Supplementary Methods

General Considerations and Strategies in Preparing Recombinant Adeno-Associated Virus Inverted Terminal Repeat–Host Genomic DNA Junction Libraries

The junction libraries were prepared using standard techniques. The first step was to select appropriate restriction enzymes for cellular DNA digestion and linker design. Several principles were applied. First, four cutter restriction enzymes with a 5′ overhang are ideal, as they cut more frequently in the host genome than the five and six cutters do and the 5′ overhangs facilitate more efficient and specific sticky end ligation. Second, since the current pyrosequencing technology allows only a 500 bp sequence coverage for each template, the cutting sites within the vector genome should be as close to the inverted terminal repeat (ITR) sequences as possible to allow more sequence read into the cellular sequences. Third, the selected enzymes should cut the remaining part of the vector genome as infrequently as possible to avoid amplification of an internal vector fragment. This could be reinforced by cutting with a second enzyme after linker ligation to remove the internal fragments. Finally, enzymes that yield ends that are difficult to ligate and have the CpG di-nucleotides in their recognition sites (CpG is rare and unevenly distributed in the cellular genome) would be avoided. On the basis of these principles, *Taq*I and *Bsr*GI were selected as primary and secondary enzymes for this study. The third enzyme *Bsr*BI with recognition sites on the plasmid backbone was selected to remove plasmid contamination since *Dpn*I digestion did not rule out some nonspecific amplifications from the plasmid backbone (data not shown). The linker was designed based on the sequence recognized by restriction enzymes *Taq*I, which is shown in Supplementary Table S1. One consideration in the linker design is that an amino-modifier group needs to be added to one 3′ end of the linker to prevent extension to the vector genome. This design ensures that PCR amplification is originated within the integrated sequences but not from the linker. The second step was to design vector-specific primers. The key factors to be considered include the hairpin loop structure of recombinant adeno-associated virus (rAAV) ITR and the necessity to start the sequencing reactions as close to the ITR as possible. We selected the sequences near the "D" region (125–145 nucleotides of ITR) in the vector genome as the annealing targets for the vector-specific primers (Supplementary Table S1). The second pair of the primers was designed for a nested PCR. One of the primers was tagged with a 454 sequencing primer-specific sequence only, and the other was tagged with a sample specific bar-coding sequence followed by a 454 sequencing primer-specific sequence (Supplementary Table S1). The PCR conditions for efficient and precise amplification of ITR–cellular DNA junctions were optimized and validated with naïve mouse genomic DNA (gDNA) spiked with different copy numbers of pAAV-TBG-*mOTC* and pAAV2.1-TBG-*LacZ* plasmid DNAs.

Once the PCR-based recovery of ITR–gDNA junctions was optimized and validated, we used total cellular DNA samples isolated from five tumor and five adjacent normal liver tissues (Bell *et al*., 2006) as the templates to generate the ITR–gDNA junction amplicon libraries. Briefly, except for 542TM (1.1 μg) and 838TM (1.6 μg), 2 μg each of gDNA from all other samples was used as the starting materials for creating the amplicon libraries. Duplicated gDNA samples with 1 μg per reaction were digested with *Taq*I (New England Biolabs, Ipswich, MA) at 65°C overnight. About 1 μg of plasmid DNA was treated under the same condition as the control to validate the completion of the digestion. Digested DNA was purified using a PCR purification kit (Qiagen). Annealed double-stranded linker adapter (Supplementary Table S1) was ligated to the purified digested DNA ends by T4 DNA ligase (New England Biolabs) at 16°C for 18 hr and then heated at 65°C for 15 min to inactivate DNA ligase. Ligated DNA samples were digested again with *Bsr*GI or *Bsr*GI plus *Bsr*BI at 37°C overnight and then the ITR–host gDNA junctions were amplified by PCR using linker primer1 and rAAV vector primer1 described previously (Supplementary Table S1). The first-round PCR products were then diluted 1:200 in nuclease-free water followed by a second-round nested PCR using linker primer2 and rAAV vector primer2 described previously (Supplementary Table S1). The TAKARA LA Taq polymerase (Fisher Scientific, Pittsburgh, PA) was used in the PCRs.

