virus 2 infection in children with non-A-E hepatitis. Nature *617*, 555–563. https://doi.org/10.1038/s41586-023-05948-2.

7. Morfopoulou, S., Buddle, S., Torres Montaguth, O.E., Atkinson, L., Guerra-Assunção, J.A., Moradi Marjaneh, M., Zennezini Chiozzi, R., Storey, N., Campos, L., Hutchinson, J.C., et al. (2023). Genomic investigations of unexplained acute hepatitis in children. Nature *617*, 564–573. https://doi.org/10.1038/s41586-023-06003-w.

8. Servellita, V., Sotomayor Gonzalez, A., Lamson, D.M., Foresythe, A., Huh, H.J., Bazinet, A.L., Bergman, N.H., Bull, R.L., Garcia, K.Y., Goodrich, J.S., et al. (2023). Adeno-associated virus type 2 in US children with acute severe hepatitis. Nature *617*, 574–580. https://doi.org/10.1038/s41586-023-05949-1.

9. Song, L., Samulski, R.J., and Hirsch, M.L. (2020). Adeno-associated virus vector mobilization, risk versus reality. Hum. Gene Ther. *31*, 1054–1067. https://doi.org/10.1089/hum.2020.118.

10. Wright, J.F. (2014). Product-related impurities in clinical-grade recombinant AAV vectors: Characterization and risk assessment. Biomedicines *2*, 80–97. https://doi.org/10.3390/biomedicines2010080.

11. Shmidt, A.A., and Egorova, T.V. (2021). PCR-based analytical methods for quantification and quality control of recombinant adeno-associated viral vector preparations. Pharmaceuticals *15*, 23. https://doi.org/10.3390/ph15010023.

12. Levy, S.E., and Myers, R.M. (2016). Advancements in next-generation sequencing. Annu. Rev. Genomics Hum. Genet. *17*, 95–115. https://doi.org/10.1146/annurev-genom-083115-022413.

13. Lecomte, E., Tournaire, B., Cogné, B., Dupont, J.-B., Lindenbaum, P., Martin-Fontaine, M., Broucque, F., Robin, C., Hebben, M., Merten, O.-W., et al. (2015). Advanced characterization of DNA molecules in rAAV vector preparations by single-stranded virus next-generation sequencing. Mol. Ther. Nucleic Acids *4*, e260. https://doi.org/10.1038/mtna.2015.32.

14. Penaud-Budloo, M., Lecomte, E., Guy-Duché, A., Saleun, S., Roulet, A., Lopez-Roques, C., Tournaire, B., Cogné, B., Léger, A., Blouin, V., et al. (2017). Accurate identification and quantification of DNA species by next-generation sequencing in adeno-associated viral vectors produced in insect cells. Hum. Gene Ther. Methods *28*, 148–162. https://doi.org/10.1089/hgtb.2016.185.

15. Lecomte, E., Saleun, S., Bolteau, M., Guy-Duché, A., Adjali, O., Blouin, V., Penaud-Budloo, M., and Ayuso, E. (2021). The SSV-Seq 2.0 PCR-free method improves the sequencing of adeno-associated viral vector genomes containing GC-rich regions and homopolymers. Biotechnol. J. *16*, 2000016. https://doi.org/10.1002/biot.202000016.

16. Marx, V. (2023). Method of the year: long-read sequencing. Nat. Methods *20*, 6–11. https://doi.org/10.1038/s41592-022-01730-w.

17. Xie, J., Mao, Q., Tai, P.W.L., He, R., Ai, J., Su, Q., Zhu, Y., Ma, H., Li, J., Gong, S., et al. (2017). Short DNA hairpins compromise recombinant adeno-associated virus genome homogeneity. Mol. Ther. *25*, 1363–1374. https://doi.org/10.1016/j.ymthe.2017.03.028.

18. Tai, P.W.L., Xie, J., Fong, K., Seetin, M., Heiner, C., Su, Q., Weiand, M., Wilmot, D., Zapp, M.L., and Gao, G. (2018). Adeno-associated virus genome population sequencing achieves full vector genome resolution and reveals human-vector chimeras. Mol. Ther. Methods Clin. Dev. *9*, 130–141. https://doi.org/10.1016/j.omtm.2018.02.002.

19. Tran, N.T., Heiner, C., Weber, K., Weiand, M., Wilmot, D., Xie, J., Wang, D., Brown, A., Manokaran, S., Su, Q., et al. (2020). AAV-genome population sequencing of vectors packaging CRISPR components reveals design-influenced heterogeneity. Mol. Ther. Methods Clin. Dev. *18*, 639–651. https://doi.org/10.1016/j.omtm.2020.07.007.

20. Tran, N.T., Lecomte, E., Saleun, S., Namkung, S., Robin, C., Weber, K., Devine, E., Blouin, V., Adjali, O., Ayuso, E., et al. (2022). Human and insect cell-produced recombinant adeno-associated viruses show differences in genome heterogeneity. Hum. Gene Ther. *33*, 371–388. https://doi.org/10.1089/hum.2022.050.

21. Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K.F. (2021). Nanopore sequencing technology, bioinformatics and applications. Nat. Biotechnol. *39*, 1348–1365. https://doi.org/10.1038/s41587-021-01108-x.

22. Radukic, M.T., Brandt, D., Haak, M., Müller, K.M., and Kalinowski, J. (2020). Nanopore sequencing of native adeno-associated virus (AAV) single-stranded DNA using a transposase-based rapid protocol. NAR Genom. Bioinform. *2*, lqaa074. https://doi.org/10.1093/nargab/lqaa074.

