Molecular Therapy
Methods & Clinical Development
Original Article

Comparison and cross-validation of long-read and short-read target-enrichment sequencing methods to assess AAV vector integration into host genome

Mark Sheehan,[1] Steven W. Kumpf,[1] Jessie Qian,[1] David M. Rubitski,[1] Elias Oziolor,[2] and Thomas A. Lanz[1]

[1]Global Discovery, Investigative and Translational Sciences, Pfizer Inc, Groton, CT 06340, USA; [2]Computational Safety Sciences, Pfizer Inc, Groton, CT 06340, USA

Evaluation of host integration profiles by adeno-associated virus (AAV) is an important component of de-risking novel AAV gene therapies. Targeted enrichment sequencing (TES) is a cost-effective and comprehensive method for assessing integration. Most published TES datasets have been generated using short-read sequencing, which enables quantitation of integration sites (ISs) and identifies patterns such as hotspots or clonal expansion. Characteristics such as IS length and recombination require longer reads to measure. The present study compared short-read to long-read TES using samples from monkeys treated with AAV and used *in vitro* lentiviral-treated samples, a stable cell line, and an engineered spike-in as controls. Both methods showed stochastic integration by both AAV and lentivirus, with most vector domains identified in ISs. More ISs were identified by short-read TES, as deeper coverage per base was achieved from a single sequencing run. AAV-treated samples showed minimal evidence of clonal expansion, in contrast to *in vitro* treated and stably transduced lentiviral cell line samples. Long-read TES revealed vector rearrangement in 4%–40% of ISs in AAV-treated animals. In summary, both methods yielded similar conclusions about relative numbers of ISs and overall patterns. Long-read TES identified fewer ISs but enabled measurement of IS length and recombination patterns.

INTRODUCTION

Decades of research into gene therapy as a mechanism for delivering a lasting means to overcome monogenic disorders has resulted in successful development of several candidate therapies being tested in patients. Adeno-associated virus (AAV) has become a popular tool for this platform, with over 2,000 gene-therapy trials registered in clinicaltrials.gov and several gene-therapy products approved by the US and/or EU. The popularity of AAV comes in large part due to its genome remaining predominantly episomal in host cells; early gene-therapy trials using integrating retroviral vectors produced several cases of leukemia.[1,2]

While several AAV gene therapies to date have proved safe in humans, a small percentage of the AAV genome (0.1%–0.5%) is known to integrate into the host genome.[3] Integration into the rodent-specific Rian locus has been implicated in AAV-induced hepatocellular carcinoma (HCC).[4,5] However, no cancer-associated integrations have been observed in other species treated with recombinant AAV, including long-term studies in cats, dogs, and non-human primates.[6] In humans, no cases of cancer have been linked to AAV gene-therapy treatment. The potential for such an event has led to US Food and Drug Administration (FDA) guidance including an assessment of vector integration in development of new AAV therapeutics.[7]

Methods for evaluating vector integration involve sequencing host genomic DNA and either bioinformatically or physically enriching for vector DNA sequences and evaluating vector-host junctions. A comparison of two enrichment methods, shearing extension primer tag selection ligation-mediated PCR (S-EPTS/LM-PCR) and targeted enrichment sequencing (TES), to each other and to whole-genome sequencing, has been described previously.[8] While whole-genome sequencing theoretically would be the most comprehensive, it would be cost-prohibitive to sequence to a sufficient depth to provide sensitivity to detect integration sites (ISs) whose frequency is very low. Targeted methods enable deeper sequencing of regions of interest. S-EPTS/LM-PCR focuses on the inverted terminal repeat (ITR) sequence as an anchor, while TES involves hundreds of sequences spanning the whole viral vector as baits for hybridization and subsequent sequencing. TES thus offers a comprehensive solution with relatively inexpensive cost per base.

Short-read TES allows vector-host junctions to be sequenced and evaluated rapidly, enabling high coverage of insertion detection across the genome, requiring less input sample, and using more ubiquitous technology available to most next-generation sequencing (NGS) labs. The read-length limitations of short-read sequencing, however, lead to bioinformatic challenges and technical limitations in the characterization

Received 5 January 2024; accepted 4 October 2024;
https://doi.org/10.1016/j.omtm.2024.101352.

Correspondence: Mark Sheehan, Global Discovery, Investigative and Translational Sciences, Pfizer Inc, Groton, CT 06340, USA.
E-mail: mark.sheehan@pfizer.com

of AAV vector and host-insertion reads, some of which may be addressed with long-read sequencing. A previous study demonstrated frequent rearrangement events in the vector genome,[9] with different characteristics in the context of AAV vector and insertion events. While short-read analysis may be able to capture rearrangement break-end locations in general, long reads can robustly differentiate between vector and host-inserted (chimeric) reads, and fully characterize rearrangements in the inserted sequences. Distinguishing between true chimeric reads and vector-only or host-only reads would reduce the rate of potential false-positive detection.

To compare the benefits and drawbacks of both sequencing modalities in the context of vector integration assessment, the present study investigated vector integration into the host genome from DNA samples derived from AAV-treated cynomolgus macaque liver, as well as lentivirus-treated *in vitro* samples and a stably transduced cell line, using short-read TES and long-read TES. By comparing the insertion detection and event characterization possible with these approaches, the results support the selection of the appropriate modality for future AAV gene-therapy studies, depending on the experimental question.

RESULTS

DNA was isolated from livers of cynomolgus monkeys (*Macaca fascicularis*; cyno) following treatment with a recombinant AAV expressing a Frataxin transgene, as well as HEK293 cells stably transfected with GFP, and HepG2 cells transduced with FusionRed containing lentivirus. Full sample descriptions are provided in Table S1. Long-read and short-read TES analyses were performed to evaluate vector integration patterns, identify possible clonal expansion, and assess functional consequences indicated by each of these modalities.

Quantification of ISs

Following long-read TES, PacBio sequences were filtered for PCR duplicates using pbmarkdup, then aligned to hybrid reference genomes generated by concatenating the relevant host and vector genomes for the given sample. After filtering alignments for host-aligned, vector-aligned, and chimeric reads with adjacent sequences aligning to both host and vector, over 1e5 aligned reads per sample were obtained. In AAV-treated cyno liver samples, the relative number of reads coming from vector-only and chimeric reads increased with higher doses of AAV, and, at the high dose, >100-fold more reads came from vector-only compared to chimeric reads (Figure 1A). Short-read data show higher counts of vector-only reads across samples, including the positive control insertion DNA spike-in where all vector sequence present in the sample is in the context of insertion (Figure 1B), supporting the notion that a subset of DNA fragments containing insertions are missed when limited to shorter read lengths. Lentivirus-treated stably integrated HEK cells showed similar vector-only and chimeric reads across samples; chimeric read counts were greater than the highest-dose AAV, and on average only 1.25-fold vector-only reads were observed compared to chimeric in the lentivirus samples. As the majority of AAV vector content is expected to be in the episome, most of the vector-only reads are likely from this source of DNA.

Read lengths for long-read data were obtained and evaluated for chimeric reads (presumably integrated vector) versus vector-only reads. Chimeric reads on average were longer than those of vector-only reads in both the highest dose AAV (27.4%) as well as lentivirus-treated samples (14.0% and 43.2% for GFP and RFP, respectively) (Figure 1C). The total number of vector-only reads was substantially lower for lentivirus compared with AAV, as is expected for a canonically integrating vs. non-integrating vector.

Short-read TES data from the same DNA were analyzed to identify vector insertion sites in the 150-bp paired-end reads. This analysis was performed using an updated version of a previously published pipeline,[8] with a notable change being that sequences were mapped first to vector, and aligned portions of reads were masked prior to alignment to host. In the current dataset, the updated pipeline reduced high-read-count integration host loci hotspots that were observed across multiple AAV-treated samples (Table S2). These hotspot counts were dose responsive (Figure S1A), consistent with both vector-only and insertion reads from long-read data. In the updated pipeline, the majority of reads from what had been identified in the initial pipeline as the top five hotspots were completely masked after vector alignment (Figure S1B), meaning they are fully explained by the vector sequence and cannot be differentiated from vector-only reads. Background integration (e.g., IS detected in animals that did not receive vector), while consistently low across pipelines here, was similarly reduced in the updated pipeline, as shown with other vector and host genomes (M.S., S.K., J.Q., and T.L., unpublished data). Thus, the updated pipeline has resulted in a lower background and reduced rate of false positives.

Compared to those observed in the long-read data (Figure 1D), the absolute numbers of identified insertion sites in short-read data (Figure 1E), with the dose-responsive relationship to insertions well maintained: vector copy number correlated with IS count (Figure S2). By each modality, counts in the untreated cyno liver sample showed limited background, while the positive control insertion DNA spike-in sample showed counts comparable to the cyno samples with lower doses of AAV.

