Molecular Therapy
Methods & Clinical Development
Original Article

The longitudinal kinetics of AAV5 vector integration profiles and evaluation of clonal expansion in mice

Ashrafali Mohamed Ismail,[1] Evan Witt,[1] Taren Bouwman,[1] Wyatt Clark,[1] Bridget Yates,[1] Matteo Franco,[2,3] and Sylvia Fong[1]

[1]BioMarin Pharmaceutical Inc., Novato, CA 94949, USA; [2]ProtaGene CGT GmbH, Heidelberg 69120, Germany; [3]ProtaGene Inc., Burlington, MA 01803, USA

Adeno-associated virus (AAV)-based vectors are used clinically for gene transfer and persist as extrachromosomal episomes. A small fraction of vector genomes integrate into the host genome, but the theoretical risk of tumorigenesis depends on vector regulatory features. A mouse model was used to investigate integration profiles of an AAV serotype 5 (AAV5) vector produced using *Sf* and HEK293 cells that mimic key features of valoctocogene roxaparvovec (AAV5-hFVIII-SQ), a gene therapy for severe hemophilia A. The majority (95%) of vector genome reads were derived from episomes, and mean (± standard deviation) integration frequency was 2.70 ± 1.26 and 1.79 ± 0.86 integrations per 1,000 cells for *Sf*- and HEK293-produced vector. Longitudinal integration analysis suggested integrations occur primarily within 1 week, at low frequency, and their abundance was stable over time. Integration profiles were polyclonal and randomly distributed. No major differences in integration profiles were observed for either vector production platform, and no integrations were associated with clonal expansion. Integrations were enriched near transcription start sites of genes highly expressed in the liver ($p = 1 \times 10^{-4}$) and less enriched for genes of lower expression. We found no evidence of tumorigenesis or fibrosis caused by the vector integrations.

INTRODUCTION

Wild-type adeno-associated viruses (AAVs) are small, single-stranded DNA parvoviruses and are considered relatively nonpathogenic.[1] Recombinant AAV vectors are derived from multiple serotypes and widely used as therapeutic agents for the delivery of transgenes,[2] in part because these vectors efficiently transduce mammalian cells and have relatively low immunogenicity.[2,3] Furthermore, AAV vectors can exhibit distinct tissue and organ tropism as there are more than 20 different identified glycan receptors for AAV serotypes.[4] AAV vectors for gene therapy are frequently manufactured using either human embryonic kidney 293 (HEK293) cell or *Spodoptera frugiperda* (*Sf*) insect cell systems.[5–8] A long-term study characterizing transgene expression from HEK293- and *Sf*-produced AAV serotype 5 (AAV5) vectors found small differences in the short-term kinetics of vector expression, but vectors produced by both manufacturing systems demonstrated a similar long-term expression profile.[9] Valoctocogene roxaparvovec utilizes a replication-incompetent AAV5 vector to deliver a B-domain-deleted factor VIII (FVIII) coding sequence regulated by a hepatocyte-selective promoter[10] and is produced using the *Sf* cell manufacturing system.[11] A single infusion of valoctocogene roxaparvovec provided therapeutic levels of endogenous FVIII expression and protection from bleeding in male participants with severe hemophilia A.[12,13] This durable hemostatic benefit was observed for up to 6 years in a phase 1/2 trial (NCT02576795) and for up to 3 years in a phase 3 trial (NCT03370913).[12–14]

The circularization and concatemerization of AAV vectors associated with episome formation are responsible for stabilizing the vector genome and support gene expression from the vector.[2,15] Whereas the majority of AAV vectors persist in host cells as episomes, a minor proportion integrate into the host genome[16] and may contribute to long-term transgene expression.[17] The integration of AAV vectors has been observed in rodents,[18–22] dogs,[23,24] nonhuman primates,[17,25–27] and humans,[22,25,28] raising concerns about the potential risk of AAV-mediated oncogenesis.[29] As the frequency of AAV vector integration into host genomes represents only a small proportion of the total administered AAVs, they are generally regarded as safe and effective vectors for gene therapy.[2,15,16] However, some studies in mice suggest there is a risk of insertional mutagenesis that can lead to tumorigenesis after AAV gene therapy. The risk is higher when the AAV vector is administered during the neonatal period or in some adult mouse models of disease, such as chronic inflammation, obesity, and diabetes,[18,30–34] if the 3′ untranslated region of the wild-type AAV2 genome with liver-specific promoter activity is left intact,[35,36] or if high doses of the vector are administered.[18] In contrast, other studies in mice illustrate how the risk of tumorigenesis can be mitigated. For example, the risk of hepatocellular carcinoma (HCC) development following AAV vector gene therapy was lower in adult mice compared with neonatal mice.[19,37] Furthermore, a

Received 28 March 2024; accepted 24 June 2024;
https://doi.org/10.1016/j.omtm.2024.101294.

Correspondence: Sylvia Fong, BioMarin Pharmaceutical Inc., Novato, CA 94949, USA.
E-mail: sylviayufong@gmail.com

large-scale study that explored a variety of AAV vectors for liver-targeted gene transfer in 695 adult mice found no increased risk of tumor formation compared with control animals.[37] Some studies also showed that the choice of transgene or promoter, and not the AAV vector per se, influences the risk of tumorigenesis.[18,38] The AAV vector dose also requires careful consideration, as demonstrated in a humanized liver mouse model of xenographic liver regeneration.[22]

Two studies have reported results on AAV vector-mediated gene therapy in a large animal model of hemophilia A.[23,24] Both studies reported findings on the frequency of AAV vector integration and its impact on clonal expansion. The first, a long-term study in a canine model of hemophilia A, using AAV8 or AAV9 vectors expressing canine FVIII, demonstrated that the AAV vector can integrate near genes associated with cell growth and cancer. These relatively rare integrations were associated with the clonal expansion of cells harboring integrated vectors, but tumor formation was not observed.[23] Of note, both vectors used in the study contained the liver-specific enhancer-promoter element present within the wild-type AAV2 inverted terminal repeat (ITR), and one of the vectors also contained the thyroxine-binding globulin promoter.[23] Each of these regulatory elements was previously associated with HCC development in mouse studies.[18,36] However, a separate independent long-term study also using a canine model of hemophilia A found that the proportion of AAV vector integrations near cancer-associated genes was not statistically different from a simulated random dataset used as a control, and integrations were not observed in the same set of cancer-associated genes. The study also found no clear evidence for dominant clonal evolution.[24,39] The AAV vector construct used in that study did not include the liver-specific enhancer-promoter element in the wild-type AAV2 ITR. However, the relatively small sample size (<10 animals) for both studies should be noted. Therefore, the association between AAV vectors and tumorigenesis remains uncertain, is likely context dependent, and warrants further study.[18]

Previously, we reported on the long-term expression profiles and mechanisms of epigenetic-mediated declines in transgene expression for an AAV5-human alpha-1 anti-trypsin (AAV5-hA1AT) vector that mimics the size and regulatory elements of valoctocogene roxaparvovec.[9] Because wild-type mice do not develop a humoral immune response to hA1AT protein, the hA1AT transgene was used in place of the FVIII coding sequence to support long-term expression of the protein to characterize mechanisms of transgene silencing over time.[9,40–42]

To complement our prior expression studies,[9] here we use an adult wild-type C57BL/6J mouse model to report longitudinal kinetics and integration profiles using the same AAV5 vector produced in two commonly used systems for clinical AAV vector manufacturing: HEK293 and *Sf* cells. Importantly, to increase the translatability of our findings to the GENEr8-1 clinical trial program, the AAV5 vector used in this study retains the same regulatory elements and approximate size of valoctocogene roxaparvovec currently being evaluated in phase 1/2 (NCT02576795) and 3 clinical trials (NCT03370913).

RESULTS

To perform our integration analysis, we used an AAV5-hA1AT vector that mimics design elements of valoctocogene roxaparvovec, as reported previously.[9] The vectors were produced using two independent manufacturing systems, HEK293 and *Sf* cells, and all experiments were performed independently using vectors produced by each manufacturing system. This study included a total of 65 mice, with 30 receiving the *Sf*-produced vector, 30 receiving the HEK293-produced vector, and 5 vehicle controls; liver tissue samples were collected 1, 3, 12, 24, and 57 weeks post-dose.

Estimation of coverage and integration frequency

While the majority of the AAV vector persists in transduced cells in its episomal form,[16] a small fraction of the vector will integrate into the host genome.[18–21] Accordingly, target enrichment sequencing (TES) was used to characterize the frequency of vector integration into mouse liver tissue at 1, 3, 12, 24, and 57 weeks after AAV vector administration (Figure 1A). The number of unique integration sites (ISs) ranged from 307 to 1,941 and 259 to 1,628 per sample for *Sf*- and HEK293 vector-treated mice, respectively. The total number of IS reads detected by TES per sample ranged from 341 to 1,961 and 273 to 1,637 for *Sf*- and HEK293 vector-treated mice, respectively. The total number of IS reads was slightly larger than the number of unique ISs because the total number accounts for ISs detected by multiple sequencing reads (i.e., a cell with a unique IS underwent a round of cellular division and the IS was subsequently detected in each cell). To support detection of the vector and calculation of integration frequencies, two additional sub-genomic regions were analyzed to serve as an internal reference standard for the analysis. To confirm successful capture of these regions, the TES analysis was performed with double capture using RNA baits 120 bp in length to enrich for the vector and for the sub-genomic regions. Based on this analysis and using the sub-genomic regions as an internal standard, the estimated average vector copy number (VCN) ranged from 2.98 to 3.67 vector genomes per cell (vg/cell) for *Sf* vector-treated mice and 2.33 to 3.49 vg/cell for HEK293 vector-treated mice. The total number of unique ISs and the average VCN were then used to approximate the average number of unique ISs per cell and per vector genome. Based on these calculations, the average integration frequency per cell in samples from *Sf* vector-treated mice was 3.62×10^{-3}, 3.42×10^{-3}, 1.92×10^{-3}, 2.31×10^{-3}, and 2.20×10^{-3} ISs/cell at weeks 1, 3, 12, 24, and 57, respectively. The average integration frequency per cell in samples from HEK293 vector-treated mice was 2.00×10^{-3}, 2.30×10^{-3}, 1.40×10^{-3}, 1.60×10^{-3}, and 1.60×10^{-3} ISs/cell at weeks 1, 3, 12, 24, and 57, respectively (Figure 1B). The integration frequency at each time point was not statistically different between *Sf*- and HEK293 vector-treated mice based on a one-way analysis of variance (ANOVA). When combining all time points, the average (± standard deviation) integration frequency was 2.70 ± 1.26 and 1.79 ± 0.86 integration events per 1,000 cells for samples from *Sf*- and HEK293

