Molecular Therapy
Methods & Clinical Development
Original Article

# Comparing molecular and computational approaches for detecting viral integration of AAV gene therapy constructs

Elias M. Oziolor,[1] Steven W. Kumpf,[2] Jessie Qian,[2] Mark Gosink,[1] Mark Sheehan,[2] David M. Rubitski,[2] Leah Newman,[2] Laurence O. Whiteley,[3] and Thomas A. Lanz[2]

[1]Global Computational Safety Sciences, Pfizer Inc., Groton, CT 06340, USA; [2]Global Discovery, Investigative and Translational Sciences, Pfizer Inc., Groton, CT 06340, USA; [3]Global Pathology, Pfizer Inc., Cambridge, MA 02139, USA

**Many current gene therapy targets use recombinant adeno-associated virus (AAV). The majority of delivered AAV therapeutics persist as episomes, separate from host DNA, yet some viral DNA can integrate into host DNA in different proportions and at genomic locations. The potential for viral integration leading to oncogenic transformation has led regulatory agencies to require investigation into AAV integration events following gene therapy in preclinical species. In the present study, tissues were collected from cynomolgus monkeys and mice 6 and 8 weeks, respectively, following administration of an AAV vector delivering transgene cargo. We compared three different next-generation sequencing approaches (shearing extension primer tag selection ligation-mediated PCR, targeted enrichment sequencing [TES], and whole-genome sequencing) to contrast the specificity, scope, and frequency of integration detected by each method. All three methods detected dose-dependent insertions with a limited number of hotspots and expanded clones. While the functional outcome was similar for all three methods, TES was the most cost-effective and comprehensive method of detecting viral integration. Our findings aim to inform the direction of molecular efforts to ensure a thorough hazard assessment of AAV viral integration in our preclinical gene therapy studies.**

## INTRODUCTION

Gene therapy offers the potential for permanent transformation of patient cells to correct or overcome disease-causing mutations. One of the most efficient methods for delivering DNA cargo into living cells is through a virus. Engineered versions of the adeno-associated virus (AAV) have become a favored vehicle for gene therapy in recent years, with over 130 active AAV clinical trials listed on the clinicaltrials.gov website and two FDA-approved AAV gene therapies currently in the United States: Luxturna (for inherited retinal disease) and Zolgensma (for spinal muscular atrophy).

One of the advantages of gene delivery through AAV over lentiviral or retroviral vectors is the maintenance of viral DNA in an episome, rather than relying on integration into host DNA for transgene expression. This should lower the risk of an oncogenic insertional mutation, as has been observed in early gene therapy trials using retroviral vectors.[1–3] While the vast majority of AAV DNA remains episomal, approximately 0.1%–0.5% of delivered AAV DNA has been found to integrate into the host cell DNA.[4] A finding of insertional mutagenesis of naturally occurring AAV2 has been found in human hepatocellular carcinoma (HCC) samples, in which integrations near five known cancer driver genes were detected.[5] Recombinant AAV, in which most of the endogenous viral DNA has been replaced with a therapeutic payload, has been evaluated for genome integration and insertional mutagenesis in several preclinical studies. A finding of HCC caused by a recombinant AAV integration into the Rian locus has been observed in mice following neonatal injection,[6] a pattern that has been replicated using other AAV serotypes with differing cargos.[7] No Rian locus integration or HCC have been reported in adult mice following recombinant AAV administration (reviewed previously[8,9]). Long-term dog hemophilia gene therapy studies tracked animals up to 10 years following treatment with recombinant AAV8/9, and while >2000 integration sites have been identified, no tumors or liver toxicities have been identified.[10–12] Similarly, cats treated with AAV for mucopolysaccharidosis type VI as juveniles were followed for 8 years and showed no signs of HCC.[13] Furthermore, an evaluation of non-human primate tissues and liver biopsies from human subjects dosed with recombinant AAV2/5 detected integration sites that lacked a preference for cancer driver genes and did not associate with HCC.[14]

While no data exist to suggest a link between recombinant AAV integration and HCC outside of neonatal mice or adult mice with nonalcoholic fatty liver disease,[15] the FDA recommends assessment of vector integration in preclinical studies for new AAV gene therapies.[16] A number of molecular methods for identifying sites of viral integration have been published. Techniques such as linear amplification-mediated or

Received 28 June 2022; accepted 28 April 2023;
https://doi.org/10.1016/j.omtm.2023.04.009.

**Correspondence:** Thomas A. Lanz, Pfizer Inc., Eastern Point Rd, Groton, CT 06340, USA.
**E-mail:** Thomas.a.lanz@pfizer.com

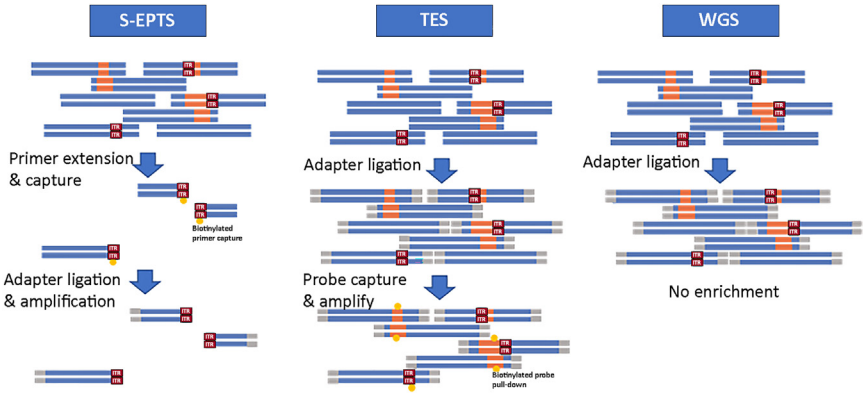


**Figure 1. DNA samples were processed for evaluation of viral integration by three different workflows**

Blue bars represent host DNA, orange bars represent viral DNA (boxes representing ITR sequence), and gray bars represent Illumina sequencing adapters.

shearing extension primer tag selection ligation-mediated PCR (S-EPTS) anchor a primer to the viral construct terminal repeat and extend a PCR product out, which can be sequenced to identify any connected genomic DNA.[17–19] An alternative approach to capturing DNA associated with multiple viral elements is targeted enrichment sequencing (TES), in which oligonucleotides designed to all portions of the vector are used to pull down DNA from fresh, frozen, or formalin fixed paraffin embedded (FFPE) samples for subsequent sequencing.[20–23] Finally, it is possible to bioinformatically identify virally associated DNA sequences in whole-genome sequencing (WGS) data.[24,25] The present study sought to compare the advantages and drawbacks of different molecular techniques of assessing viral integration sites to support ongoing and future AAV gene therapy studies. An AAV construct bearing a codon-optimized transgene was administered to cynomolgus monkeys and mice. Tissues from treated animals were evaluated for genomic integration by S-EPTS, TES, and WGS using a common analytical pipeline. Given the flexibility and sensitivity of TES, additional experiments were performed to evaluate FFPE samples, and optimizations were made to increase the signal:noise ratio.

## RESULTS

### Viral integration workflows

DNA was isolated from cyno and mouse gene therapy studies for an experimental AAV construct 6–8 weeks after dosing (n = 2 per tissue per dose group). Biodistribution data measured by qPCR confirmed dose-dependent expression of vector DNA in both species (Figure S2). The highest dose in cyno was 10-fold higher than the highest dose in mice, and this is reflected in the vector DNA content between the species. Two samples were selected at two dose levels for both mouse and cyno for analysis of viral integration in liver, as this tissue retained the greatest concentration of AAV DNA (up to 4.1 x $10^7$ copies per µg genomic DNA). Two heart samples from cyno were also selected to assess the sensitivity of integration techniques, as expression in the chosen samples ranged from 1.7 x $10^3$ to 2.8 x $10^5$ copies per µg genomic DNA.