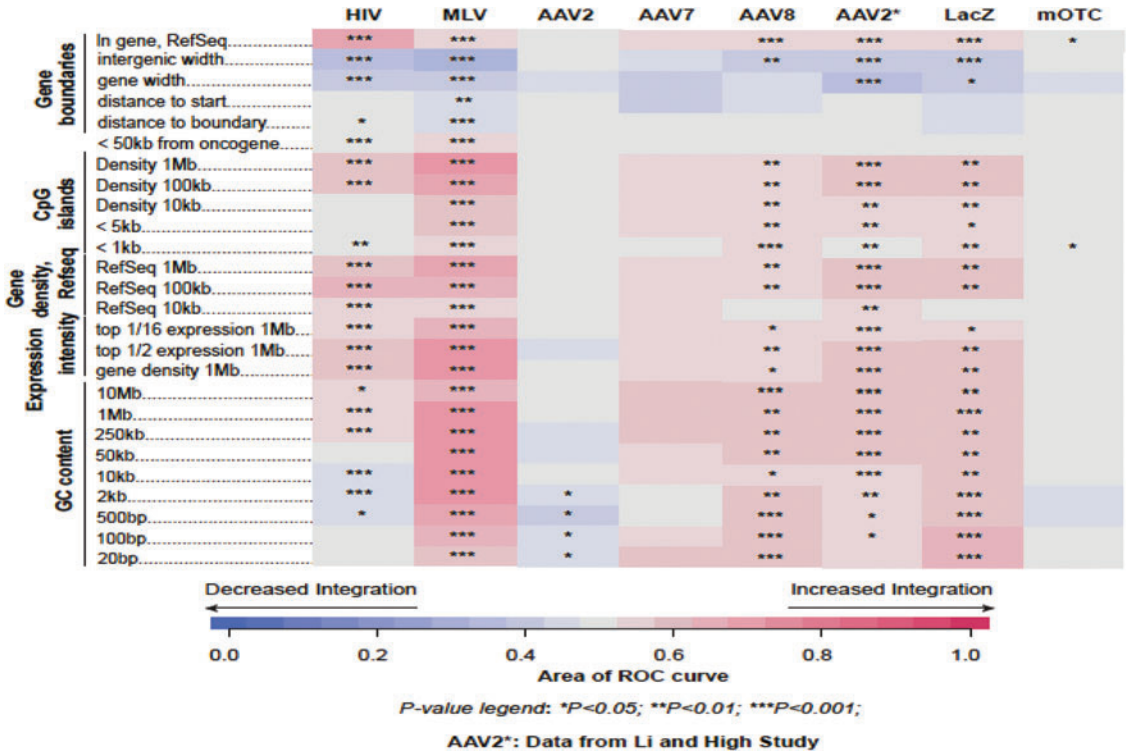
Cloning of rAAV ITR–Host DNA Junctions in Bacterial Plasmids and Verification by Sanger Sequencing

The most abundant junctions between rAAV and host DNA from tumors were verified where possible using DNA cloning and Sanger sequencing. Eleven integration sites were analyzed from three tumors (the remaining two could not be studied because of limitations of DNA availability) and one flanking normal liver tissue. A total of 120 pairs of primers were designed and used in 439 PCRs. PCR products were cloned into TOPO cloning vector pCR4-TOPO using the TOPO cloning kit following manufacturer's instructions (Invitrogen, Carlsbad, CA) and subjected to Sanger sequencing by Eurofins MWG Operon (Huntsville, AL). The sequencing data were analyzed using NCBI blast (www.ncbi.nlm.nih.gov). Sequence analysis of 282 of the Topo-clones allowed verification of 9 of the sites (Supplementary Table S3b).