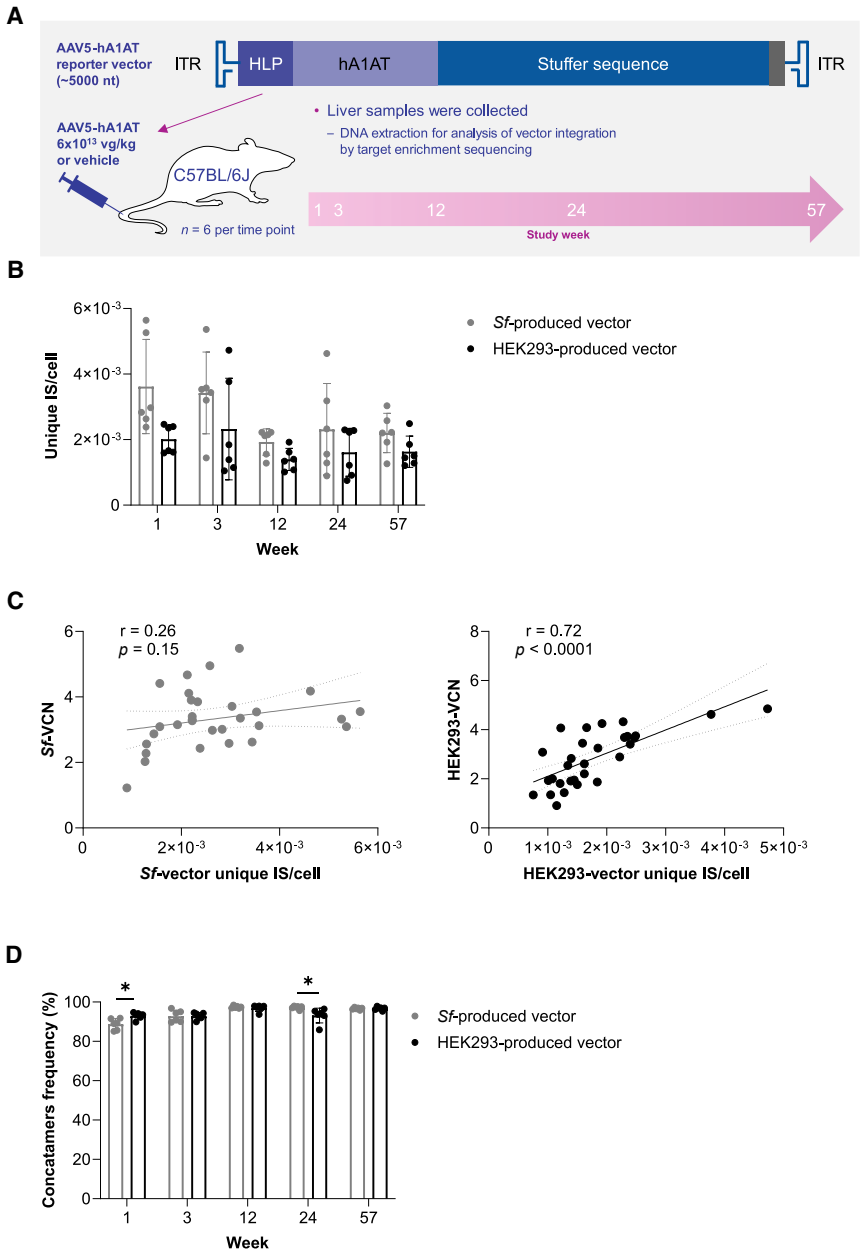

Figure 1. Dynamics of vector integration in long-term vector-treated mice

(A) Study design. (B) Integration frequency in vector-transduced mouse livers determined by TES. (C) Correlation between VCN and integration frequency. (D) Frequency of episomal vector genomes determined by TES. (B and D) Comparisons between *Sf*- and HEK293-produced vector were not significant unless noted. The integration frequency is the measure of vector-host genome reads, and concatamer frequency is the measure of vector-vector genome reads obtained from the TES results. Data are presented as the mean ± SD. $^{*}p < 0.05$ using a one-way ANOVA with Tukey's multiple comparison test. $n = 6$ per time point, with each dot representing a single liver sample from each mouse (average of technical duplicate). AAV5-hA1AT, adeno-associated virus serotype 5-human alpha-1 anti-trypsin; ANOVA, analysis of variance; HEK293, human embryonic kidney 293; HLP, human liver-specific promoter; IS, integration site; ITR, inverted terminal repeat; nt, nucleotide; SD, standard deviation; *Sf*, *Spodoptera frugiperda*; TES, target enrichment sequencing; VCN, vector copy number.

vector-treated mice, respectively. With the average VCN per cell as a reference, the calculated frequency for ISs per vector genome (ISs/vg) from the *Sf*-produced vector would then be 1.19×10^{-3}, 1.11×10^{-3}, 5.40×10^{-4}, 6.71×10^{-4}, and 6.30×10^{-4} ISs/vg at weeks 1, 3, 12, 24, and 57, respectively. For HEK293-produced vector, the calculated frequency for ISs/vg would be 5.82×10^{-4}, 5.44×10^{-4}, 7.29×10^{-4}, 9.22×10^{-4}, and 5.13×10^{-4} ISs/vg at weeks 1, 3, 12, 24, and 57, respectively. Whereas there was a strong correlation between VCN and the number of unique ISs/cell for HEK293-produced vector ($p \leq 0.0001$), this relationship was not observed for *Sf*-produced vector ($p = 0.15$; Figure 1C).

Importantly, the measurement of VCN does not distinguish between episomal and integrated forms of the vector. Therefore, we used the TES results to estimate vector genomes present as episomes (approximated as reads that contained a junction between two vector fragments) vs. vector genomes that integrated into the host genome (a vector fragment followed by a host genome sequence). For vector produced from either manufacturing system, on average 95% (range, 85.2%–98.5%) of the vector genomes persisted in the episomal form throughout the 57-week observation period (Figure 1D), and the circular episomes, measured by droplet digital polymerase chain reaction (ddPCR), were detected as early as 1 week post-dose (Figure S1), consistent with our previously published findings.[9,43,44] Collectively, these results suggest that, for an AAV5-based vector produced with either the *Sf* or HEK293 manufacturing systems, vector integrations occur primarily within the first week after vector administration and the frequency of vector integration remains stable for up to 1 year.

Characterizing individual IS frequency

From the TES results, we can also estimate clonal abundance or the frequency of each unique integration event. The relative frequency of ISs was measured by determining the number of reads corresponding to each individual IS divided by the total number of IS reads in the same sample. When comparing the samples collected at 1 and 57 weeks post-dose, the integration profiles showed no evidence of

clonal expansion, as the top 10 most abundant ISs were still detected by only 1 to 4 sequence reads regardless of vector manufacturing platform (Tables S1 and S2). Similar results were also observed at weeks 3, 12, and 24 (Tables S3 and S4). A site in the gene *chromodomain helicase DNA binding protein 1* (*Chd1*) was the IS detected with the highest read count. *Chd1* corresponded to 1.25%, or 6 total reads, of all the IS sequence counts detected in sample 6 at the 24-week time point (Table S3). However, aside from *Chd1*, across all samples, the remainder of the top 10 contributing ISs all had sequence reads of 4 or less, and the maximum number of sequence reads did not increase with time up to 57 weeks post-dose. These sequencing results suggested a polyclonal integration profile, in the absence of clonal expansion.

Evaluation of polyclonality

The Polyclonal-Monoclonal Distance (PMD) tool[45] was used to further characterize the possibility of clonal enrichment after AAV5-hA1AT treatment. This tool creates a relative profile for each sample based on the distribution of sequencing reads across detectable ISs. Clonality is then estimated by measuring the distance between richness (represented by the total number of ISs within a sample) and evenness (represented by the relative abundance, or frequency, of each IS). The *Sf*- and HEK293 vector-treated samples, regardless of the time post-dose, clustered in the polyclonal region (Figures 2A and 2B). Samples present in the polyclonal region of the graph reflect the minor occurrence of some ISs that are detected by multiple read counts, as seen in the TES analysis above. The PMD indices were consistent across biological replicates and did not change with time.

Histology

The PMD analysis suggested there were no clonal expansions associated with vector integration at a molecular level, and these results were also examined histologically. Liver samples were taken at 57 weeks post-dose and prepared with hematoxylin and eosin for histopathological analysis. No signs of tumors were observed in the livers of *Sf*- or HEK293 vector-treated mice. In addition, there were no signs of dysplasia, inflammation, fibrosis, or cellular stress detected. However, varying levels of steatosis were detected in both the vehicle-, *Sf*-, and HEK293 vector-treated groups (Figure S2). These histologic results suggest a lack of negative effects on the livers of vector-treated mice and confirmed the absence of clonal outgrowth presented in the PMD analysis.

Common IS analysis

Multiple preclinical and clinical studies have documented vector insertions into the host genome.[47–49] This includes clusters of integration in small genomic regions, defined as common ISs (CISs), or vector integration hotspots. We used CIS analysis to identify IS accumulations that are statistically unlikely to occur by chance. This analysis applies a CIS order to reflect the total number of unique ISs across a 50-kb genomic window or CIS region. When the CIS order is <5, the vector integration into that location has an increased probability of occurring by chance and is thus less likely to have any biological significance.[50] When the CIS order is ≥ 5, the number of vector integrations in the CIS region is greater than would be expected by chance. From all the *Sf* vector-treated mice samples combined ($N = 30$; 5 time points), 19,273 ISs (69.3%) were detected in a total of 6,533 CISs with an order ranging from 2 to 37. For the ISs detected in CISs, approximately 88.3% had an order <5 and only 11.7% had a CIS order ≥ 5. From all the HEK293 vector-treated mice samples combined ($N = 30$; 5 time points), 10,376 ISs (56.0%) were detected in a total of 3,998 CISs with an order ranging from 2 to 25. For the ISs detected in CISs, approximately 94.8% had an order <5 and only 5.2% had a CIS order ≥ 5. The relatively minor proportion of CISs with an order ≥ 5 indicates the vector produced from either manufacturing system has a poor ability to target specific genomic regions. Furthermore, the low degree of recurring integrations into specific genomic regions demonstrates a mostly random IS distribution pattern across the host genome, and the majority of the top 10 CISs in all protein-coding genes (Table 1) occurred near highly expressed genes in the liver (Figure S3).

Mapping ISs near genomic features

Whereas the CIS analysis did not identify integration hotspots, the majority of the top 10 CISs were still in highly expressed genes in the liver, suggesting that genomic characteristics could influence where integrations are more likely to occur. Therefore, we used assay for transposase-accessible chromatin using sequencing (ATAC-seq) analysis to determine if integrations were enriched in regions of open chromatin. To support interpretation of the results, two different methods were used to analyze the ATAC-seq data that rely on overlap simulations and distance simulations. The overlap simulations identify the proportion of ISs near genomic features, such as open chromatin, and estimate how much more frequently the ISs appear next to the genomic feature compared with a randomly generated dataset (i.e., chance). The distance simulations support the overlap simulations by then showing how much closer (i.e., the genomic distance in base pairs) the observed ISs appear to the genomic feature vs. a randomly generated dataset.

In brief, the overlap simulations show a distribution of simulated proportions for a sample of ISs the size of the observed dataset (i.e., the total number of simulated ISs per sample is equal to the observed ISs for the same sample). Across all vector-treated samples, the proportion of ISs present within regions of open chromatin was approximately 1.9-fold greater than would be expected by chance for vector produced by both manufacturing platforms ($q < 0.05$; Figure 3A). This apparent affinity for regions of open chromatin was observed regardless of the open chromatin's proximity to a protein-coding gene's transcription start site (TSS) (Figure 3B).

The second method, using distance simulations, compared the distribution of distances from an IS to the closest genomic feature for the observed ISs and a simulated dataset. When combining all vector-treated samples, the median distance from an IS to a region of open chromatin was 2,935 bp for the observed ISs and 6,080 bp for the simulated ISs ($p \leq 0.0001$; Figure S4A). Nearly identical results were found regardless of the vector manufacturing platform. The

