bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Nanopore sequencing of native adeno-associated virus single-stranded DNA using a transposase-based rapid protocol

**Marco T. Radukic[1,†], David Brandt[2,†], Markus Haak[2], Kristian M. Müller[1,*] and Jörn Kalinowski[2,*]**

[1] Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany; [2] Center for Biotechnology (CeBiTec), Bielefeld University, 33594 Bielefeld, Germany

[†] These authors contributed equally to this work

[*] To whom correspondence should be addressed: Jörn Kalinowski; Tel: +49 521 106 8756; Fax: +49 521 106 89041; Email: joern@cebitec.uni-bielefeld.de or Kristian M. Müller; Tel: +49 521 106 6323; Fax: +49 521 106 156318; Email: kristian@syntbio.net

## ABSTRACT

**Monitoring DNA integrity and DNA contaminants in adeno-associated virus (AAV) gene therapy vectors is of major interest, because of clinical applications with increasing therapeutic doses. We here report direct, amplification-free nanopore sequencing of single-stranded AAV DNA using a rapid transposase-based protocol. Direct sequencing of bacteriophage M13 single-stranded DNA supports the finding that single-stranded DNA in general is amenable to direct transposase-based library generation, albeit with increased insertion bias. Sequencing AAV DNA from purified viral particles readily covered the otherwise notoriously difficult to sequence inverted terminal repeats and revealed single-nucleotide variants across the transgene cassette. Significant methylation of the packaged DNA was not identified. Furthermore, nanopore sequencing provided long reads up to full genome coverage and enabled detection of *a priori* unknown packaged DNA, which sets it apart from short read techniques or qPCR. Long reads directly revealed packaging of two fused genomes and fusions of a genome to the plasmid backbone. Preferred packaging of distinct forms of backbone DNA from producer plasmids, caused by a so far unknown mechanism, were uncovered. The findings promote direct nanopore sequencing as a fast and versatile platform for AAV vector characterization in research and clinical settings even on single-stranded DNA viruses.**

## INTRODUCTION

The adeno-associated virus (AAV) is a small, non-enveloped virus with a single-stranded DNA (ssDNA) genome. Recombinant AAV (rAAV) are preferred gene therapy vectors, with currently two drug approvals in the United States (Luxturna and Zolgensma). Their clinical success is driven by a low immunogenic profile and extrachromosomal stability of its genomes. AAV vectors are produced in eukaryotic cell culture by plasmid transfection. For production in mammalian HEK-293 cells, the wild-type AAV genome is separated onto two plasmids such that one plasmid carries the AAV genes *rep* and *cap* (pRepCap) and the other carries the gene of interest to be packaged into viral capsids (pITR). The gene of interest is flanked by the AAV inverted terminal repeat (ITR) sequences, which mediate genome replication and packaging. A third plasmid provides adenoviral helper functions (pHelper) (1, 2) and can be combined with pRepCap to one AAV helper plasmid (3). Based on AAV biology, AAV vectors harbour a single-stranded DNA, but this DNA can be designed to be self-complementary (4).

For quality control of AAV vectors in a clinical context, the state of the encapsulated AAV genome must be tightly monitored. Vector genome quality issues arise from falsely packaged contaminating DNA, which was initially identified by Southern hybridization and quantitative PCR methods to be *rep* and *cap* sequences (5) or sequences from the bacterial plasmid backbone (6). These can make up 0.5% to 6.1% of the cargo DNAs of AAV vectors, dependent on the plasmids used for AAV production (7). The same study found that the amount of contaminating DNA is even higher in self-complementary vectors. These contaminants should be avoided, as the transfer of bacterial sequences is linked to inflammatory response and gene silencing (8, 9) and *cap*-positive vectors have been shown to express AAV capsid proteins, potentially leading to an increased immune response to the vector and thereby impeding its efficacy (10).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


While DNA probe-based methods enable investigation of known contaminants, they only allow for a partial view of the sample. In search for unbiased methods to assess contaminations, next-generation sequencing (NGS) protocols have been developed. A first approach to AAV single-molecule sequencing relied on the Helicos HeliScope sequencer and identified low levels of contaminating plasmid DNA, but was deemed too expensive for routine quality control (11). An advancement in this field was an Illumina platform-based method for single-stranded AAV vectors (SSV-Seq), which identified—next to the known contaminations—randomly packaged host cell sequences and AAV purification-specific DNA impurities, as well as helper plasmid-derived impurities (12). AAV self-complementary vectors on the other hand are particularly amenable to NGS by the single-molecule real time sequencing (SMRT) approach, which revealed human DNA-vector chimeras, but requires double-stranded substrates (13). Illumina and SMRT are sequencing-by-amplification methods and, in general, require extensive sample preparation.

The rapid transposase-based protocol provided by Oxford Nanopore Technologies offers the advantage of amplification-free direct sequencing, thereby simplifying the sample preparation and potentially eliminating additional sources of bias. In regular rapid library generation, double-stranded samples are fragmented by a transposase and adapters are ligated to the sample fragments as part of the transposase reaction. The sample can then be directly used for nanopore sequencing. We report here the application of this convenient protocol for direct AAV single-stranded vector sequencing and sequencing of bacteriophage M13 single-stranded DNA with results obtainable within one working day. In addition, we demonstrate possibilities of large-scale single virus genome analysis.

## MATERIAL AND METHODS

**AAV production and ssDNA extraction.** AAV vectors were produced by the calcium phosphate triple transfection method in adherent HEK-293 cells. Cells were co-transfected by plasmids pRepCap (a plasmid encoding for the replicases of AAV serotype 2 and the capsid of AAV serotype 9) and pHelper (Agilent Technologies) which provide AAV adenoviral helper functions, and pITR (encoding fluorescence reporter mKate2 under control of a CMV promoter and human growth hormone polyadenylation signal), which provides the gene of interest to be packaged into viral capsids (Supplementary Figure S8-10). Three days after transfection, cells were harvested by scraping and lysed by three freeze-thaw cycles. Free nucleic acids in the soluble lysate were then digested with 60 units/ml Benzonase Nuclease (Merck Millipore) for 30 or 60 minutes at 37°C. Remaining short nucleic acids were removed by subsequent ammonium sulphate precipitation of the lysate and the culture media. The pellet of the ammonium sulphate precipitation was dissolved in PBS and the solution was run through a bed of Poros CaptureSelect AAVX (Thermo Scientific) affinity resin at 1 ml/min. The affinity material was washed with PBS containing 0.05% Tween 20 using at least ten times the bed volume. AAV vectors were eluted with 100 mM citric acid at pH 2.0 and the eluate was immediately brought to a neutral pH with 1M Tris, pH 8.8. AAV vectors were finally rebuffered to PBS containing a total of 180 mM NaCl and 0.005% Pluronic-F68 and stored at -80°C. A typical yield was $10^{13}$ DNaseI-resistant particles from five TC150 cell culture dishes.

Vector DNA was extracted by capsid disruption and subsequent silica-affinity based DNA purification. For this, the vector stock was first brought to 100 mM guanidine and 50 mM EDTA from a six-fold stock solution (Tris-buffered pH 8.0). Proteinase K (New England Biolabs) was added to a final concentration of 4 units/ml. Then, the mixture was incubated at 37°C for one hour and at 95°C for 20 minutes. Vector DNA was extracted from this solution by the NucleoSpin Gel Extraction and PCR Cleanup Kit (Macherey Nagel) as per the manufacturer's instruction.

