nature genetics

AAV serotype 2 vectors preferentially integrate into active genes in mice

Hiroyuki Nakai[1], Eugenio Montini[2,4], Sally Fuess[1], Theresa A Storm[1], Markus Grompe[2,3] & Mark A Kay[1]

Recombinant adeno-associated virus serotype 2 (rAAV2) is a promising vector for gene therapy because it can achieve long-term stable transgene expression in animals and human subjects after direct administration of vectors into various target tissues[1]. In the liver, although stable transgene expression primarily results from extrachromosomal vector genomes[2], a series of experiments has shown that vector genomes integrate into host chromosomes in hepatocytes[3–5] at a low frequency[2]. Despite the low integration efficiency, recent reports of retroviral insertional mutagenesis in mice[6] and two human subjects[7,8] have raised concerns about the potential for rAAV2-mediated insertional mutagenesis. Here we characterize rAAV2-targeted chromosomal integration sites isolated from selected or non-selected hepatocytes in vector-injected mouse livers. We document frequent chromosomal deletions of up to 2 kb at integration sites (14 of 14 integrations, 100%; most of the deletions were <0.3 kb) and preferred integration into genes (21 of 29 integrations, 72%). In addition, all of the targeted genes analyzed (20 of 20 targeted genes, 100%) were expressed in the liver. This is the first report to our knowledge on host chromosomal effects of rAAV2 integration in animals, and it provides insights into the nature of rAAV2 vector integration into chromosomes in quiescent somatic cells in animals and human subjects.

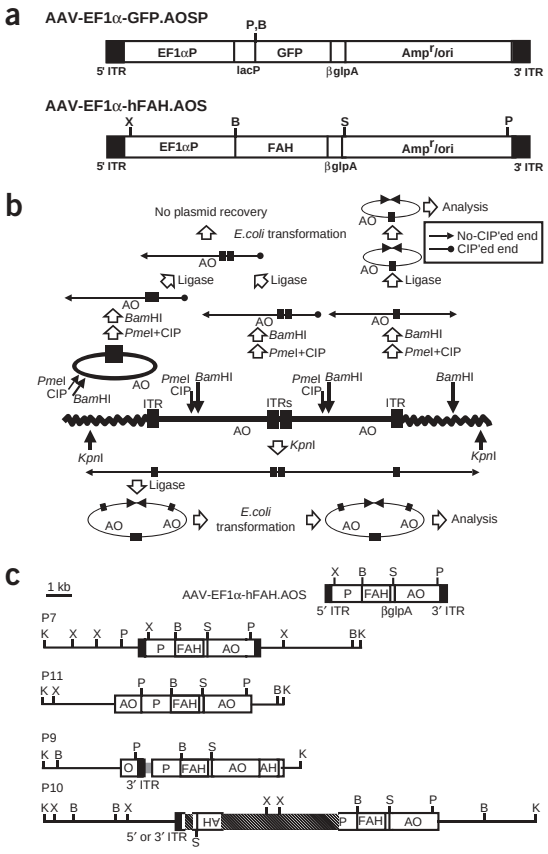
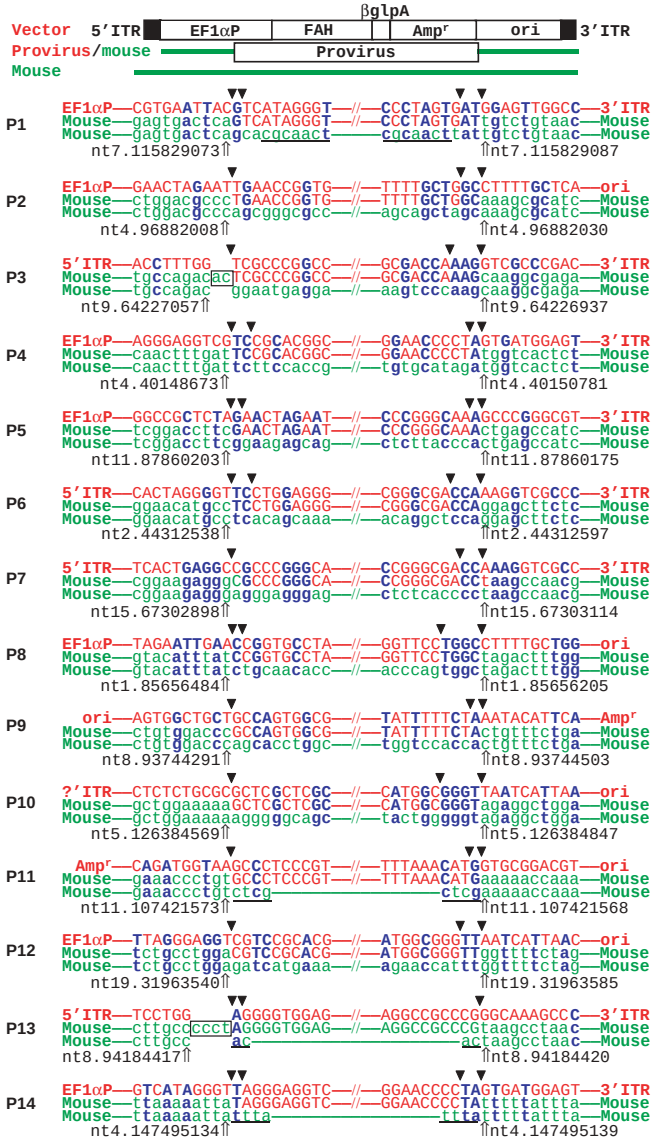