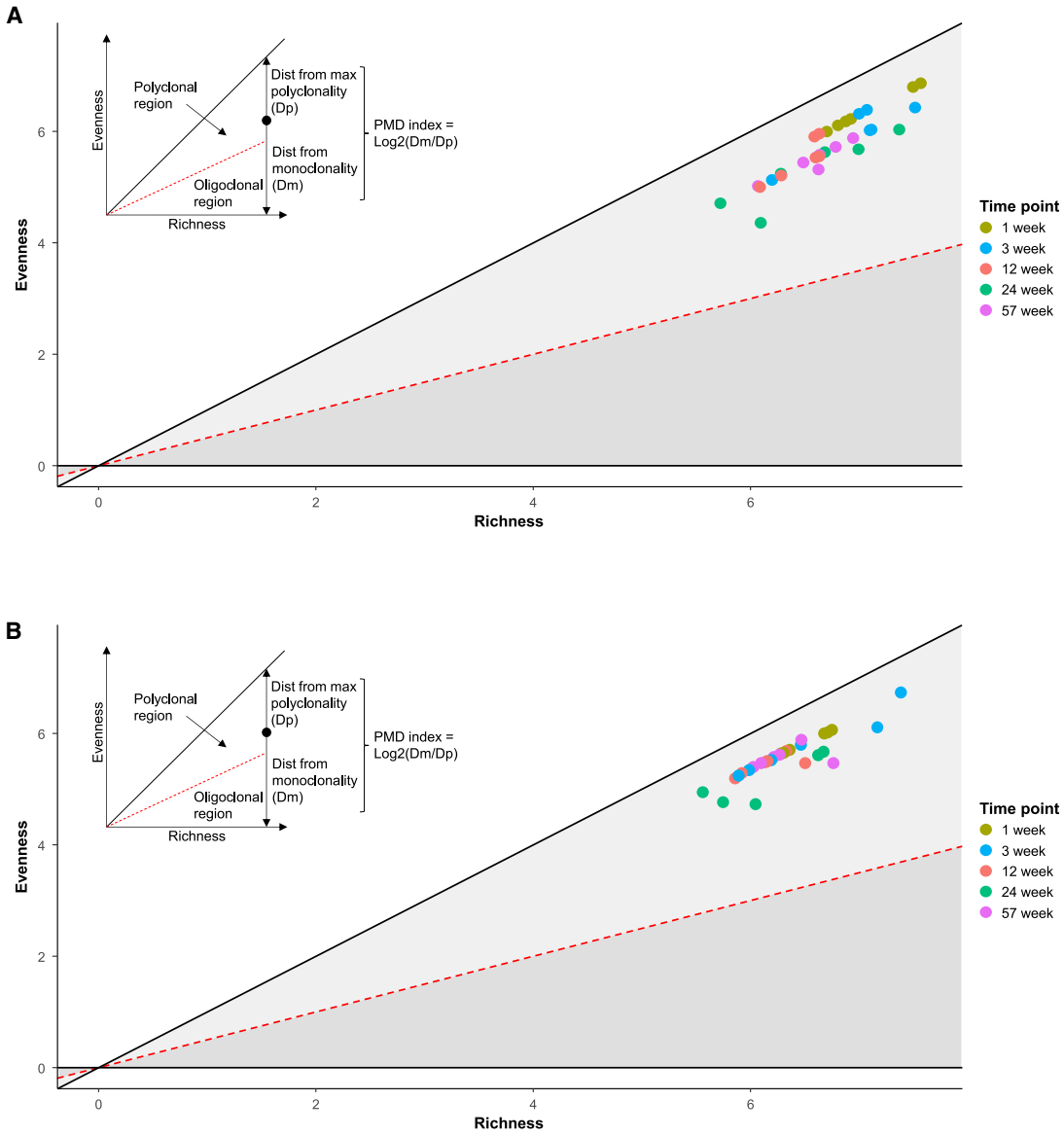

Figure 2. PMD clonal framework-based evaluation of clonality

(A) *Sf*- and (B) HEK293 vector-treated samples. Based on the properties of Rényi entropies,[46] a clonality plane was constructed based on two extreme components of diversity: richness (the total number of integration sites) and evenness (the relative number of reads of each integration site). Richness represents the upper bound for evenness, meaning that evenness can never be higher than richness. For this reason, when evenness equals richness, the sample is considered perfectly polyclonal, and it will sit on the line that bisects the first quadrant. On the other hand, the lower the evenness, the closer the sample will be to the x axis. A monoclonal sample is defined by having an evenness of 0. The PMD index reports the ratio of the distance from the theoretical threshold for maximal polyclonality and monoclonality (see figure insert on top-left side of the figure for further details).[45,] Dist, distance; Dm, distance from monoclonality; Dp, distance from maximum polyclonality; HEK293, human embryonic kidney 293; max, maximum; PMD, polyclonal-monoclonal distance; *Sf*, *Spodoptera frugiperda*.

median distance from an IS to a region of open chromatin for *Sf* vector-treated samples was 2,953 and 6,123 bp for observed and simulated ISs, respectively (Figure S4B). For the HEK293 vector-treated samples, the median distances for the observed and simulated ISs were 2,913 and 6,052 bp, respectively (Figure S4C). The small bias for vector integration near regions of open chromatin was also similar for vector produced from either manufacturing platform regardless of the open chromatin's proximity to a TSS (Figure S5) and the time period or size of the genomic window used for the analysis (Figure S6). Collectively, these data suggest that ISs are enriched near regions of open chromatin at frequencies greater than would be expected by chance. Furthermore, the degree of enrichment near regions of open chromatin was similar for both *Sf*- and HEK293-produced vector.

Table 1. Top 10 CISs detected in liver samples from *Sf*- and HEK293 vector-treated mice

CIS rank	CIS order	Chr	Average position	Dimension (nt)	Gene	Normalized entropy	Contributing samples[a]
Sf-produced vector							
Top 1	33	17	39845964	5,609	*Rn45s*	0.895	24 of 30
Top 2	27	18	12679116	126,239	*Cabyr/Ttc39c*	0.801	17 of 30
Top 3	23	5	90470587	53,459	*Alb/Afp*	0.749	14 of 30
Top 4	19	2	98664806	5,061	*Lrrc4c*	0.694	12 of 30
Top 5	19	9	46037434	338,841	*Sik3*	0.670	11 of 30
Top 6	17	9	121913038	38,574	*1700048O20Rik*	0.671	11 of 30
Top 7	17	14	31127780	373,509	*Dnah1*	0.761	14 of 30
Top 8	15	1	67200726	210,088	*Cps1*	0.613	9 of 30
Top 9	15	2	26486672	274,663	*Notch1*	0.715	12 of 30
Top 10	14	1	88231197	142,390	*Trpm8*	0.659	10 of 30
HEK293-produced vector							
Top 1	25	18	12738790	218,322	*Cabyr/Ttc39c*	0.722	13 of 30
Top 2	18	5	90505939	85,066	*Alb/Afp*	0.705	12 of 30
Top 3	14	17	39845783	5,351	*Rn45s*	0.659	10 of 30
Top 4	12	17	25867962	221,709	*Rab40c*	0.663	10 of 30
Top 5	11	11	16839959	161,289	*Egfr*	0.668	10 of 30
Top 6	11	4	46015636	182,868	*Tmod1*	0.631	9 of 30
Top 7	11	16	14120571	258,581	*Myh11*	0.668	10 of 30
Top 8	10	5	50017197	102,688	*Gpr125*	0.636	9 of 30
Top 9	10	5	125308660	159,321	*Scarb1*	0.677	10 of 30
Top 10	10	11	117080443	197,728	*Mgat5b*	0.580	8 of 30

The CIS order reflects the total number of unique ISs across a 50-kb genomic window or CIS region. When the CIS order is ≥ 5, the number of vector integrations in the CIS region is greater than would be expected by chance. Dimension is the distance between the most proximal and distal ISs within a specific CIS region. For normalized entropy, values of 0 indicate that there is a single contributing sample to that CIS region, and values close to 1 indicate that all the different samples contributed equally.
Chr, chromosome; CIS, common integration site; HEK293, human embryonic kidney 293; IS, integration site; nt, nucleotide; *Sf*, *Spodoptera frugiperda*.
[a]Not including vehicle-treated controls.

Mapping ISs near cancer-associated genes

To determine if ISs also cluster around genomic regions near cancer-associated genes, we evaluated the frequency of ISs present within 100 kb of a TSS of cancer-associated genes as identified in the Catalogue of Somatic Mutations in Cancer (COSMIC)-annotated database (v.90), also known as the Cancer Gene Census database, maintained by the Wellcome Sanger Institute.[51] Since there is no cancer gene database for mice, the IS positions from the mouse genome were lifted over to the human genome so that the ISs from mice could be analyzed for their proximity to homologous human genes within the human Cancer Gene Census database.

For *Sf* vector-treated mice, 1,231 ISs were found within 100 kb of a TSS for a cancer-associated gene (9.53% of the total 12,919 ISs carried over to the human genome). Similar results were found for HEK293 vector-treated mice with 783 ISs (9.22% of the total 8,849 ISs carried over to the human genome) found within 100 kb of a TSS for a cancer-associated gene. The strongest relative contribution for *Sf* vector-treated mice was found in sample number 2 from the 24-week treatment group for cancer-associated platelet-derived growth factor receptor alpha (*PDGFRA*). Even so, *PDGFRA* constituted only 0.6% of all the IS read counts, which corresponds to two sequence reads from the sample (data not shown; Sequence Read Archive [SRA] accession number SRA: PRJNA1076258). The strongest relative contribution for HEK293 vector-treated mice was found in sample number 6 from the 57-week treatment group for cancer-associated terminal nucleotidyltransferase 5C (*TENT5C*) and U2 small nuclear RNA auxiliary factor 1 (*U2AF1*). Each gene constituted only 0.46% of all IS read counts (corresponding to two sequence reads for each gene, data not shown; SRA accession number SRA: PRJNA1076258). Given that the strongest relative contributions from cancer-associated genes corresponded to detection by only two sequence reads, the results suggest that clonal expansion was not triggered as a result of the vector integrations near the *PDGFRA*, *TENT5C*, or *U2AF1* locus. Of the 6,533 observed CISs in *Sf* vector-treated mice, only nine CISs (0.14% of the total) were located in or near genes associated with severe adverse effects, such as clonal outgrowth, identified in clinical gene therapy studies with retroviral or lentiviral vectors.[52–59] The CIS order for two of the CISs was ≥ 5 (Table 2), and the remaining seven CISs had an order <5 (Table S5). Of the 3,998 CISs observed in HEK293 vector-treated

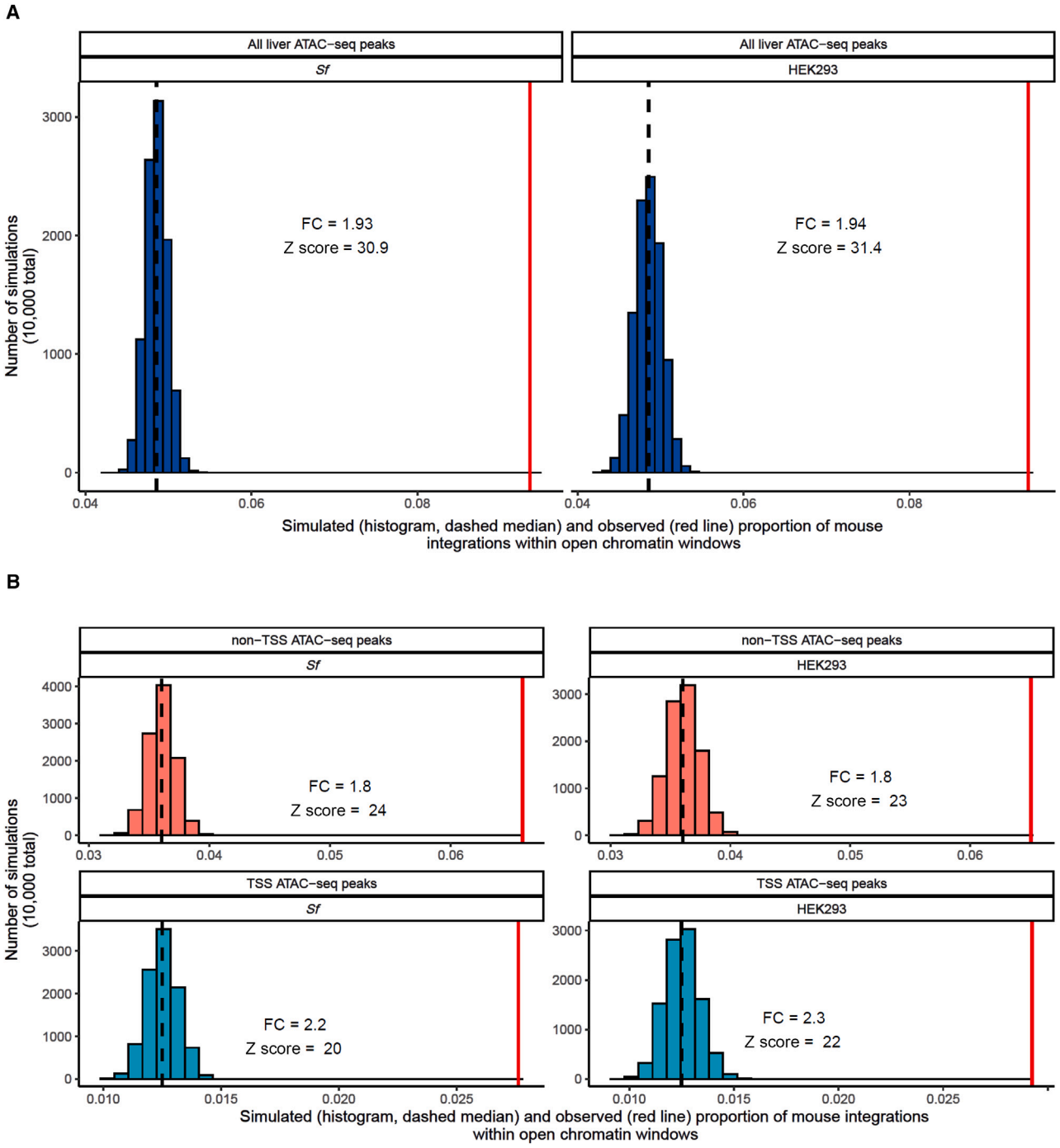

Figure 3. Observed and expected fraction of integrations for *Sf*- and HEK293 vector-treated mice near regions of open chromatin

(A) all regions of open chromatin and (B) open chromatin in close proximity to a TSS. The red line represents the proportion of observed integrations that fall within 20 kb of an ATAC-seq peak, and the histogram represents the distribution that would be expected by chance based on the median simulated value. The observed integrations are a single value because they represent the total number of ISs observed in all samples combined found near open chromatin windows. The simulated values are represented by a histogram to reflect the distribution of each individual simulation from the 10,000 total simulations used in the analysis. ATAC-seq, assay for transposase-accessible chromatin with sequencing; FC, fold change; HEK293, human embryonic kidney 293; IS, integration site; *Sf*, *Spodoptera frugiperda*; TSS, transcription start site.