**Bacteriophage M13 production and ssDNA extraction.** M13mp18-phagemid and the corresponding ssDNA were purchased from New England Biolabs. M13KO7 helper phage was produced as described in the supplementary information.

**qPCR analysis.** Quantitative PCR analysis was performed on a Roche LightCycler 480 II using the Promega GoTaq qPCR Master Mix. Primers, annealing temperatures and qPCR qualification data are given in the supplementary information.

**Nanopore sequencing sample preparation.** 9 µl or up to 400 ng equivalent of the vector DNA and 1 µl fragmentation mix were used for preparation of barcoded libraries using the Oxford Nanopore Rapid Barcoding Kit (RBK-004) and sequenced with R9.4.1 MinION flow cells on the Oxford Nanopore GridION sequencing machine. Non-barcoded libraries were prepared using the RAD-004 kit. For the sequencing of the commercially obtained M13mp18 DNA, 400 ng (according to the manufacturers' concentration measurements) of each dsDNA and ssDNA were used.

**Data evaluation.** Basecalling was carried out using ont-guppy-for-gridion (v3.0.6) with the high accuracy model (dna_r9.4.1_450bps_hac.cfg). Porechop (v0.2.4) was used for adapter-trimming and demultiplexing. Reads were mapped to the reference sequences using minimap2 (14) (v2.10-r761) with the map-ont preset. Per-base read coverage was calculated using BEDTools genomecov (15) (v2.27.1) separately for both strands. Assignment of reads to

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


the respective subject sequence was done using BLASTn (16) (E-value threshold: 1e-25). BLASTn results were analyzed using a custom python script, counting high-scoring segment pairs (HSPs) to each subject and in the case of multiple HSPs of a single query making a subject assignment based on the highest bitscore. Detection of CpG methylation was carried out by realignment of the Nanopore raw data against the respective reference sequence using the re-squiggle algorithm of Tombo (v1.5) (17) and subsequent analysis using DeepSignal (18) with standard parameter settings and the supplied CpG model (model.CpG.R9.4_1D.human_hx1.bn17.sn360). Single-nucleotide variants (SNVs) were called using Longshot (19) (v0.3.5), with a strand bias p-value cutoff of 0.01 and a maximum coverage of 500,000.

Determination of transposition sites was carried out using a custom python script. Untrimmed reads with more than 500 nt length were mapped to the reference genome using minimap2 and the map-ont preset, excluding secondary alignments and alignments shorter than 100 nt. For each read, the alignment closest to the read start was selected and the start coordinate on the reference sequence (the end coordinate for mappings against the negative strand) was taken as an estimate for the insertion site. Each remaining read was realigned against a set of 31 reference sequences, each consisting of the transposase adapter sequence (supplementary information) and 75 nt chromosomal sequence downstream of each position within 15 nt proximity to the estimated insertion site. If the highest scoring realignment was longer than 100 nt and comprised at most 3 gaps and/or insertions in a sequence window of 10 nt around the start of the genomic sequence, it was considered as a transposition site.

To assess possible dsDNA conformation of ssDNA, the propensity of nucleotide regions in single stranded genomes to be double-stranded during transposase-based library preparation was estimated by calculating ss-counts, where a ss-count is the number of times a base is single stranded in a group of predicted foldings. Calculations were based on one hundred predicted ssDNA folding structures using mfold (20) (v3.6) with parameter settings "W=10", "T=25" and "LC=circular" in case of circular genomes.

## RESULTS

The intention of this study was to find a general and fast protocol for AAV ssDNA genome sequencing for quality control of virus batches. We chose nanopore sequencing and reasoned that a convenient way of library creation would be a transposase-based protocol, in which a transposase randomly cleaves the DNA and ligates the fragments to sequencing adapters. If desired, DNA barcodes for sample assignment in multiplexed sequencing could also be added. Tagmentation with Tn*5* transposase has been used for Illumina dye sequencing of randomly primed AAV ssDNA before (21). However, the presented approach relied on a multi-step sample preparation to gain a double-stranded tagmentation substrate. Direct adapter ligation is therefore desirable. Transposases used for rapid library creation in next-generation sequencing are, to the best of our knowledge, not known to use single-stranded DNA (ssDNA) as transposition substrate. We considered methods to obtain double-stranded substrates such as priming the genome at the inverted terminal repeats and subsequently generating the complementary strand with a polymerase. On the other hand, we assume that the AAV inverted terminal repeat (ITR) sequences located at both AAV genome termini are probably already present as dsDNA and could suffice for transposase fragmentation and adapter ligation. Furthermore, AAV packages one of both DNA strands of its genome with equal probability, with the minor exception of some ITR-modified variants that package only a single-polarity genome (22). DNA extracted from AAV vector stocks might therefore already be in a partly double-stranded state, which should enable direct library creation without prior second-strand synthesis. We followed the two routes of either ITR priming and second-strand synthesis or direct library creation with $10^{11}$ vector genomes in both cases. Indeed, sequencing reads of comparable quality were obtained from both samples (data not shown). The overall read count in this initial test was low, which we attributed to the low DNA input. At this point, we saw the potential for direct sequencing of AAV ssDNA as a convenient characterization tool. Direct sequencing of the ssDNA genome is a preferred method, because hands-on time and thereby additional sources of bias are reduced. We did not observe insertion bias towards the ITRs in this initial experiment and therefore wondered, if ssDNA in general might be a valid substrate for the transposase reaction. At this point of course, we could not rule out strand hybridization as the cause of successful library generation.

### Bacteriophage M13 ssDNA is amenable to direct nanopore sequencing

We tested our hypothesis of generalized transposase-based sequencing of ssDNA by sequencing of M13 phage DNA, which is a commonly used ssDNA reference. Unlike AAV, the bacteriophage M13 packages only one circular strand referred to as the (+) strand during propagation. DNA prepared from this phage therefore is uniform and double strand

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


formation is unlikely. We obtained commercial M13mp18 ssDNA and corresponding dsDNA phagemids. The direct preparation of the transposase-based library from M13 ssDNA and nanopore sequencing was then carried out as before without prior second-strand synthesis. Conforming with our hypothesis, the M13 ssDNA sample was readily sequenced. Sequencing yielded 5,841 reads with an N50 of 6,887 bp for the single-stranded M13 DNA. 5,704 of the total 5,841 reads mapped to the reference sequence of M13mp18. Thereof, 5,591 reads mapped to the (+) strand and 113 reads mapped to the (-) strand, which corresponds to a ssDNA purity of 98%. Furthermore, 3,165 (+) reads and 42 (-) reads passed the filtering criteria to estimate transposase insertion sites. (Figure 1 A).

From the M13mp18 dsDNA sample, 384,091 reads with an N50 of 7,224 bp were generated and in contrast to the ssDNA sample, reads of the phagemid sample mapped to both strands with near-equal distribution. Of 382,079 total mapped reads, 191,446 reads mapped to the (+) strand and 190,633 reads mapped to the (-) strand, respectively. For the estimation of transposase insertion sites, 110,994 (+) reads and 105,783 (-) reads passed the filtering process (Figure 1 B).