DNA samples from each subject were split into three aliquots to contrast three methods of measuring viral integration: S-EPTS, TES, and WGS. Simplified workflows are presented in Figure 1A. All three workflows started with DNA fragmentation. S-EPTS used sonication, and TES and WGS used enzymatic fragmentation. For S-EPTS, biotinylated primers designed against the 5′ inverted terminal repeat (ITR) sequence extend out into any genomic DNA attached to those regions. These sequences are enriched, ligated to Illumina adapters, and amplified. For TES, a panel of biotinylated oligonucleotides was designed against the entire sequence of the AAV vector. Following ligation of Illumina adapters to the DNA fragments, the oligonucleotide baits are used to pull down only DNA that contains some of the AAV sequence, and these fragments are amplified and sequenced. For WGS, fragmented DNA is simply prepared for Illumina sequencing. No enrichment is performed prior to sequencing; all identification of viral integration happens during bioinformatic analysis.

For all three workflows, viral integration site analysis depended upon one of two types of evidence of hybrid viral-host DNA: paired-end reads or soft-clipped reads. For paired-end calls, one read for a given transcript maps to the host genome, while its read pair maps to the recombinant AAV genome. For soft-clipped calls, a single read mapped to both AAV genome and host genome. Due to the high sequence similarity of the transgene to the endogenous gene, this portion of the construct was ignored for these mappings.

### Insertional event origination site

Detected viral insertion events were sorted by their start site within the AAV sequence and compared across the three methods in cyno tissue samples (Figure 2A). For S-EPTS, insertional events were primarily restricted to the 5′ ITR, as this was the only sequence used to enrich viral DNA. The presence of insertion site(s) (IS) starting near the 3′ polyA signal may be indicative of integrated concatemer. TES and WGS datasets showed that integration events began at locations across the viral genome. The distribution differed between TES and WGS, perhaps attributable to differential probe sensitivity in the TES panel. Even the WGS data, however, are not uniform across the viral genome, so particular sequence motifs may play a role in integration. While the WGS method may be the most unbiased method for identifying integration sites, more IS were detected by TES in the current study, despite almost 100 times deeper sequencing of the WGS samples. The enrichment provided by S-EPTS and TES significantly reduces the sequencing depth and analysis time required to identify IS. The start site location within the viral genome that mapped to the host genome integration location for TES data is shown in Figure 2B, which demonstrates a fairly random distribution of IS across

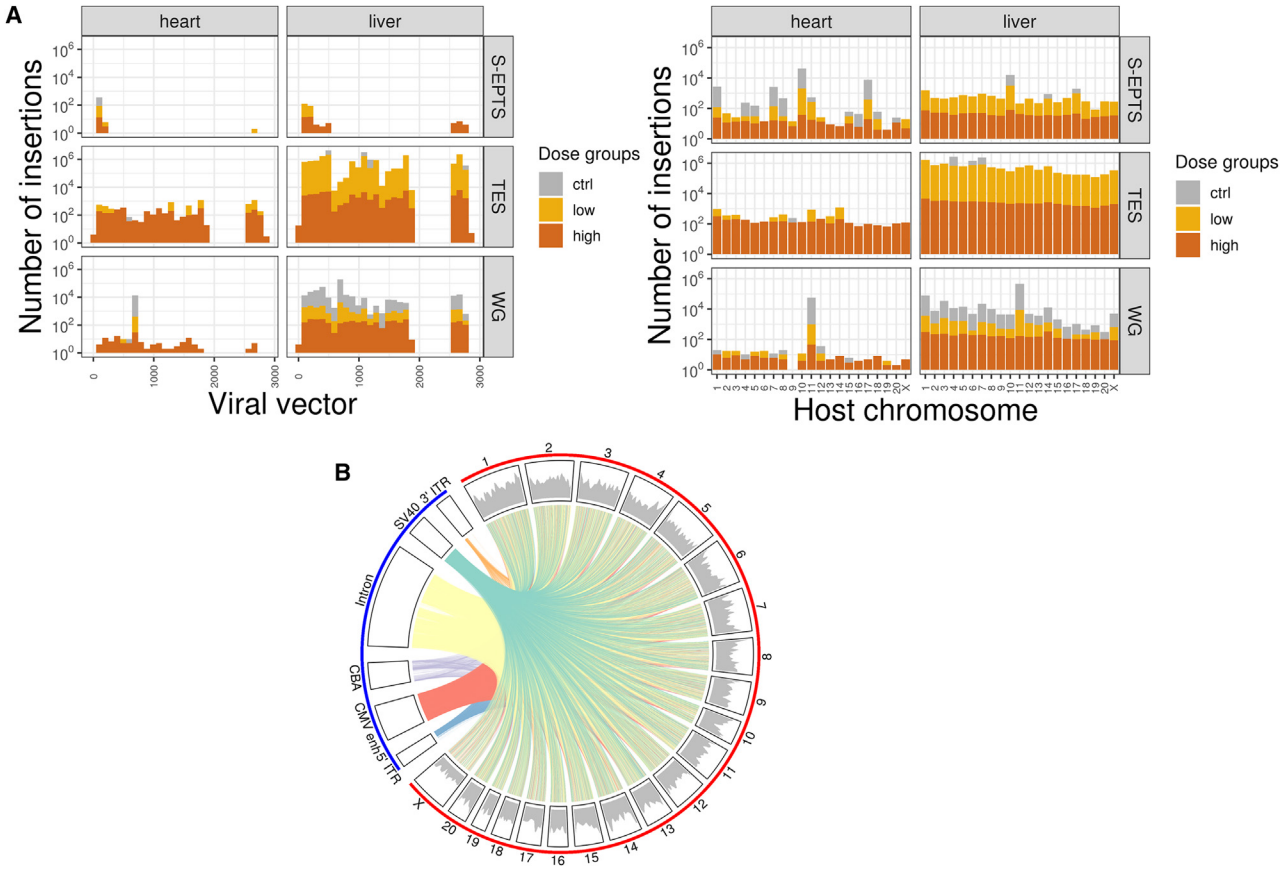


**Figure 2. Quantitation of insertion sites as a function of viral origin or host location**

(A) Insertional start sites in cyno are presented as a histogram of aggregated sample values for each method, with the y axis representing number of insertional events, and x axis representing position on the viral genome (left) or host chromosome (right). The transgene portion was omitted from analysis due to high sequence similarity to the endogenous gene. IS count is shown in dark orange for high dose samples, light orange for low dose, and gray for vehicle control samples. (B) An example (sample 191107330) of the rAAV source locus and ultimate host location of insertion. Line colors represent the different segments of the rAAV origin and edges represent the origin and destination of insertion.

the host genome. Results in mouse (Figure S2) show similar results to the cyno. IS were found in every chromosome, with no genomic locus standing out as accepting a majority of IS. A chi-squared comparison of control adjusted insertional site frequencies in chromosomal length adjusted frequency expectations revealed no significant over-representation of insertional sites across chromosomes ($p > 0.05$) for all methods, doses, and tissues (Table S1).

### Quantification of insertional events

The number of insertional events detected by each method in each tissue is shown in Figure 3 (a full list of all IS associated with genes is found in Table S2). In general, the frequency of insertion correlates with AAV copy numbers previously measured by qPCR (Figure S3). While all three methods were successful in discriminating high dose from vehicle, WGS exhibited a higher background and could not distinguish low dose from vehicle in cyno. In both cyno and mouse samples, the WGS appeared to have a lower dynamic range than the S-EPTS or TES techniques. Consistent with higher dose and vector copy number, the total number of IS identified in cyno at the high dose was greater than that identified in mice.

As a positive control for integration, the ITR-spanning transgene construct was inserted into a lentiviral vector and dosed into HT1080 cells at MOI ranging from 3 to 30 MOI, and DNA was extracted and analyzed by TES. Over 1,000 IS were measured at each MOI and exhibited a dose-responsive increase (Figure S4A). While LTR sequences were not captured by the TES, integration start sites were distributed throughout the vector, much like the *in vivo* AAV data (Figure S4B).