Location of insertions, host and vector

Using the host-aligned sequences from chimeric reads, we can identify the location of captured insertion sites in the host genome as well as the location within the vector of the inserted sequence. The insertion counts per chromosome show lack of preference for specific chromosomal locations between long-read (Figure 2A) and short-read (Figure 2B) modalities, with insertions distributed across the host genome, and no consistent enrichment for individual chromosomes observed between samples or modalities in AAV-treated samples. The ratio of observed insertions to expected insertions adjusted for chromosome length did not show significant correlation between long- and short-read data for the majority of samples (Figure S3; Table S3). The control spike-in sample is one exception, with reads coming only from chromosome 1, as the sequence for this standard was engineered based on the FMO1 sequence on chromosome 1.

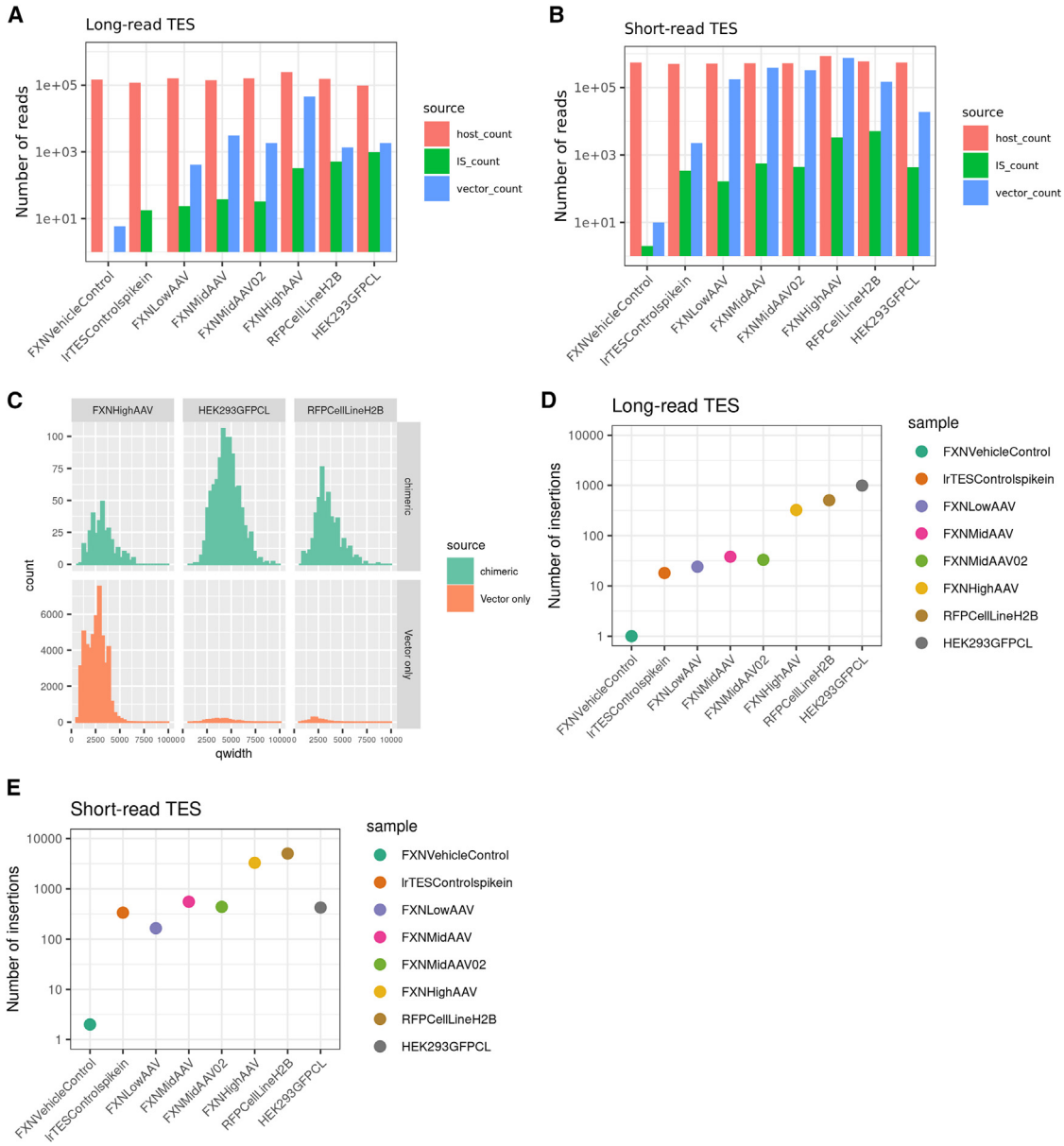

Figure 1. Characterization of reads from short- and long-read TES

(A) Read count from each source per sample from long-read data. (B) Read count unambiguously from each source per sample from short-read data. (C) Histogram of read sequence lengths from chimeric and vector-only reads from long-read data. (D) Total insertion read counts from long-read data. (E) Total insertion read counts from short-read datahost_count, reads aligning entirely to host genome; IS_count, reads showing vector insertion in host; vector_count, reads aligning entirely to vector genome; qwidth, query read sequence width.

The other exception is the lentivirus-treated stably integrated HEK cells, where clonality (more than one cell with the exact same integration event) is expected. Further, the AAV samples do not show significant correlation in chromosomal enrichment between samples, within either modality individually (Table S4).

In both the long- and short-read datasets, we can evaluate the breakend (host-vector junction) position information across the vector genome (Figures 2C–2F). In the breakend data, similar patterns of inserted vector sequences are observed between the two modalities (Figure S4), with decreased capture between ∼1,000 and 1,900 bp in the AAV sequence (containing CAG promoter and intron sequences), and additional decrease after ∼3,250 bp (within the 3′ ITR). Additionally, by both modalities, the vector genome segments observed in ISs appear to be stochastic in the GFP cell line.

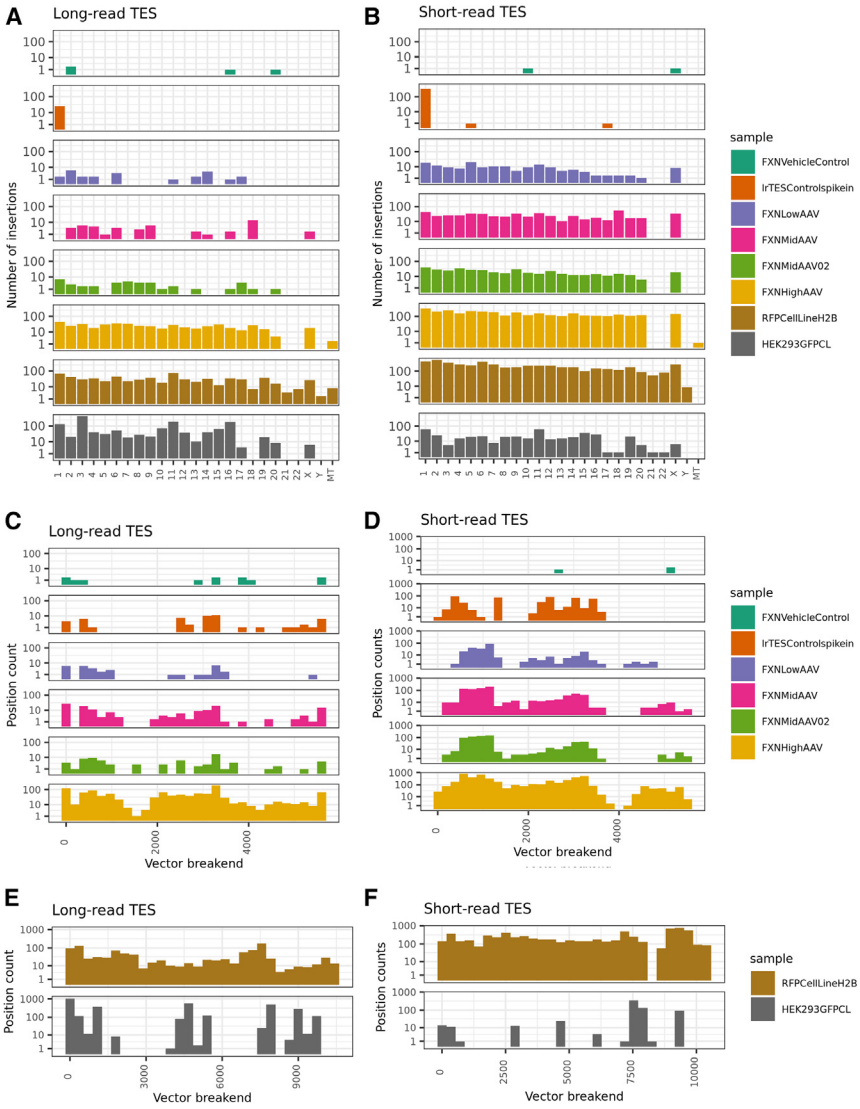

Figure 2. Host and vector breakend location data from identified insertions in short- and long-read TES data

(A) Insertion counts per chromosome in long-read data. (B) Insertion counts per chromosome in short-read data. (C) Counts of vector breakend locations in the AAV vector sequence from chimeric reads in long-read data. (D) Counts of vector breakend locations in the AAV vector sequence from chimeric reads in short-read data. (E) Counts of vector breakend locations in the lentiviral vector sequence from chimeric reads in long-read data. (F) Counts of vector breakend locations in the lentiviral vector sequence from chimeric reads in short-read data.