Regarding reactivity, we found that our ssDNA samples gave a significantly reduced output compared to the dsDNA phagemid sample. On first sight, the mapped reads were evenly distributed over the reference sequence for all data sets, however when we plotted the corresponding transposition sites, hot spots were apparent only for the ssDNA sample (Figure 1 A and B) and 18% of total reads started at these positions. There was no clear correlation of these hot spots to the substrates ss-count in mfold (23), which would indicate transposase preference towards dsDNA stretches (Supplementary Figure S1). We therefore searched for local mismatched hairpins within the M13mp18 sequence with EMBOSS software suite (24) but we found no hairpins (gap penalty: 6) and also no palindromes (mismatches allowed: 5) that correlate with read start hot spots (not shown), where palindromes may be indicative of intermolecular base pairing. We repeated the experiment with M13KO7 helper phage propagated in our lab and obtained a similar result (Supplementary Figure S2). We deduce from this that the transposase has indeed enough activity on ssDNA substrates to be applicable for direct sequencing.

It can be deduced from the patent literature that the transposase in the Oxford Nanopore protocol is the MuA transposase (25). The mechanism of this transposase is complex, and ssDNA has been shown to be a cleavage Mu-end substrate (26), but not a target for a transposition event. Whether mechanistically, the transposase truly acts on ssDNA or whether the activity is due to spontaneous (self-) annealing on short stretches of a few bases is not clear from the experiment, although the fact that we did not observe insertion bias towards the ITRs in the preliminary experiment and the missing correlation between insertion hot spots and DNA fold hints on the former mechanism. For AAV vector quality control by direct library generation and nanopore sequencing, these results mean that the presence of both strands in the sample is not a necessity and that also ssDNA contaminations are accessible by this method.

**Nanopore ssDNA sequencing allows for direct, amplification-free sequencing of AAV vectors**

As we had observed a relatively low AAV read count in our initial test, we next optimized ssDNA extraction from AAV to gain more reads. In the end, we settled with an AAV purification protocol based on Benzonase nuclease digest of the producer cell lysate, ammonium sulphate precipitation and subsequent Poros Capture Select AAVX affinity chromatography. Residual Benzonase inactivation and capsid disruption for ssDNA release was then performed with 50 mM EDTA, 100 mM guanidine and proteinase K at slightly basic pH. Afterwards, ssDNA purification from this solution was achieved by silica-adsorption chromatography with a commercial kit and the eluate was used for the transposase reaction. In the end, using this protocol, we performed two independent sample preparations with a time delay in between of three months starting from individual cryo-cultures of producer cells, with five TC150 cell culture dishes each. These samples will be referred to hereafter as sample 1 and sample 2 with their sequencing runs being run 1 and run 2. We obtained about $10^{13}$ DNase I-resistant particles after affinity chromatography from both cultures. From these AAV particles we were able to obtain 50 µl DNA solutions with optical densities of $OD_{260,\ 10mm}$ = 0.40 and $OD_{260,\ 10mm}$ = 0.85, corresponding to an equivalent total of 1.0 µg and 2.1 µg dsDNA. Since DNA in these samples might be partially single-stranded and double-stranded, and since these forms have different absorption coefficients at 260 nm, we prefer to use volumes and optical densities for indications of quantities. Agarose gel electrophoresis of these two

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


**Table 1.** BLASTn read assignments and qPCR results for two independently produced and sequenced rAAV samples (sample 1 and 2)

**A Nanopore BLAST bins as percent of total hits**

| | Run 1 (sample 1) | | Run 2 (sample 2) | |
|---|---|---|---|---|
| **Group/ Threshold** | **>500 nt** | **>1000 nt** | **>500 nt** | **>1000 nt** |
| rAAV genome | 97.06% | 97.38% | 97.95% | 98.03% |
| pITR | 0.69% | 0.86% | 0.53% | 0.71% |
| pRepCap | 0.96% | 1.01% | 0.70% | 0.82% |
| pHelper | 0.12% | 0.13% | 0.10% | 0.10% |
| hg38 | 1.18% | 0.68% | 0.72% | 0.34% |

**B qPCR results as percent of total (measurable) with 95% confidence interval**

| Primer | Sample 1 | Sample 2 |
|---|---|---|
| bla | 2.0% ± 0.3% | 2.9% ± 0.4% |
| Rep | 0.22% ± 0.04% | 0.24% ± 0.04% |
| E4 | 0.062% ± 0.009% | 0.08% ± 0.01% |

**A**: Total contamination levels in both samples are independent on the read quality thresholds tested here, however the individual share of contaminations shifts towards higher amounts of human genomic sequences for the lower threshold. **B**: qPCR results lay in comparable ranges to the sequencing results, although a larger discrepancy is seen for the second sample in terms of *bla* and for *rep*-sequences in general.

samples, directly after extraction and after sample freeze-thaw, showed distinct bands attributable to the rAAV genome in single-stranded, hybridized and aggregated states (Supplementary Figure S3).

We used 9 µl of each of the two AAV samples for the transposase reaction and sequenced sample 1 on a pre-used flow cell with sample assignment by barcodes (run 1). A fresh flow cell was used for sample 2 without barcoding (run 2). Again, as expected, both samples were readily sequenced, but gave vastly different read counts that passed our initial length quality threshold of ≥1,000 bases read length (22,174 reads for run 1 versus 291,036 reads for run 2). We performed a first mapping analysis of these reads and found that the vast majority of these raw reads mapped to the reference genome (Table 1 A). Coverage steadily increased until it reached a stable plateau at about half the genome length. At the ITRs however, a sudden decrease in coverage was observed (Figure 1 C displays run 2). Nonetheless, ITR coverage is still 222,712-fold in run 2 and ITR sequences are thus accessible by nanopore sequencing despite their known tendency to form secondary structures. In the transposase insertion site analysis of reads longer than 500 nt of AAV sample 2, 220,075 (+) and 189,481 (-) reads passed filtering, revealing that strand-specific hot spots were again apparent, although overall, most reads started throughout the genome (Figure 1 C). These hot spots again did not correlate with the DNA fold and correlations to the GC content were minimal (Supplementary Figure S4). Interestingly, the read start pattern seems to be a combination of the patterns observed for M13 ssDNA and dsDNA.

### Direct nanopore AAV ssDNA sequencing reveals single-base heterogeneity and methylation status

Comparison of the assembled genome to the reference sequence revealed single-nucleotide variants, as seen before for rAAV (12). The high coverage enabled these conclusions despite the relatively low base accuracy of about 93%, which is an intrinsic property of the current nanopore sequencing technology. SNVs were located within ITRs in the short hairpins and were transversions as well as transitions with an individual abundance of about 20%. We were able to link ITR SNVs to the two possible ITR configurations in FLIP and FLOP, so that the found ITR SNVs are in the end expected to arise. On the other hand, prominent SNVs across the transgene cassette were mostly transitions with a hot spot located in the polyadenylation signal and throughout the CMV promoter with an abundance up to 30% (Figure 1 D). Raw reads were also analyzed for methylated CpG sites separately for both strands using a custom software workflow with Tombo and DeepSignal. The studied rAAV genome contains 129 CG dinucleotides, 123 of which have been mapped in run 2. When we compared reads from run 2 to reads of *in vitro* amplified rAAV genomes, no substantial methylation was identified. However, these results are based on the current algorithms used by the applied software and need to be verified by additional experiments.