Figure 1 rAAV2 shuttle vectors, strategy for isolating vector-cellular DNA junctions and structures of rAAV2 proviruses isolated from mouse livers. (**a**) rAAV2 vector maps. ITR, AAV2 inverted terminal repeat; EF1αP, the enhancer-promoter of the gene encoding eukaryotic translation elongation factor 1 α1 (*EEF1A1*); lacP, bacterial *lac* operon promoter; βglpA, the poly(A)$^+$ of human β-globin gene; FAH, *FAH* cDNA; B, *Bam*HI; P, *Pme*I; S, *Swa*I; X, *Xba*I. (**b**) Plasmid rescue strategies for isolating proviral genomes from rAAV2-transduced mouse livers. A rAAV2 proviral genome integrated in a head-to-tail tandem array is indicated with straight (vector genome) and zigzag (mouse genome) thick lines. A representative form of extrachromosomal circular monomers, which are abundantly present in non-selected livers but not in *in vivo* selected livers, is indicated with a thick-lined circle. The strategies for isolation of junctions from non-selected hepatocytes transduced with AAV-EF1α-GFP.AOSP and from *in vivo* selected hepatocytes transduced with AAV-EF1α-hFAH.AOS are shown above and below the provirus genome, respectively. Incorporation of a *Pme*I digestion followed by calf intestinal alkaline phosphatase (CIP) treatment greatly facilitated isolation of junctions from non-selected hepatocytes, but only one side of the junctions could be isolated[4]. AO, Amp^r/ori; ITR, AAV2 inverted terminal repeat. (**c**) Structures of proviral genomes isolated from *in vivo* selected hepatocytes transduced with AAV-EF1α-hFAH.AOS. A representative structure of monomer provirus (P7) and three proviral genomes with various complicated structures (P9, P10 and P11) are shown, with a unit length vector genome to the upper right. Thin lines represent cellular DNA sequences. P9 contained a portion of the rAAV2 vector plasmid backbone sequence (shown with a gray box in the provirus). The internal structure of P10 (indicated with hatched boxes) was undetermined owing to the complexity. P, EF1αP; FAH, *FAH* cDNA; AO, Amp^r/ori; AH, a 3′ portion of FAH; O, 3′ portion of AO; B, *Bam*HI; K, *Kpn*I; P, *Pme*I; S, *Swa*I; X, *Xba*I.

[1]Departments of Pediatrics and Genetics, Stanford University School of Medicine, 300 Pasteur Dr. Rm G305A, Stanford, California 94305, USA. Departments of [2]Molecular and Medical Genetics and [3]Pediatrics, Oregon Health & Science University, Portland, Oregon 97201, USA. [4]Present address: Instituto per la cura e la ricerca del Cancro, Candiolo, Torino 10060, Italy. Correspondence should be addressed to M.A.K. (markay@stanford.edu).

Despite recent advances towards understanding the mechanisms of rAAV2 vector transduction *in vivo* (in animals), host chromosomal effects of vector integration and target site selection in quiescent somatic cells in animal tissues are not known, owing to the lack of an efficient system that allows for isolation of the infrequently integrated proviruses from the large number of extrachromosomal vector genomes in transduced tissues. Most previous studies analyzing rAAV2 integration used *in vitro* (in cell culture dishes) systems[9–11], in which genetically unstable cell lines were transduced and clonally selected against a marker gene product. We previously established a plasmid rescue technique to retrieve rAAV2 vector-cellular DNA junction sequences as plasmids in bacteria (**Fig. 1a**,**b** and **Supplementary Note** online) and successfully isolated 18 vector-cellular DNA junctions from rAAV2 vector–injected normal C57BL/6 mouse livers without selection[4]. We reported the cellular sequences flanking the rAAV2 proviral genomes[4] but could not annotate them owing to the lack of satisfactory information of the mouse genome at that time. Now, with the availability of the public mouse genome database[12], the precise integration sites in cellular chromosomes can be determined, to clarify the specificity of integration target site selection.