23. Namkung, S., Tran, N.T., Manokaran, S., He, R., Su, Q., Xie, J., Gao, G., and Tai, P.W.L. (2022). Direct ITR-to-ITR nanopore sequencing of AAV vector genomes. Hum. Gene Ther. *33*, 1187–1196. https://doi.org/10.1089/hum.2022.143.

24. Oscorbin, I.P., and Filipenko, M.L. (2021). M-MuLV reverse transcriptase: Selected properties and improved mutants. Comput. Struct. Biotechnol. J. *19*, 6315–6327. https://doi.org/10.1016/j.csbj.2021.11.030.

25. Green, M.R., and Sambrook, J. (2021). Alkaline agarose gel electrophoresis. Cold Spring Harb. Protoc. *2021*, 458–462. https://doi.org/10.1101/pdb.prot100438.

26. Zhang, J., Yu, X., Chrzanowski, M., Tian, J., Pouchnik, D., Guo, P., Herzog, R.W., and Xiao, W. (2024). Thorough molecular configuration analysis of noncanonical AAV

genomes in AAV vector preparations. Mol. Ther. Methods Clin. Dev. *32*, 101215. https://doi.org/10.1016/j.omtm.2024.101215.

27. Zhang, T., Li, H., Ma, S., Cao, J., Liao, H., Huang, Q., and Chen, W. (2023). The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. Appl. Environ. Microbiol. *89*, e0060523. https://doi.org/10.1128/aem.00605-23.
28. Bull, R.A., Adikari, T.N., Ferguson, J.M., Hammond, J.M., Stevanovski, I., Beukers, A.G., Naing, Z., Yeang, M., Verich, A., Gamaarachchi, H., et al. (2020). Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. Nat. Commun. *11*, 6272. https://doi.org/10.1038/s41467-020-20075-6.
29. Delahaye, C., and Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. PLoS One *16*, e0257521. https://doi.org/10.1371/journal.pone.0257521.
30. Hörer, M., Sonntag, F., and Kober, R. (2022). Plasmid system. Eur. Pat. EP3722434B1.
31. Chadeuf, G., Ciron, C., Moullier, P., and Salvetti, A. (2005). Evidence for encapsidation of prokaryotic sequences during recombinant adeno-associated virus production and their *in vivo* persistence after vector delivery. Mol. Ther. *12*, 744–753. https://doi.org/10.1016/j.ymthe.2005.06.003.
32. Brimble, M.A., Cheng, P.-H., Winston, S.M., Reeves, I.L., Souquette, A., Spence, Y., Zhou, J., Wang, Y.-D., Morton, C.L., Valentine, M., et al. (2022). Preventing packaging of translatable P5-associated DNA contaminants in recombinant AAV vector preps. Mol. Ther. Methods Clin. Dev. *24*, 280–291. https://doi.org/10.1016/j.omtm.2022.01.008.
33. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. https://doi.org/10.1038/nbt.1754.
34. Janik, J.E., Huston, M.M., and Rose, J.A. (1981). Locations of adenovirus genes required for the replication of adenovirus-associated virus. Proc. Natl. Acad. Sci. USA *78*, 1925–1929. https://doi.org/10.1073/pnas.78.3.1925.
35. Nayak, R., and Pintel, D.J. (2007). Adeno-associated viruses can induce phosphorylation of eIF2α via PKR activation, which can be overcome by helper adenovirus type 5 virus-associated RNA. J. Virol. *81*, 11908–11916. https://doi.org/10.1128/JVI.01132-07.
36. Knezevic, I., Stacey, G., and Petricciani, J.; WHO Study Group on cell substrates (2008). WHO Study Group on cell substrates for production of biologicals, Geneva, Switzerland, 11–12 June 2007. Biologicals *36*, 203–211. https://doi.org/10.1016/j.biologicals.2007.11.005.
37. Hebben, M. (2018). Downstream bioprocessing of AAV vectors: industrial challenges & regulatory requirements. Cell Gene Ther. Insights *4*, 131–146. https://doi.org/10.18609/cgti.2018.016.
38. Tam Tran, N., and Wl Tai, P. (2024). Profiling AAV vector heterogeneity & contaminants using next-generation sequencing methods. Cell Gene Ther. Insights *09*, 1565–1583. https://doi.org/10.18609/cgti.2023.206.
39. Fuentes, C., Staudhammer, J., Wright, J.F., Paulk, N., and Cross, S. (2023). Beyond empty and full: Understanding heterogeneity in rAAV products and impurities (Dark Horse Consulting Group). https://www.darkhorseconsultinggroup.com/post/heterogeneity-in-raav-vector-genomes-etc.
40. Brimble, M.A., Zhou, J., Morton, C., Meagher, M., Nathwani, A.C., Gray, J.T., and Davidoff, A.M. (2016). 547. AAV preparations contain contamination from DNA sequences in production plasmids directly outside of the ITRs. Mol. Ther. *24*, S218–S219. https://doi.org/10.1016/S1525-0016(16)33355-X.
41. Miao, C.H., Ohashi, K., Patijn, G.A., Meuse, L., Ye, X., Thompson, A.R., and Kay, M.A. (2000). Inclusion of the hepatic locus control region, an intron, and untranslated region increases and stabilizes hepatic Factor IX gene expression *in vivo* but not *in vitro*. Mol. Ther. *1*, 522–532. https://doi.org/10.1006/mthe.2000.0075.
42. Steinig, E., and Coin, L. (2022). Nanoq: ultra-fast quality control for nanopore reads. J. Open Source Softw. *7*, 2991. https://doi.org/10.21105/joss.02991.
43. Rausch, T., Hsi-Yang Fritz, M., Korbel, J.O., and Benes, V. (2019). Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. Bioinformatics *35*, 2489–2491. https://doi.org/10.1093/bioinformatics/bty1007.
44. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
45. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
46. Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.