Our nucleotide reference database so far contained only the rAAV genome. We next extended this database to include the human genome build hg38 (GCF_000001405.39) as well as the utilized producer plasmids. Now, of all reads that passed the length quality threshold of ≥1,000 bases, 96.73% (run 1) and 99.92% (run 2) gave a BLASTn-hit with our database. Of the reads not assigned to our database, 17% (101 reads, run 1) and 13% (33 reads, run 2) gave a hit against the NCBI Nucleotide database. We performed

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


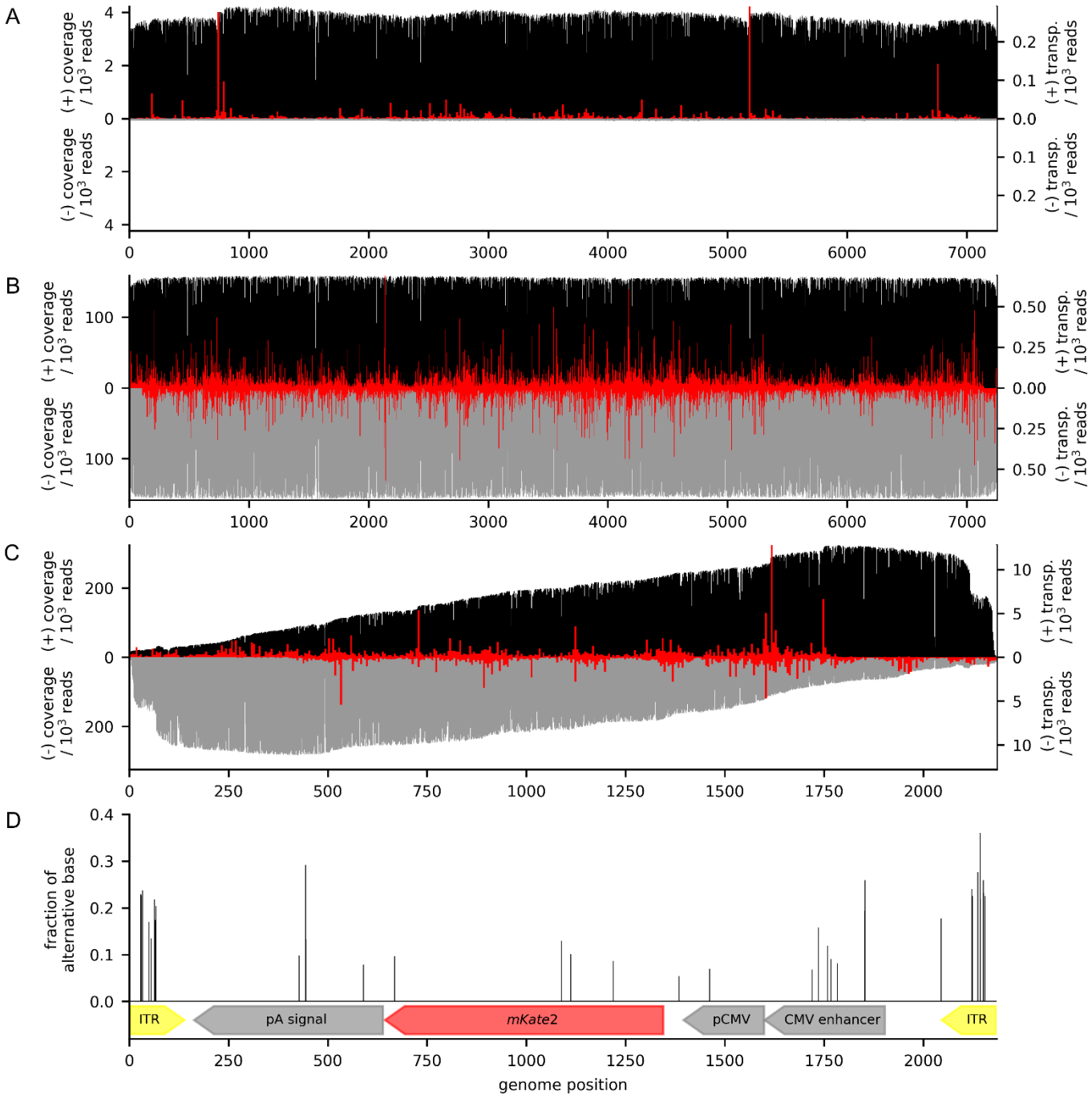


**Figure 1.** Strand-specific sequence coverage of M13mp18 samples and the rAAV genome, as well as its unveiled single-nucleotide variants and respective transposition sites. **A**: Sequence coverage for M13mp18 ssDNA (black and grey for (+) and (-) strand) is near-constant, which is expected for a circular molecule and long reads. Overlaid transposition sites (red, 15 nt bin width) reveal that most reads stem from three distinct starting points. **B**: Coverage for M13mp18 phagemid dsDNA is homogenous throughout the sequence and both strands. Transposition sites (red, 15 nt bin width) had no hot spots, indicating that the mode of action of the transposase is inherently different between ssDNA and dsDNA. **C**: Coverage of the rAAV genome (black and grey for (+) and (-) strand) is constantly increasing towards the 3'-end, as expected for a linear substrate, until it reaches a plateau and suddenly halves within the ITRs. Both strands are covered, as AAV packages both strands during production. Overlaid transposition sites (red, 5 nt bin width) exhibit an even distribution across the genome with few hot spots. The lack of transposition sites towards the ends of the linear sequence is attributable to the applied read length cut-off. Furthermore, a 3 bp discrepancy to the theoretical sequence is uncovered near the 3' ITR. **D**: Fraction of alternative bases at identified sites of single-nucleotide variants (SNVs).

further analysis only with reads assigned to our database. The assignments calculated as percent-contaminants are summarized in Table 1 A. Of the contaminants, *rep-cap* genes showed the highest prevalence, representing 1.0% (run 1) and 0.8% (run 2) of total reads attributable to our database. They

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


were followed by the backbone of the pITR, representing 0.8% and 0.7% of all reads. The proportion of reads that map to the hg38 showed the biggest variance between the two samples with 0.7% and 0.3% each. 0.1% of attributable reads in both runs were assigned to the adenoviral helper genes.

We performed additional qPCR analysis with primer sets that allow the amplification of the rAAV genome, *rep* gene, *E1* gene of the adenoviral helper plasmid as well as the ampicillin resistance gene *bla,* which is present on all three producer plasmids. Results are given as percentage of the combined absolute copy number of all four measurements (Table 1 B). As expected for the contaminants, the *bla* gene was present with the highest proportion of 2.0% ± 0.3% (95 % confidence interval) in sample 1 and 2.9% ± 0.4% in sample 2. This result can be compared to the nanopore reads that map to one of the three *bla* containing producer plasmids, which had a combined share of 1.9% and 1.6% of all referenced reads. The qPCR result for *bla* is thereby slightly higher than expected from the nanopore analysis. On the other hand, the proportion of *rep* genes found by qPCR was three- to five-fold lower than the proportion of pRepCap-derived sequences observed by nanopore sequencing. In contrast, qPCR and nanopore sequencing gave comparable results for the adenoviral *E1* gene. We wondered at this point, whether certain contaminations are present in the capsid as small fragments below 1,000 nt and if our initial length quality threshold of ≥1,000 nt would cause the deviations between qPCR and nanopore sequencing. We therefore re-analyzed our datasets and included all reads above 500 nt and by doing this, the accepted read count for run 2 increased from 291,036 reads to 647,246 reads. The results of this analysis showed that, while the overall share of contaminants within the sample stays roughly the same regardless of the thresholds tested, the share of individual contaminants shifts. We found that the proportion of pHelper-derived contaminants remained constant for both analyses, whereas the proportion of human genomic contaminants doubles for the lower threshold and the proportion of pITR backbone- and pRepCap-derived contaminants decreases accordingly. Clearly, at this point, a more descriptive data evaluation tool was needed to find the source of this disparity.