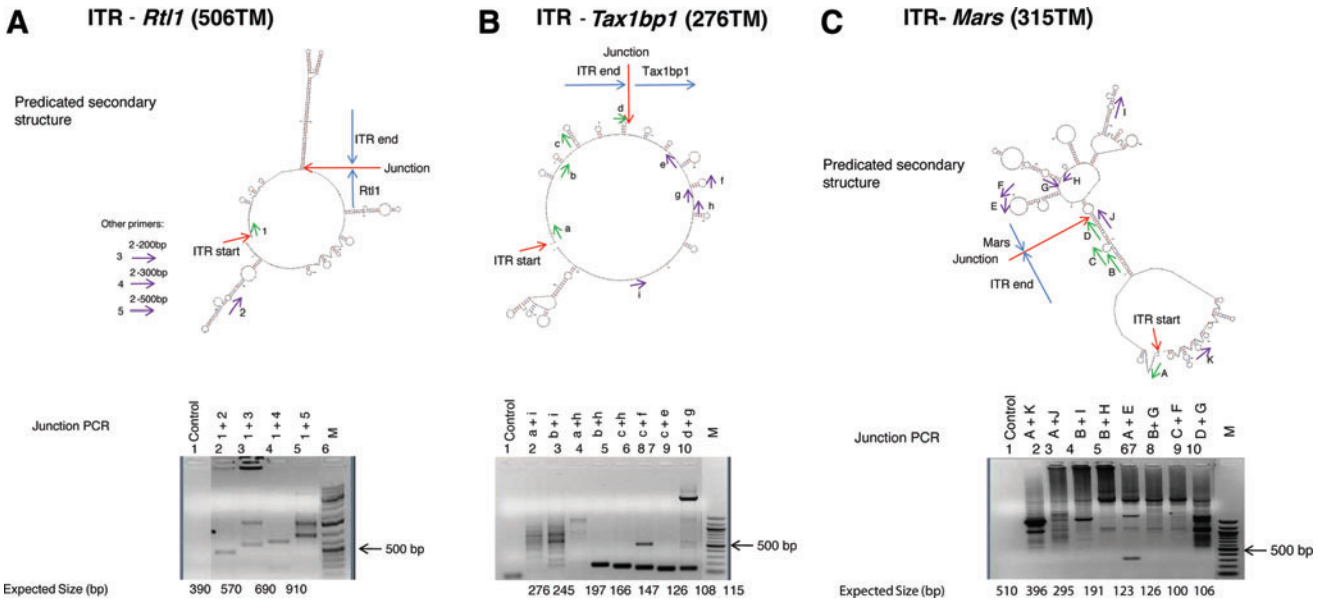
PCR Quantification of rAAV ITR–gDNA Junctions

The rAAV integration site junctions in hepatocellular carcinoma and adjacent normal liver gDNA samples were quantified using a StepOnePlus real-time PCR system (Applied Biosystems) and SYBR Green GoTaq quantitative PCR (qPCR) master mix (Promega, Madison, WI). Vector- and host genome-specific primers were designed to amplify through ITR–host genome junctions to quantify the rAAV integrants at the nine confirmed loci by using PCR-cloned junction plasmids as standards (Supplementary Table S3a). To optimize and validate SYBR Green qPCR, 100 ng each of normal mouse liver gDNA was spiked with different copy numbers (10–10^8 genome copies) of ITR–gDNA junction