Table 2. CIS order ≥5 within or near genes associated with severe adverse effects detected in liver samples from *Sf* vector-treated mice

CIS rank	CIS order	Chr	Integration locus	Dimension (nt)	Normalized entropy	Genes
Top 191	7	5	111262381	160,140	0.572	*Mn1*, *Pitpnb*, *Ttc28*
Top 250	6	6	127076511	80,132	0.527	*9330179D12Rik*, *9630033F20Rik*, *Ccnd2*, *Fgf23*

The CIS order reflects the total number of unique ISs across a 50-kb genomic window or CIS region. When the CIS order is ≥ 5, the number of vector integrations in the CIS region is greater than would be expected by chance. Dimension is the distance between the most proximal and distal ISs within a specific CIS region. For normalized entropy, values of 0 indicate that there is a single contributing sample to that CIS region, and values close to 1 indicate that all the different samples contributed equally. There were no CISs ≥ 5 identified within or near genes associated with severe adverse effects in liver samples from HEK293 vector-treated mice.
Chr, chromosome; CIS, common integration site; HEK293, human embryonic kidney 293; IS, integration site; nt, nucleotide; *Sf*, *Spodoptera frugiperda*.

mice, only 3 CISs (0.05% of the total) were in or near genes associated with severe adverse effects, and the CIS order of all the CISs was <5 (Table S5). Since all but two of the CISs had an order <5, the data suggest that the minor proportion of integrations that occurred in or near genes associated with severe adverse effects have an increased probability of occurring by chance.

The following genes were linked to severe adverse effects based on prior studies: *CCND2*, *HMGA2*, *LMO2*, *MECOM*, and *MN1*.[52–59] For vector-treated mice, the observed CISs detected in or near genes linked to severe adverse effects had low numbers of unique ISs regardless of the vector manufacturing platform (Table 3). For example, in the *Sf* vector-treated mice, the CISs with the highest number of combined unique ISs, across all samples ($N = 30$), contained only 7 independent integration events spanning approximately 160 kb. The majority were detected by a single sequence read, suggesting there were no hotspots or clonal expansion associated with these integrations. Furthermore, whereas previous studies have shown the potential for AAV integration into the *Rian* locus to contribute to tumorigenesis,[18,29,30,34] across our entire study, including all samples and time points, there was only a single IS detected in exon 2 of the *Rian* locus 1 week after vector administration in sample 4 from the HEK293 vector-treated mice.

Integration enrichment around TSSs

Previous studies have suggested that the majority of AAV vector integrations occur in transcriptionally active regions and near the TSS of genes.[20,60] Therefore, after examining vector integration within 100 kb of the TSS of cancer-associated genes, we next determined the relative proximity of these integrations within 10 kb of a TSS for both protein-coding and cancer-associated genes. Again, we used overlap simulations as a benchmark for a truly random integration profile (Figure 4A). In *Sf* vector-treated mice, we found that integrations occurred within 10 kb of a TSS more than would be expected by chance (enrichment = 1.5×, $p = 1 \times 10^{-4}$, Z score = 25; Figure 4B), and the fold enrichment of the integrations was greater near genes highly expressed in the liver (enrichment = 1.9×, $p = 1 \times 10^{-4}$, Z score = 16; Figure 4B). Comparable results were observed in HEK293 vector-treated mice, with a 1.5-fold enrichment of ISs within 10 kb of a TSS for all genes ($p = 1 \times 10^{-4}$, Z score = 19; Figure 4C) and a greater enrichment near highly expressed genes (enrichment = 1.8×, $p = 1 \times 10^{-4}$, Z score = 12; Figure 4C). Distance simulations supported these results and found that the median distance from an IS to a TSS observed across all samples (84,109 bp) was approximately 1.8-fold less than would be expected by chance (151,761 bp; Figure S7A) and that this was true for both *Sf* vector-treated mice (Figure S7B) and HEK293 vector-treated mice (Figure S7C). In addition to the TSS analysis, similar trends were also found for IS enrichment within gene bodies (Figure S8), and distance simulations showed that the median distance from an IS to a gene body observed across all samples (36,640 bp) was approximately 2.1-fold less than would be expected by chance (76,848 bp; Figure S9A). This was true for both *Sf* vector-treated mice (Figure S9B) and HEK293 vector-treated mice (Figure S9C).

We then explored integration enrichment across different gene sets. Since a small fraction of integrations did occur near 10 kb of cancer-associated genes and integration enrichment is associated with levels of gene expression, we asked whether cancer-associated genes were significantly more expressed than other protein-coding genes in the liver. Using the Genotype-Tissue Expression (GTEx) project 2017 median tissue transcripts per million dataset,[61] we found a significant difference in the expression level of the Cancer Gene Census database-annotated cancer genes[51] compared with that of non-cancer-associated genes in the liver (Wilcoxon $p < 2.2 \times 10^{-16}$; Figure S10). Therefore, genes were grouped based on their expression patterns from our mouse liver RNA sequencing data and their cancer-associated status according to the Cancer Gene Census database. For *Sf* vector-treated mice, when comparing non-cancer-associated genes and cancer-associated genes, we found that integrations occurred within 10 kb of a TSS more frequently than would be expected by chance for both gene sets (enrichment = 1.5 vs. 1.7, $p = 1 \times 10^{-4}$; Z score = 24 vs. 7.3), and cancer-associated genes showed a larger fold enrichment compared with the median simulated frequency (Figure 4B). Similar results were found for HEK293 vector-treated mice (enrichment = 1.5 vs. 1.6, $p = 1 \times 10^{-4}$; Z score = 19 vs. 4.8; Figure 4C). We then examined whether the magnitude of integration enrichment correlated with gene expression for cancer- and non-cancer-associated gene sets. In brief, the results demonstrated that the integration fold enrichment and Z scores were greater among highly expressed genes compared with those of unexpressed genes in the liver (Figure 4). Of note, when comparing genes with similar expression levels, the Z scores for non-cancer-associated genes were higher than for cancer-associated genes, suggesting that the cancer-associated genes are not a hotspot for vector integration regardless of the vector manufacturing platform. In addition, a significant difference was not observed in Z scores for cancer-associated genes between week 1 and 57 (Figure S11). Integration enrichment near a

Table 3. Individual integrations within or near genes associated with severe adverse effects detected in liver samples from *Sf*- and HEK293 vector-treated mice

Week	Sample	SAE gene	Distance to TSS (nt)	Chr	Integration locus	Relative contribution (%)	Total sequence count	Sequence count contribution
Sf-produced vector								
1	1	*Ccnd2*	14,464	12	4293264	0.121	824	1
1	1	*Hmga2*	2,266	12	65830267	0.121	824	1
1	2	*Mn1*	71,747	22	27873503	0.097	1,033	1
1	3	*Hmga2*	−77,988	12	65746143	0.055	1,833	1
1	3	*Lmo2*	53,054	11	33945130	0.055	1,833	1
1	3	*Mecom*	23,564	3	169292802	0.055	1,833	1
1	5	*Mecom*	−30,991	3	169238247	0.109	920	1
3	4	*Lmo2*	8	11	33870537	0.078	1,280	1
3	3	*Mecom*	−28,837	3	169240401	0.053	1,900	1
3	3	*Mecom*	−52,519	3	169611099	0.053	1,900	1
12	3	*Mecom*	3,543	3	169151277	0.133	752	1
12	5	*Mecom*	−27,072	3	169242166	0.126	792	1
24	3	*Mecom*	−32,831	3	169089841	0.117	854	1
24	5	*Mecom*	58,575	3	169327813	0.059	1,706	1
24	6	*Mecom*	−23,894	3	169639724	0.208	480	1
24	3	*Mn1*	−93,904	22	27697979	0.117	854	1
57	4	*Mn1*	−892	22	27796004	0.120	834	1
57	5	*Ccnd2*	−5,984	12	4267778	0.107	938	1
HEK293-produced vector								
3	4	*Lmo2*	−13,175	11	33878901	0.207	483	1
3	3	*Mecom*	93,059	3	169362297	0.061	1,637	1
3	3	*Mecom*	−85,787	3	169577831	0.061	1,637	1
3	4	*Mecom*	−37,877	3	169625741	0.207	483	1
3	4	*Mecom*	−97,187	3	169025485	0.207	483	1
24	2	*Ccnd2*	26,674	12	4305474	0.114	874	1
24	4	*Ccnd2*	1,882	12	4280682	0.366	273	1
57	1	*Mecom*	898	3	169148632	0.109	920	1

Chr, chromosome; HEK293, human embryonic kidney 293; nt, nucleotide; SAE, severe adverse effect; *Sf*, *Spodoptera frugiperda*; TSS, transcription start site.

TSS did not change in a meaningful way regardless of the size of the genomic window used to define the TSS for *Sf*- or HEK293 vector-treated mice (Tables S6 and S7).

Collectively, our analysis using random simulations demonstrated that integrations occur preferentially near genes in the liver associated with higher expression and open chromatin. However, despite the elevated expression of cancer-associated genes in the mouse liver, we did not observe an integration bias in these cancer-associated genes compared with other protein-coding genes of similar expression levels.

Comparability of vector genome integration features with different manufacturing systems

Since *Sf*-produced vector contains more fragmented vector DNA compared with HEK293-produced vector,[9] the frequency of vector regions that integrate into the host genome was further characterized. For *Sf* vector-treated mice, the vector fragments that integrated into the host genome were derived from regions throughout the vector sequence (mean from all time points: 21.94%, 27.54%, 17.78%, and 32.74% for the promoter, transgene, polyadenylation [poly(A)], and ITR regions, respectively; Figure 5; Table S8). The vector fragment profiles in HEK293 vector-treated mice were not significantly different from *Sf* vector-treated mice based on a one-way ANOVA (mean from all time points: 25.28%, 26.88%, 20.54%, and 27.31% for the promoter, transgene, poly(A), and ITR regions, respectively; Figure 5; Table S9). Thus, despite the previously published differences in the size distribution of *Sf*- vs. HEK293-produced vectors, the profiles and frequency of integrated vector fragments were similar (Figure 5).