While the read-length limitations of short-read data yield primarily this breakend position information, the more complete sequence information across the insertion available with long-read data allows further characterization of vector sequence representation in insertions (Figures 3A and 3C). Here, there remains decreased representation from the CAG promoter and intron sequences in AAV-treated samples, suggesting reduced insertion of these sequences in general rather than simply reduced probability of insertion starting from these sequences. However, we observe a more uniform distribution across the vector backbone beyond ~3,500 bp, where there is reduced breakend representation by both modalities.

Long-read data additionally yield information on positions represented in the vector-only sequences captured (Figures 3B and 3D). Here, we observe a pattern in the AAV sequences captured similar to that seen in the chimeric reads, with decreased representation of sequences between ~1,000 and 1,900 bp.

Clonal expansion evaluation and nearby genes

To further understand the nature of integration of AAV into the host genome, we evaluated integration patterns to look for genomic regions over-represented with ISs.

The breakpoint junctions between vector and host genome observed in chimeric reads within the long-read data show an apparently random distribution of insertions across the host genome, while the positive control insertion DNA spike-in sample shows the expected engineered breakpoint between the vector genome and the FMO1 locus on chromosome 1 (Figure 4A). While the overall captured insertions are higher in the short-read data, the distributions appear similarly random (Figure 4B). Two positions in the vector genome, corresponding to the 5′ and 3′ ITR, show higher frequency breakends across samples in long- and short-read data, and these are associated with insertions observed across the host genome. While at lower frequencies, both methods also highlighted breakends falling outside the ITR-spanning region, indicating that unintended vector backbone packaged into the AAV can be found in IS.

The overlap in nearby genes within 50 kb of insertion sites by long- vs. short-read analysis is limited, with the majority of genes limited to one or the other modality, consistent with the observed random distribution of insertions throughout the genome. The overall number of potentially impacted genes (Figure 5A) increases with increased AAV dose (and correspondingly increased insertions), and the proportion of overlapping genes (Figure 5B) between modalities correspondingly increases but remains quite low even at the high dose. The proportions of overlapping genes were consistent with the observed

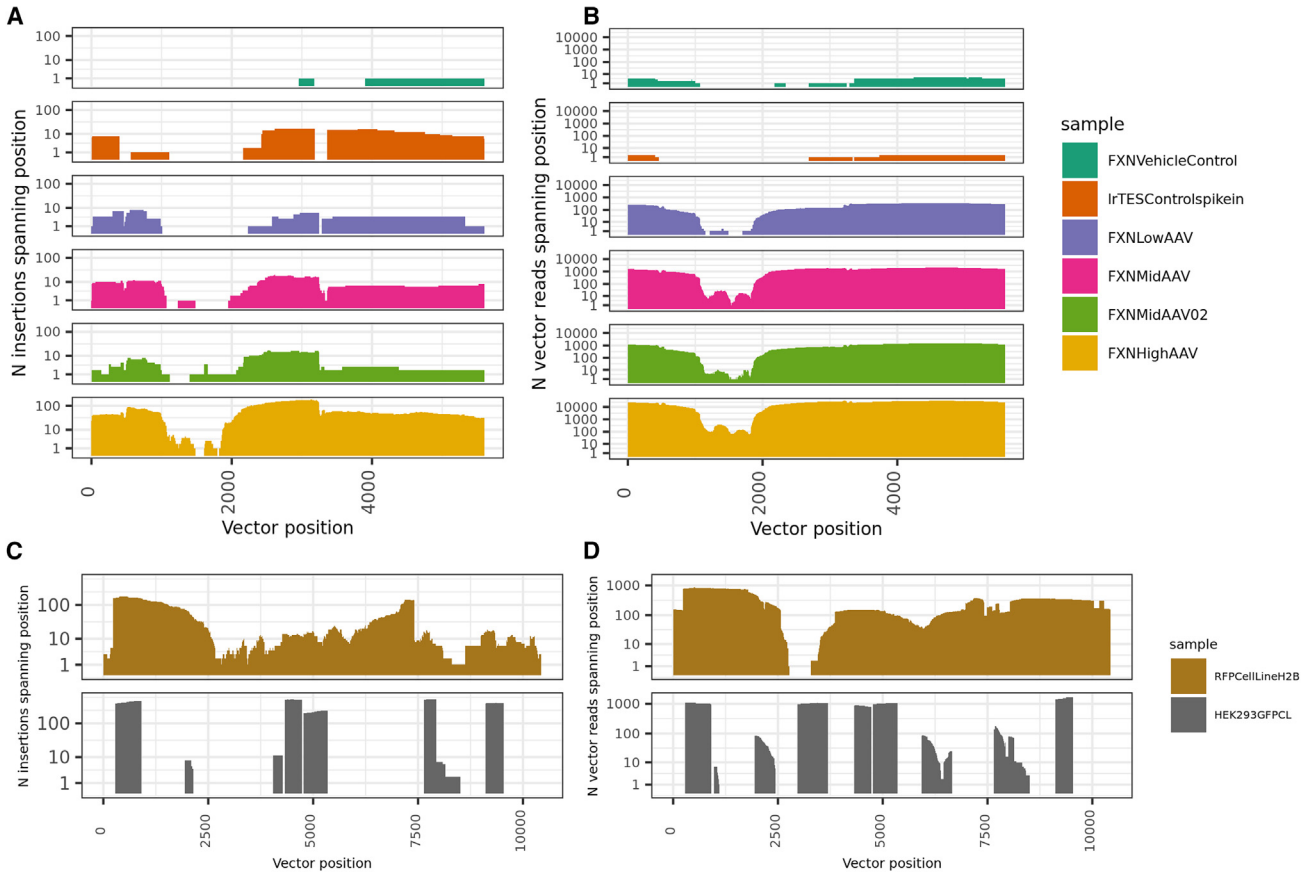

Figure 3. Vector positional data from identified insertions and vector-only reads in long-read TES data

(A) Total counts of AAV vector positions captured in chimeric reads from long-read data. The ITR-spanning sequence ranges from 416 to 3,328. (B) Total counts of AAV vector positions captured in vector-only reads from long-read data. The ITR-spanning sequence ranges from 416 to 3,328. (C) Total counts of lentiviral vector positions captured in chimeric reads from long-read data. (D) Total counts of lentiviral vector positions captured in vector-only reads from long-read data.

distribution of proportions when simulating insertions randomly across the genome (Figure S5; Table S5). The sole exception was the control spike-in sample, as the sequence for this standard was engineered based on the FMO1 sequence, with this gene detected by both modalities.

While the AAV-treated samples did not appear to show any strong insertion patterns at the whole-genome scale, we further investigated potential clonal expansion or integration hotspots using loci showing multiple independent fragments. Positive control insertion DNA spike-in and lentivirus-treated *in vitro* samples showed a subset of loci with high numbers of distinct insertion site reads; in AAV-treated samples, the maximum distinct fragment numbers were lower than observed in these controls, with some apparent dose-related increase (Figure 6).

Visualizing loci with multiple independent fragments across the host genome by each modality (Figure 7), the positive control insertion DNA spike-in sample shows the same engineered locus in short- and long-read data. The lentivirus stably integrated HEK cells show 10 multi-fragment loci shared between short- and long-read data. The high-dose AAV sample has lower peak counts of distinct fragments at any individual locus and does not have any such loci overlapping between modalities. Thus, any potential clonal expansion or hotspot identified in the AAV samples was likely a false positive. Only the positive control DNA spike-in sample and HEK cell line show any overlap between short- and long-read data, and these overlaps are statistically significant by Fisher exact test (Table S6).