The bioinformatic information obtained from our previous study allowed for characterization of only one side of the vector-cellular DNA junctions owing to the nature of the plasmid rescue strategy (**Fig. 1b**). Obtaining the whole proviral vector genome together with both the 5′ and 3′ junction sequences is particularly important to establish chromosomal deletions or rearrangements associated with rAAV2 integration in animals. To this end, we used an *in vivo* hepatocyte selection system based on a hereditary tyrosinemia type I (HTI) mouse model. HTI is an inherited fatal metabolic hepatorenal disease caused by deficiency of fumarylacetoacetate hydrolase (FAH; ref. 13). Because there is a selective repopulation of stable genetically modified FAH-positive hepatocytes in HTI mouse livers[14], it allows for *in vivo* clonal selection of hepatocytes with integrated rAAV2 vector genomes expressing FAH[5], resulting in dilution of all the extrachromosomal circular rAAV2 vector genomes by cell division. With injection of a human FAH-expressing rAAV2 vector (**Fig. 1a**) into HTI mouse livers followed by *in vivo* selection, we isolated a total of 14 whole proviral vector genomes with both junction sequences, using a plasmid rescue strategy (**Fig. 1b**). Detailed restriction enzyme mapping and sequencing of proviral genomes identified complicated structures in three cases (**Fig. 1c**) and frequent partial and sometimes complete deletions of the viral ITRs (**Table 1**). By aligning the vector sequences obtained from our past and present studies with the sequences of the integration sites obtained from the mouse genome database, we characterized each integration event (**Fig. 2** and **Table 1**). We identified deletions of cellular genomes at all the integration sites (14 of 14 integrations, 100%), ranging from 2 bp to ~0.3 kb in most cases. A deletion of 2.1 kb occurred in an integration event, but there was no large rearrangement or translocation of chromosomes. We observed nucleotide insertions of 1–4 bp of unknown origin in three cases. We found no important homologies between vector and cellular DNA but frequently found patch homologies up to 4 bp around the integration sites. The number of nucleotides shared by both vector and cellular DNA sequences within the 10-bp region inside from the breakpoints was significantly higher than the expected number calculated from a random model (a two-tailed binomial test, $P = 0.0002$), confirming that rAAV2 integration was influenced by microhomology. The G/C contents in a 200-bp window around each breakpoint ranged from 33% to 61% (average 48%, s.d. 7%), showing no general trend compared with the mean G/C content of the mouse genome (42%; ref. 12).

With the information from a total of 29 integration sites identified in rAAV2-injected mouse livers without selection[4] and *in vivo*-selected rAAV2-transduced hepatocytes, we began to elucidate how rAAV2 vectors might select target sites for integration in genetically stable somatic cells in animals. The results are summarized in **Table 1**.

Figure 2 Sequences of rAAV2 vector-mouse cellular DNA junctions. Sequences of rAAV2 vector (top line), provirus with flanking mouse genomic DNA (middle line) and the mouse genome (bottom line) around 5′ and 3′ vector-cellular DNA junctions are aligned. Red upper-case letters represent vector genome and green lower-case letters represent mouse genome. The locations of each junction are indicated with an arrowhead and nucleotide positions in the mouse genome (nt; the first number indicates a chromosome number, followed by a nucleotide position number obtained from the NCBI database). Blue letters indicate nucleotides shared by vector and mouse genomic sequences. Nucleotides in a box in P3 and P13 show a nucleotide insertion at the junction. Two arrowheads at a junction indicate that the breakpoint should be located between the arrowheads but the exact location cannot be determined because of microhomology. Underlined sequences show an overlap between 5′ and 3′ junctions. The origin of the left ITR of P10 was not determined.