### Pattern of integration: Hotspot analysis, genomic features, and clonality assessment

As integration is expected at low frequency with AAV, the pattern of integration is more likely to have functional consequences than the total number of IS. Overlap in genes associated with integration events was scarce between methods (<2%; Figure S5). To probe

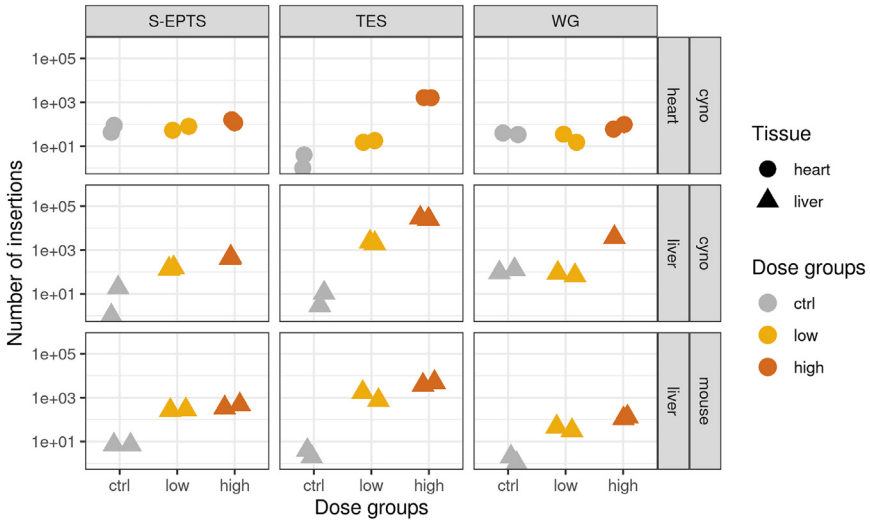


**Figure 3. Number of regions found with insertions by each method in cyno (top panels) and mouse samples (bottom panels)**

Points are colored by dose, and symbols denote different tissues.

further, a hotspot analysis was conducted to determine whether certain genomic loci were more prone to viral integration. Five or fewer hotspots were detected in all conditions (Figure 4A). Hotspots were most frequently detected in cyno liver, which were the samples with the highest AAV content and thus highest IS content. The base size of integration was highly variable, ranging from <10 to >$10^6$ (Figure 4B). Overlap in genes identified under hotspots between methods is shown in Figure 4C, with overlap in cancer genes shown in Figure 4D.

We also examined the IS using the HOMER (hypergeometric optimization of motif enrichment) package to determine if IS were enriched with any general type of genomic features. Across the different methods, the WGS approach yielded the least bias in the regions harboring apparent IS (Figure 5). Both the TES and the S-EPTS approaches showed some preferences for certain genomic features (e.g., ribosomal RNA, potentially reflecting regions of frequently accessible chromatin) with the S-EPTS approach displaying the most bias.

To further characterize repeated integrations into the same site, a clonality analysis was performed based on looking at the number of different fragment lengths of a particular IS. Unique fragment lengths with the same IS are suggestive of clonal expansion of cells following an insertion event. Such clones were identified in cyno samples across all three methods (Figure 6). In several cases, an IS with putative expanded clones was found in multiple individuals (represented by connector lines). The vast majority of IS were represented by just a single fragment size. As fewer IS were found in mice than cyno, fewer putative expanded clones were detected (Figure 6). Given the minimal evidence of clonal expansion, no loci were followed up for additional analysis.

#### Additional optimization of TES method

As TES provided the best balance between comprehensive coverage of the vector and cost, additional experiments were performed to better understand application to FFPE samples and to increase signal:noise. One difference between fresh-frozen and FFPE-derived DNA is reduced DNA integrity in FFPE. For the SureSelect library preparation, the manual suggests additional fragmentation time and additional PCR cycles for FFPE-derived samples. To evaluate these changes, FFPE and fresh-frozen derived DNA were each run through both the standard library prep conditions and the modified FFPE conditions. The FFPE protocol resulted in cDNA libraries that were on average 50 bp smaller than the standard protocol (likely due to increased fragmentation time), and regardless of the source of the starting sample, the FFPE protocol yielded more IS than the fresh-frozen protocol (Figure 7A).

It has been previously suggested that artifactual IS can be created during the library preparation process.[26] Since vehicle controls in the present study showed some level of IS by all three methods, an additional control was run in which viral DNA was spiked into host DNA at a concentration equivalent to a high-dose treatment condition, and IS were measured by TES. In agreement with the prior finding using a PCR-based method, control samples spiked with AAV resulted in nearly 1,000 regions with IS detected (Figure 7B). As ligation is the most likely step where this artifact may be produced, tagmentation was used as an alternative method for adding sequencing adapters, as this method employs a transposase that simultaneously cleaves the DNA and adds an adapter sequence. The result was a near elimination of IS detection in the control samples by TES. All three protocols used adapter ligation during library preparation, so this is likely the major contributor to detection of IS.

## DISCUSSION

The present study compared three methods for measuring AAV insertion into the host genome using two different species, using samples dosed with vehicle and different dose levels of a recombinant AAV. These methods were targeted to the ITR sequence (S-EPTS), the entire AAV vector sequence (TES), or untargeted (WGS). IS could be measured by all three techniques in cyno and mouse tissues, though WGS could not distinguish low dose from background in cyno, and it had the lowest dynamic range of the three methods. Under the sequencing parameters used in this study, TES identified the greatest number of IS and exhibited the largest dose-responsive dynamic range. TES and WGS revealed that integration is not ITR dependent and can start at any point along the vector genome. Thus the S-EPTS technique only captured a fraction of the total