DISCUSSION

We present a molecular characterization of vector integration profiles produced with two clinically relevant manufacturing platforms (*Sf*

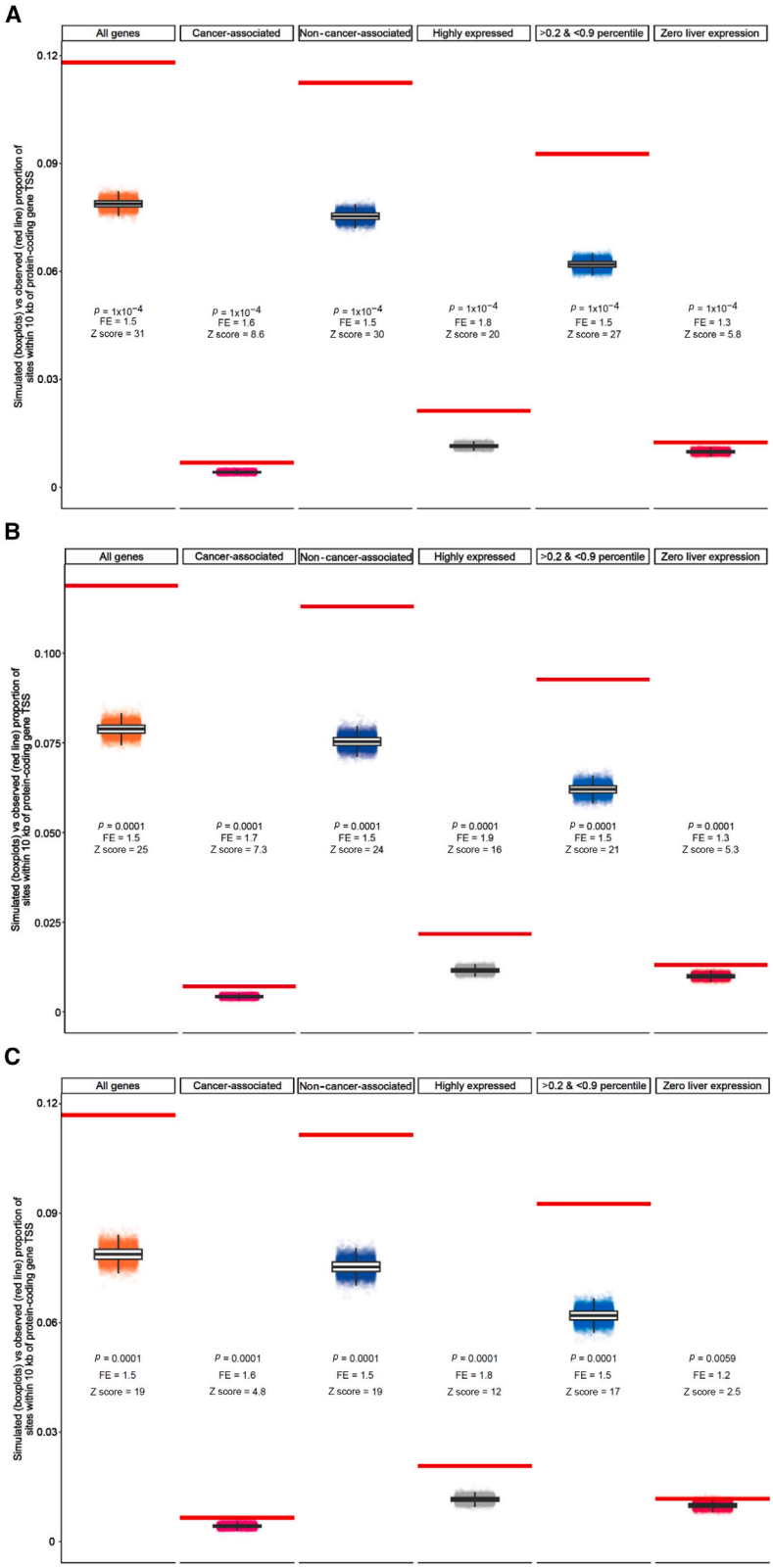

Figure 4. Observed and expected fraction of integrations near genes of different expression levels

(A) all samples combined and for samples from (B) *Sf*- and (C) HEK293 vector-treated mice. The red line represents the proportion of observed integrations that fall within 10 kb of a TSS for a protein-coding gene, and the boxplots represent the distribution of integrations that would be expected by chance based on the median simulated value. Genes with the highest expression profile in the liver (90th percentile) according to the GTEx liver data are classified as highly expressed. *p* values represent 1 minus the percentile of the observed value within the expected distribution, and the z scores were used as a proxy for enrichment. Boxes represent the interquartile range, whiskers represent the range up to 1.5× the interquartile range, black horizontal lines represent the median. FE, fold enrichment; GTEx, genotype-tissue expression; HEK293, human embryonic kidney 293; *Sf*, *Spodoptera frugiperda*; TSS, transcription start site.

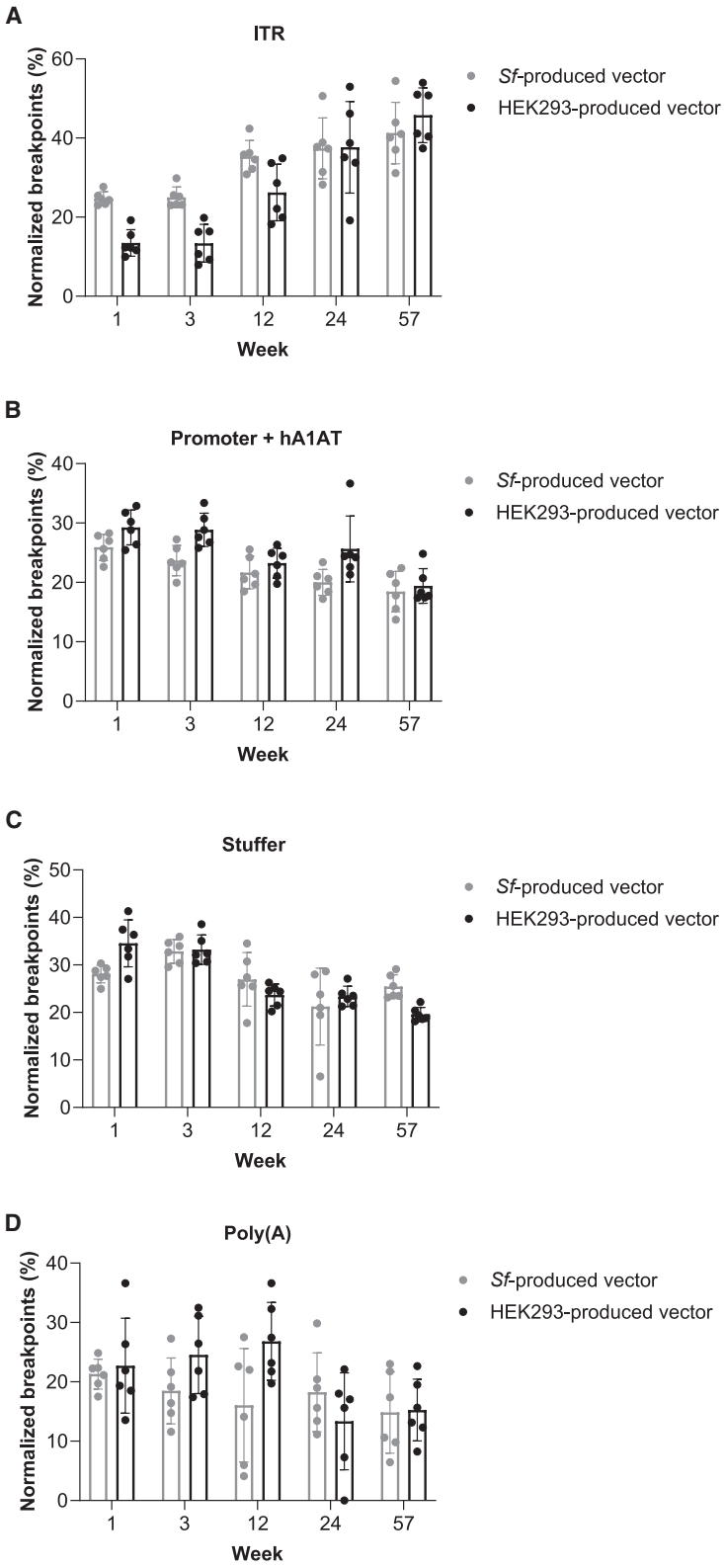

Figure 5. Comparison of integrated vector genome fragments between *Sf*- and HEK293 vector-treated mice

(A) ITR, (B) promoter and hA1AT, (C) stuffer sequence, and (D) poly(A) regions. Data are presented as the mean ± SD with $n = 6$ per time point. Each dot represents a single liver sample from each mouse, and the data were analyzed using a one-way ANOVA with Tukey's multiple comparison test. ANOVA, analysis of variance; hA1AT, human alpha-1 anti-trypsin; HEK293, human embryonic kidney 293; ITR, inverted terminal repeat; poly(A), polyadenylation; SD, standard deviation; *Sf*, *Spodoptera frugiperda*.

and HEK293 cells) using a mouse model and an AAV5 vector that mimics the size and regulatory elements of valoctocogene roxaparvovec, a US Food and Drug Administration- and European Medicines Agency-approved gene therapy for severe hemophilia A.[62,63] In our analysis, we found no adverse events related to treatment with AAV5-hA1AT vector produced by either vector manufacturing system, including no evidence of clonal expansion by molecular or histologic analysis for up to 1 year after vector administration.

Our data demonstrate that the majority of vector genomes persist in their episomal form. While a minor proportion of the vector genomes do integrate into the host genome, they do so with low frequency, within 1 week of vector dosing, and the number of unique integrations remains generally stable over time. The frequency of vector integrations observed in our study with an AAV5 vector produced by *Sf* and HEK293 cells aligns with those reported previously using a similar vector administered in a severe hemophilia A dog model (9.55×10^{-4} ISs/cell by TES and 4.5×10^{-4} ISs/cell by linear amplification-mediated PCR [LAM-PCR])[64] and those reported for 5 clinical trial participants after valoctocogene roxaparvovec infusion (3.97×10^{-3} ISs/cell by TES).[65] Our results are also consistent with the frequency of vector integration detected in nonhuman primate and human livers (2.00×10^{-4} and 1.17×10^{-3} ISs/cell, respectively, by LAM-PCR) after infusion of a recombinant AAV2/5 vector[25] or with rates of integration reported for natural AAV infection and orders of magnitude lower than the rate of somatic mutations reported for humans.[66,67]

Given that linear and circular DNA have different capacities for integration into the host genome and since there are differences in the characteristics of vectors produced by *Sf* and HEK293 cell manufacturing systems, the frequency of vector integration and the corresponding vector fragments that integrate into the host genome were evaluated. For example, *Sf*-produced vectors may contain increased amounts of truncated vector genome fragments compared with HEK293-produced vector.[68] Furthermore, since one manufacturing system is mammalian derived and the other is insect derived, there are also inherent differences in the post-translational modifications of the vector genomes.[69] Despite these differences, the overall number of vector integrations and the integration profiles were similar with AAV vector produced from either manufacturing system. In addition, like the *Sf*-produced vector, the HEK293-produced vector had some overlap in the top 10 CIS genes, such as albumin, likely reflecting the small integration bias observed for highly expressed genes of the liver.

Histology of the livers from vector-treated mice found no evidence of liver tumors. The histology findings are not surprising, given that the molecular analysis also found no evidence of clonal expansion. While the PMD index tool demonstrated the vector-treated samples did not perfectly align with the theoretical maximum for polyclonality, these results are consistent with an expected polyclonal integration profile from liver tissue, where a small subset of cells divide as part of the normal homeostatic process of tissue maintenance.[70] These results would also be expected based on the low-level detection of proliferating/mitotic cells with Ki-67 and phosphorylated histone 3 staining previously observed in the liver.[9]

We did not find substantial differences across time points, biological replicates, or vector manufacturing systems when evaluating vector integration frequencies. Collectively, the integration frequency across all treatment groups highlights the rare nature of these integrating events and are aligned with previously published reports.[23,25,71,72] While we did find integrations near a small fraction of the COSMIC Cancer Gene Census database genes, these integrations were associated with the respective genes' high expression in the liver. The ATAC-seq analysis supported these findings and suggested that vector integrations are enriched in regions of open chromatin. Despite this, vector integrations were not disproportionately enriched near cancer-associated genes when compared with genes of similar expression levels.

The study design presented here differs from some previous reports linking AAV-based gene therapy with the risk of tumorigenesis. However, our goal was not to reproduce those study designs but instead to assess the integration profiles of a vector with clinically relevant regulatory elements at a dose used in clinical trials. Our results presented here are in agreement with other studies suggesting that, with careful consideration of the vector capsid (e.g., AAV serotype and hepatotropism), regulatory elements of the vector (e.g., promoter strength), and the study design (e.g., age at administration and vector dose), AAV vector-mediated gene therapy carries minimal risk of tumorigenesis.[18,19,25,28,37,38] Future studies are needed to evaluate if administering the same AAV5 vector used here in adult mice at higher doses or in neonates increases the risk of tumorigenesis. This will be especially relevant to assess the risk of AAV-based gene therapy in a pediatric population. Importantly, whether gene therapy in humans carries additional risk of tumorigenesis when administered in the context of underlying liver disease, as shown in diabetic and obese mouse models, requires further study.[33,34] Although there is debate whether wild-type AAV oncogenic integration in humans may be associated with the development of a specific subgroup of HCC that occurs in the absence of other etiologies,[35] this association appears linked with AAV2 and, more specifically, a liver-specific promoter found within the 3′ untranslated region of the AAV2 genome.[35,36] However, this region is absent from the AAV5 vector used in our study.