Table 1 Structures of rAAV2 proviral genomes and host chromosomal effects

ID	Selection[a]	Provirus			Chromosome			Microhomology[d]	G/C content (%) at breakpoints[e]		Target
		Structure	Deletion (bp)[b]		Number/ band	Deletion (bp)	Insertion[c] (bp)				
			5′ junction	3′ junction					5′	3′	
P1	Yes	Monomer	Δ173	Δ129 (ITR)	7 / F3	Δ13	0	8 / 20	45	47	Gene, intron, reversed
P2	Yes	Monomer	Δ221	Δ230	4 / C6	Δ21	0	7 / 20	52	53	Intergenic
P3	Yes	Monomer	Δ77 (ITR, flop)	Δ71 (ITR, flop)	9 / C	Δ119	+2	7 / 20	55	52	Gene, intron, reversed
P4	Yes	Monomer	Δ194	Δ134 (ITR)	4 / A5	Δ2107	0	6 / 20	39	38	2 genes:(i) complete deletion, forward; (ii) intron & exon, reversed
P5	Yes	Monomer	Δ211	Δ72 (ITR, flip)	11 / C	Δ27	0	9 / 20	48	48	Gene, intron, reversed
P6	Yes	Monomer	Δ141 (ITR)	Δ74 (ITR, flop)	2 / C1	Δ58	0	7 / 20	50	51	Gene, intron, forward
P7	Yes	Monomer	Δ41 (ITR, flop)	Δ75 (ITR, flop)	15 / D3	Δ215	0	7 / 20	61	51	Intergenic (hit a Uni Gene cluster[i])
P8	Yes	Monomer	Δ225	Δ241	1 / C5	Δ278	0	7 / 20	46	48	Gene, intron, reversed
P9	Yes	Complicated	Δ4399	Δ1972	8 / C5	Δ211	0	7/ 20	51	49	Intergenic (hit a Uni Gene cluster[i])
P10	Yes	Complicated	Δ21 (ITR, flip)[f]	Δ160	5 / F	Δ277	0	9 / 20	49	57	Gene, intron, forward/reversed
P11	Yes	Complicated	Δ3882	Δ190	11 / E1	Δ4	0	3 / 8	43	44	Gene, intron, forward
P12	Yes	Monomer	Δ192	Δ159	19 / C3	Δ44	0	7 / 20	38	35	Intergenic
P13	Yes	Monomer	Δ147	Δ78 (ITR, flip)	8 / C5	Δ2	+4	1 / 4	49	49	Intergenic
P14	Yes	Monomer	Δ183	Δ134 (ITR)	4 / E2	Δ4	0	3/ 8	33	33	Gene, intron, forward
J16	No	ND		Δ141 (ITR)[k]	11 / B1		(0)	4 / 10		43	Gene, intron, forward
J104	No	Complicated[k]	Δ557[k]		11 / A1		(0)	1 / 10	54		Gene, intron, forward/reversed
J121	No	ND		Δ114 (ITR)[k]	10 / C1		(0)[k]	2 / 10		57	Gene, intron, reversed[k]
J134	No	ND		Δ76 (ITR, flop)[k]	6 / E2		(0)	4 / 10		39	Intergenic
J166	No	ND		Δ135 (ITR)[k]	14 / E4		(0)	6 / 10		39	Gene, exon, reversed
J175	No	Complicated[k]		Δ2 (ITR, flip)[k]	4 / E2		(0)	3 / 10		55	Gene, intron, forward
J192	No	ND		Δ78 (ITR, flop)[k]	NA[g]		(0)[k]	2 / 10		51	Gene, transcribed region, reversed[k]
J216	No	ND		Δ118 (ITR)[k]	6 / C3		(+1)	3 / 10		50	Intergenic
J236	No	ND		Δ124 (ITR)[k]	11 / E1		(0)	3 / 10		49	Gene, intron, reversed
J270	No	ND		Δ175[k]	7 / A3		(0)	5 / 10		54	Gene, exon, forward
J278	No	ND		Δ77 (ITR, flip)[k]	NA[g]		(0)	3 / 10		50	Intergenic
J288	No	ND		Δ74 (ITR, flop)[k]	4 / B1		(0)	3 / 10		36	Gene, intron, reversed
J299	No	ND		Δ106 (ITR)[k]	14 / B		(0)	3 / 10		52	Gene, intron, forward
J305	No	ND		Δ107 (ITR)[k]	15 / E1		(0)	4 / 10		54	Gene, intron, reversed
J313	No	ND		Δ106 (ITR)[k]	2 / F3		(0)	2 / 10		55	Gene, exon, reversed
Total (mean ± s.d.)								136 / 400[h]		(48 ± 7%)	Selection (Yes), hit genes: 9/14[j]
											Selection (No), hit genes: 12/15[j]
											Total, hit genes: 21/29

[a]Proviral genomes were isolated from *in vivo* selected HTI mouse hepatocytes (Yes) or from C57BL/6 mouse hepatocytes without selection (No). [b]The number of nucleotides that were deleted at each end of the vector genome is shown. (ITR, flop) and (ITR, flip) indicate that a portion of the ITR sequence remained at the junction and was identified with flip or flop orientation. (ITR) indicates that a portion of the ITR sequence remained, but the ITR orientation could not be determined because the length of the ITR remnant was less than 64 bp. [c]The number in parentheses represents nucleotide insertion at only one junction; therefore, it does not necessarily mean the actual amount of nucleotide insertion at each integration site. [d]The number of nucleotides that were shared by both vector and cellular DNA sequences within a 10-bp stretch inside of each breakpoint (the deleted side of the cellular sequences) was counted. When information for both 5′ and 3′ junctions were available, we combined them. [e]G/C contents of a 200-bp window around 5′ and 3′ breakpoints of cellular genomes are shown. [f]This may be another 3′ junction, undetermined because of complicated head-to-head provirus structure. [g]NA, not applicable because of the sequence redundancy of the target region in the mouse genome. J192 and J278 targeted the 45s pre rRNA gene and its intergenic spacer, respectively. [h]The number of shared nucleotides is higher than expected with a statistical significance (a two-tailed binomial test, $P = 0.0002$). [i]When rAAV2 integration targeted an EST that belonged to a UniGene cluster but was not identified as a gene by either the NCBI Map Viewer or the Ensembl browsers, we separately described it in parentheses. [j]The P values against a random integration model calculated by a two-tailed binomial test are $P = 0.0001$ and $P = 0.002$ under non-selective and selective conditions, respectively, with a predicted probability of hitting a gene as 0.25, and $P = 0.003$ and $P = 0.1$ under non-selective and selective conditions, respectively, with a probability of hitting a gene as 0.41. A comparison of frequency of hitting a gene by χ^2 test and Fisher's exact probability test showed no statistical difference between non-selective and selective conditions ($\chi^2 = 1.451 < \chi_1^2(0.05) = 3.841$; Fisher's two-tailed probability, $P = 0.385 > 0.05$). [k]These data have been published previously[4]. Three of 18 junction sequences in the previous study[4] did not match with sequences in the mouse genome database. ND, not determined.