**Direct sequencing reveals the molecular state of the genome and its contaminants**

As we use a direct sequencing approach, each read represents a single 3'-end ssDNA fragment of a natively packaged nucleic acid, presuming that it was fragmented only once by a transposase. This makes the fragments' GC content a calculatable (from the known sequence) as well as measurable (from sequencing) quantity for a given fragment length, at least for the recombinant AAV genome. Conclusions on the molecular state of the genome and its contaminants can then be drawn from a %GC *versus* read length plot showing reads selected based on the BLAST assignment. In these plots—and assuming no sequence preference of the transposase—reads of originally circular molecules ideally appear as single points, as these have a constant %GC content and the same read length, independent of the cut site (for singularly cut genomes). Accordingly, reads of linear fragments will produce a vertical line (|), if the GC content is constant along the DNA and a slanting line (/ or \), if the GC content increases towards one end. The same fragment will result in an upper-case lambda (Λ) structure when both strands are present, because both directions are sequenced. In such a plot and according to expectations, the M13 reads group around the theoretical GC content and length with conical tailing towards shorter reads. We suspect the latter to arise from double transposase fragmentations and premature sequencing breakoffs (Supplementary Figure S5).

In the AAV sequencing runs on the other hand, reads that gave a BLAST hit with the rAAV genome showed a more complex mirrored lower-case lambda-like pattern. The lambda pattern becomes easier to spot in the large data volume when reads are binned in a 2D histogram, as shown in Figure 2 A for run 2 (refer to Supplementary Figure S6 for run 1). The histogram further shows that most read lengths are below the theoretical genome length, which is 2.2 kb. The shape of the data distribution in the plot is a function of the fragment's nucleic acid composition. We therefore simulated the transposase reaction for the rAAV genome and found that in the plot the measured data are shifted slightly towards lower GC content compared to that of the simulation (Figure 2 A, green line and supplementary information for the simulation script). Looking again at the genome coverage, it appears that roughly half of the reads will miss a large part of the GC-rich containing ITR, which might explain the shift. We performed the simulation again with

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


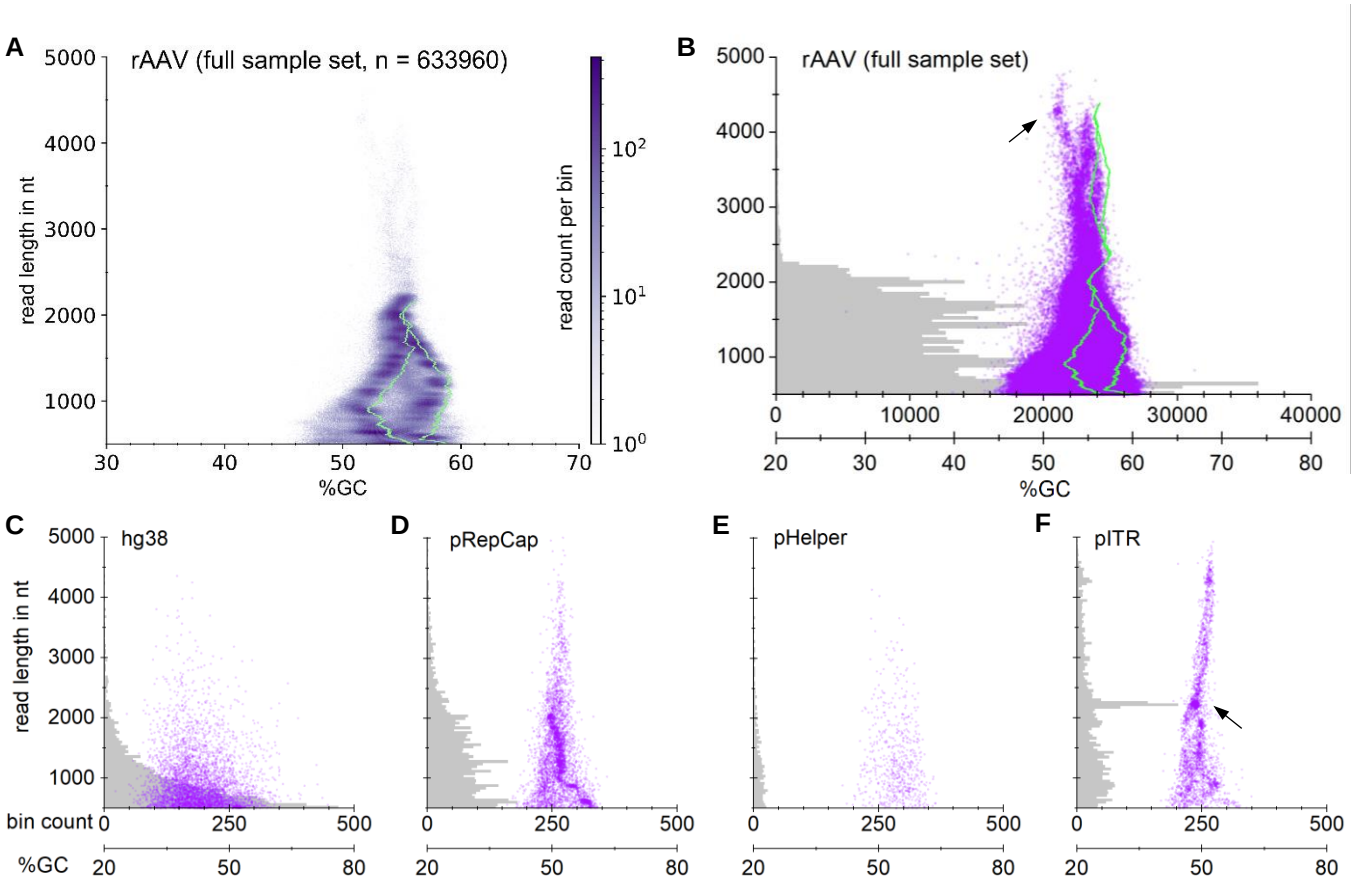


**Figure 2.** Molecular state of the recombinant AAV genome and its nucleic acid contaminants in %GC vs. read length plots. Grouping by BLAST assignments for run 2. **A**: A 2D histogram with hexagonal bins and logarithmic scale reveals a distinct underlying structure. The structure is a function of the genome's nucleic acid sequence and can be predicted by simulating one transposase fragmentation reaction per genome (green line). The data set is shifted towards lower %GC compared to the simulation, because many reads miss part of the 3'-ITR sequence. **B-F**: Each transparent purple dot represents one individual read. A histogram of read length distribution is underlaid in grey (histogram bin size of 40 nt). **B**: Display of the full sample set in the rAAV genome bin. The histogram illustrates that most reads are genome monomers. The dot plot reveals that larger forms of the genome are packaged in the capsid as well. The simulation (green line) and single-read investigation unveil these as covalent genome head-to-tail dimers. The arrow indicates genome-backbone heterodimers. **C**: Reads in the hg38 bin show no pattern in their %GC content and an exponentially decaying size distribution, indicating packaging of fully random fragments in favour of shorter ones. **D**: A lambda-shaped point cloud for reads in the pRepCap bin indicates random packaging of fragmentated plasmid DNA, however an elongated structure hints on preferred packaging of a distinct fragment, which we found to be of plasmid-backbone origin. **E**: Reads in the pHelper bin did not show signs of packaging of a distinct fragment and tended to be of shorter size. **F**: pITR binned reads on the other hand were found with all possible lengths within the AAV packaging limit and appeared to originate from one non-random source. The arrow indicates a hot spot of reads that we were able to assemble into a circular sequence constituted of the plasmid backbone expanding to the ITRs.