Like valoctocogene roxaparvovec, etranacogene dezaparvovec is an AAV5-based gene therapy with a liver-specific promoter that delivers a factor IX transgene to hepatocytes and was evaluated in the HOPE-B phase 3 clinical trial (NCT03569891) for individuals with hemophilia B.[73] At 1 year of follow-up, a single case of HCC was detected in a trial participant, and it represented the first observed clinical case of HCC after liver-directed AAV-based gene therapy.[73] Examination of the case and the participant's tissue sample by an independent laboratory determined there was no evidence of AAV vector involvement in the HCC.[74] The AAV vector integration events

were rare, with only 0.027% of cells in the sample containing an IS, and there was no evidence of clonal expansion. Furthermore, whole-genome sequencing revealed multiple genetic mutations, independent of the vector insertions, that increase the risk of HCC.[74] The participant also had several risk factors for HCC, such as a history of hepatitis C and hepatitis B, and evidence of non-alcoholic fatty liver disease changes was found in the biopsy.[73]

There are several limitations to the presented analysis. First, a mouse transgene was not used and, therefore, homology-driven effects unique to FVIII cannot be assessed by the hA1AT transgene. We also cannot rule out the possibility that the TES results overestimate the episomal vector frequency due to detection of integrated concatemers or other integrated, recombined forms of the vector that would still appear as vector-vector junctions. In addition, distinct integration patterns in the host genome cannot be ruled out in regions not sampled by the TES analysis. In addition, some studies have found evidence of AAV-based gene therapy-induced HCC risk in mice followed for greater than 1 year.[18,31] Therefore, although we did not find any evidence of clonal expansion at the molecular level, we cannot definitively state that the risk of tumorigenesis in our study would not change over a longer time horizon. In this regard, it is worth noting that we did not observe differences in the rates of clonality between week 1 and 57 post-infusion samples. Finally, this study uses a mouse model to assess dynamics of vector integration. The risk of tumorigenesis in mice appears highly context dependent.[18,30–32,35,36] While this risk has not been observed in larger animal models or humans, there is still a comparatively small amount of data regarding vector integration in humans. Therefore, additional studies from nonhuman primates or biopsy samples from gene therapy participants would help clarify the translatability of risks specific to the vector and study design from the mouse model to human gene therapy. However, it is worth noting that the average vector integration frequency reported in mice for our study (2.70×10^{-3} ISs/cell for *Sf*-produced vector and 1.79×10^{-3} ISs/cell for HEK293-produced vector) is in line with those reported previously and discussed above.[25]

This longitudinal analysis represents an important step forward in evaluating the safety of vector integration following AAV-based gene therapy. This study assesses the risk of clonal outgrowth after AAV vector administration using vector mimicking features from a clinically relevant gene therapy, produced using two different vector production platforms. The integration profiles were examined by TES to characterize, in high resolution, the frequency of vector integration beyond 1 year of follow-up in 60 mice. This approach offers the capacity to detect, with resolution down to the cellular level, clonal outgrowth with high sensitivity.[19,37] The results presented here confirm that AAV5 vectors have low integration rates regardless of the vector production platform. Of the integration events that did occur, most ISs were detected by only one to two sequencing reads, and they were defined by a polyclonal integration profile with no evidence of integration hotspots or clonal expansion, even in animals followed beyond 1 year after administration of *Sf*- and HEK293-produced vector. The low degree of recurring insertions demonstrates that the vector from either production platform has a poor ability to target specific genomic regions.

MATERIALS AND METHODS

AAV reporter vector construction

For this study, a previously described AAV5-hA1AT reporter vector was used instead of valoctocogene roxaparvovec.[9] Since hA1AT is non-immunogenic in wild-type mice, using this alternative reporter enabled long-term studies in immune-competent animals while still maintaining the ITR region and regulatory elements of valoctocogene roxaparvovec.[42] AAV5-hA1AT is a replication-incompetent AAV5 vector containing a single-stranded DNA encoding an hA1AT reporter controlled by a hybrid liver-specific promoter with double-stranded DNA ITRs at both the 5′ and 3′ ends. AAV5-hA1AT also contains a stuffer sequence to match the approximate size of valoctocogene roxaparvovec. The vector was manufactured using HEK293 or *Sf* cell systems. The vector from the HEK293 cell system was manufactured by SAB Tech (Philadelphia, PA), and the vector from the *Sf* cell system was manufactured by BioMarin Pharmaceutical. The *Sf*- and HEK293-produced vectors used here were characterized previously.[9] In brief, the HEK293-produced vector contained higher viral protein 1 (VP1) capsid content and more homogeneous encapsidated DNA compared with *Sf*-produced vector.[9]

Study design

Male 8-week-old C57BL/6J mice were purchased from Jackson Laboratory and used in this study. Mice received a vehicle control or 6×10^{13} vg/kg of either the HEK293- or the *Sf*-produced vector. Mice were separated into cohorts so that one cohort from each treatment group could be euthanized, and tissue was collected at 1 of 5 time points (1, 3, 12, 24, or 57 weeks post-dose). Treated cohorts consisted of 7–10 mice per group, and samples from 6 of the mice were sent for TES analysis. The control cohorts consisted of 5 mice per group. All mouse experiments were performed in accordance with the institutional guidelines under protocols approved by the Institutional Animal Care and Use Committee of the Buck Institute and the BioMarin Animal Resource Committee.

On the day of treatment, mice were weighed and given a single bolus via intravenous injection of the tail vein with the respective treatment using an injection volume of 4 μL/g of body weight. For tissue collection, mice were anesthetized and exsanguinated. The median liver lobe was fixed and processed for histologic analysis. The remaining liver was split into two samples, snap-frozen using liquid nitrogen, and stored at −80°C until all the samples were collected and processed for molecular or biochemical analysis. For each animal, one liver fragment, approximately 10–20 mg in weight, was homogenized using a stainless steel bead and 500 μL RLT buffer (QIAGEN, Hilden, Germany) using a TissueLyser II instrument (QIAGEN). The genomic DNA and total RNA were then extracted from the same homogenate using the DNA/RNA AllPrep kit (QIAGEN) following the manufacturer's instructions.

Measurement of vector genomes in mouse livers by ddPCR

Quantitative measurements to determine the vector genome forms were performed by treating the extracted liver DNA using various combinations of DNA digestion enzymes followed by ddPCR assays and various primer sets using ddPCR methods described previously.[43] In this study, the SQ primer set was utilized to quantify total vector genome copies. Plasmid-Safe ATP-Dependent DNase (PS-DNase; Petaluma, CA) sample treatment hydrolyzes linear DNA (host or vector derived), and only circular episomal forms of the vector remain, allowing quantitation by ddPCR. However, during DNA extraction, some episomal vector DNA can be sheared and linearized, which in turn would also be hydrolyzed by the DNase treatment.[43] Therefore, the fraction of episomal vector genomes could be underestimated. Also, given the presence of concatemeric genome forms of the vector, this approach would underestimate the number of vector genome units (genome units defined as a vector genome containing a promoter, transgene, and poly(A) tail). To enable the detection of concatemeric vector genomes, KpnI restriction enzyme digests were performed. This treatment separates vector genome units within the concatemeric form, enabling a true quantification of vector genome units.

Integration analysis

The integration profile analysis was performed by ProtaGene CGT GmbH (Heidelberg, Germany). In brief, the IS detection was performed using TES and next-generation deep sequencing to identify AAV vector genomes that integrated into the mouse liver tissue. TES was performed by double capture using an RNA bait set 120 bp in length, designed based on 8× tiling. The vector-specific baits were designed to be homologous to the entire vector sequence, and baits targeted to the ITR regions were overrepresented 10-fold compared with the internal vector baits to compensate for the high guanine and cytosine content and secondary structure of the ITR region. Baits for two sub-genomic regions of the mouse genome were designed and included in the bait set (chr7:38,153,607-38,154,606 and chr15:70,650,524-70,651,523). The biotinylated baits were used with magnetic capture to enrich the vector sequences. This method enriches vector-vector as well as vector-genome junctions, and the sub-genomic baits allow estimation of VCN in each sample. Each sample was analyzed in duplicate using 1,000 ng of DNA per replicate. The DNA was sheared to approximately 500 bp in length using an ultrasonicator according to the manufacturer's instructions, and fragment length was verified with TapeStation (Agilent, Santa Clara, CA). Libraries were constructed using the Agilent SureSelect HS2 kit according to the manufacturer's instructions. The libraries were sequenced by 2 × 250-bp symmetric paired-end reads on the Illumina MiSeq platform. For the IS analysis, the TES-derived data were then analyzed using GENE-IS.[75] Raw sequence data were filtered using barcode identity and trimmed according to sequence quality using a Phred score of 20. For each sample, the replicates were analyzed individually. Sequencing reads were aligned to both the vector genome and murine reference genome (mm10) for IS analysis. The number of total, sorted, and IS reads are reported in Tables S10–S12. The data discussed in this article have been deposited in the National Institutes of Health (NIH) SRA accession number SRA: PRJNA1076258.

Vector coverage for each replicate was analyzed using an R script developed by ProtaGene CGT GmbH. The VCN for the samples was estimated using the average vector coverage normalized by the average coverage of the sub-genomic regions. The average number of uniquely mapped ISs per cell was estimated using the number of unique ISs per 1,000 ng of DNA divided by the number of cells in 1,000 ng of DNA. This calculation uses the assumption that 1,000 ng of mouse genomic DNA corresponds to 172,000 cells. The number of unique ISs per vector genome was then estimated by dividing the unique number of ISs by the average VCN per sample.

Diversity measurements to examine clonality

To examine clonality, a clonality plane was constructed based on the two extreme components of diversity, richness and evenness, using a previously published method.[45,46] In brief, richness defines the number of ISs present within a sample, and evenness represents the equal distribution of those ISs within the sample. When considering richness and evenness, there are theoretical maximums for polyclonality and monoclonality. Richness represents the upper bound for evenness, meaning that evenness can never be higher than richness. For this reason, when evenness equals richness, the sample is considered perfectly polyclonal, and it will sit on the line that bisects the first quadrant. On the other hand, the lower the evenness, the closer the sample will be to the x axis. A monoclonal sample is defined by having an evenness of 0. For each sample, the ratio of the distance from the theoretical maximum possible for polyclonality and monoclonality is determined and represents the clonality plane.[46]

Liver histology

Formalin-fixed paraffin-embedded liver sections of 5 μm in thickness were stained with hematoxylin and eosin for histopathological analysis. Slides were digitally scanned and reviewed by a board-certified pathologist.

Common IS analysis

A systems biology framework, based on graphs, was used to identify biologically relevant CISs that were unlikely to occur by chance using a previously published method.[76] In brief, the CISs are represented by graphical networks that are constructed based on the maximal distance between two unique ISs. CISs were defined based on the following steps. Each IS found within a sample was associated with a node that contained the location of the ISs in a graph specific to that sample. If the distance between two nodes within the graph was less than the threshold distance of 50 kb, then the two nodes were connected and considered a CIS.

ATAC-seq

The livers from 40 *Sf*- and HEK293 vector-treated mice at 12, 24, and 57 weeks post-dose, six to seven mice per time point, were used for the ATAC-seq analysis. The data presented in this publication have been deposited in the NIH's SRA accession number SRA: PRJNA1076258.

Regions of open chromatin were identified using ATAC-seq. Reads were aligned to the GRCm38.p6 reference genome using Burrows-Wheeler Alignment-MEM Tool,[77] and narrow peaks were called with MACS2,[78,79] with a q value cutoff of 0.05. Peaks were classified as "TSS" or "non-TSS" based on whether they were within 20 kb of a canonical TSS for a protein-coding gene in the GENCODE M25[80] annotation. The ATAC-seq data were then analyzed using overlap and distance simulations.

In the overlap simulations, for 10,000 iterations, genomic positions equal to the number of detected TES junction sites were sampled. For each sampled site, the distance to the closest ATAC-seq window was found using valR.[81] For each simulation, the median distance from all sites was recorded to create a distribution to compare with the median distance for the observed junction sites. Fold change was defined as the grand median simulated distance divided by the median distance from the observed junctions. This analysis was performed separately for TSS-associated and non-TSS-associated ATAC-seq peaks to assess whether junctions tend to land closer to open chromatin windows with or without associated gene expression. The *p* value is calculated by comparing the single observed proportion of ISs for a sample with the distribution of simulated proportions of ISs by fitting an estimator of the cumulative distribution to the simulated data to determine where the observed value falls within it.

In the distance simulations, 100,000 random genomic positions were selected. For each position, the distance from the site and the closest open chromatin window (in the liver) was calculated with valR::bed_closest(). Then the distribution profile of the 100,000 simulated values was compared with the observed distribution profile. *p* values were calculated using a two-sided Wilcoxon test.