Vector characterization

The additional read length afforded by PacBio sequencing relative to our short-read sequencing allows us to further characterize the vector sequences, both in vector-only reads as well as in the context of insertion events (Figure 8A). A large proportion of vector-only sequences detected in each of the AAV-treated samples show evidence of rearrangements, with only 11%–15% of reads showing no rearrangement, and more than half of reads showing 2+ rearrangements. In chimeric reads, the majority of insertion events showed no rearrangement,

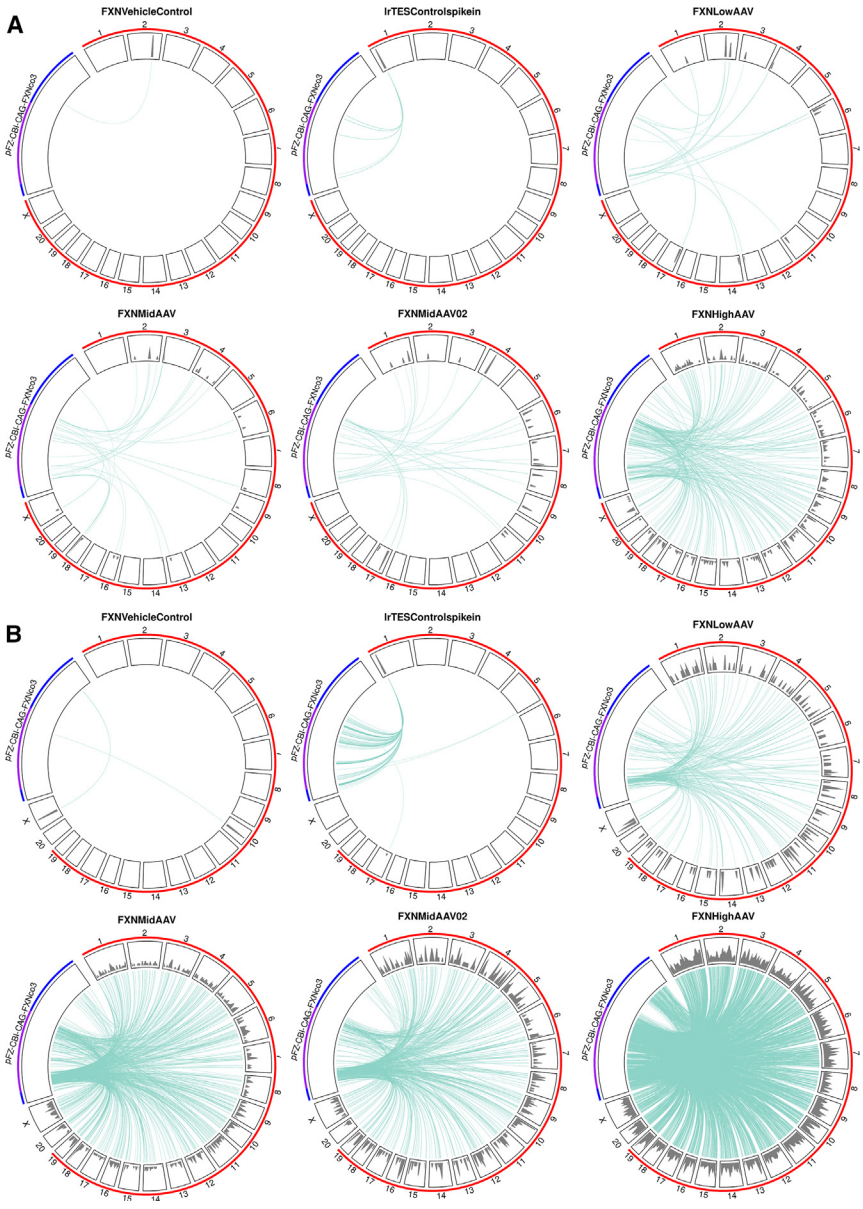

Figure 4. Vector source locus and host location of insertion sites in long- and short-read data

(A) Long-read circos plots of vector and host genomes, with arcs linking observed host-vector breakpoints. (B) Short-read circos plots of vector and host genomes, with arcs linking observed host-vector breakpoints. The purple highlighted portion of the vector sequence indicates the ITR-spanning region, while blue signifies the rest of the plasmid backbone.

although 4%–40% of reads had at least one rearrangement. This may be in part due to reduced readthrough across rearrangements in the context of chimeric reads where a portion of the read is taken up by host sequence, but truncation of vector-only reads to randomly sampled observed lengths of the vector components from chimeric reads only partially reduced this discrepancy in observed rearrangements between vector-only and chimeric reads (Figure S6).

To get additional information on the vector domains involved in rearrangement events, we counted the appearance of each domain across the vector sequence in the captured vector-vector breakends (Figure 8B). In both chimeric and vector-only reads showing rearrangements, we observed 5′ and 3′ ITR involvement in 51%–72% of chimeric reads and 59%–77% of vector-only reads. Analysis of the initial AAV dosing material showed a similar pattern of recombination to what was observed in the chimeric reads (Figure S7), suggesting that much of the recombination observed in ISs occurred during vector production.

DISCUSSION

The present study compared target-enrichment sequencing using long-read and short-read modalities in cyno liver samples treated with an AAV vector and *in vitro* samples treated with lentiviral vectors. Each modality yielded similar conclusions in terms of increasing AAV insertion events with increasing dose and vector copy number and of overall similar patterns of integration breakpoint locations from the host and vector genomes.

A greater number of total individual insertion events were identified with the greater sequencing depth afforded by short-read sequencing. However, the AAV dose-responsiveness in cyno liver insertion numbers was consistent between sequencing modalities, meaning that each modality was suitable to capture relative insertion frequencies between samples. The lower cost per base associated with short-read sequencing, and correspondingly greater depth, may make this modality more suitable for situations when a comprehensive set of insertion sites is required, while either modality may be suitable for capturing differences in insertion frequency between samples.

The distribution of observed host breakends from insertions is similarly randomly distributed across the host genome with AAV treatment by each modality. Particularly with higher doses, insertions were observed with no clear pattern between or within chromosomes. Restricting to loci with multiple distinct insertion events in a given sample by either modality, the high-dose AAV sample does not show consistent high-insertion loci between modalities, indicative of either an insertion hotspot in the genome or clonal expansion of

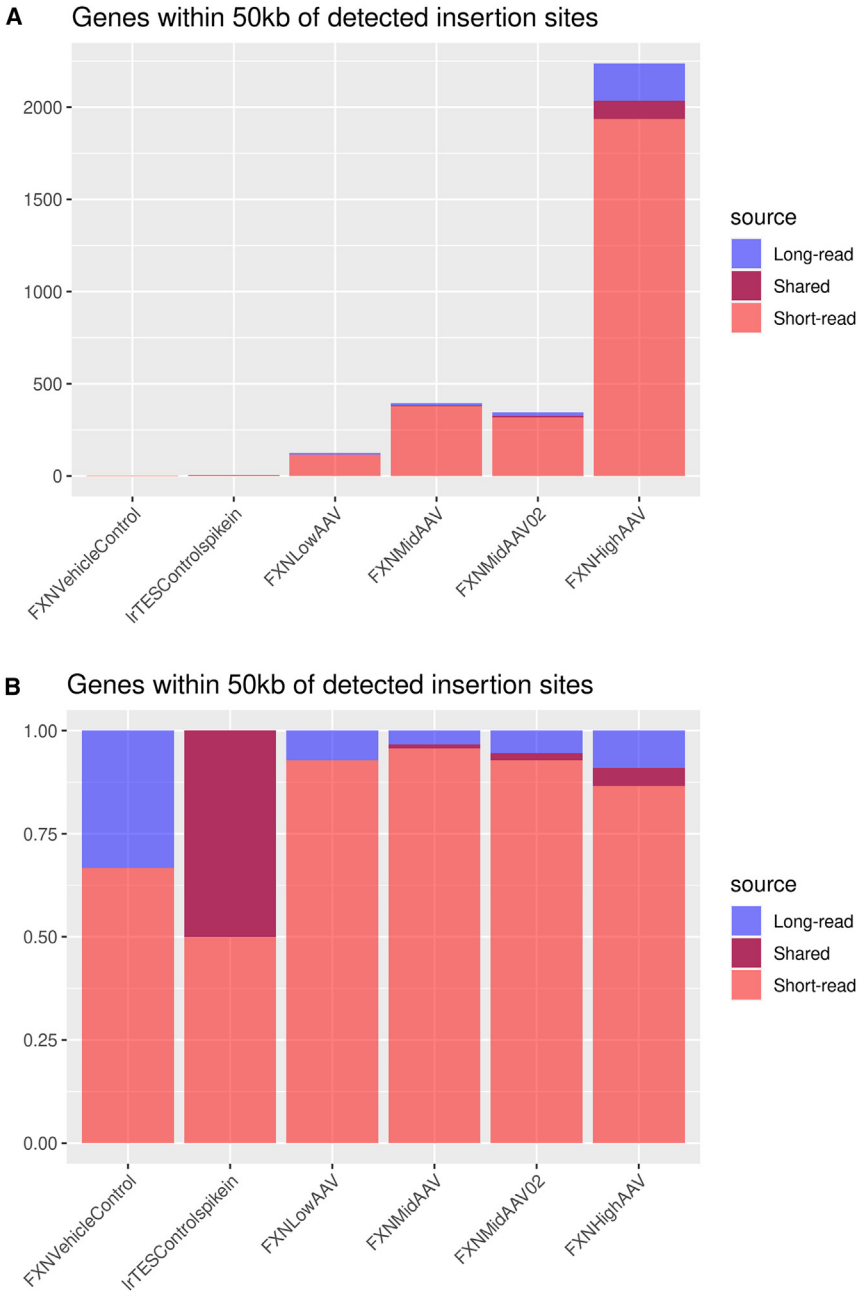

Figure 5. Genes nearby insertion sites in long- and short-read data

(A) Counts of genes within 50 kb of observed insertions in short-read data (red), long-read data (blue), or both modalities (maroon). (B) Within-sample proportion of genes within 50 kb of observed insertions in short-read data, long-read data, or both modalities.