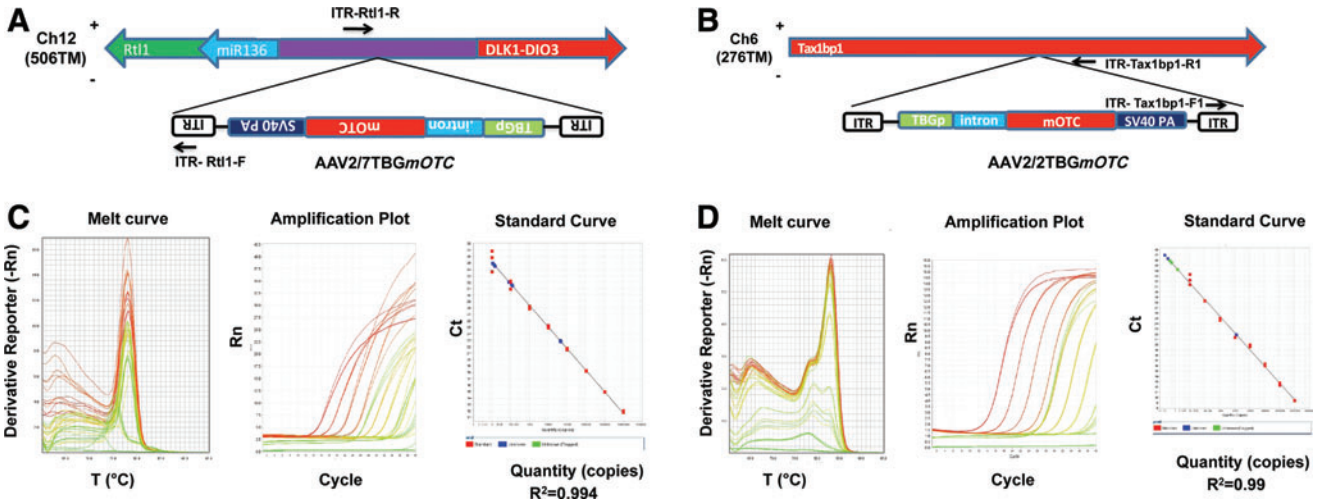
plasmids and subjected to qPCR to screen 6–7 pairs of primers per integration locus. By analyzing melt curves, amplification plots, and standard curves, the best-performed 1–2 pairs of primers for each junction were selected for PCR quantification of nine rAAV-ITR–gDNA junctions in tumor and adjacent normal liver DNA samples (Supplementary Table S3a and Supplementary Fig. S3). The numbers of rAAV integrants per cell for all nine loci are shown in Supplementary Table S3b. The primer pairs used for quantification of the monoclonal integrants in the *Rtl1* in the *Dlk1–Dio3* region and the *Tax1bp1* region are bolded in Supplementary Table S3b.



SUPPLEMENTARY FIG. S1. Global integration site distributions: comparison of integration site patterns for the different rAAV serotypes and transgenes. Origins of datasets are shown as columns; genomic features are shown in rows of the heatmap. The experimental integration sites in each dataset were compared with matched random controls to assess the relative frequency of integration near the indicated genomic feature. Biases are expressed using the ROC area using the color scale at the bottom of the heat map. The increasing density of blue indicates disfavored integration compared with random; red indicates favored integration, with respect to the genomic feature indicated to the left (e.g., smaller genes, shorter distance to transcription start sites and boundaries of genes, or decrease of integration chance). An ROC area below 0.5 indicates disfavored compared with random; an area above 0.5 indicates favored. Comparisons to genomic features were as described in previous publications (Berry *et al.*, 2006; Li *et al.*, 2011). Asterisks represent the statistical significance of departures from random distributions by comparison to ROC area = 0.5 (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$). Note that sample sizes are small for some of the sets, so that apparent differences in rAAV integration site distributions generally do not achieve significance. rAAV, recombinant adeno-associated virus; ROC, receiver operating characteristic.



