OPEN ACCESS Freely available online

PLoS PATHOGENS

Integration Preferences of Wildtype AAV-2 for Consensus Rep-Binding Sites at Numerous Loci in the Human Genome

Daniela Hüser[1], Andreas Gogol-Döring[2], Timo Lutter[1], Stefan Weger[1], Kerstin Winter[1], Eva-Maria Hammer[1], Toni Cathomen[1,3], Knut Reinert[2], Regine Heilbronn[1]*

1 Institute of Virology, Campus Benjamin Franklin, Charité-Universitätsmedizin Berlin, Berlin, Germany, **2** Institute for Computer Science, Freie Universität Berlin, Berlin, Germany, **3** Department of Experimental Hematology, Hannover Medical School, Hannover, Germany

Abstract

Adeno-associated virus type 2 (AAV) is known to establish latency by preferential integration in human chromosome 19q13.42. The AAV non-structural protein Rep appears to target a site called AAVS1 by simultaneously binding to Rep-binding sites (RBS) present on the AAV genome and within AAVS1. In the absence of Rep, as is the case with AAV vectors, chromosomal integration is rare and random. For a genome-wide survey of wildtype AAV integration a linker-selection-mediated (LSM)-PCR strategy was designed to retrieve AAV-chromosomal junctions. DNA sequence determination revealed wildtype AAV integration sites scattered over the entire human genome. The bioinformatic analysis of these integration sites compared to those of *rep*-deficient AAV vectors revealed a highly significant overrepresentation of integration events near to consensus RBS. Integration hotspots included AAVS1 with 10% of total events. Novel hotspots near consensus RBS were identified on chromosome 5p13.3 denoted AAVS2 and on chromsome 3p24.3 denoted AAVS3. AAVS2 displayed seven independent junctions clustered within only 14 bp of a consensus RBS which proved to bind Rep *in vitro* similar to the RBS in AAVS3. Expression of Rep in the presence of *rep*-deficient AAV vectors shifted targeting preferences from random integration back to the neighbourhood of consensus RBS at hotspots and numerous additional sites in the human genome. In summary, targeted AAV integration is not as specific for AAVS1 as previously assumed. Rather, Rep targets AAV to integrate into open chromatin regions in the reach of various, consensus RBS homologues in the human genome.

Citation: Hüser D, Gogol-Döring A, Lutter T, Weger S, Winter K, et al. (2010) Integration Preferences of Wildtype AAV-2 for Consensus Rep-Binding Sites at Numerous Loci in the Human Genome. PLoS Pathog 6(7): e1000985. doi:10.1371/journal.ppat.1000985

Editor: Bryan R. Cullen, Duke University Medical Center, United States of America

Received December 24, 2009; **Accepted** June 3, 2010; **Published** July 8, 2010

Funding: The study was supported by intramural funds of Charité Medical School, Berlin to RH and by grant PERSIST (no. 222878) of the European Commission 7th Framework Program to TC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: RH is an inventor of patents related to rAAV technology and owns equity in a gene therapy company that is commercializing AAV for gene therapy applications.

* E-mail: regine.heilbronn@charite.de

Introduction

The family of adeno-associated virus (AAV) represents defective, helper-dependent viruses that need to establish latency to ensure persistence in their primate hosts [1]. Upon natural infections in humans AAV genomes were shown to persist mainly as episomes and integrated AAV genomes were rarely detected [2]. The molecular mechanisms leading to integration have only been characterized for AAV type 2 that prefers integration near a site on human chromosome 19q13.42, called AAVS1 [3]. The specificity of AAV integration is mediated by the large regulatory AAV proteins, Rep78/68 [4]. During productive AAV replication in the presence of either adeno- or herpesvirus as a helper virus, Rep78/68 is required for AAV gene expression and DNA replication. The AAV origins of DNA replication reside in the 145 bp inverted terminal repeats (ITRs) that flank the 4.7 kb single-stranded AAV genome. Rep78 and/or Rep68 are expressed from the AAV p5 promoter and were shown to bind to the Rep-binding site (RBS) within the AAV-ITRs [5]. Rep unwinds the DNA and introduces a single-strand nick at the adjacent terminal resolution site (*trs*) [6]. The AAV-ITRs also serve as *cis* elements for chromosomal integration [4]. A RBS homologue present in the AAV p5 promoter was shown to mediate AAV integration in the absence of the ITRs [7]. DNA sequences homologous to the RBS and a nearby *trs* element were also found in AAVS1 [8,9] and, *in vitro*, ternary complex formation of Rep68 with the AAV-ITR and AAVS1 was shown [10]. A 33 bp sequence of AAVS1 spanning the RBS and the *trs* element was sufficient to mediate AAV integration *in vivo* [4,11]. AAV integrated at variable distances from the RBS in AAVS1 and sequence rearrangements were frequently found at AAV-chromosome junctions [8,9,12,13, 14,15]. Quantitative real-time PCR analysis of AAVS1-specific AAV-2 integration within hours after AAV-2 infection and at increasing MOIs showed that 10 to 20% of infected cells displayed AAV integration within a 4 kb region of AAVS1 on chromosome 19q13.42 [16,17]. In AAV-infected and subsequently selected cell clones up to 80% of AAVS1-specific integration had been described before [18].