cell(s) harboring integration. Conversely, the positive control insertion DNA spike-in sample shows a robust shared hotspot in the designed location on chromosome 1, suggesting that both modalities are capable of capturing such hotspots/clonal expansions where present. Similarly, many of the loci showing multiple distinct insertions in the GFP lentivirus-treated cell line, which were of lower overall frequency than in AAV- or RFP-lentivirus-treated samples, are conserved between modalities. One caveat to the estimates of clonal abundance is the lack of clear limit of detection versus a positive control. While the use of lentiviral samples and an engineered spike-in in the present experiment provide evidence that the clonal abundance algorithm is working as designed, a tumor sample with true clonal expansion would be the ideal positive control. A similar clonal abundance analysis has previously been shown to identify clonal expansion in dogs in the absence of tumors.[10] As clonal expansion associated with a tumor might be expected to exceed the level of naturally occurring expansion, these data suggest that the algorithm does have sufficient sensitivity to detect functionally relevant clonal expansion.

The distributions of insertion breakends across the vector genomes were generally similar between modalities here, albeit with better capture in lower-frequency locations with the greater depth from short-read sequencing. The higher frequency of ITR involvement in the breakends is evident in even low-dose AAV-treated cyno liver by long and short read, while the profiles in the high-dose sample show consistent locations of peaks and troughs in vector breakend involvement. Thus, most of the ISs are derived from the ITR-spanning sequence, and no specific vector component was driving integration. The long-read data, sequencing through the insertion event, gave greater resolution into the full vector sequences present in the host genome following insertion. In the AAV-treated samples, the resulting profiles remained largely consistent with what was observed in the breakend data. These additional data give greater confidence that the reduced representation of CAG promoter and intron sequences from the short-read results is not simply a function of decreased frequency of involvement of these elements in the breakpoint of the insertion.

In addition to better measurement of vector sequence representation across insertions, the long-read data enable additional information about vector rearrangements, in the context of both inserted and vector DNA. High rates of rearrangement were observed in the AAV vector in both instances, consistent with a published long-read study of AAV integration[9] and prior publications using other methods of measuring integration.[11–13] Evaluation of the AAV

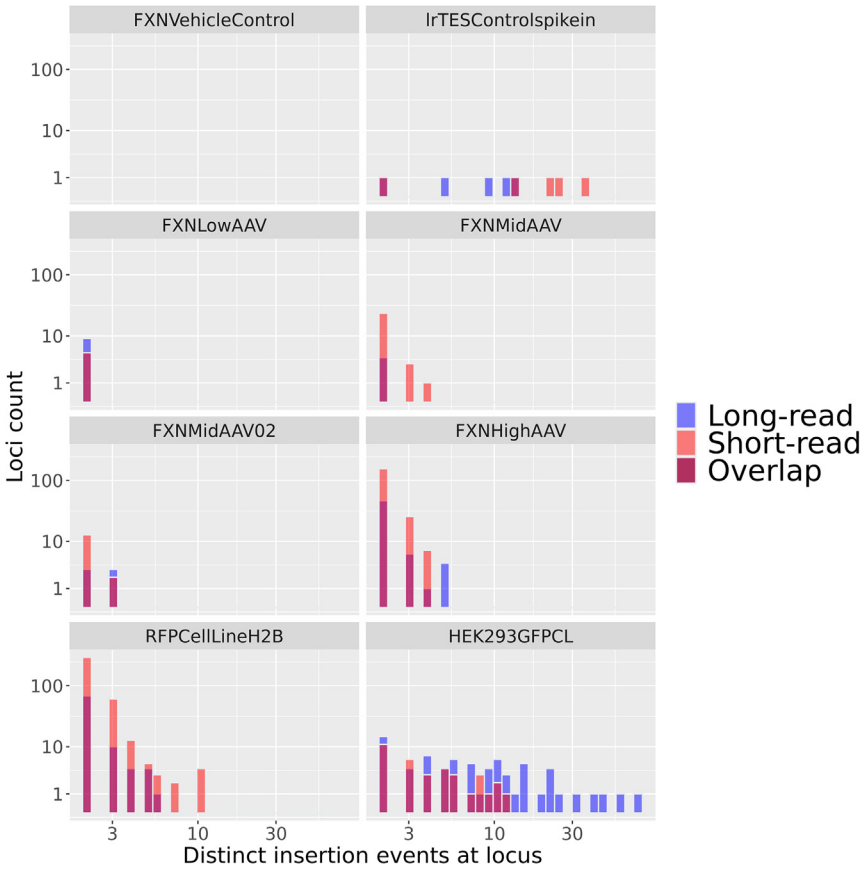

Figure 6. Interrogating possible clonal expansion by distinct fragment counts at single IS loci in short- and long-read data

Counts of loci showing a given number of distinct insertion events in long-read (blue) and short-read (red) data (maroon indicating overlapping distributions).

dosing material revealed that the majority of rearrangements originated during vector production. Furthermore, contaminants such as vector backbone could be identified in ISs; thus, the IS profile is reflective of all elements of AAV packaging. The functional significance of vector rearrangement is unclear, as, even 10 years following transduction in dogs, transgene expression was preserved despite a high level of rearrangements detected. The frequencies of vector sequences showing increasing rearrangement events in the present study were distinct between those coming from chimeric vs. vector-only reads, with lower rearrangement frequency in chimeric reads and a mode at two rearrangements in vector-only reads. This general pattern was consistent across the dose range. In the observed rearranged vector sequences, there is an over-representation of ITR sequences at the rearrangement breakends, both in chimeric and in vector-only reads, again consistent with previous observations.[9] Validation of common rearrangements by another technique such as droplet digital PCR (ddPCR) in future studies would both verify the sequence and enable more accurate quantification of the proportion of total vector containing specific rearrangements. While the short-read data, with pipeline modifications, may be able to capture vector domain involvement in rearrangement breakends, it will likely miss the apparent differences in the profiles between integrated and episomal DNA.

Using an updated version of the previously published short-read insertion identification pipeline,[8] as well as long-read data, the present study confirms previous findings that AAV vector DNA integrates quasi-stochastically into the host genome, and no specific component of the AAV vector is required for integration. While updates to the computational analysis pipeline reduced the rate of false positives, putative ISs were still detected in vehicle-treated animals, suggesting that false positives may still be detected. The true false-positive rate is difficult to quantify, but inclusion of appropriate negative controls helps to understand the baseline IS rate detected by any particular method. Understanding the false-positive/negative rate would be a useful area of future investigation. Simulations or advanced data modeling may help to define these parameters and assess the performance of the IS and rearrangement analyses. The conclusions derived from long- and short-read TES data here overlap in several key areas, including relative insertion frequency between doses, general patterns of host integration loci and evidence of clonality, and identification of multiple vector elements in insertion breakends. The greater depth that is more feasible with short-read sequencing results in a more comprehensive profile of insertions present in the sample, including host loci with multiple independent insertions that occur at lower frequencies. On the other hand, the long-read sequencing gives greater information about the vector sequences present in the sample, both from inserted and episomal DNA, where internal rearrangement events appear to be common. The appropriate modality for studying vector integration in the context of AAV gene-therapy safety will ultimately depend on the experimental question.