Cancer genes and TSS

To evaluate whether ISs accumulate near the TSS of cancer-related genes, the integration events ±100 kb from the TSS were identified. Cancer-related genes were identified using the COSMIC Cancer Gene Census database.[51] Since this is a human genome-specific database, the ISs were lifted over from the mouse genome to the human genome. The ISs detected within the ±100-kb window of a TSS for a cancer-related gene were then examined for the frequency of occurrence among the total number of reads.

Random simulation to examine global enrichment of integrations near TSSs and cancer-related genes

Random simulations were conducted to determine whether the observed number of ISs that occur near genomic elements, such as the TSSs of all genes or cancer-related genes, are present at a level statistically higher than would be expected by chance. To support the calculation of *p* values and *Z* scores, a null distribution of the expected number of integrations to fall within a defined distance of the TSS was determined using a numeric simulation. The null distribution was then used to estimate the statistical significance of the observed IS data. The analysis was performed by carrying out 10,000 independent simulations, where the total number of ISs identified by TES (46,375) were sampled randomly from the genome. After each simulation, the number of ISs within 10 kb of a TSS of a gene were recorded. Some analyses used previously published mouse RNA sequencing results[82] to bin genes by liver expression. ISs near genes used as bait sequences to normalize for genomic copy number were excluded. Identical ISs found between replicates of the same animal or ISs that had a one nucleotide offset were considered as the same IS and merged. The results of the simulations were then compared with the observed proportion of ISs found within 10 kb of a TSS. To determine statistical significance, one-sided *p* values were calculated by taking 1 minus the percentile of the observed proportion found within the simulated distribution. The *p* value for observations that occurred outside the range of the simulated distribution was set to 1×10^{-4}. Enrichment represents the ratio of the observed fraction of ISs meeting the criteria above and the median value observed in the corresponding simulation. *Z* scores were used as a proxy for enrichment and calculated to compare the observed fractions of ISs to the simulated distributions. To calculate the *Z* scores, the following equation was used:

$$Z = \frac{(\text{observed proportion} - \text{mean}(\text{simulated proportion}))}{\text{standard deviation}\,(\text{simulated proportion})}$$

DATA AND CODE AVAILABILITY

Materials and protocols will be distributed for noncommercial, academic purposes upon reasonable request. The data presented in this publication have been deposited in the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) accession number GEO: GSE254503 or deposited in the NIH's SRA accession number SRA: PRJNA1076258.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtm.2024.101294.

ACKNOWLEDGMENTS

Medical writing support was provided by Tony Sallese, PhD, of AlphaBioCom, a Red Nucleus company, and funded by BioMarin Pharmaceutical Inc. ProtaGene CGT GmbH was contracted by BioMarin Pharmaceutical Inc. to perform the integration profile analyses. The presented work was funded by BioMarin Pharmaceutical Inc.

AUTHOR CONTRIBUTIONS

S.F. and A.M.I. contributed to the study design. A.M.I. and T.B. performed the molecular and biochemical analyses. B.Y. performed the histologic analysis. E.W., W.C., and M.F. performed the integration profile analyses. All authors substantively contributed to interpretation of the data and critically reviewed the manuscript.

DECLARATION OF INTERESTS

A.M.I., E.W., and B.Y. are employees and stockholders of BioMarin Pharmaceutical Inc. T.B.,W.C., and S.F. are former employees and

potential stockholders of BioMarin Pharmaceutical Inc. M.F. is an employee of ProtaGene US Inc.

REFERENCES

1. Srivastava, A., Lusby, E.W., and Berns, K.I. (1983). Nucleotide sequence and organization of the adeno-associated virus 2 genome. J. Virol. *45*, 555–564. https://doi.org/10.1128/JVI.45.2.555-564.1983.
2. Wang, D., Tai, P.W.L., and Gao, G. (2019). Adeno-associated virus vector as a platform for gene therapy delivery. Nat. Rev. Drug Discov. *18*, 358–378. https://doi.org/10.1038/s41573-019-0012-9.
3. Samulski, R.J., and Muzyczka, N. (2014). AAV-mediated gene therapy for research and therapeutic purposes. Annu. Rev. Virol. *1*, 427–451. https://doi.org/10.1146/annurev-virology-031413-085355.
4. Srivastava, A. (2016). *In vivo* tissue-tropism of adeno-associated viral vectors. Curr. Opin. Virol. *21*, 75–80. https://doi.org/10.1016/j.coviro.2016.08.003.
5. Chahal, P.S., Schulze, E., Tran, R., Montes, J., and Kamen, A.A. (2014). Production of adeno-associated virus (AAV) serotypes by transient transfection of HEK293 cell suspension cultures for gene delivery. J. Virol. Methods *196*, 163–173. https://doi.org/10.1016/j.jviromet.2013.10.038.
6. Kondratov, O., Marsic, D., Crosson, S.M., Mendez-Gomez, H.R., Moskalenko, O., Mietzsch, M., Heilbronn, R., Allison, J.R., Green, K.B., Agbandje-McKenna, M., and Zolotukhin, S. (2017). Direct head-to-head evaluation of recombinant adeno-associated viral vectors manufactured in human versus insect cells. Mol. Ther. *25*, 2661–2675. https://doi.org/10.1016/j.ymthe.2017.08.003.
7. Kotin, R.M., and Snyder, R.O. (2017). Manufacturing clinical grade recombinant adeno-associated virus using invertebrate cell lines. Hum. Gene Ther. *28*, 350–360. https://doi.org/10.1089/hum.2017.042.
8. Kurasawa, J.H., Park, A., Sowers, C.R., Halpin, R.A., Tovchigrechko, A., Dobson, C.L., Schmelzer, A.E., Gao, C., Wilson, S.D., and Ikeda, Y. (2020). Chemically defined, high-density insect cell-based expression system for scalable AAV vector production. Mol. Ther. Methods Clin. Dev. *19*, 330–340. https://doi.org/10.1016/j.omtm.2020.09.018.
9. Handyside, B., Ismail, A.M., Zhang, L., Yates, B., Xie, L., Sihn, C.R., Murphy, R., Bouwman, T., Kim, C.K., De Angelis, R., et al. (2022). Vector genome loss and epigenetic modifications mediate decline in transgene expression of AAV5 vectors produced in mammalian and insect cells. Mol. Ther. *30*, 3570–3586. https://doi.org/10.1016/j.ymthe.2022.11.001.
10. Bunting, S., Zhang, L., Xie, L., Bullens, S., Mahimkar, R., Fong, S., Sandza, K., Harmon, D., Yates, B., Handyside, B., et al. (2018). Gene therapy with BMN 270 results in therapeutic levels of FVIII in mice and primates and normalization of bleeding in hemophilic mice. Mol. Ther. *26*, 496–509. https://doi.org/10.1016/j.ymthe.2017.12.009.
11. Rangarajan, S., Walsh, L., Lester, W., Perry, D., Madan, B., Laffan, M., Yu, H., Vettermann, C., Pierce, G.F., Wong, W.Y., and Pasi, K.J. (2017). AAV5-factor VIII gene transfer in severe hemophilia A. N. Engl. J. Med. *377*, 2519–2530. https://doi.org/10.1056/NEJMoa1708483.
12. Ozelo, M.C., Mahlangu, J., Pasi, K.J., Giermasz, A., Leavitt, A.D., Laffan, M., Symington, E., Quon, D.V., Wang, J.D., Peerlinck, K., et al. (2022). Valoctocogene roxaparvovec gene therapy for hemophilia A. N. Engl. J. Med. *386*, 1013–1025. https://doi.org/10.1056/NEJMoa2113708.
13. Pasi, K.J., Laffan, M., Rangarajan, S., Robinson, T.M., Mitchell, N., Lester, W., Symington, E., Madan, B., Yang, X., Kim, B., et al. (2021). Persistence of haemostatic response following gene therapy with valoctocogene roxaparvovec in severe haemophilia A. Haemophilia *27*, 947–956. https://doi.org/10.1111/hae.14391.
14. Mahlangu, J., Kaczmarek, R., von Drygalski, A., Shapiro, S., Chou, S.C., Ozelo, M.C., Kenet, G., Peyvandi, F., Wang, M., Madan, B., et al. (2023). Two-year outcomes of valoctocogene roxaparvovec therapy for hemophilia A. N. Engl. J. Med. *388*, 694–705. https://doi.org/10.1056/NEJMoa2211075.
15. Duan, D., Sharma, P., Yang, J., Yue, Y., Dudus, L., Zhang, Y., Fisher, K.J., and Engelhardt, J.F. (1998). Circular intermediates of recombinant adeno-associated virus have defined structural characteristics responsible for long-term episomal persistence in muscle tissue. J. Virol. *72*, 8568–8577. https://doi.org/10.1128/jvi.72.11.8568-8577.1998.
16. Nakai, H., Yant, S.R., Storm, T.A., Fuess, S., Meuse, L., and Kay, M.A. (2001). Extrachromosomal recombinant adeno-associated virus vector genomes are primarily responsible for stable liver transduction *in vivo*. J. Virol. *75*, 6969–6976. https://doi.org/10.1128/jvi.75.15.6969-6976.2001.
17. Greig, J.A., Martins, K.M., Breton, C., Lamontagne, R.J., Zhu, Y., He, Z., White, J., Zhu, J.X., Chichester, J.A., Zheng, Q., et al. (2023). Integrated vector genomes may contribute to long-term expression in primate liver after AAV administration. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-01974-7.
18. 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.
19. Li, H., Malani, N., Hamilton, S.R., Schlachterman, A., Bussadori, G., Edmonson, S.E., Shah, R., Arruda, V.R., Mingozzi, F., Wright, J.F., et al. (2011). Assessing the potential for AAV vector genotoxicity in a murine model. Blood *117*, 3311–3319. https://doi.org/10.1182/blood-2010-08-302729.
20. Nakai, H., Montini, E., Fuess, S., Storm, T.A., Grompe, M., and Kay, M.A. (2003). AAV serotype 2 vectors preferentially integrate into active genes in mice. Nat. Genet. *34*, 297–302. https://doi.org/10.1038/ng1179.
21. Wu, P., Phillips, M.I., Bui, J., and Terwilliger, E.F. (1998). Adeno-associated virus vector-mediated transgene integration into neurons and other nondividing cell targets. J. Virol. *72*, 5919–5926. https://doi.org/10.1128/jvi.72.7.5919-5926.1998.
22. 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.
23. 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.
24. Batty, P., Mo, A.M., Hurlbut, D., Ishida, J., Yates, B., Brown, C., Harpell, L., Hough, C., Pender, A., Rimmer, E.K., et al. (2022). Long-term follow-up of liver-directed, adeno-associated vector-mediated gene therapy in the canine model of hemophilia A. Blood *140*, 2672–2683. https://doi.org/10.1182/blood.2021014735.
25. 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.
26. Pañeda, A., Lopez-Franco, E., Kaeppel, C., Unzu, C., Gil-Royo, A.G., D'Avola, D., Beattie, S.G., Olagüe, C., Ferrero, R., Sampedro, A., et al. (2013). Safety and liver transduction efficacy of rAAV5-cohPBGD in nonhuman primates: a potential therapy for acute intermittent porphyria. Hum. Gene Ther. *24*, 1007–1017. https://doi.org/10.1089/hum.2013.166.
27. 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.
28. Kaeppel, C., Beattie, S.G., Fronza, R., van Logtenstein, R., Salmon, F., Schmidt, S., Wolf, S., Nowrouzi, A., Glimm, H., von Kalle, C., et al. (2013). A largely random AAV integration profile after LPLD gene therapy. Nat. Med. *19*, 889–891. https://doi.org/10.1038/nm.3230.
29. 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 integration: an integrated perspective. Mol. Ther. *30*, 2646–2663. https://doi.org/10.1016/j.ymthe.2022.06.004.
30. 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.
31. Donsante, A., Vogler, C., Muzyczka, N., Crawford, J.M., Barker, J., Flotte, T., Campbell-Thompson, M., Daly, T., and Sands, M.S. (2001). Observed incidence of

tumorigenesis in long-term rodent studies of rAAV vectors. Gene Ther. *8*, 1343–1346. https://doi.org/10.1038/sj.gt.3301541.

32. Walia, J.S., Altaleb, N., Bello, A., Kruck, C., LaFave, M.C., Varshney, G.K., Burgess, S.M., Chowdhury, B., Hurlbut, D., Hemming, R., et al. (2015). Long-term correction of Sandhoff disease following intravenous delivery of rAAV9 to mouse neonates. Mol. Ther. *23*, 414–422. https://doi.org/10.1038/mt.2014.240.