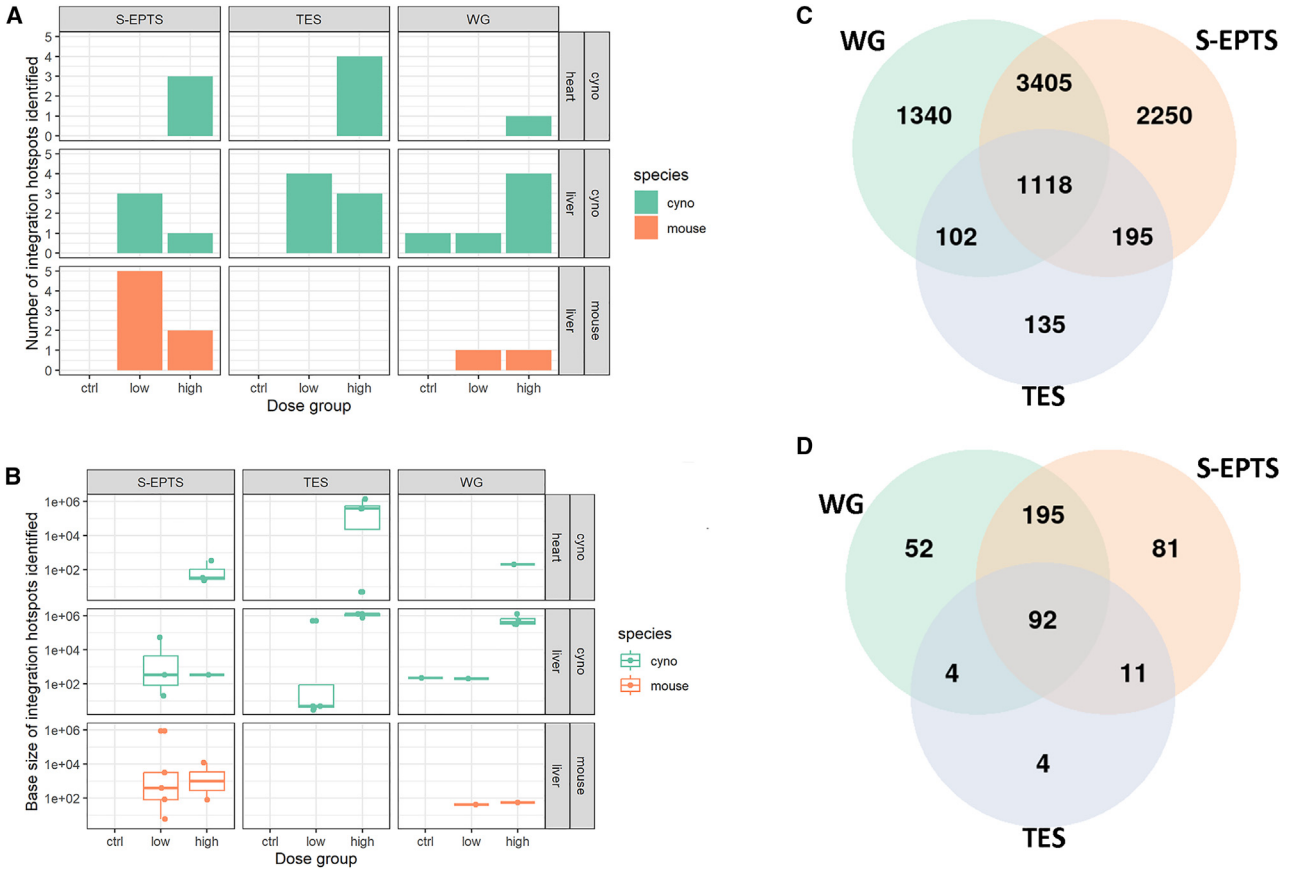


**Figure 4. Integration site distribution was analyzed for hotspots in the genome**

(A) Total number of integration hotspots is listed for each technique (columns) across treated cyno (blue) and mouse (orange) samples. (B) Base size of the integration hotspot is shown for each hotspot identified. (C) Venn diagram showing the overlap of the insertion sites within hotspot regions between the different techniques. (D) Overlap of cancer implicated genes under statistical hotspots between different techniques.

integration events, though hotspot, clonality, and functional analyses resulted in similar conclusions from all three methods: integration was not biased toward any particular part of the genome and did not result in significant clonal expansion in either heart or liver.

The quasi-stochastic nature of AAV integration into host genome has been observed by several other studies. While wild-type AAV2 shows a hotspot of integration into AAVS1 locus in human chromosome 19, likely through a Rep-dependent mechanism, this has not been observed with recombinant AAV.[27,28] Rep and cap sequences found in wild-type AAV are removed from the vector sequence when designing therapeutic vectors to eliminate the possibility of replication. Yet AAV integration sites are not completely without a pattern. Hotspot analysis in the current dataset showed some enrichment in transcriptional units, consistent with past observations. AAV integration has shown a preference for accessible chromatin.[28] Reducing DNA methylation in cells has also been shown to increase integration frequency.[29] Double-stranded DNA breaks provide an opportunity for AAV DNA to integrate, and chemical or radiation treatments that increase these breaks increase integration rate.[30] CRISPR-Cas9 is another mechanism of producing double-stranded breaks and has likewise shown an increase in AAV DNA integration.[31] AAV delivery of CRISPR cargo is being pursued to enable targeted gene editing for a number of therapeutic targets.[32] Thus a number of factors in the cell may direct viral integration to a subset of the whole genome.

Data from the present study support the concept that specific elements of the AAV vector sequence are not required for initiation of an integration event. The TES and WGS analyses showed that integration start sites were found throughout the vector, as has been described for other recombinant vectors.[12,26,33] Even the S-EPTS method picked up a proportion of IS that initiated beyond the 5′ ITR. These sites could be the result of vector re-arrangement *in vivo*, which has been observed previously.[12,33] Truncation and re-arrangement of vector elements likely negatively impacted the number of such sequences captured by the S-EPTS and TES techniques. The TES oligonucleotides, for example, are 120-bp oligonucleotides based on the original vector sequence. Re-arrangements within the sequence of a probe would limit the ability of that probe to bind. Concatemerization would be difficult to accurately assess using short-read

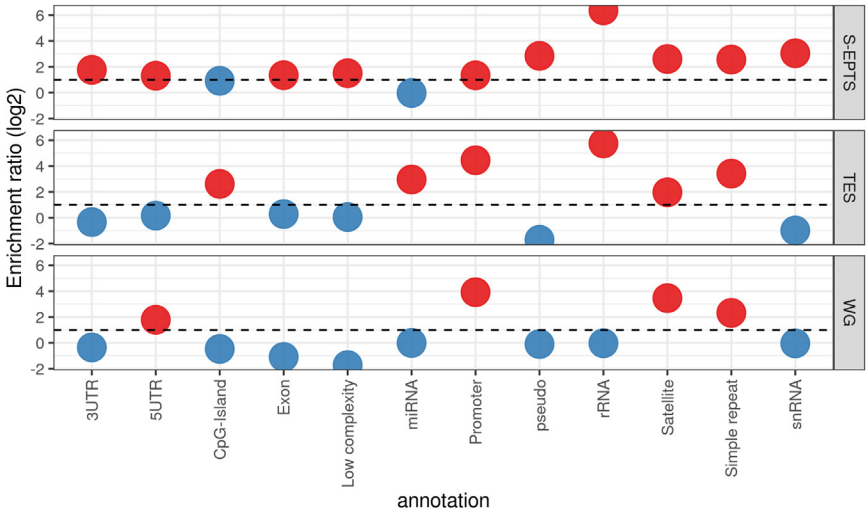


**Figure 5. A motif analysis was performed on the integration to look for over-enriched genomic DNA features across the three methods**

The y axis displays the log2 of the ratio of reads observed between the listed method and a random distribution for different gene regions. The dashed line signifies the significance threshold, with points passing this threshold colored in red.

sequencing, though estimation of concatemer frequency by identifying junctions of 3′ and 5′ ITR could be a possible next step with the present datasets, though long-read sequencing would provide a fuller picture of such events. The higher number of IS measured in the FFPE DNA, which resulted in libraries that were smaller than non-FFPE DNA samples, could reflect exposure of a greater pool of re-arranged vector sequences due to DNA cleavage interrupting a re-arrangement. The potential for fixation-related artifactual cross-linking that was not properly reversed during DNA extraction cannot be ruled out as a partial contributor; however, altering DNA fragmentation time and library prep conditions demonstrated that these factors can have a significant impact on total IS captured. FFPE conditions also increase IS detected in spike-in samples, demonstrating that artifactual IS can be created during library preparation. The switch from ligation to tagmentation resulting in fewer IS detected in all samples, including controls and spike-in samples. The potential for library prep-associated artifacts, however, underscores the need for additional follow-up for any IS identified that confer potential risk.

It is possible that technical limitations of the sequencing technology resulted in a subset of integration events being overlooked by all three techniques. The histogram of integration start sites (Figure 2) showed some portions of the AAV vector with relatively low abundance, such as the middle of the chimeric intron. As the pattern at this site was observed with both TES and WGS, it cannot be attributed to lack of coverage by the baits. With Illumina sequencing, highly repetitive elements can be difficult to sequence accurately. Sequences of high similarity between vector genome and host genome can also present a challenge to distinguish IS from purely endogenous host sequences. The sequence similarity to the transgene led to masking this sequence from analysis to prevent detection of false positives. The presence of false positives for IS in control samples suggests that further approaches in sequence similarity masking of host and vector sequence are necessary to avoid over-estimation of IS. Prior exposure to wild-type AAV may be another source of false positives, and this is perhaps the reason why the cyno control had a higher number of apparent IS than the mouse control group, as cyno are more likely to have been exposed to a wild-type AAV than an inbred mouse line.