depleted ITRs and observed a left shift for the simulation as well (not shown). Hot spots of transposition sites are also apparent in this plot in accordance with Figure 1 C. Notably, a larger proportion of AAV-assigned reads are shorter than 1,000 nt, which hints on double fragmentations. Plotting all 633,960 mapped reads directly reveals additional reads of a distinct distribution, which are longer than the theoretical rAAV genome (Figure 2 B), although of all reads in the rAAV genome bin, only 0.9% are longer than 2,300 nt. We reasoned that these could be genome head-to-tail fusions. Indeed, a simulation of such a fusion overlays well with the oversized reads. Contrary, head-to-head or tail-to-tail fusions (self-complementary genomes) were not found. Single read investigation furthermore revealed that some oversized rAAV genomes are genome-backbone fusions (black arrow in Figure 2 B).

Similar plots for the other BLAST bins reveal differing molecular states of the individual contaminants. Reads assigned to hg38 appear to be of completely random human origin, with an exponentially decaying size distribution (Figure 2 C). Reads assigned to pRepCap

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


appear to be randomly fragmented and packaged as indicated by the overall triangle shape of the point cloud. However, a slanting (\) data group up to 2,000 nt in length indicates preferred nonrandom packaging of one distinct linear fragment (Figure 2 D). We mapped this subgroup to the pRepCap reference and found that it mostly comprises reads of the plasmid bacterial origin of replication (Supplementary Figure S7). Reads assigned to pHelper do not show a sign of preferred packaging of one fragment, although the sample size may be too small for a definite conclusion (Figure 2 E). Finally, reads assigned to the pITR backbone appear to mostly originate from one distinct fragment. This bin also showed the most even size distribution, except for a cumulation around 2.2 – 2.3 kb read length (Figure 2 F). Cumulations in this plot might indicate transposase insertion bias on linear DNA resulting in fragments of similar length, although the prevalence of this 2.2 kb point cloud is more prominent than other hot spots for the rAAV genome. Another explanation might be a circular fragment. Most reads had a similar read start, which indicates transposase bias. The majority of the reads from this hot spot map to the reverse strand of the plasmid backbone, starting at a conserved position in the bacterial ori and ending at the adjacent ITR.

## DISCUSSION

Nucleic acid contaminants in AAV vector stocks for gene therapy are gaining attention alongside the increase of therapeutic doses from $10^{12}$ viral genomes per kg in the first authority approved product Glybera to recently approved $10^{14}$ viral genomes per kg for Zolgensma, both single systemic applications (27, 28). Potentially, even higher doses in multi-administration therapies, like cancer gene therapy, are conceivable. The United States Food and Drug Administration recommends for a vaccine dose that residual cell-substrate DNA should be ≤10 ng and the median DNA size should be of 200 bp or lower (29). Vector manufacturers take extensive precautionary measures to ensure a homogenous product and preempt tighter AAV-specific regulations. These measures include the use of bacterial backbone-depleted circular plasmid-derivates (7, 30), or plasmid insertions of uncritical stuffer DNA beyond the AAVs packaging limit, to avoid packaging of the bacterial backbone and antibiotic-resistance gene (31). Monitoring of contaminants is a routine task in vector manufacturing and new time-saving techniques with reduced hands-on time are appreciated.

We report here the direct transposase-based library generation and nanopore sequencing of AAV ssDNA and ssDNA in general as a convenient and versatile tool for characterization of AAV packaged DNA. We present proof for direct ssDNA sequencing by use of bacteriophage M13mp18 ssDNA as control (Figure 1 A). Use of a transposase for library generation was originally designed for dsDNA tagmentation (NExtera library preparation using Tn*5* transposase) and sequencing on the Illumina sequencing platform (32) and it was adapted for direct dsDNA sequencing on nanopores by Oxford Nanopore Technologies with MuA transposase (25, 33). MuA forms a homotetrameric synaptic complex around paired phage Mu genome ends and then catalyzes strand transfer, leaving behind nicks that act as replication primers in the wild type (34). A transposome consisting of MuA and end substrates (mini-Mu) is sufficient for *in vitro* transposition (35) and it shows slight target DNA bias towards a 5'-CYSRG pentamer (36, 37). Hyperactive MuA variants with also low target bias have been reported (38, 39). However, we were unable to find previous reports of MuA (or other DDE transposase superfamily members) activity on ssDNA targets. Our data confirms the relatively low insertion bias of MuA on dsDNA (Figure 1 B), but not on M13 ssDNA, where hot spots of insertions are seen, and activity is reduced. We observed three especially prominent hot spots of transposition sites for M13 ssDNA.

At first thought, this result could be explained by MuA action on transient hairpins within the ssDNA target. To fit into the MuA target binding pocket, hairpins require a stem of at least 23 – 25 nt (34, 37). These hairpin insertion sites should be easily predicted using software tools but an ssDNA probability score (ss-count) calculated by mfold software (23) does not correlate with the most prominent insertion sites (Supplementary Figure S1). One possible explanation might be that MuA exhibits increased activity on mismatched targets (40) and the extent to which mismatches are tolerated has yet not been investigated. However, when we searched for mismatched hairpins, no correlations to the read start hot spots were apparent. We concluded that, while mismatched hairpins might explain the background of transposition sites we observed for the M13 ssDNA, they are likely not responsible for the most prominent hot spots so that we favor a model of transposase action on true ssDNA. Compared to the M13 ssDNA sample, the AAV samples exhibited an overall relatively even distribution of transposition sites on

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


both strands with a remarkable symmetry. It thereby closely resembles the dsDNA M13 sample. This effect is probably due to hybridization of two ssDNA genomes of AAV to one dsDNA, which we also observed in agarose gel electrophoresis. Overlaid, we further find hot spots of transposition sites on one strand only. This is probably a different effect of transposase action on ssDNA, as seen for the M13 ssDNA. Given that hot spots lay on one strand only, we deduce that the effect is sequence-specific rather than conformation-specific, because of the inherent symmetry of hairpins when both strands are present. Further work will be required to elucidate the MuA transposase action on ssDNA targets.