33. Cheng, Y., Zhang, Z., Gao, P., Lai, H., Zhong, W., Feng, N., Yang, Y., Yu, H., Zhang, Y., Han, Y., et al. (2023). AAV induces hepatic necroptosis and carcinoma in diabetic and obese mice dependent on Pebp1 pathway. EMBO Mol. Med. *15*, e17230. https://doi.org/10.15252/emmm.202217230.

34. 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.

35. La Bella, T., Imbeaud, S., Peneau, C., Mami, I., Datta, S., Bayard, Q., Caruso, S., Hirsch, T.Z., Calderaro, J., Morcrette, G., et al. (2020). Adeno-associated virus in the liver: natural history and consequences in tumour development. Gut *69*, 737–747. https://doi.org/10.1136/gutjnl-2019-318281.

36. Logan, G.J., Dane, A.P., Hallwirth, C.V., Smyth, C.M., Wilkie, E.E., Amaya, A.K., Zhu, E., Khandekar, N., Ginn, S.L., Liao, S.H.Y., et al. (2017). Identification of liver-specific enhancer-promoter activity in the 3' untranslated region of the wild-type AAV2 genome. Nat. Genet. *49*, 1267–1273. https://doi.org/10.1038/ng.3893.

37. Bell, P., Wang, L., Lebherz, C., Flieder, D.B., Bove, M.S., Wu, D., Gao, G.P., Wilson, J.M., and Wivel, N.A. (2005). No evidence for tumorigenesis of AAV vectors in a large-scale study in mice. Mol. Ther. *12*, 299–306. https://doi.org/10.1016/j.ymthe.2005.03.020.

38. Bell, P., Moscioni, A.D., McCarter, R.J., Wu, D., Gao, G., Hoang, A., Sanmiguel, J.C., Sun, X., Wivel, N.A., Raper, S.E., et al. (2006). Analysis of tumors arising in male B6C3F1 mice with and without AAV vector delivery to liver. Mol. Ther. *14*, 34–44. https://doi.org/10.1016/j.ymthe.2006.03.008.

39. Batty, P., Fong, S., Franco, M., Sihn, C.R., Swystun, L.L., Afzal, S., Harpell, L., Hurlbut, D., Pender, A., Su, C., et al. (2024). Vector integration and fate in the hemophilia dog liver multiple years after AAV-FVIII gene transfer. Blood *143*, 2373–2385. https://doi.org/10.1182/blood.2023022589.

40. Chiuchiolo, M.J., Kaminsky, S.M., Sondhi, D., Hackett, N.R., Rosenberg, J.B., Frenk, E.Z., Hwang, Y., Van de Graaf, B.G., Hutt, J.A., Wang, G., et al. (2013). Intrapleural administration of an AAVrh.10 vector coding for human α1-antitrypsin for the treatment of α1-antitrypsin deficiency. Hum. Gene Ther. Clin. Dev. *24*, 161–173. https://doi.org/10.1089/humc.2013.168.

41. De, B.P., Heguy, A., Hackett, N.R., Ferris, B., Leopold, P.L., Lee, J., Pierre, L., Gao, G., Wilson, J.M., and Crystal, R.G. (2006). High levels of persistent expression of alpha1-antitrypsin mediated by the nonhuman primate serotype rh.10 adeno-associated virus despite preexisting immunity to common human adeno-associated viruses. Mol. Ther. *13*, 67–76. https://doi.org/10.1016/j.ymthe.2005.09.003.

42. Zimmerer, J.M., Pham, T.A., Sanders, V.M., and Bumgardner, G.L. (2010). CD8+ T cells negatively regulate IL-4-dependent, IgG1-dominant posttransplant alloantibody production. J. Immunol. *185*, 7285–7292. https://doi.org/10.4049/jimmunol.1001655.

43. Sihn, C.R., Handyside, B., Liu, S., Zhang, L., Murphy, R., Yates, B., Xie, L., Torres, R., Russell, C.B., O'Neill, C.A., et al. (2022). Molecular analysis of AAV5-hFVIII-SQ vector-genome-processing kinetics in transduced mouse and nonhuman primate livers. Mol. Ther. Methods Clin. Dev. *24*, 142–153. https://doi.org/10.1016/j.omtm.2021.12.004.

44. Zhang, L., Yates, B., Murphy, R., Liu, S., Xie, L., Handyside, B., Sihn, C.R., Bouwman, T., Galicia, N., Tan, D., et al. (2022). Young mice administered adult doses of AAV5-hFVIII-SQ achieve therapeutic factor VIII expression into adulthood. Mol. Ther. Methods Clin. Dev. *26*, 519–531. https://doi.org/10.1016/j.omtm.2022.08.002.

45. Afzal, S., Gil-Farina, I., Gabriel, R., Ahmad, S., von Kalle, C., Schmidt, M., and Fronza, R. (2019). Systematic comparative study of computational methods for T-cell receptor sequencing data analysis. Briefings Bioinf. *20*, 222–234. https://doi.org/10.1093/bib/bbx111.

46. Rényi, A. (1961). On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed. (University of California Press), pp. 547–561.

47. Biffi, A., Bartolomae, C.C., Cesana, D., Cartier, N., Aubourg, P., Ranzani, M., Cesani, M., Benedicenti, F., Plati, T., Rubagotti, E., et al. (2011). Lentiviral vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection. Blood *117*, 5332–5339. https://doi.org/10.1182/blood-2010-09-306761.

48. Lund, A.H., Turner, G., Trubetskoy, A., Verhoeven, E., Wientjens, E., Hulsman, D., Russell, R., DePinho, R.A., Lenz, J., and van Lohuizen, M. (2002). Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. Nat. Genet. *32*, 160–165. https://doi.org/10.1038/ng956.

49. Mikkers, H., Allen, J., Knipscheer, P., Romeijn, L., Hart, A., Vink, E., and Berns, A. (2002). High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. Nat. Genet. *32*, 153–159. https://doi.org/10.1038/ng950.

50. Wu, X., Luke, B.T., and Burgess, S.M. (2006). Redefining the common insertion site. Virology *344*, 292–295. https://doi.org/10.1016/j.virol.2005.08.047.

51. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.

52. 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.

53. Cavazzana-Calvo, M., Payen, E., Negre, O., Wang, G., Hehir, K., Fusil, F., Down, J., Denaro, M., Brady, T., Westerman, K., et al. (2010). Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. Nature *467*, 318–322. https://doi.org/10.1038/nature09328.

54. Deichmann, A., Hacein-Bey-Abina, S., Schmidt, M., Garrigue, A., Brugman, M.H., Hu, J., Glimm, H., Gyapay, G., Prum, B., Fraser, C.C., et al. (2007). Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. J. Clin. Invest. *117*, 2225–2232. https://doi.org/10.1172/jci31659.

55. 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.

56. Hacein-Bey-Abina, S., von Kalle, C., Schmidt, M., Le Deist, F., Wulffraat, N., McIntyre, E., Radford, I., Villeval, J.L., Fraser, C.C., Cavazzana-Calvo, M., and Fischer, A. (2003). A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. N. Engl. J. Med. *348*, 255–256. https://doi.org/10.1056/nejm200301163480314.

57. Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M.P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C.S., Pawliuk, R., Morillon, E., et al. (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science *302*, 415–419. https://doi.org/10.1126/science.1088547.

58. 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.

59. Ott, M.G., Schmidt, M., Schwarzwaelder, K., Stein, S., Siler, U., Koehl, U., Glimm, H., Kühlcke, K., Schilz, A., Kunkel, H., et al. (2006). Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. Nat. Med. *12*, 401–409. https://doi.org/10.1038/nm1393.

60. 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.

61. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.

62. Blair, H.A. (2022). Valoctocogene Roxaparvovec: First Approval. Drugs *82*, 1505–1510. https://doi.org/10.1007/s40265-022-01788-y.

63. VandenDriessche, T., Pipe, S.W., Pierce, G.F., and Kaczmarek, R. (2022). First conditional marketing authorization approval in the European Union for hemophilia "A" gene therapy. Mol. Ther. *30*, 3335–3336. https://doi.org/10.1016/j.ymthe.2022.09.020.

64. Batty, P., Fong, S., Franco, M., Gil-Farina, I., Mo, A.M., Harpell, L., Hough, C., Hurlbut, D., Pender, A., Sardo Infirri, S., et al. (2020). Frequency, location and nature of AAV vector insertions after long-term follow up of FVIII transgene delivery in a hemophilia A dog model. Res. Pract. Thromb. Haemost. *4*.

65. Eggan, K. (2022). Exploratory analyses of healthy liver biopsies and a single case of parotid acinar cell carcinoma do not identify a role for valoctocogene roxaparvovec vector insertion in altering cell growth. In Presented at the World Federation of Hemophilia (WFH) Congress (Montréal).

66. Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., and Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. Nat. Commun. *8*, 15183. https://doi.org/10.1038/ncomms15183.

67. Qin, W., Xu, G., Tai, P.W.L., Wang, C., Luo, L., Li, C., Hu, X., Xue, J., Lu, Y., Zhou, Q., et al. (2021). Large-scale molecular epidemiological analysis of AAV in a cancer patient population. Oncogene *40*, 3060–3071. https://doi.org/10.1038/s41388-021-01725-5.

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

69. Rumachik, N.G., Malaker, S.A., Poweleit, N., Maynard, L.H., Adams, C.M., Leib, R.D., Cirolia, G., Thomas, D., Stamnes, S., Holt, K., et al. (2020). Methods matter: standard production platforms for recombinant AAV produce chemically and functionally distinct vectors. Mol. Ther. Methods Clin. Dev. *18*, 98–118. https://doi.org/10.1016/j.omtm.2020.05.018.

70. Chang, M., Parker, E.A., Muller, T.J.M., Haenen, C., Mistry, M., Finkielstain, G.P., Murphy-Ryan, M., Barnes, K.M., Sundaram, R., and Baron, J. (2008). Changes in cell-cycle kinetics responsible for limiting somatic growth in mice. Pediatr. Res. *64*, 240–245. https://doi.org/10.1203/PDR.0b013e318180e47a.

71. Inagaki, K., Piao, C., Kotchey, N.M., Wu, X., and Nakai, H. (2008). Frequency and spectrum of genomic integration of recombinant adeno-associated virus serotype 8 vector in neonatal mouse liver. J. Virol. *82*, 9513–9524. https://doi.org/10.1128/jvi.01001-08.

72. 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.

73. Schmidt, M., Foster, G.R., Coppens, M., Thomsen, H., Cooper, D., Dolmetsch, R., Sawyer, E.K., Heijink, L., and Pipe, S.W. (2021). Liver safety case report from the phase 3 HOPE-B gene therapy trial in adults with hemophilia B [abstract]. Res. Pract. Thromb. Haemost. *5*.

74. (2021). uniQure announces findings from reported case of hepatocellular carcinoma (HCC) in hemophilia B gene therapy program. https://www.globenewswire.com/news-release/2021/03/29/2200653/0/en/uniQure-Announces-Findings-from-Reported-Case-of-Hepatocellular-Carcinoma-HCC-in-Hemophilia-B-Gene-Therapy-Program.html.

75. Afzal, S., Wilkening, S., von Kalle, C., Schmidt, M., and Fronza, R. (2017). GENE-IS: time-efficient and accurate analysis of viral integration events in large-scale gene therapy data. Mol. Ther. Nucleic Acids *6*, 133–139. https://doi.org/10.1016/j.omtn.2016.12.001.

76. Fronza, R., Vasciaveo, A., Benso, A., and Schmidt, M. (2016). A graph based framework to model virus integration sites. Comput. Struct. Biotechnol. J. *14*, 69–77. https://doi.org/10.1016/j.csbj.2015.10.006.

77. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at ArXiv. https://doi.org/10.48550/arXiv.1303.3997.

78. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

79. Find Peaks Using MACS2. Accessed July 12, 2023. https://chipster.csc.fi/manual/macs2.html.

80. Genome reference consortium mouse build 38 patch release 6 (GRCm38.p6). https://www.gencodegenes.org/mouse/release_M25.html.

81. Riemondy, K.A., Sheridan, R.M., Gillen, A., Yu, Y., Bennett, C.G., and Hesselberth, J.R. (2017). valr: reproducible genome interval analysis in R. F1000Res. *6*, 1025. https://doi.org/10.12688/f1000research.11997.1.

82. Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., et al. (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep. *3*, 2179–2190. https://doi.org/10.1016/j.celrep.2013.05.031.