Although the integration sites seemed to be distributed on mouse chromosomes with no significant bias, rAAV2 integration was favored in genes regardless of whether they were analyzed in normal liver or after selective repopulation *in vivo* (12 of 15 (80%) under a non selective condition; 9 of 14 (64%) under a selective condition; 21 of 29 (72%) overall frequency). This bias was statistically significant (a two-tailed binomial test, $P = 0.0000001$ and $P = 0.0006$, with a predicted probability of hitting a gene as 0.25 and 0.41, respectively; see **Supplementary Note** online for details). Both exons and introns were disrupted with rAAV2 integrations, with a higher incidence in introns. There seemed to be no bias for orientation of rAAV2 proviral genomes relative to gene transcription. Notably, web-based public databases[15,16] and our RT–PCR analysis confirmed that 20 of 20 (100%) target genes that we analyzed were expressed in the liver, and expression of approximately half of these genes was upregulated in the liver (**Table 2** and **Fig. 3**).

Although the number of integration events analyzed was relatively small and the results may be somewhat biased by the procedures for provirus isolation, our study showed that rAAV2 preferentially integrated into active genes when delivered directly

Table 2 Expression of rAAV2-targeted genes in the liver

ID	Targeted gene		Expression in the liver						
	Name	UniGene ID (other IDs[a])	Web-based database			RT–PCR[d]			Summary[e]
			Liver as a cDNA source	READ[b]	Gene Expression Atlas[c] (median)	BALB/c	C57BL/6	HTI	
P1	RIKEN cDNA 2410027J01	Mm.26928	Yes	NI	NI	3.92	3.95	4.24	Expressed (Up)
P3	LOC208011	(XM164986)	NI	NI	NI				ND
P4	RIKEN cDNA 2010003O02	Mm. 1103	Yes	0.345	548 (~740)				Expressed
P4	Expressed sequence AW105885	Mm.39006	Yes	NI	NI	0.88	0.92	0.80	Expressed
P5	RIKEN cDNA 1200011M11	Mm.23257	NI	0.654	NI	0.39	0.43	0.56	Expressed
P6	Kynureninase (L-kynurenine hydrolase)	Mm.105278	Yes	NI	NI	Upregulated	Upregulated	Upregulated	Expressed (Up)
P8	DNA segment, Chr 1, ERATO Doi 757, expressed	Mm.27888	NI	0.616 1.261	NI				Expressed (Up)
P10	Epimorphin	Mm.3003	Yes	NI	20 (20)				Expressed
P11	Testis expressed gene 2	Mm.245663	Yes	0.429 1.260	NI				Expressed (Up)
P14	RIKEN cDNA 1300002F13	Mm.21679	Yes	NI	3985 (~200)				Expressed (Up)
J16	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 9	Mm.11360	NI	NI	74 (~80)	0.31	0.20	0.53	Expressed
J104	Oxoglutarate dehydrogenase (lipoamide)	Mm.30074	Yes	0.413	707.5 (~1100)				Expressed
J121	Procollagen, type XVIII, alpha 1	Mm.4352	Yes	0.602	2574.5 (20) 691.5 (20)				Expressed (Up)
J166	Similar to Eukaryotic translation initiation factor 4B (eIF-4B)	(XM139255)	NI	NI	NI				ND
J175	Period homolog 3 (*Drosophila*)	Mm.10723	NI	NI	1641.5 (~950)				Expressed
J192	45s pre rRNA, 28s rRNA transcribed region	(X82564)	Known to be universally expressed						Expressed
J236	ATP-binding cassette, sub-family A (ABC1), member 8a	Mm.138955	Yes	1.669	NI				Expressed (Up)
J270	RIKEN cDNA 1700023M09	Mm.41511	Yes	0.277	213.5 (~250)				Expressed
J288	Aldolase 2, B isoform	Mm.218862	Yes	0.258	5634 (~400)				Expressed (Up)
J299	RIKEN cDNA A130034K24	Mm.212365	NI	NI	306 (~450)				Expressed
J305	RIKEN cDNA 1810044A24	Mm.31995	NI	0.488	NI				Expressed
J313	Inosine triphosphatase (nucleoside triphosphate pyrophosphatase)	Mm.21399	Yes	–0.265	133.5 (~190)				Expressed

[a]In the case where rAAV2-targeted transcripts do not belong to UniGene clusters, other available IDs are listed in parentheses: XM164986 and XM139255 (RefSeq) and X82564 (GenBank accession number). [b]See ref. 15. Values are log-transformed ratios. Negative values indicate downregulated and positive values indicate upregulated expression, compared to a reference RNA sample (mRNA from E17.5 embryos). [c]Gene Expression Atlas of Genomics Institute of the Novartis Research Foundation[16]. Values are average differences. Median values are indicated in parentheses. [d]Values are ratios of the amount of RT–PCR product from a target gene to that from *Gapd*, normalized with the ratio from a universal reference RNA. Values <1 and 1> represent downregulated and upregulated expression of the target gene, respectively, compared to the reference RNA sample normalized with *Gapd* expression. In one case, upregulation of a target gene transcript was obvious without coamplification with *Gapd* transcript, indicated as 'Upregulated'. [e]Target genes that are apparently upregulated in the liver are indicated with 'Up'.

Two target UniGene clusters that NCBI Map Viewer or the Ensembl browser do not consider as genes (Mm.12505 in P7 and Mm.136889 in P9) are expressed in the liver based on the fact that the ESTs were retrieved from mouse livers. ND, not determined; NI, no information is available in web-based public databases.