Albeit the given insertion bias of the transposon on ssDNA, nanopore long reads compensated for this and still enabled full coverage. We achieved a 356,009-fold coverage (run 2), which also enabled investigation of single-nucleotide variants despite the lower base accuracy of nanopore sequencing compared to other NGS methods. The sudden halving of coverage within the ITRs might be explained by difficulties in sequencing, either stemming from difficulties of the helicase with the strong ITR secondary structure, or as a result of back-folding of ITRs after passing through the transmembrane pore. A previous AAV NGS study found SNVs within the rAAV genome, mostly located within one region in the coding sequence and both ITRs (12). Regarding the ITRs, we found as well variants that locate in the ITR B-C hairpins (according to the ITR naming convention). However, these are attributable to the two possible states of ITRs which arise from AAV genome replication: FLIP and FLOP, where FLIP ITRs harbor the inverse complement C'-B' hairpins compared to FLOP ITRs while the rest of the ITR sequence stays the same. Our producer plasmid pITR encodes FLOP ITRs on both sides and FLIP-specific SNVs appeared with a 20% frequency. We note that only those ITR conformation-specific (FLIP) SNVs were called, that lay in the outer arms of the respective hairpins, and we are so far unable to explain this finding. We also observed SNVs within the coding sequence with frequencies up to 30%. Given that variant calling from nanopore data is a relatively recent technique, we suggest re-cloning of AAV DNA and Sanger sequencing of individual clones to confirm these high SNV abundancies. Nonetheless, our finding is overall in agreement with the previous study (12) which found SNVs with an abundance up to 15%. The punctually high SNV abundance raises the question where these variants come from and what their implication for vector quality is. We find it unlikely that these SNVs are already present on producer plasmid level, as this would render cloning in *E. coli* in general impractical to impossible and is also not in accordance with our frequent Sanger sequencing of pITR plasmids after cloning steps. Much rather, a somewhat error-prone AAV genome replication during virus production may be responsible and it would be very interesting to compare different producer cell lines (different mammalian, insect and yeast) and wild-type AAV under this aspect.

Nanopore sequencing also offered us the convenient opportunity to investigate CpG methylations from raw reads. In a previous study, bisulfite PCR sequencing for packaged AAV2 wild type genomes showed little to no methylation with a maximum share of 1.7% methylated CpG dinucleotides, but revealed hypermethylation of integrated genomes (41). We used recently published deep learning tools to investigate CpG methylations from nanopore raw data, but we did not observe significant methylation above an unmethylated reference. The finding supports the previous study, which used AAV wild-type and highlights the similarity of wild-type and recombinant genome replication.

As we present a direct sequencing method, the sample input is higher compared to other NGS methods. We find that extracted DNA from $10^{13}$ DNase I-resistant particles is enough for about five sequencing reactions. We also saw that a critical step in sample preparation is the Benzonase digest of the producer cell lysate. When performed for one hour with the given concentration, no fragments beyond the AAV packaging limit of about 5 kb are seen (Figure 2). Digestion for only 30 minutes on the other hand led to emergence of longer reads in small proportions (Supplementary Figure S6) and since we performed virus precipitation and antibody-based affinity chromatography for sample preparation, we attribute these to overlength fragments protruding the capsid and otherwise capsid-associated DNA. In the future, this incomplete (or omitted) digest could be used as a method to investigate rAAV genome replication and packaging intermediates directly. Also, there does not seem to be a linear correlation between sample DNA input and total read output, and we recommend using samples of $OD_{260,\ 10mm}$ = 0.8 or higher for library preparation. Furthermore, the incubation time of the transposase can be optimized to yield longer fragments. In multiplexing we observed overspill and a lower read count, which may also be attributed to the

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


differences in library preparation for multiplexing. We therefore recommend non-multiplexed sequencing for quality control settings.

We assigned reads to single entries of our reference database by BLASTn bit score and compared the results to our qPCR measurements. The share of contaminants in general was in the same range between the two methods. However, we found that the long reads of the nanopore platform offered a much deeper view into the sample. Long reads enabled us to draw conclusions on the state of packaged DNA, but at the same time complicate comparison of contamination rates to qPCR and Illumina sequencing results. This is due to the unique read assignments which assigns one read to only the most prominent hit in the genome database and cannot assign genome-contaminant fusions to both parental sequence sources, which leads to overrepresentation of the rAAV genome bin compared to the others. This might explain the discrepancy between sequencing and *bla* specific qPCR. We also found preferred packaging of pRepCap backbone sequences and not Rep coding sequences. While both end up in the pRepCap bin, the *rep*-specific qPCR will underestimate contaminations stemming from this plasmid and short Illumina reads might not suffice to distinguish between different backbones, since many features between backbones are shared. After all, more advanced data evaluation for the nanopore reads like *in silico* read fragmentation and subsequent assignment might bring nanopore sequencing, qPCR and Illumina sequencing results closer together. However, this comes at the cost of losing a fascinating layer of information or doubling the computing time for data evaluation and should ultimately be a case by case decision. Concerning the length thresholding of reads for the BLAST analysis, we find that a threshold of >1000 nt represents a good trade-off between depth of analysis, computing time and the emergence of double-fragmented sequences. Furthermore, seeing the likely ssDNA sequence bias of the transposase, the question arises, if direct transposase-based library generation can be a quantitative tool rather than a qualitative vector characterization. As described, most discrepancies between qPCR and direct nanopore sequencing are due to long reads and our unique read assignment. Long reads can on the other hand compensate for ssDNA contaminations that might harbor only a few preferred transposase targets and further work will be needed to qualify this process as a quantitative tool. Since insertion bias can be identified by read start analysis, the highly frequent reads could be compensated for in more advanced analyses algorithms to aid quantitative comparisons between different sequences already from the presented data.

To further characterize AAV packaged DNA, a %GC *vs.* read length plot proved to be a very convenient visualization for our nanopore data, as both parameters are computable quantities for uniquely fragmented sequences. When we plotted individual BLAST bins like this, a diverse picture emerged. Firstly, all bins looked vastly different, showing that the read assignment works as intended. Secondly, read length histogram analyses revealed that more than 99% of the rAAV genomes are of the expected size. Even though each read represents a fragmented genome, we were able to draw this conclusion, because both strands are equally likely packaged and independently sequenced, so that the strand-unspecific coverage in total is roughly constant along the genome length (Figure 1 C). We were also able to simulate these unique fragmentations and thereby confirmed our assumption of unique fragmentations. Further analysis showed that genome head-to-tail fusions are packaged in the capsid to a larger extend. These species might emerge from the AAV rolling circle-like replication but are unexpected, because the postulated mechanism for AAV genome replication suggests resolution of head-to-head and tail-to-tail fusions as replication intermediates (42).

In terms of expected contaminations, human genomic sequences were found to be randomly packaged and of exponentially decaying size distribution, which hints at the involvement of enzymatic digestion. This raises the question, where these fragments originate from. We find it unlikely that the host cell genome is highly fragmented during virus production, however, host cell DNA is treated with Benzonase during virus purification. We wonder whether residual helicase activity of the AAV replicases during Benzonase treatment of producer cell lysate is responsible for randomly packaged host cell DNA. Work towards a specific replicase inhibitor that can be added during virus purification might be a chance to further improve vector quality. We also found a distinct fragment of about 2 kb stemming from the bacterial backbone of the pRepCap helper plasmid (Figure 2 D) and furthermore present evidence for a previously undescribed contamination stemming from the pITR plasmid bacterial backbone sequence that comprises only a single strand. Mechanistically, one of both distinct contaminations fits into the capsid together with one of the rAAV genomes. We suggest that

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


exhausting the AAV packaging limit helps to avoid these distinct contaminations, as well as packaging of genome-genome fusions and genome-backbone fusions.

In conclusion, we present here unprecedentedly deep nanopore sequencing of packaged ssDNA in recombinant AAV with the possibility to expand the application range to other single-stranded viruses or bacteriophages. While the technique dramatically simplifies sample preparation and reduces turnover times compared to other NGS characterization methods, the information content of the results increases. The platform revealed single-nucleotide variants within the coding sequence and allowed insights on the CpG methylation status. The long nanopore reads further gave direct proof that a substantial amount of contaminating bacterial backbone DNA is fused with the rAAV genome and that the other contaminants may also not be of completely random origin. The present study highlights the necessity to further understand the AAV basic biology to gain high-transducing vectors with homogeneous payloads for gene therapy applications. Analytical procedures must keep pace with new diagnostic developments, and we foresee that quantitative PCR will lose its status as the gold standard, as unbiased next-generation sequencing protocols become cheaper and readily available.