SUPPLEMENTARY FIG. S2. Impact of the predicated secondary structures of ITR–gDNA junctions on the PCR amplification through the junctions in tumor DNA samples. The secondary structures of ITR–*Rtl1* **(A**, top**)** (506TM), ITR–*Tax1bp1* **(B,** top**)** (276TM), and ITR–*Mars* **(c,** top**)** (315TM) junctions were predicated by Mfold (http://mfold.rna.albany.edu/). DNAs were amplified by PCR using the indicated primer pairs (1–5, a–I, and A–K, respectively). About 5 μl each of the PCR products were examined by 1.5% agarose gel electrophoresis (bottom). **(A, B)** The ITR–Rtl1 and ITR–Tax1bp1 junctions did not contain complex secondary structures and were PCR amplified efficiently by four out of four and five out of eight sets of primer combinations tested, respectively. The fidelities of the PCR products were confirmed by Topo-cloning and Sanger sequencing. **(C)** The complex secondary structure in the ITR–Mars junction prevented PCR amplifications by all eight sets of primer combinations tested. ITR, inverted terminal repeat; gDNA, genomic DNA; PCR, polymerase chain reaction.



SUPPLEMENTARY FIG. S3. Validation of SYBR Green qPCR and documentation of monoclonal integration for the ITR–*Rtl1*junction in 506 tumor DNA and the ITR–*Tax1bp1* junction in 276 tumor DNA. Locations of the PCR primers used for amplifications of the ITR–*Rtl1* junction **(A)** and the ITR–*Tax1bp1* junction **(B)** are schematically presented. To validate SYBR green qPCR using those primers, 100 ng each of normal mouse liver gDNA was spiked with different copy numbers (10–10^8 genome copies) of the PCR-cloned *Rtl1*–ITR and *Tax1bp1*–ITR junction plasmids and subjected to qPCR. Melt curves, amplification plots, and standard curves were presented for the best-performed PCR primer pairs that target *Rtl1*–ITR **(C)** and *Tax1bp1*–ITR **(D)** junctions. Using the optimized primers and PCR conditions, DNAs from tumors and adjacent normal liver tissues were quantified by the SYBR Green qPCR method. The copies of rAAV integrants in the *Rtl1* site of the *Dlk1–Dio3* region (left) and *Tax1bp1* site (right) were compared **(E)**. qPCR, quantitative PCR; LV, adjacent normal liver; TM, tumor; Mock, no DNA template control.

Supplementary Table S1. Sequences of Oligonucleotides Used in the Inverted Terminal Repeat Junction Library Preparation

Oligos	*sequences*
TaqI linker+ (generate linker)	GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC
TaqI linker− (generate linker)	PO_4-CGGTCCCTTAAGCGGAG-AmC7-Q
TaqI linker primer 1 (PCR1)	GTAATACGACTCACTATAGGGC
AAV ITR primer 1 (PCR1)	AGGATCTTCCTAGAGCATGGCTACGTAG
TaqI linker primer 2 (PCR2)	GCCTCCCTCGCGCCATCAGAGGGCTCCGCTTAAGGGAC
AAV ITR primer 2	GCCTTGCCAGCCCGCTCAG-Barcode-GTAGATAAGTAGCATGGCGGGTTAATC

Supplementary Table S2. Abundance of Persisted rAAV Genomes and rAAV Integration Events Detected in All Study Samples

Animal ID/genotype	*rAAV*	*Age at injection/ necropsy/study duration (days)*	*Samples*	*Vector copies/cell*	*Total reads*	*Unique sites (% of total reads)*
276/spf^{ash}	2/2TBGmOTC	157/530/373	LV	3	109244	286 (0.26)
			TM	1.9	100909	335 (0.33)
315/spf^{ash}	2/8TBGLacZ	126/489/363	LV	21	134561	165 (0.12)
			TM	1.6	123960	60 (0.05)
506/spf	2/7TBGmOTC	93/451/358	LV	1.6	117229	132 (0.11)
			TM	1.4	38359	13 (0.03)
542/spf	2/8TBGLacZ	92/450/358	LV	49	82570	176 (0.21)
			TM	23	76912	84 (0.11)
838/spf	2/8mOTC	126/387/261	LV	23	124591	213 (0.17)
			TM	6.3	91107	116 (0.13)
Total					999442	1580 (0.16)
LV					568195	972 (0.17)
TM					431247	608 (0.14)

rAAV, recombinant adeno-associated virus; LV, liver; TM, tumor.