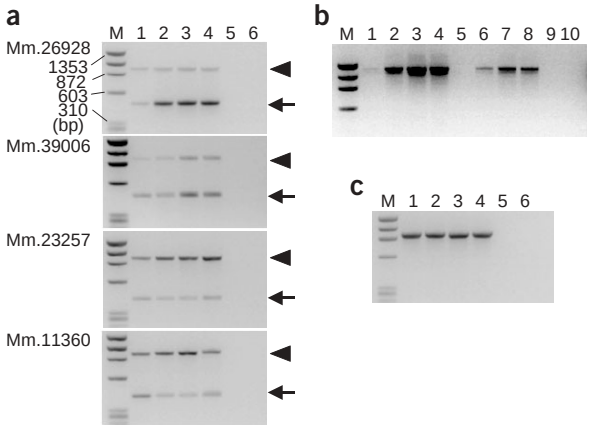


Figure 3 RT–PCR analysis for expression of rAAV2-targeted genes in mouse livers. Total liver RNA was extracted from adult BALB/c, C57BL/6 and HTI mouse strains, and expression of each targeted gene was analyzed by RT–PCR. We separated 15 μl of the RT–PCR products on a 2.0% agarose gel and stained it with ethidium bromide. (**a**) Coamplification of each targeted gene transcript and *Gapd* transcript. Lane 1, universal reference RNA; lane 2, BALB/c mouse liver RNA; lane 3, C57BL/6 mouse liver RNA; lane 4, HTI mouse liver RNA; lane 5, RT-minus negative control containing 0.1 μg universal reference RNA; lane 6, template-minus negative control. The positions of target transcript and *Gapd* transcript are indicated with an arrow and an arrowhead, respectively. (**b**) RT–PCR amplification of Mm.105278. Lanes 1 and 5, universal reference RNA; lanes 2 and 6, BALB/c; lanes 3 and 7, C57BL/6; lanes 4 and 8, HTI mouse; lane 9, RT-minus negative control containing 0.5 μg universal reference RNA; lane 10, template-minus negative control. RT products corresponding to 0.5 μg and 0.05 μg RNA are used for lanes 1–4 and lanes 5–8, respectively. (**c**) RT–PCR amplification of *Gapd* transcript. Lanes are the same as in **a**. M, *Hae*III-digested ΦX 174 DNA fragments. The gel images are inverted.

into experimental animals, regardless of how the proviruses were isolated. Similar unpredictable imperfect structures of rAAV2 proviral genomes and chromosomal deletions at integration sites in G418-selected, genetically unstable HeLa cells were recently reported[11], but there was no statistical analyses to establish if there was preference for integration into intragenic regions. There is considerable evidence that target site selection by retroviruses and retrotransposons is non-random. Although there are several conflicting results, local DNA structures and surrounding environment including chromatin structures and transcription factors influence the choice for a target site[17–24]. It has been recently shown that HIV-1 selectively integrates into active genes[25]. It is important to note that retroelements and rAAV2 proviral sequences integrate by different mechanisms. Retroelements use their encoded integrase to catalyze integration, whereas rAAV2 vector DNA integration is totally dependent on host cellular proteins. Thus the possibility remains that rAAV2 vectors preferentially integrate into chromosomal regions that are already broken[11]. Nevertheless, the preferred integration into active genes may be a common propensity of certain kinds of integrating elements including rAAV2. Although rAAV2 vectors integrate at a low efficiency, the current results will need to be considered in risk/benefit considerations until the consequences for vector integration are more fully understood.

METHODS

rAAV2 shuttle vectors. We produced rAAV2 shuttle vectors AAV-EF1α-GFP.AOSP and AAV-EF1α-hFAH.AOS (**Fig. 1a**) based on plasmids pAAV-EF1α-GFP.AOSP and pAAV-EF1α-hFAH.AOS2 as described in **Supplementary Note** online. Both vectors carried the bacterial gene encoding β-lactamase (ampicillin resistance gene or Amp^r) and the ColE1 plasmid origin of replication (ori), allowing for retrieval of vector genome sequences in bacteria.

Strains of mice and animal husbandry. All the animal experiments were done according to the guidelines for animal care at Stanford University and Oregon Health & Science University. We purchased female C57BL/6 mice 6–8 weeks old from Jackson Laboratory. HTI mice were the FAH$^{\Delta exon5}$ strain previously described[13] and inbred at the Department of Animal Care, Oregon Health & Science University. We gave HTI mice drinking water containing 2-(2-nitro-4-trifluoro-methylbenzoyl)-1,3-cyclohexanedione (NTBC; Swedish Orphan AB) at a concentration of 7.5 mg l^{-1}. FAH-negative hepatocytes accumulate the toxic metabolite fumarylacetoacetate (FAA) and die in a cell-autonomous manner, but oral administration of NTBC reduces FAA, allowing normal hepatic function and preventing hepatocellular damage. For *in vivo* selection of HTI hepatocytes with integrated rAAV2 vector genomes, we withdrew NTBC from drinking water.