Detection of IS and analysis for any patterns that would infer a risk of oncogenesis will be an important component of development for new gene therapy therapeutics relying on delivery by AAV. The present data showed that three different methods could achieve this task adequately, and all showed lack of enrichment in cancer genes. Non-human primate liver TES data were presented recently for a hemophilia A gene therapy in development (valoctocogene roxaparvovec); like the present dataset, it showed that integration sites were more common near actively transcribed genes but showed no enrichment near cancer genes.[22] Most published integration data to date have focused on liver, as both the tissue that may accumulate the greatest concentration of AAV following systemic delivery and the only tissue associated with causative tumor formation (albeit only in mice). The finding of integration sites in heart in the present dataset was not surprising given detection of vector DNA in that tissue by qPCR. As the number of AAV insertions correlates with biodistribution of the vector, serotypes or delivery methods that focus on maximizing vector dose in a particular tissue must expect a larger burden of IS in that tissue as well.

Clonal abundance can be estimated in the sequencing data based on DNA fragmentation pattern. When the genomic DNA is isolated, DNA is randomly fragmented, resulting in differing chromosomal DNA sizes going into library preparation for sequencing. Any PCR amplification during library preparation would yield identical DNA fragment sizes attached to the viral DNA, whereas if a specific integration was found in multiple cells, it would result in multiple host DNA fragment sizes attached to the same sequence of viral DNA. Thus measuring the number of differing fragment sizes at an IS can yield a rough estimate of clonal abundance. This analysis is not conclusive of expansion. False positives can occur if the location is a hotspot for insertion, as multiple fragment sizes could be reflecting different cells with independent IS, rather than a single IS that was passed along through clonal expansion. In addition, expansion itself can occur in non-carcinogenic processes (i.e., tissue regeneration after wound healing and as part of the normal aging process).[34,35] These normal

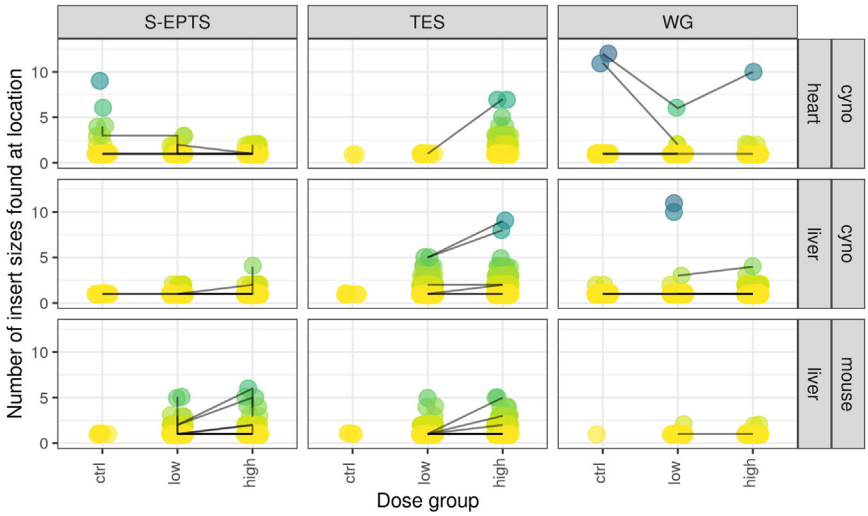


**Figure 6. Clonality assessment was performed for cyno (A) and mouse (B) samples**

The y axis represents the number of distinct fragments identified at the same IS, indicative of potential clonal expansion. Points represent fragment (clone) number, and dashed lines connect the same IS clones identified across samples.

processes can also lead to polyploidy,[36] which cannot be distinguished from low-level clonal expansion in an estimation based on DNA alone. These methods have been published in a long-term AAV dog study that identified potential sites of clonal expansion of over 100 cells at a given site in multiple animals with no associated tumors or adverse events.[12] Thus the detection of 15 or fewer IS fragment sizes in both tissues in the present study suggests minimal to no clonal expansion following integration.

Regardless of tissue target, the TES method offers more comprehensive vector coverage compared with S-EPTS, and it is more efficient and cost-effective than WGS. While WGS is theoretically the most comprehensive method, it would require over 1,000 times the sequencing depth per sample compared with TES due to the need to sequence the entire genome to capture events present at a relatively small frequency. The increased data burden comes with significantly increased computational time and storage costs. FFPE compatibility offers additional flexibility for use of TES on biopsy samples or sections commonly retained from preclinical toxicology studies, though FFPE samples should not be mixed with fresh-frozen without careful cross-validation, and each new TES panel should be independently validated for performance with the type of DNA to be tested. Furthermore, tagmentation is superior to ligation for adding adapters, due to potential ligation-mediated artifacts. As a positive control for the TES assay development, we constructed a lentivirus that contained the AAV ITR-spanning sequence to force integration into a human cell line. Such a control can be useful in assessing assay sensitivity and reproducibility prior to using human samples. Establishing methods for integration site analysis and a framework for interpretation of results will be an important step in advancing gene therapy treatments.

## MATERIALS AND METHODS

### Sample preparation

The samples used to evaluate viral integration were derived from *in vivo* studies in C57Bl6/J mice (7 weeks old at onset of dosing) and normal cynomolgus monkey (*Macaca fascicularis*; cyno; 2–2.5 years old) dosed with a recombinant AAV9 construct (Figure S1). Sequencing of the vector prep confirmed 77% full-length sequence, with 21% of reads coming from plasmid backbone and <2% derived from helper plasmids or rep/cap. Cynos received a single intravenous injection of vehicle (saline) or AAV at 1 x $10^{12}$ or 1 x $10^{14}$ vg/kg, and tissues were collected 6 weeks post-dose. Mice received a single intravenous injection of vehicle (saline) or AAV at 1 x $10^{12}$ or 1 x $10^{13}$ viral genomes per kg body weight (vg/kg), and tissues were collected 8 weeks post-dose. The dose of this vector at 1 x $10^{14}$ had previously been shown to result in liver toxicity in a mouse line on C57Bl6/J background in mice,[37] hence the down-shifted top dose for mice. For both species, 30- to 50-mg portions of liver were collected into DNAse-free microcentrifuge tubes, snap-frozen on dry ice, and stored at −80°C. Heart samples were collected in the same manner for cyno. Additional samples of cyno liver were formalin-fixed and embedded in paraffin. All procedures performed on animals were in accordance with regulations and established guidelines and were reviewed and approved by Pfizer's Institutional Animal Care and Use Committee.

Positive control samples for integration using lentivirus were prepared in HT-1080 cells. Cloning was performed by Genscript Biotech. The sequence from recombinant AAV was placed in a lentivector from System Biosciences (SBI, cat.# CD822A-1) while removing approximately 3.7 kb of sequence between the cPPT and 3′ LTR. Lentivirus was generated with SBI pPACKH1 (cat.# LV500A-1) packaging plasmids using Gibco LV-MAX Production System (cat.# A35684) following the protocol for a 125-mL shaker flask. 50 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, cat.# 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 (cat.# 631280). HT-1080 cells were transduced at three different MOIs. Six well plates were seeded with 2 x $10^5$ cells per well in Gibco DMEM (cat.# 11995-040), 10% FBS (cat.# 16140-071), and penicillin/streptomycin (cat.# 15070-063). 24 hours after plating, cells were treated with lentivirus in medium containing 5 µg/ml polybrene (Millipore cat.# TR-1003) at 30, 10, and 3 MOI. 72 hours post transduction, medium was removed and cells were washed with PBS and then collected in Gibco Cell Dissociation Buffer (cat.# 13151-014).