### SUPPLEMENTARY DATA

Supplementary Data are available at BioRxiv.

### FUNDING

MR and KM acknowledge funding by the European Regional Development Fund (EFRE) and the State of North-Rhine Wesphalia (Project “ATIVAA – next gene biologics”; grant number: EFRE-0400293). DB is funded by a grant from the European Commission (project Virus-X: Viral Metagenomics for Innovation Value; grant number: 685778).

### CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCES

1. Matsushita T, et al. (1998) Adeno-associated virus vectors can be efficiently produced without helper virus. *Gene Ther* 5(7):938–45.
2. Xiao X, Li J, Samulski RJ (1998) Production of high-titer recombinant adeno-associated virus vectors in the absence of helper adenovirus. *J Virol* 72(3):2224–32.
3. Grimm D, Kern A, Rittner K, Kleinschmidt JA (1998) Novel tools for production and purification of recombinant adenoassociated virus vectors. *Hum Gene Ther* 9(18):2745–60.
4. McCarty DM, Monahan PE, Samulski RJ (2001) Self-complementary recombinant adeno-associated virus (scAAV) vectors promote efficient transduction independently of DNA synthesis. *Gene Ther* 8(16):1248–54.
5. Nony P, Chadeuf G, Tessier J, Moullier P, Salvetti A (2003) 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* 77(1):776–81.
6. Chadeuf G, Ciron C, Moullier P, 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(4):744–53.
7. Schnödt M, et al. (2016) DNA Minicircle Technology Improves Purity of Adeno-associated Viral Vector Preparations. *Mol Ther - Nucleic Acids* 5(8):e355.
8. Bauer S, et al. (2001) Human TLR9 confers responsiveness to bacterial DNA via species-specific CpG motif recognition. *Proc Natl Acad Sci U S A* 98(16):9237–42.
9. Hyde SC, et al. (2008) CpG-free plasmids confer reduced inflammation and sustained pulmonary gene expression. *Nat Biotechnol* 26(5):549–51.
10. Halbert CL, Metzger MJ, Lam S-L, Miller AD (2011) Capsid-expressing DNA in AAV vectors and its elimination by use of an oversize capsid gene for vector production. *Gene Ther* 18(4):411–7.
11. Kapranov P, et al. (2012) Native molecular state of adeno-associated viral vectors revealed by single-molecule sequencing. *Hum Gene Ther* 23(1):46–55.
12. Lecomte E, et al. (2015) Advanced characterization of DNA molecules in rAAV vector preparations by single-stranded virus next-generation sequencing. *Mol Ther - Nucleic Acids* 4(10):e260.
13. Tai PWL, et al. (2018) Adeno-associated Virus Genome Population Sequencing Achieves Full Vector Genome Resolution and Reveals Human-Vector Chimeras. *Mol Ther Methods Clin Dev* 9(June):130–141.
14. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18):3094–3100.
15. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–2.
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–10.
17. Stoiber AM, et al. (2016) De novo Identification of DNA

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.27.885319; this version posted December 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


Modifications Enabled by Genome-Guided Nanopore Signal Processing. doi:10.1101/094672.

18. Ni P, et al. (2019) DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* 35(22):4586–4595.
19. Edge P, Bansal V (2019) Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat Commun* 10(1):4660.
20. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31(13):3406–15.
21. Maynard LH, et al. (2019) Fast-Seq, a simple method for rapid and inexpensive validation of packaged ssAAV genomes in academic settings. *Hum Gene Ther Methods*:hum.2019.110.
22. Ling CC, et al. (2015) Enhanced transgene expression from recombinant single-stranded D-sequence-substituted adeno-associated virus vectors in human cell lines in vitro and in murine hepatocytes in vivo. *J Virol* 89(2):952–61.
23. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31(13):3406–15.
24. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276–277.
25. Marion L, White J (2015) Method for characterising a double stranded nucleic acid using a nano-pore and anchor molecules at both ends of said nucleic acid.
26. Saariaho A-H, Savilahti H (2006) Characteristics of MuA transposase-catalyzed processing of model transposon end DNA hairpin substrates. *Nucleic Acids Res* 34(10):3139–49.
27. European Medicines Agency Committee for Medicinal Products for Human EC (2012) Assessment Report: Glybera. Available at: https://www.ema.europa.eu/en/documents/assessment-report/glybera-epar-public-assessment-report_en.pdf [Accessed November 12, 2019].
28. U.S. Food and Drug Administration (2019) Zolgensma. Available at: https://www.fda.gov/vaccines-blood-biologics/zolgensma [Accessed December 19, 2019].
29. U.S. Food and Drug Administration (2010) *Guidance for Industry Cell Characterization and Qualification of Cell Substrates and Other Biological Materials Used in the Production of Viral Vaccines for Infectious Disease Indications* Available at: https://www.fda.gov/downloads/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/Vaccines/UCM202439.pdf.
30. Karbowniczek K, et al. (2017) Doggybone™ DNA: an advanced platform for AAV production. *Cell Gene Ther Insights* 3(9):731–738.
31. Hauck B, et al. (2009) Undetectable transcription of cap in a clinical AAV vector: implications for preformed capsid in immune responses. *Mol Ther* 17(1):144–52.
32. Adey A, et al. (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* 11(12):R119.
33. White J (2016) WO 2016/059363 Al.
34. Montaño SP, Pigli YZ, Rice PA (2012) The μ transpososome structure sheds light on DDE recombinase evolution. *Nature* 491(7424):413–7.
35. Savilahti H, Rice PA, Mizuuchi K (1995) The phage Mu transpososome core: DNA requirements for assembly and function. *EMBO J* 14(19):4893–4903.
36. Haapa S, Taira S, Heikkinen E, Savilahti H (1999) An efficient and accurate integration of mini-Mu transposons in vitro: A general methodology for functional genetic analysis and molecular biology applications. *Nucleic Acids Res* 27(13):2777–2784.
37. Haapa-Paananen S, Rita H, Savilahti H (2002) DNA transposition of bacteriophage Mu. A quantitative analysis of target site selection in vitro. *J Biol Chem* 277(4):2843–2851.
38. Kim YC, Morrison SL (2009) N-terminal domain-deleted mu transposase exhibits increased transposition activity with low target site preference in modified buffers. *J Mol Microbiol Biotechnol* 17(1):30–40.
39. Rasila TS, et al. (2018) Mu transpososome activity-profiling yields hyperactive MuA variants for highly efficient genetic and genome engineering. *Nucleic Acids Res* 46(9):4649–4661.
40. Fuller JR, Rice PA (2017) Target DNA bending by the Mu transpososome promotes careful transposition and prevents its reversal. *Elife* 6:1–20.
41. Tóth R, et al. (2019) Methylation Status of the Adeno-Associated Virus Type 2 (AAV2). *Viruses* 11(1):38.
42. King JA, Dubielzig R, Grimm D, Kleinschmidt JA (2001) DNA helicase-mediated packaging of adeno-associated virus type 2 genomes into preformed capsids. *EMBO J* 20(12):3282–3291.