Portal vein injection, *in vivo* selection and hepatocyte transplantation. We carried out portal vein injection of AAV-EF1α-GFP.AOSP into C57BL/6 mice and sample collection as previously described[4]. We injected adult male HTI mice on NTBC with 3.0×10^{11} particles of AAV-EF1α-hFAH.AOS into the portal vein ($n = 8$) and then divided the mice into two groups ($n = 4$ each). After being kept on NTBC for six weeks (enough time to establish stable hepatocyte transduction with rAAV2), we withdrew NTBC from the mice in Group 1, but continued to give the mice in Group 2 water containing NTBC for an additional eight weeks. To further select for integrated vector genomes and dilute non-integrated vector genomes, after eight weeks with (Group 1) or without (Group 2) *in vivo* selection (14 weeks after injection), we isolated hepatocytes from vector-injected mice by a two-step collagenase perfusion and injected one million hepatocytes in 100 μl of appropriate medium into the portal vein of recipient HTI mice on NTBC as previously described[14]. We used liver DNA from a Group 1 recipient isolated after a 7-month *in vivo* selection to isolate rAAV2 proviral genomes. The outcomes of hepatocyte transplantation into recipient HTI mice in Groups 1 and 2 are summarized in **Supplementary Note** online.

Isolation of proviruses and mouse genomes around integration sites. Vector-cellular DNA junctions from AAV-EF1α-GFP.AOSP-injected C57BL/6 mouse livers and detailed procedures for the isolation were previously reported[4] and are concisely explained in **Figure 1b**. The strategy for isolating whole proviral vector genomes together with 5′ and 3′ vector-cellular DNA junctions from AAV-EF1α-hFAH.AOS-transduced HTI mouse hepatocytes was basically the same as our previously published method with minor modifications[26]. The detailed procedures are found in **Supplementary Note** online.

Construction of restriction enzyme maps of isolated plasmids containing the whole proviral genome. We digested each rescued plasmid with *Kpn*I, *Xba*I, *Pme*I or *Bam*HI, alone or in any possible combination, to draw draft maps (see **Fig. 1a**). A *Kpn*I site does not exist in the vector but must reside only once in each rescued plasmid. Combining the sequence information, we identified that 11 of 14 proviral genomes resulted from rAAV2 monomer integration with various terminal deletions of the vector genome. The remaining three proviral genomes required additional restriction enzyme digestion and sequencing of subcloned fragments in pBluescript II KS$^-$. When we obtained plasmid clones that were identical based on their restriction maps and sequencing data, we considered those as a single integration event, not individual different integration events.

Sequencing of junctions. We carried out sequencing using an ABI PRISM 377 DNA Sequencer (PE Applied Biosystems). The detailed methods for sequencing are available in **Supplementary Note** online.

Bioinformatics. Isolated mouse cellular DNA sequences were BLAST searched against the public mouse genome database through the National Center for Biotechnology Information (NCBI) and Ensembl browsers. We did a targeted gene search based on the chromosomal localization of each integration site using NCBI Map Viewer and Mouse Contig View of Ensembl Mouse Genome

Browser. We assessed the transcriptional activity of each targeted gene using web-based public databases and browsers: NCBI UniGene, SOURCE[27], READ[15] and Gene Expression Atlas of Genomics Institute of the Novartis Research Foundation (GNF)[16]. Additional information about the bioinformatic analyses is available in **Supplementary Note** online.

RT–PCR of rAAV2-tageted gene transcripts. Among 22 rAAV2-targeted genes, we analyzed expression of 5 genes by RT–PCR. We extracted total liver RNA from an adult female C57BL/6 mouse and an adult male HTI mouse. We purchased total liver RNA from BALB/c mice and a mouse universal reference total RNA from Clontech. We coamplified each target transcript and the gene encoding glyceraldehyde-3-phosphate dehydrogenase (*Gapd*) in the same tube. We calculated the ratio of the PCR product from a target transcript to that from *Gapd* for each sample and normalized it to the ratio from the reference RNA. Thus, the normalized ratio of the universal reference RNA is always 1.0 regardless of the target transcripts, and the normalized ratio of each target transcript in samples can be used to assess as an increase (the ratio >1.0) or a decrease (the ratio <1.0) in the amount of the target transcript in the liver compared to that in the reference sample. Additional information about the RT–PCR analysis is available in **Supplementary Note** online.

Statistics. We assessed the statistical significance of the bias for or against preferential integration into genes by a two-tailed binomial test. In humans, transcription units are estimated to account for about 25–33% of the genome[28,29]. Based on the difference in the length of the genome and the number of transcripts between the human and the mouse, we estimated the proportion of transcription units to the whole genome (gene density) in the mouse as 0.25–0.41 (see **Supplementary Note** online). We assessed the influence of the presence of *in vivo* selective pressure on integration target site selection by rAAV2 using the χ^2 test and Fisher's exact probability test. We excluded from the analyses two integrations that targeted the 45s pre rRNA gene (J192) and the 45s pre rRNA gene intergenic spacer (J278) because of the redundancy of the ribosomal RNA genes in the genome. The number of shared nucleotides between aligned rAAV2 vector and mouse genomic sequences is one of the indicators for microhomologies. We tested a null hypothesis that there is no bias for or against base sharing by a two-tailed binominal test (see **Supplementary Note** online).