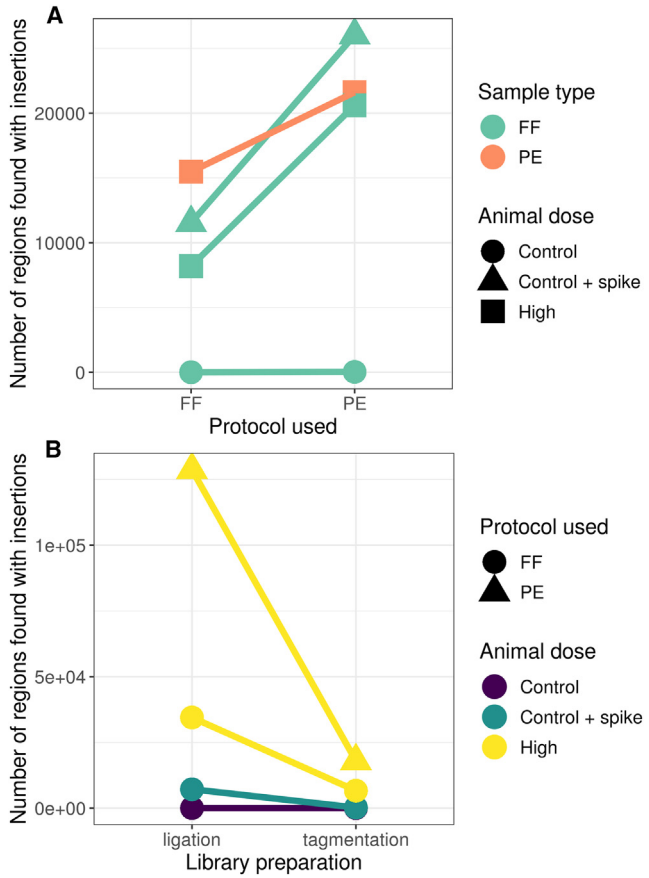


**Figure 7. Optimization of TES library prepration conditions**

(A) Comparison of fresh-frozen (FF) to FFPE (PE) samples. Points represent samples of FF or PE origin run through the FF or PE protocol. Samples were from control (untreated, circles), high dose animals (square), or viral spike-in controls (triangles). (B) Ligation was compared with tagmentation to determine if the latter would reduce the IS identified in control or spike-in samples.

DNA was isolated from all tissues or cells using DNeasy Blood and Tissue kits following homogenization in Buffer ATL in a TissueLyser (Qiagen). 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. For a subset of samples, DNA was isolated from FFPE liver sections using the QIAamp FFPE tissue kit (Qiagen) according to manufacturer's instructions. DNA concentrations were quantified by the Qubit Broad Range DNA assay (Life Technologies), and aliquots were split between multiple downstream processing methods.

### Vector copy number

A custom TaqMan assay for the transgene construct was designed, and standard curves were created with the plasmid DNA, ranging from $5 \times 10^0$ to $5 \times 10^9$. 100 ng DNA per sample was tested in triplicate in 96-well PCR plates with TaqMan Universal Master Mix II for a 20-µL reaction volume run on a ViiA7 Real-Time PCR Systems (Life Technologies). Viral genome copy number was interpolated for each sample using the plasmid standard curve.

### S-EPTS

The S-EPTS technique was performed at GeneWerk (Heidelberg, Germany) as previously described.[18] Briefly, 500 ng DNA was fragmented to a median length of 500 bp by sonication, followed by biotinylated primer extension and purification. Illumina adapters were ligated, and 250 x 50 bp paired-end sequencing was performed on a MiSeq (Illumina) at a depth that resulted in an average of approximately 270,000 reads per sample.

### Whole-genome sequencing

DNA library preparation was performed using the NEBNext Ultra II kit (New England Biolabs). Briefly, DNA samples were enzymatically fragmented, followed by end repair and ligation of Illumina adapter sequences. Libraries were sequenced on a HiSeq (Illumina) using 150-bp paired-end sequencing. Target genome coverage was 100x (approximately 2.3 billion paired-end reads per sample).

### Targeted enrichment sequencing

The SureSelect custom capture library was designed by Agilent Technologies.[38] Probes were designed for the capture of DNA sequences from the AAV vector. A panel of 825 120-mer oligonucleotide probes was used, for a total probe library size of 2 kbp. 200 ng total of gDNA per sample was loaded into the SureSelect target enrichment library preparation workflow (Agilent, Santa Clara, CA). The library preparations were performed according to the SureSelect XT HS2 Target Enrichment System for DNA Library Preparation and Target Enrichment for Illumina Paired-End Multiplexed Sequencing protocol (Version A0, Jan. 2020). First, all DNA samples were fragmented using the SureSelect Enzymatic Fragmentation Kit (Agilent) with a 10-min hold at 37°C for high-quality DNA samples and 15 min at 37°C for FFPE DNA samples. Further library preparation steps were followed per protocol with 8 cycles of amplification for high-quality DNA samples and 11 cycles of amplification for FFPE DNA samples. For a subset of samples, fragmentation and adapter addition were achieved via transpose instead of ligation using a tagmentation kit (Illumina) according to manufacturer instructions.

DNA libraries were purified using AMPure XP beads (Beckman Coulter, Brea, CA, USA) following amplification. Quality and quantity of libraries were determined by TapeStation using a D1000 ScreenTape (Agilent). Next, 1 µg of each library was hybridized with the custom SureSelect capture library overnight at 65°C. The hybridized libraries were purified with Dynabeads MyOne Streptavidin T1 magnetic beads (Thermo Fisher Scientific, Waltham, MA), and then the beads with captured DNA were washed once with wash buffer 1 and six times with wash buffer 2 kept at 70°C to remove non-specific binding. After all wash steps, the beads were suspended in 25 µL of nuclease free water. All 25 µL of the DNA libraries, bound to streptavidin beads, was amplified by PCR using SureSelect post capture primer mix and Herculase II Fusion DNA polymerase. The cycling conditions were as follows: 98°C for 2 min; followed by

16 cycles of 98°C for 30 s, 60°C for 30 s, and 72°C for 1 min and a final extension at 72°C for 5 min. After PCR, streptavidin beads were removed using a magnet stand, and the PCR products were further purified with AMPure XP beads. High-quality libraries were identified with an Agilent TapeStation using High Sensitivity D1000 ScreenTape and then equimolarly pooled for sequencing based on QuBit High Sensitivity DNA readings. Sequencing of SureSelect enriched libraries was performed on an Illumina NextSeq 500 (Illumina) using 150-bp paired-end reads, for an average of approximately 30 million reads per sample (Table S3).

## Analysis

### *Insertion site identification*

*Alignment to host genome.* We downloaded the Ensembl genomes (v. 99) for cyno (*Macaca fascicularis*) and mouse (*Mus musculus*) and produced index files using both BWA version 0.7.17[39] and samtools v. 1.9.[40] We built a custom bash pipeline to process all three sequencing data streams (https://github.com/eoziolor/insertional_mutagenesis_public). If a sample is paired, we interleave the fastq files for processing through the pipeline. We use cutadapt v. 1.9.1[41] to trim reads for quality and Illumina adapter sequences with minimum length threshold of 40 and Phred quality of 30. We align the resulting reads using BWA MEM, filter the alignment for minimum mapping quality of 30, apply samtools fixmate, and remove duplicates using samtools markdup.

To identify IS, we used two sources of read information: (1) mate read (reads in which one of the pairs maps to host genome and the other maps to viral genome) and (2) soft-clipped reads (reads in which a portion of the read maps to host genome and another portion maps to the viral genome).

#### *Read extraction*

*Mate read* From the alignment files for each sample, we extracted the reads that did not properly align to the host genome using samtools view (-h -f 4 -F 8) and converted them back to fastq files using bedtools.[42] We also extracted the reads, which mapped properly, but whose pair did not map properly to the host genome using samtools view (-f 8 -F4 -q 30).

*Soft-clipped read* From the alignment files we extracted reads, which had a minimum of 30 base pairs properly mapped to the host genome and a minimum of 30 bases that are soft-clipped for lack of alignment. We used a modified version of the Perl script samclip,[43] which we call samclip2_Pfmod and the code for which is available in our GitHub repository. We then converted these bam files to fastq files using bedtools.

*Alignment to viral genome.* We separately aligned the resulting unmapped and soft-clipped reads to a hard-masked version of the AAV construct. The regions hard-masked were pre-identified high sequence similarity regions to the host genome, which include the transgene and portions of the CMV promoter and intron. We aligned the reads using BWA MEM, filtered for proper alignment with samtools view (-F 4 -q 30), removing reads with fewer than 30 bases properly mapped and more than 120 bases clipped using sampclip2_Pfmod. We then removed duplicated reads using samtools fixmate and markdup.