Generating and interpreting host genome integration data from recombinant vectors will continue to be an important consideration in the development of novel gene-therapy vectors. Integration data in humans have been published in absence of tumors,[12] and, in two cases in which a patient developed a tumor, AAV integration was evaluated, but insertional mutagenesis was not determined to be the initiating event.[14,15] A recent publication showed a similar stochastic AAV integration profile between non-human primate and humans,[16] increasing confidence that pre-clinical evaluation of recombinant AAV can provide translational value to understanding the expected integration

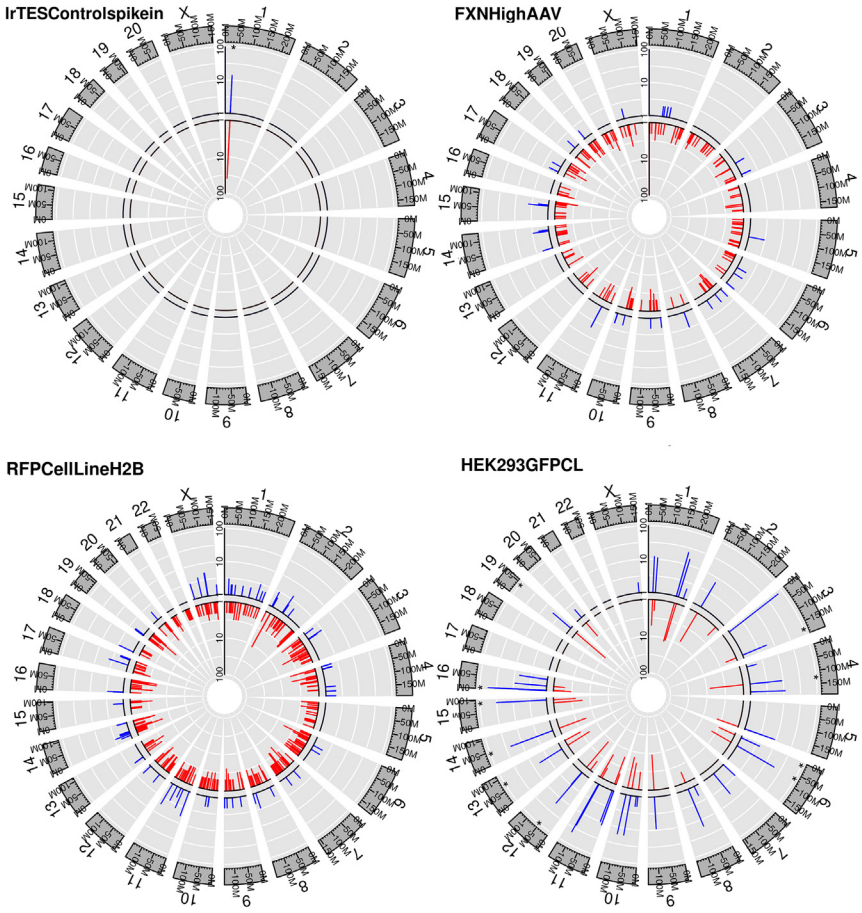

Figure 7. Comparing loci with multiple distinct fragments between long- and short-read data

Distinct insertion event counts and host genome positions in long-read (blue, outer) and short-read (red, inner) data. Bar position indicates host genomic position of the breakend, and height corresponds to the number of distinct insertion events at the locus, with counts in log-scale and radial gridlines at half-log intervals. Asterisks indicate identical loci in long and short read.

profile in humans. Continued monitoring and publication of patient integration data, along with pre-clinical integration assessments, will be critical in fully understanding the relative risk of insertional mutagenesis for gene therapies. The present manuscript and associated data and code present refined analytical pipelines for TES analysis along with methods for long-read TES for cases where a more complete understanding of vector integration structure is warranted.

MATERIALS AND METHODS

Sample preparation

Positive control samples for integration using lentivirus were prepared in HepG2 cells. Cloning of CD822A-1_CAG-H2B-FusionRed_PGK-Puro_WPRE was performed by GenScript Biotech. GenScript synthesized the nucleic acid sequence containing the AAV 5′ ITR before the CAG promoter driving expression of human histone H2B tagged with FusionRed followed by the PGK promoter in front of puromycin-N-acetyltransferase (PuroR) with the sequence for the woodchuck hepatitis virus posttranscriptional regulatory element (WPRE) before the AAV 3′ ITR sequence. This sequence was placed in a lentivirus vector from System Biosciences (SBI, catalog # CD822A-1) while removing approximately 3.7 kb of sequence between the cPPT and 3′ LTR. Lentivirus was generated with SBI pPACKH1 (catalog # LV500A-1) packaging plasmids using Gibco LV-MAX Production System (catalog # A35684) following the protocol for a 125-mL shaker flask. Fifty hours post transfection, medium containing virus was collected and filtered through a 0.45-μm membrane and virus was concentrated using PEG-it Virus Precipitation Solution (SBI, catalog # LV810A-1). After precipitation, virus was re-suspended in PBS, aliquoted, and stored at −80°C. A titer was determined using Takara Lenit-X GoStix Plus (catalog # 631280). HepG2 cells were transduced at three different MOIs. Six-well plates were seeded with 5×10^5 cells per well in Gibco DMEM (catalog # 11995-040), 10% FBS (catalog # 16140-071), and penicillin/streptomycin (catalog # 15070-063). Twenty-four hours after plating, cells were treated with lentivirus in media containing 5 μg/mL polybrene (Millipore catalog # TR-1003) at 30, 10, and 3 MOI. The IncuCyte SX5 was used to image red fluorescent nuclei to verify a successful transduction. Seventy-two hours post transduction, medium was removed and cells washed with PBS and then collected in Gibco Cell Dissociation Buffer (catalog # 13151-014).

DNA was isolated from cells using DNeasy Blood and Tissue kits following the manufacturers protocol for purification of total DNA from animal blood or cells using a spin column. Homogenates were incubated overnight with proteinase K and applied to QIAshredder columns (Qiagen). Downstream DNA isolation was performed using the Qiagen DNeasy blood and tissue protocol on the Qiacubes according to manufacturer's instructions.

The long-read spike-in control was constructed as an artificial gene-block designed by alternating 1 kb of Cyno FMO1 genomic DNA and FXN AAV DNA sequence into a custom gene synthesis order from GenScript, as presented in Figure S8. The synthesized DNA fragment was delivered double stranded and then spiked into a cyno vehicle gDNA background control sample at a target concentration of 1.80e7 copies per ug of cyno genomic DNA.

In addition to the *in vitro* control samples, DNA from AAV-treated male cynomolgus monkey liver samples was used from a study described previously.[8] All procedures performed on animals were

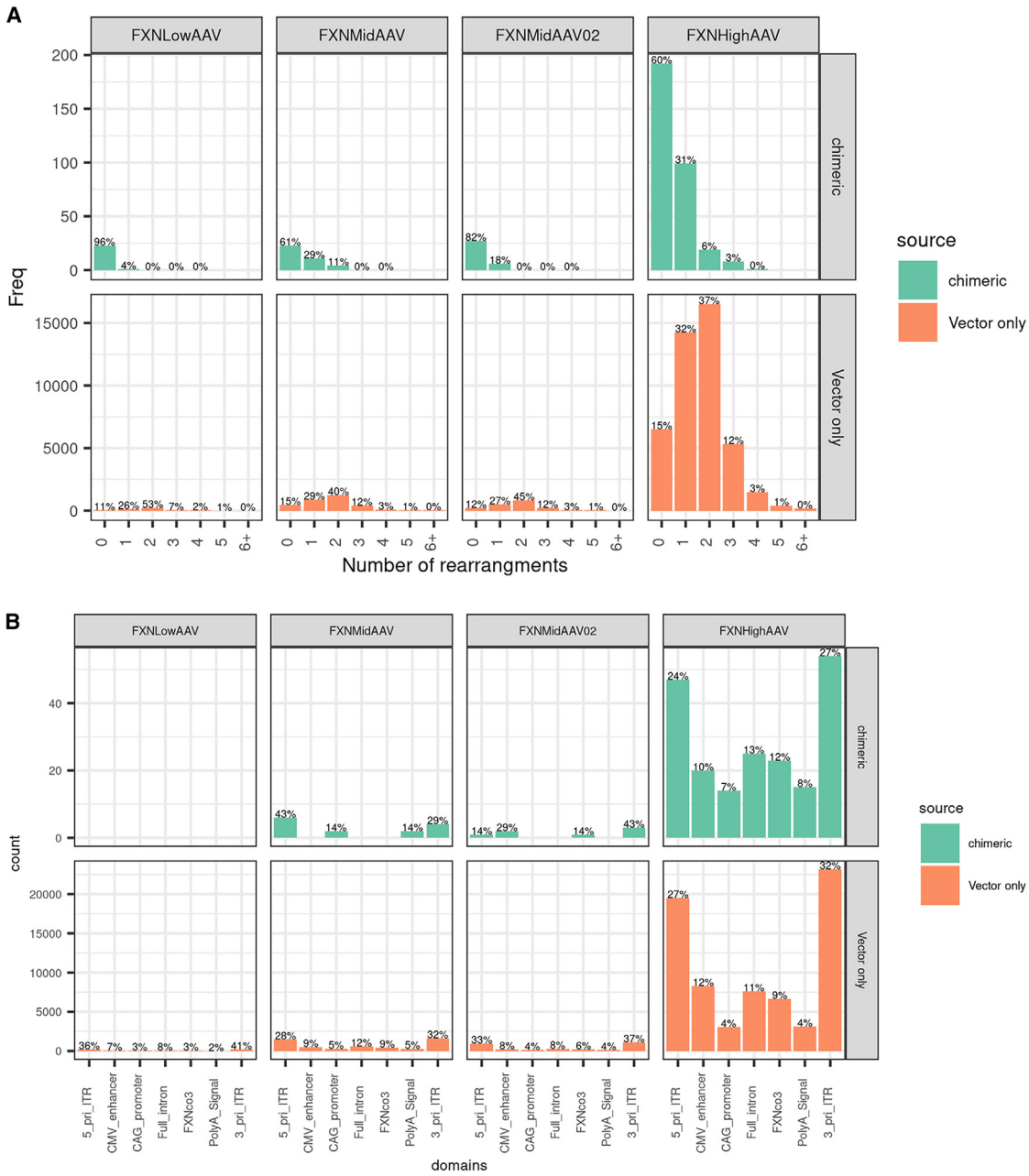

Figure 8. Vector rearrangements observed in long-read TES data

(A) Frequencies of vector rearrangement counts in chimeric (green) and vector-only (orange) reads. (B) Representation of vector domains in vector-vector breakpoints in chimeric and vector-only reads showing one or more rearrangements.

in accordance with regulations and established guidelines and were reviewed and approved by Pfizer's Institutional Animal Care and Use Committee.