Supplementary Table S3a. Sequences of Oligonucleotides Used for Endpoint PCR to Quantify the Number of rAAV Integrants per Cell at the Confirmed Integration Loci

	Primers	*Sequences*
276TM	**ITR-Tax1bp1—F1/R1**	**5′-AGGATCTTCCTAGAGCATGGCTACGTAG/5′-GGCATGTACGAACCTATCTGA**
	ITR-Tax1bp1—F2/R2	**5′-CAAGGAACCCCTAGTGATGG/5′-CTGGGATGGCTTTCATTCAT**
	ITR-Cdk11b—F1/R1	5′-TAGCATGGCGGGTTAATCAT/5′-ACCTGCTCCTTAGCGACCTT
	ITR-Cdk11b—F2/R2	5′-TAGCATGGCGGGTTAATCAT/5′-ATGGCGAGAGAACATTCCAG
	ITR-HA Arhgap42—F1/R1	5′-TAGCATGGCGGGTTAATCAT/5′-TCCGAAGCACTTCTCTTTTCA
	ITR-HA Arhgap42—F2/R2	5′-TAGCATGGCGGGTTAATCAT/5′-TAAGAGGCCAGAGTCCGAAG
	ITR-LA Arhgap42—F1/R1	5′-GAGTTGGCCACTCCCTCTCT/5′-TCTCATACTGAGACCAAGTGGATT
	ITR-LA Arhgap42—F2/R2	5′-TAGCATGGCGGGTTAATCAT/5′-ACGTTAACCATTGCCTTTCC
	ITR-Rian-miR341—F1/R1	5′-CAAATGTGGTAAAATCGATAAGGA/5′-ACCGACCGACTGACTGACA
	ITR-Rian-miR341—F2/R2	5′-CTCGCTCGCTCACTGATGT/5′-TGCAGTTCGAAGACAGGA
506TM	**ITR-Rtl1—F/R**	**5′-CAGAGAGGGAGTGGCCTTTT/5′-CTCCCAGATTTAAAACGTTGC**
	ITR-HA Rian—F/R	5′-CAAGGAACCCCTAGTGATGG/5′-CCCACCTCATCCTCTTTTGA
	ITR-LA Rian—F1/R1	5′-ATGCTGCTGTTTGGGGTTAG/5′-TCACTGAGGCCGGGTTATAC
	ITR-LA Rian—F2/R2	5′-GCTGAGTCGCTCATTGCAT/5′-GAGTTGGCCACTCCCTCTCT
276LV	ITR-LV Rian—F1/R1	5′-TAGCATGGCGGGTTAATCAT/5′-TCCTCATCTTTTGCACTGGTT
	ITR-LV Rian—F2/R2	5′-ACCCCTAGTGATGGAGTTGG/TGCACTGGTTTGAAGCTAATTC

PCR, polymerase chain reaction.

Supplementary Table S3b. The Number of rAAV Integrants Verified by Topo-PCR Cloning and Quantified by Endpoint PCR

	Copy number (GC/cell)					
	Tax1bp1	*Cdk11b*	*HA Arhgap42*	*LA Arhgap42*	*Rian-miR341*	*Rian-276LV*
276LV	0.0001 ± 0.0001	0.000001 ± 0.000001	0.00001 ± 0.000008	0.00009 ± 0.00006	0.0	0.003 ± 0.00006
276TM	0.8 ± 0.2	0.0002 ± 0.0001	0.001 ± 0.0003	0.002 ± 0.0006	0.0007 ± 0.0002	0.0008 ± 0.00001
	Rtl1	*HA Rian*	*LA Rian*			
506LV	0.003 ± 0.001	0.00002 ± 0.00002	0.00002 ± 0.00001			
506TM	1.2 ± 0.07	0.0001 ± 0.0001	0.003 ± 0.003			
	Mars	*Efha1*				
315LV	—	—				
315TM	junction clone not isolated					

GC, genome copy.