Additionally, we mapped the entire raw fastq files to the viral genome with the same parameters as above to assess the number of reads that map to any viral sequence (not just ones that also map to host genome).

#### *Extracting IS-determining reads*

*Mate reads* We re-sorted the extracted reads, whose mates did not properly align to the host genome. We then parsed the alignment file for reads that did not properly map to the host genome but properly mapped to the viral genome to extract their read names using samtools and a custom bash script. We used the names of these reads to extract their properly aligned mates using a python script.[44] The resulting reads were the most proximate genomic location of an IS with evidence coming from one pair mapped to host genome and one mapped to viral genome.

*Soft-clipped reads* We extracted the names of soft-clipped reads properly aligned to viral genome using the same technique as above and extracted those reads from the alignments to host genome with the same python script as above. Thus, we obtained the location in the host genome where there is evidence of an IS from a read that maps partially to both host and viral genome.

We then merged reads that identified host genomic locations from both sets of evidence (mate and soft-clipped) using samtools merge and re-sorted them by name (Table S4).

The pipeline above was written as a fully executable set of four steps as bash scripts available under 1.1.vi_pipe.Rmd in our GitHub repository.

### Depth of sequencing simulation

To determine to what extent the depth of sequencing is influencing IS identification, we subsampled the FFPE TES samples for the high-dose cyno livers at random to depths between 10% and 90% of the original sample, in intervals of 10% using seqtk.[45] We put the resulting subsampled files through the pipeline described above, treating them as separate unique samples.

### Hotspot identification

To identify regions of the genome that have statistical over-representation of IS compared with random expectation, we adapted a method from Persson et al.[46] In brief, we used the *Z* score threshold method, which tallies IS events in windows and converts their counts into *Z* scores. Then it applies a threshold to identify regions that contain higher than expected by random IS density, given the total number of detected IS. We combined IS from both samples in each dose range to focus on regions that may have shared hotspots of integration across the genome. The code for analysis can be found in 2.3.hotspot_analysis.Rmd in our GitHub repository (Table S5).

### Annotation of IS genomic regions

IS locations reported by each method were analyzed using the annotatePeaks.pl tool from the HOMER (v4.11.1) package.[47] To create the bed file format of IS required by the annotatePeaks.pl tool, chromosome and start locations were extracted from the sam alignment files for each read. The end location was determined by adding the match width from the SAM CIGAR string to the start location. Annotations were performed using the mouse genomics regions for mm10 from the HOMER package.

### Clonality analysis

Potential expansion of individual clones was analysis using the SonicAbundance approach of Sherman et al.[48] Briefly, random shearing of DNA prior to NGS library preparation results in a range of fragment sizes. Fragment size polymorphism at a specific IS is taken as evidence of different cells with the same insertion likely arising from a single progenitor. We utilized samtools and developed an R script to parse the bam files above to identify the specific base-pair genomic location of AAV insertion. The number of differing fragment sizes at each location is counted and plotted (Figure 6). Exact IS locations appearing in independent samples are connected with a dotted line.

## DATA AVAILABILITY

After masking proprietary sequences, remaining raw data have been deposited in SRA submission SRA: PRJNA927280 (ID 927280 - BioProject - NCBI [nih.gov]). Full raw data may be obtained following the establishment of a confidentiality disclosure agreement. Interested parties may reach out to the corresponding author for more details. We acknowledge that reproduction of our analysis will be impossible with the limited dataset we have provided. We apologize for this limitation of our ability to share our primary data and hope that our findings can still inform interested parties in their decision on molecular methods for IS identification.

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtm.2023.04.009.

## ACKNOWLEDGMENTS

The authors would like to acknowledge Jon Cook and Matt Martin for support and scientific input.

## AUTHOR CONTRIBUTIONS

E.O. led the data integration and analysis and co-wrote the manuscript. S.K. and J.Q. performed the integration experiments. M.G. contributed to bioinformatic analysis, and M.S. performed QC and updated the code. D.R. developed positive control material. L.N. performed biodistribution analysis on samples. L.W. contributed to conceptualization and project administration. T.L. led the team and wrote the manuscript.

## DECLARATION OF INTERESTS

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

## REFERENCES

1. Hacein-Bey-Abina, S., Garrigue, A., Wang, G.P., Soulier, J., Lim, A., Morillon, E., Clappier, E., Caccavelli, L., Delabesse, E., Beldjord, K., et al. (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J. Clin. Invest. *118*, 3132–3142. https://doi.org/10.1172/JCI35700.

2. 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.

3. Howe, S.J., Mansour, M.R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., Brugman, M.H., Pike-Overzet, K., Chatters, S.J., de Ridder, D., et al. (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J. Clin. Invest. *118*, 3143–3150. https://doi.org/10.1172/JCI35798.

4. 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.

5. Nault, J.C., Datta, S., Imbeaud, S., Franconi, A., Mallet, M., Couchy, G., Letouzé, E., Pilati, C., Verret, B., Blanc, J.F., et al. (2015). Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. Nat. Genet. *47*, 1187–1193. https://doi.org/10.1038/ng.3389.

6. 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.

7. Chandler, R.J., LaFave, M.C., Varshney, G.K., Trivedi, N.S., Carrillo-Carrasco, N., Senac, J.S., Wu, W., Hoffmann, V., Elkahloun, A.G., Burgess, S.M., and Venditti, C.P. (2015). Vector design influences hepatic genotoxicity after adeno-associated virus gene therapy. J. Clin. Invest. *125*, 870–880. https://doi.org/10.1172/JCI79213.

8. 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.

9. 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.

10. Everett, J.K., Nguyen, G.N., Raymond, H., Roche, A., Kafle, S., Wood, C., Leiby, J., Merricks, E.P., Kazazian, H.H., Nichols, T.C., et al. (2020). AAV integration analysis after long term follow up in hemophilia a dogs reveals the genetic consequences of AAV-mediated gene correction. Mol. Ther. *28*. https://doi.org/10.1016/j.ymthe.2020.04.019.

11. Nichols, T.C., Whitford, M.H., Arruda, V.R., Stedman, H.H., Kay, M.A., and High, K.A. (2015). Translational data from adeno-associated virus-mediated gene therapy of hemophilia B in dogs. Hum. Gene Ther. Clin. Dev. *26*, 5–14. https://doi.org/10.1089/humc.2014.153.

12. 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.

13. Ferla, R., Alliegro, M., Dell'Anno, M., Nusco, E., Cullen, J.M., Smith, S.N., Wolfsberg, T.G., O'Donnell, P., Wang, P., Nguyen, A.D., et al. (2021). Low incidence of hepatocellular carcinoma in mice and cats treated with systemic adeno-associated viral vectors. Mol. Ther. Methods Clin. Dev. *20*, 247–257. https://doi.org/10.1016/j.omtm.2020.11.015.

14. 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.

15. Dalwadi, D.A., Torrens, L., Abril-Fornaguera, J., Pinyol, R., Willoughby, C., Posey, J., Llovet, J.M., Lanciault, C., Russell, D.W., Grompe, M., and Naugler, W.E. (2021).

Liver injury increases the incidence of HCC following AAV gene therapy in mice. Mol. Ther. *29*, 680–690. https://doi.org/10.1016/j.ymthe.2020.10.018.