URLs. NCBI mouse genome BLAST search, http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html; NCBI Map Viewer, http://www.ncbi.nlm.nih.gov/mapview/static/MVstart.html; Ensembl mouse genome database, http://www.ensembl.org/Mus_musculus; NCBI UniGene http://www.ncbi.nlm.nih.gov/UniGene; SOURCE, http://source.stanford.edu/; READ http://read.gsc.riken.go.jp/; Gene Expression Atlas of Genomics Institute of the Novartis Research Foundation, http://expression.gnf.org/cgi-bin/index.cgi.

Accession numbers. The five genes that we analyzed by RT–PCR are UniGene cluster IDs Mm.26928, Mm.39006, Mm.23257, Mm.105278 and Mm.11360.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank J. Park for data analysis. This work was supported by grants from the National Heart, Lung, and Blood Institute of the US National Institutes of Health to M.A.K.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 15 January; accepted 16 May 2003
Published online 1 June 2003; doi:10.1038/ng1179

1. Kay, M.A. *et al.* Evidence for gene transfer and expression of factor IX in haemophilia B patients treated with an AAV vector. *Nat. Genet.* **24**, 257–261 (2000).
2. Nakai, H. *et al.* Extrachromosomal recombinant adeno-associated virus vector genomes are primarily responsible for stable liver transduction *in vivo*. *J. Virol.* **75**, 6969–6976 (2001).
3. Miao, C.H. *et al.* The kinetics of rAAV integration in the liver. *Nat. Genet.* **19**, 13–15 (1998).
4. Nakai, H., Iwaki, Y., Kay, M.A. & Couto, L.B. Isolation of recombinant adeno-associated virus vector-cellular DNA junctions from mouse liver. *J. Virol.* **73**, 5438–5447 (1999).
5. Chen, S.J., Tazelaar, J., Moscioni, A.D. & Wilson, J.M. *In vivo* selection of hepatocytes transduced with adeno-associated viral vectors. *Mol. Ther.* **1**, 414–422 (2000).
6. Li, Z. *et al.* Murine leukemia induced by retroviral gene marking. *Science* **296**, 497 (2002).
7. Marshall, E. Clinical research. Gene therapy a suspect in leukemia-like disease. *Science* **298**, 34–35 (2002).
8. Marshall, E. Gene therapy. Second child in French trial is found to have leukemia. *Science* **299**, 320 (2003).
9. Rutledge, E.A. & Russell, D.W. Adeno-associated virus vector integration junctions. *J. Virol.* **71**, 8429–8436 (1997).
10. Yang, C.C. *et al.* Cellular recombination pathways and viral terminal repeat hairpin structures are sufficient for adeno-associated virus integration *in vivo* and *in vitro*. *J. Virol.* **71**, 9231–9247 (1997).
11. Miller, D.G., Rutledge, E.A. & Russell, D.W. Chromosomal effects of adeno-associated virus vector integration. *Nat. Genet.* **30**, 147–148 (2002).
12. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
13. Grompe, M. *et al.* Loss of fumarylacetoacetate hydrolase is responsible for the neonatal hepatic dysfunction phenotype of lethal albino mice. *Genes Dev.* **7**, 2298–2307 (1993).
14. Overturf, K. *et al.* Hepatocytes corrected by gene therapy are selected *in vivo* in a murine model of hereditary tyrosinaemia type I. *Nat. Genet.* **12**, 266–273 (1996).
15. Miki, R. *et al.* Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci. USA* **98**, 2199–2204 (2001).
16. Su, A.I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* **99**, 4465–4470 (2002).
17. Vijaya, S., Steffen, D.L. & Robinson, H.L. Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. *J. Virol.* **60**, 683–692 (1986).
18. Sandmeyer, S.B., Hansen, L.J. & Chalker, D.L. Integration specificity of retrotransposons and retroviruses. *Annu. Rev. Genet.* **24**, 491–518 (1990).
19. Scherdin, U., Rhodes, K. & Breindl, M. Transcriptionally active genome regions are preferred targets for retrovirus integration. *J. Virol.* **64**, 907–912 (1990).
20. Chalker, D.L. & Sandmeyer, S.B. Ty3 integrates within the region of RNA polymerase III transcription initiation. *Genes Dev.* **6**, 117–128 (1992).
21. Pryciak, P.M. & Varmus, H.E. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **69**, 769–780 (1992).
22. Muller, H.P. & Varmus, H.E. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13**, 4704–4714 (1994).
23. Pruss, D., Bushman, F.D. & Wolffe, A.P. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl. Acad. Sci. USA* **91**, 5913–5917 (1994).
24. Leclercq, I. *et al.* Host sequences flanking the human T-cell leukemia virus type 1 provirus *in vivo*. *J. Virol.* **74**, 2305–2312 (2000).
25. Schroder, A.R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
26. Nakai, H. *et al.* Helper-independent and AAV-ITR-independent chromosomal integration of double-stranded linear DNA vectors in mice. *Mol. Ther.* **7**, 101–111 (2003).
27. Diehn, M. *et al.* SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.* **31**, 219–223 (2003).
28. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
29. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).