Short-read TES

The Twist Bio custom capture panel was a standard 1× design covering all target sequences designed by Twist Bioscience (South San Francisco, CA). Probes were designed for the capture of DNA sequences from multiple packaging plasmids and AAV vector sequence, as presented in Figure S9. A panel of 473,120-mer oligonucleotide probes was used, for a total probe library design size of 69,741 bp, and a total of 100 ng of gDNA per sample was loaded into the Illumina DNA Prep with Enrichment Kit (Illumina, San Diego, CA). Targeted enrichment libraries were then generated according

to manufacturers' instructions. Enrichment bead-linked transposomes were used to tagment DNA in a process that fragments and tags the DNA with adapter sequences followed by post tagmentation cleanup, tagmented DNA amplification, and a double-sided bead purification to purify the amplified libraries. The quality of the libraries was determined by Agilent Tapestation 4200 using a D1000 ScreenTape (Agilent) and quantities were determined on the Qubit 4 Fluorometer (Thermo Fisher Scientific) using the Qubit dsDNA BR Assay kit. Samples were then equimolar pooled by mass (based upon their Qubit readings), with 5 μg of human Cot DNA, which has been previously shown to reduce non-specific binding during hybridization for TES.[17] Twist Bio custom probe panel hybridization occurred overnight followed by a streptavidin magnetic-bead enrichment process used to capture probes hybridized to the target regions of interest. Following three high-temperature washes, the enriched library was amplified by PCR and subjected to AMPure bead cleanup. Finally, the library quality was confirmed with an Agilent Tapestation 4200 using High Sensitivity D1000 ScreenTape, and then equimolarly pooled to a final loading concentration of 1,000 pM for paired-end sequencing on a NextSeq 2000 (Illumina).

Long-read TES

A TES method compatible with PacBio sequencing was adapted from a previous publication.[18] Approximately 1.5 μg of genomic DNA was diluted with nuclease-free water to a total volume of 150 μL and fragmented to a target fragment length of 10 kb with a g-tube (#520079, Covaris, USA). DNA fragmentation was done by spinning the g-tube two times for 60 s at 6,000 rpm in an Eppendorf 5424 centrifuge. From the g-tube, 150 μL of fragmented sample was recovered and transferred to a new tube, where it was cleaned up with the Qiagen QIAquick PCR Purification kit (#28104) and eluted in a smaller volume. End-repair, A-tailing, and adapter ligation of universal PacBio adapters was performed by following the Twist Bio Long Read Library Preparation and Standard Hyb v2 Enrichment workflow with the Twist Standard Hyb and Wash Kit v2. Prepared samples (N = 8) were cleaned up and equimolar pooled at 187.5 ng each. Pooled samples were then hybridized overnight with the custom Twist Bio biotinylated capture panel, blockers, and 5 μg of human Cot DNA (#100285, Twist, USA) for 18 h at 70°C. Biotinylated probes were captured with M-270 streptavidin beads (#65306, Thermo Fisher). Captured fragments were amplified with KOD Xtreme Hot Start Polymerase in a 17-cycle long-range PCR (initial denaturation, 120 s at 94°C; denaturation, 10 s at 98°C; annealing, 30 s at 58.8°C; extension, 600 s at 68°C; final extension, 600 s at 68°C). The PCR product was size selected using 1.0× volume of Twist Bio DNA Purification beads and QCed via HS DNA Tapestation.

PacBio library preparation and SMRT sequencing

The 08-plexed, barcoded, and enriched pool was used as the input of the SMRTbell Prep Kit 3.0 (#102-182-700, PacBio, USA), which was completed according to the manufacturer's instructions. Prepared libraries were sequenced on one SMRT Cell 8M (#101-820-200, PacBio, USA) on the PacBio Sequel II sequencing system with a video length of 30 h.

Dosing vector libraries were following the PacBio AAV SMRTbell library protocol (102-126-400) with few exceptions in DNA extraction summarized below. Single-stranded genomic DNA was extracted from the AAV capsids via heat denaturation and hybridized into double-stranded DNA through a slow thermal annealing. Hybridized DNA was purified with the Qiagen QIAquick PCR Purification kit and prepared for sequencing using the SMRTbell Prep Kit 3.0. Prepared libraries were sequenced on the PacBio Sequel IIe using one SMRT Cell 8M with a video length of 30 h.

Long-read TES analysis

We downloaded the Ensembl genomes (v. 99) for cyno (*M. fascicularis*) and human (*Homo sapiens*). The appropriate vector genome sequences for each sample were concatenated with the host genome and index files produced using BWA (v. 0.7.17)[19] to construct custom hybrid reference genomes.

Custom bash and R scripts (https://github.com/sheehanmarkj/insertional_mutagenesis_long-read_public) were written to process the long-read fastq data to identify and characterize insertion and vector-only reads. Reads were first aligned using BWA MEM and sorted and indexed using samtools (v. 1.9).[20] Subsequently, the long-read structural variant detection tool cuteSV (v. 2.0.3)[21] was run on the sorted BAM, with report-id flag on and min-support of 0, to capture all possible insertion reads.

The resulting cuteSV VCF files were processed by a custom R script to pull BND reads involving host-vector breakends. A custom bash pipe pulled read names for alignments mapping to single contigs from the hybrid reference for downstream host and vector-only read characterization. Vector-only reads were extracted from the BAM using a Python script.[22] A final R script processed vector-only BAM and insertion reads of interest, calculating additional metrics and reconstructing the read structure for downstream characterization.

Distinct reads indicating the same insertion event were further investigated by identifying identical breakend junctions in the reconstructed read data to capture potential clonal expansion or integration hotspots. Vector rearrangements in chimeric and vector-only reads were characterized by analyzing vector-vector breakend junctions within reconstructed reads from each data source.

Short-read TES analysis

Alignment to vector

Short-read data were analyzed using an updated version (https://github.com/sheehanmarkj/insertional_mutagenesis_short-read_public) of a previously published pipeline.[8] Paired raw fastq files were interleaved and passed to cutadapt v. 1.9.1,[23] trimming reads for quality and Illumina adapter sequences with minimum length threshold of 40 and Phred quality of 30. BWA MEM was used to align the reads to the BWA-indexed vector genome. The resulting BAM files were used to generate a BED file for portions of each read that aligned to the viral genome, using a custom R script. Alignments were

subsequently filtered for minimum mapping quality of 30, and a modified version of the Perl script samclip[24] was used to filter to reads with 30–120 bases properly mapped to the viral genome. Finally, PCR duplicates were removed using samtools fixmate and markdup.

Alignment to host

The BED files generated from the unfiltered viral alignment were used to mask vector-aligned sequence to N in the cutadapt-trimmed reads using seqtk seq.[25] The resulting reads were aligned to the appropriate BWA-indexed host genome using BWA MEM and filtered for minimum mapping quality of 30. In parallel, unmasked reads were aligned to the host genome and similarly filtered. This unmasked alignment was used to identify PCR duplicates using samtools fixmate and markdup, which were subsequently removed from the masked alignment.

To identify ISs, two sources of read information were used: (1) mate read (reads in which one of the pairs maps to host genome and the other maps to viral genome); (2) softclipped reads (reads in which a portion of the read maps to host genome and another portion maps to the viral genome).

Insertion read identification: Mate reads

From the masked and unmasked alignment files for each sample, reads were extracted that did not properly align to the host genome, using samtools view (-h -f 4 -F 8). Reads that mapped properly but whose pair did not map properly to the host genome were also extracted using samtools view (-f 8 -F 4 -q 30). The intersection of masked and unmasked reads from the above procedure was kept.

Insertion read identification: Softclipped reads

A modified version of the Perl script samclip was used to identify reads in the masked alignment with a minimum of 30 bp properly mapped to the host genome and a minimum of 30 bases that are softclipped for lack of alignment. In parallel, reads were identified in the unmasked alignment with a minimum of 30 bp properly mapped to the host genome and fewer than 30 bases softclipped for lack of alignment. These will include reads that can be fully explained by the host genome, which may partially align to the viral genome due to sequence similarity. Only reads unique to the output from the filtered masked alignment were kept.