16. Guidance for Industry. Long Term Follow-Up After Administration of Human Gene Therapy Products. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/long-term-follow-after-administration-human-gene-therapy-products.
17. Brady, T., Roth, S.L., Malani, N., Wang, G.P., Berry, C.C., Leboulch, P., Hacein-Bey-Abina, S., Cavazzana-Calvo, M., Papapetrou, E.P., Sadelain, M., et al. (2011). A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. Nucleic Acids Res. *39*, e72. https://doi.org/10.1093/nar/gkr140.
18. Schmidt, M., Hoffmann, G., Wissler, M., Lemke, N., Müssig, A., Glimm, H., Williams, D.A., Ragg, S., Hesemann, C.U., and von Kalle, C. (2001). Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. Hum. Gene Ther. *12*, 743–749. https://doi.org/10.1089/104303401750148649.
19. Schmidt, M., Schwarzwaelder, K., Bartholomae, C., Zaoui, K., Ball, C., Pilz, I., Braun, S., Glimm, H., and von Kalle, C. (2007). High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat. Methods *4*, 1051–1057. https://doi.org/10.1038/nmeth1103.
20. 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.
21. Miyazato, P., Katsuya, H., Fukuda, A., Uchiyama, Y., Matsuo, M., Tokunaga, M., Hino, S., Nakao, M., and Satou, Y. (2016). Application of targeted enrichment to next-generation sequencing of retroviruses integrated into the host human genome. Sci. Rep. *6*, 28324. https://doi.org/10.1038/srep28324.
22. Sullivan, L., Zahn, M., Gil Farina, I., Kasprzyk, T., O'Neill, C.A., Eggan, K., Zoog, S.J., Veres, G., Schmidt, M., and Vettermann, C. (2021). Rare genomic integrations of AAV5-hFVIII-SQ occur without evidence of clonal activation or gene-specific targeting. Mol. Ther. *29*, 425. (ASGCT Annual Meeting Abstracts). https://doi.org/10.1016/j.ymthe.2021.04.019.
23. McEllin, B., Searle, B.C., DePledge, L., Sun, G., Cobbs, C., and Karimi, M. (2021). Detection of human papillomavirus integration in brain metastases from oropharyngeal tumors by targeted sequencing. Viruses *13*, 1536. https://doi.org/10.3390/v13081536.
24. Nguyen, N.P.D., Deshpande, V., Luebeck, J., Mischel, P.S., and Bafna, V. (2018). ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. Nucleic Acids Res. *46*, 3309–3325. https://doi.org/10.1093/nar/gky180.
25. Chandler, R.J., Mullikin, J.C., Program, N.C.S., and Venditti, C.P. (2021). DNA sequence analysis of recombinant adeno-associated viral integrations events recovered from hepatocellular carcinomas in mice reveals enhancer insertion as the mechanism of vector genotoxicity. Mol. Ther. *29*, 131. (ASGCT Annual Meeting Abstracts). https://doi.org/10.1016/j.ymthe.2021.04.019.
26. Cogné, B., Snyder, R., Lindenbaum, P., Dupont, J.B., Redon, R., Moullier, P., and Leger, A. (2014). NGS library preparation may generate artifactual integration sites of AAV vectors. Nat. Med. *20*, 577–578. https://doi.org/10.1038/nm.3578.
27. Weitzman, M.D., Kyöstiö, S.R., Kotin, R.M., and Owens, R.A. (1994). Adeno-associated virus (AAV) Rep proteins mediate complex formation between AAV DNA and its integration site in human DNA. Proc. Natl. Acad. Sci. USA *91*, 5808–5812. https://doi.org/10.1073/pnas.91.13.5808.
28. Hüser, D., Gogol-Döring, A., Chen, W., and Heilbronn, R. (2014). Adeno-associated virus type 2 wild-type and vector-mediated genomic integration profiles of human diploid fibroblasts analyzed by third-generation PacBio DNA sequencing. J. Virol. *88*, 11253–11263. https://doi.org/10.1128/JVI.01356-14.
29. Chanda, D., Hensel, J.A., Higgs, J.T., Grover, R., Kaza, N., and Ponnazhagan, S. (2017). Effects of cellular methylation on transgene expression and site-specific integration of adeno-associated virus. Genes *8*, 232. https://doi.org/10.3390/genes8090232.
30. Miller, D.G., Petek, L.M., and Russell, D.W. (2004). Adeno-associated virus vectors integrate at chromosome breakage sites. Nat. Genet. *36*, 767–773. https://doi.org/10.1038/ng1380.
31. Hanlon, K.S., Kleinstiver, B.P., Garcia, S.P., Zaborowski, M.P., Volak, A., Spirig, S.E., Muller, A., Sousa, A.A., Tsai, S.Q., Bengtsson, N.E., et al. (2019). High levels of AAV vector integration into CRISPR-induced DNA breaks. Nat. Commun. *10*, 4439. https://doi.org/10.1038/s41467-019-12449-2.
32. He, X., Urip, B.A., Zhang, Z., Ngan, C.C., and Feng, B. (2021). Evolving AAV-delivered therapeutics towards ultimate cures. J. Mol. Med. *99*, 593–617. https://doi.org/10.1007/s00109-020-02034-2.
33. Nakai, H., Wu, X., Fuess, S., Storm, T.A., Munroe, D., Montini, E., Burgess, S.M., Grompe, M., and Kay, M.A. (2005). Large-scale molecular characterization of adeno-associated virus vector integration in mouse liver. J. Virol. *79*, 3606–3614. https://doi.org/10.1128/JVI.79.6.3606-3614.2005.
34. Brunner, S.F., Roberts, N.D., Wylie, L.A., Moore, L., Aitken, S.J., Davies, S.E., Sanders, M.A., Ellis, P., Alder, C., Hooks, Y., et al. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature *574*, 538–542. https://doi.org/10.1038/s41586-019-1670-9.
35. Passman, A., Haughey, M., Carlotti, E., Williams, M., Cereser, B., Lin, M., Devkumar, S., Gabriel, J., Russo, F., Hoare, M., et al. (2021). Clonal dynamics of normal hepatocyte expansions in homeostatic human livers and their association with the biliary epithelium. Preprint at bioRxiv. https://doi.org/10.1101/2021.07.02.450704.
36. Donne, R., Saroul-Aïnama, M., Cordier, P., Celton-Morizur, S., and Desdouets, C. (2020). Polyploidy in liver development, homeostasis and disease. Nat. Rev. Gastroenterol. Hepatol. *17*, 391–405. https://doi.org/10.1038/s41575-020-0284-x.
37. Huichalaf, C., Perfitt, T.L., Kuperman, A., Gooch, R., Kovi, R.C., Brenneman, K.A., Chen, X., Hirenallur-Shanthappa, D., Ma, T., Assaf, B.T., et al. (2022). In vivo overexpression of frataxin causes toxicity mediated by iron-sulfur cluster deficiency. Mol. Ther. Methods Clin. Dev. *24*, 367–378. https://doi.org/10.1016/j.omtm.2022.02.002.
38. Cai, W., Nunziata, S., Rascoe, J., and Stulberg, M.J. (2019). SureSelect targeted enrichment, a new cost effective method for the whole genome sequencing of Candidatus Liberibacter asiaticus. Sci. Rep. *9*, 18962. https://doi.org/10.1038/s41598-019-55144-4.
39. 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.
40. 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.
41. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. next generation sequencing. EMBnet. j. *17*, 10. https://doi.org/10.14806/ej.17.1.200.
42. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
43. Seeman, T.. samclip: filter SAM file for soft and hard clipped alignments. https://github.com/tseemann/samclip.
44. Stuart, T.. extract_reads.py. https://gist.github.com/timoast/2264a79f93b3f1cb3aac.
45. Li, H.. seqtk. https://github.com/lh3/seqtk.
46. Presson, A.P., Kim, N., Xiaofei, Y., Chen, I.S., and Kim, S. (2011). Methodology and software to detect viral integration site hot-spots. BMC Bioinf. *12*, 367. https://doi.org/10.1186/1471-2105-12-367.
47. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.
48. Sherman, E., Nobles, C., Berry, C.C., Six, E., Wu, Y., Dryga, A., Malani, N., Male, F., Reddy, S., Bailey, A., et al. (2017). INSPIIRED: a pipeline for quantitative analysis of sites of new DNA integration in cellular genomes. Mol. Ther. Methods Clin. Dev. *4*, 39–49. https://doi.org/10.1016/j.omtm.2016.11.002.