Extracting reads

The names of candidate mate reads from the host alignment were used to extract their properly aligned mates in the vector alignment using a Python script, keeping only vector alignments with the appropriate host-aligned read unmapped. The names of candidate softclipped reads properly aligned to host genome were extracted from the viral genome using the same approach.

These final high-confidence insertion reads were extracted from the masked host-aligned BAM. Reads that identified host genomic locations from both sets of evidence (mate and softclipped) were merged using samtools merge and re-sorted by name. Relevant information for host and vector location information was pulled from these final BAMs for softclipped-vector, mate-vector, and host alignments.

DATA AND CODE AVAILABILITY

Raw data have been deposited in SRA submissions (SRA: PRJNA1060711) (ID: 1060711 BioProject, NCBI [nih.gov]) (cyno) and PRJNA1060724 (ID: 1060724 BioProject, NCBI [nih.gov]) (HEK293 and HepG2). Analysis code for long and short read data is available at the GitHub repositories referenced in their respective methods sections.

ACKNOWLEDGMENTS

The authors would like to acknowledge Megan Leander for generating PacBio libraries and data for the AAV.

The authors also thank Yale Center for Genome Analysis and Keck Microarray Shared Resource at Yale University for providing the necessary PacBio sequencing services, which is funded in part by the National Institutes of Health instrument grant 1S10OD028669-01.

Graphical abstract created in BioRender. Sheehan, M. (2024) BioRender.com/p27u784.

AUTHOR CONTRIBUTIONS

M.S. led the data integration and analysis and co-wrote the manuscript. S.W.K. and J.Q. performed the integration experiments. D.M.R. developed positive control material. E.O. helped design and contributed to bioinformatic analysis. T.A.L. led the team and co-wrote the manuscript.

DECLARATION OF INTERESTS

This work was funded by Pfizer, Inc., and all authors are Pfizer employees.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtm.2024.101352.

REFERENCES

1. Braun, C.J., Boztug, K., Paruzynski, A., Witzel, M., Schwarzer, A., Rothe, M., Modlich, U., Beier, R., Göhring, G., Steinemann, D., et al. (2014). Gene therapy for Wiskott-Aldrich syndrome–long-term efficacy and genotoxicity. Sci. Transl. Med. *6*, 227ra33. https://doi.org/10.1126/scitranslmed.3007280.
2. Cavazzana, M., Six, E., Lagresle-Peyrou, C., André-Schmutz, I., and Hacein-Bey-Abina, S. (2016). Gene Therapy for X-Linked Severe Combined Immunodeficiency: Where Do We Stand? Hum. Gene Ther. *27*, 108–116. https://doi.org/10.1089/hum.2015.137.
3. McCarty, D.M., Young, S.M., Jr., and Samulski, R.J. (2004). Integration of adeno-associated virus (AAV) and recombinant AAV vectors. Annu. Rev. Genet. *38*, 819–845. https://doi.org/10.1146/annurev.genet.37.110801.143717.
4. Chandler, R.J., Sands, M.S., and Venditti, C.P. (2017). Recombinant Adeno-Associated Viral Integration and Genotoxicity: Insights from Animal Models. Hum. Gene Ther. *28*, 314–322. https://doi.org/10.1089/hum.2017.009.
5. Donsante, A., Miller, D.G., Li, Y., Vogler, C., Brunt, E.M., Russell, D.W., and Sands, M.S. (2007). AAV vector integration sites in mouse hepatocellular carcinoma. Science *317*, 477. https://doi.org/10.1126/science.1142658.
6. Sabatino, D.E., Bushman, F.D., Chandler, R.J., Crystal, R.G., Davidson, B.L., Dolmetsch, R., Eggan, K.C., Gao, G., Gil-Farina, I., Kay, M.A., et al. (2022). Evaluating the State of the Science for Adeno-Associated Virus (AAV) Integration:An Integrated Perspective. Mol. Ther. *30*, 2646–2663. https://doi.org/10.1016/j.ymthe.2022.06.004.
7. (2020). Guidance for Industry: Long Term Follow-Up After Administration of Human Gene Therapy Products.
8. Oziolor, E.M., Kumpf, S.W., Qian, J., Gosink, M., Sheehan, M., Rubitski, D.M., Newman, L., Whiteley, L.O., and Lanz, T.A. (2023). Comparing molecular and computational approaches for detecting viral integration of AAV gene therapy

constructs. Mol. Ther. Methods Clin. Dev. *29*, 395–405. https://doi.org/10.1016/j.omtm.2023.04.009.

9. Dalwadi, D.A., Calabria, A., Tiyaboonchai, A., Posey, J., Naugler, W.E., Montini, E., and Grompe, M. (2021). AAV integration in human hepatocytes. Mol. Ther. *29*, 2898–2909. https://doi.org/10.1016/j.ymthe.2021.08.031.

10. Nguyen, G.N., Everett, J.K., Kafle, S., Roche, A.M., Raymond, H.E., Leiby, J., Wood, C., Assenmacher, C.-A., Merricks, E.P., Long, C.T., et al. (2021). A long-term study of AAV gene therapy in dogs with hemophilia A identifies clonal expansions of transduced liver cells. Nat. Biotechnol. *39*, 47–55. https://doi.org/10.1038/s41587-020-0741-7.

11. Nowrouzi, A., Penaud-Budloo, M., Kaeppel, C., Appelt, U., Le Guiner, C., Moullier, P., von Kalle, C., Snyder, R.O., and Schmidt, M. (2012). Integration frequency and intermolecular recombination of rAAV vectors in non-human primate skeletal muscle and liver. Mol. Ther. *20*, 1177–1186. https://doi.org/10.1038/mt.2012.47.

12. Gil-Farina, I., Fronza, R., Kaeppel, C., Lopez-Franco, E., Ferreira, V., D'Avola, D., Benito, A., Prieto, J., Petry, H., Gonzalez-Aseguinolaza, G., and Schmidt, M. (2016). Recombinant AAV Integration Is Not Associated With Hepatic Genotoxicity in Nonhuman Primates and Patients. Mol. Ther. *24*, 1100–1105. https://doi.org/10.1038/mt.2016.52.

13. Nguyen, G.N., Everett, J.K., Kafle, S., Roche, A.M., Raymond, H.E., Leiby, J., Wood, C., Assenmacher, C.-A., Merricks, E.P., Long, C.T., et al. (2021). A long-term study of AAV gene therapy in hemophilia A dogs identifies clonal expansions of transduced liver cells. Nat. Biotechnol. *39*, 47–55.

14. Retson, L., Tiwari, N., Vaughn, J., Bernes, S., Adelson, P.D., Mansfield, K., Libertini, S., Kuzmiski, B., Alecu, I., Gabriel, R., and Mangum, R. (2023). Epithelioid neoplasm of the spinal cord in a child with spinal muscular atrophy treated with onasemnogene abeparvovec. Mol. Ther. *31*, 2991–2998. https://doi.org/10.1016/j.ymthe.2023.08.013.

15. Schmidt, M., Foster, G.R., Coppens, M., Thomsen, H., Dolmetsch, R., Heijink, L., Monahan, P.E., and Pipe, S.W. (2023). Molecular evaluation and vector integration analysis of HCC complicating AAV gene therapy for hemophilia B. Blood Adv. *7*, 4966–4969. https://doi.org/10.1182/bloodadvances.2023009876.

16. Martins, K.M., Breton, C., Zheng, Q., Zhang, Z., Latshaw, C., Greig, J.A., and Wilson, J.M. (2023). Prevalent and Disseminated Recombinant and Wild-Type Adeno-Associated Virus Integration in Macaques and Humans. Hum. Gene Ther. *34*, 1081–1094. https://doi.org/10.1089/hum.2023.134.

17. Duncavage, E.J., Magrini, V., Becker, N., Armstrong, J.R., Demeter, R.T., Wylie, T., Abel, H.J., and Pfeifer, J.D. (2011). Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. J. Mol. Diagn. *13*, 325–333. https://doi.org/10.1016/j.jmoldx.2011.01.006.

18. Steiert, T.A., Fuß, J., Juzenas, S., Wittig, M., Hoeppner, M.P., Vollstedt, M., Varkalaite, G., ElAbd, H., Brockmann, C., Görg, S., et al. (2022). High-throughput method for the hybridisation-based targeted enrichment of long genomic fragments for PacBio third-generation sequencing. NAR Genom. Bioinform. *4*, lqac051. https://doi.org/10.1093/nargab/lqac051.

19. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

20. 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.

21. Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. *21*, 189. https://doi.org/10.1186/s13059-020-02107-y.

22. Stuart, T. (2015). extract_reads.py.

23. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

24. Seeman, T. (2020). samclip: filter SAM file for soft and hard clipped alignments.

25. Li, H. (2023). Seqtk.