Although AAV has not been associated with disease in humans, it is well established that AAV Rep78/68 induces DNA damage, cell cycle arrest [19] and apoptosis [20]. In addition, AAV Rep interferes with helper adenovirus- [21] herpes simplex virus

Author Summary

This is the first unbiased genome-wide analysis of wildtype AAV integration combined with a thorough bioinformatic analysis of preferred genomic motifs and patterns in the neighbourhood of the integration sites identified. The preference of Rep-dependent AAV integration near multiple consensus Rep-binding sites was lost in the case of AAV vector integration in the absence of Rep expression. Our findings challenge the commonly accepted notion of site-specific AAV targeting to AAVS1 on chromosome 19q13.42. Although AAVS1 contains a canonical Rep-binding site, numerous additional sites including the newly identified hotspots AAVS2 on chromosome 5p13.3 and AAVS3 on chromosome 3p24.3 harbour functional Rep-binding sites suitable for AAV integration. AAV vectors are quickly moving forward in the clinic and Rep-dependent vector targeting strategies are being actively pursued. Detailed information of AAV wildtype versus recombinant AAV vector integration sites and preferences are needed to evaluate the safety profile of AAV vectors in gene therapy.

replication [22]. AAV holds much promise as a vector for gene therapy. As a rule, recombinant AAV vectors persist as non-integrated, nuclear episomes. AAV vectors lack the integration promoting *rep* gene and therefore only occasionally integrate into the host cell genome. The preferred integration of wildtype AAV-2 in chromosome 19q13.42 is unique and is commonly viewed as a specifically evolved virus-encoded targeting mechanism. Multiple attempts were published that aim to exploit Rep-mediated targeting specificity for chromosome 19q13.42 for the specific integration of gene therapy vectors [23,24,25,26,27,28]. Yet chromosome 19q13.42 is not the only target region. The presence of alternative integration sites has long been postulated and *in silico* analysis detected numerous consensus Rep-binding sites in the human genome. Many of these bound Rep *in vitro* [29] but their *in vivo* accessibility for AAV integration has not been explored so far. From an evolutionary standpoint the assumption that AAV latency is ensured by more than one target site or mechanism appeared reasonable.

This study was designed to close the knowledge gap between AAVS1-specific and assumedly non-AAVS1-specific wildtype AAV integration and to compare the identified genomic sites to those preferred upon AAV vector transduction. An open survey of chromosomal integration preferences for wildtype AAV-2 was conducted and complemented by the bioinformatic analysis of genomic motifs and patterns in the genomic regions surrounding the integration loci.

Results

General strategy of LSM-PCR

The genomic structure of latent AAV in infected cells is highly variable. Wildtype AAV-2 was shown to integrate into the host cell genome, as well as persist as extrachromosomal, nuclear episomes [2,30]. In either case multicopy, concatemeric structures predominate and often lead to unpredictable rearrangements involving the 145 bp inverted terminal repeats (ITRs). Therefore the retrieval of AAV-chromosome junctions suffers from the inherent problem of inefficient PCR reads through the hairpin ITR into the adjacent chromosomal sequences. This leads to a predominance of rearranged AAV genomes lacking chromosomal junctions in previous PCR-based studies [31,32,33]. Furthermore, previously cloned junctions often displayed unknown intervening sequences of varying lengths between AAV and the identified chromosomal sequence [12,15,16,27,34,35,36]. Therefore, unambiguous assignment of the AAV-derived and chromosome-derived parts of junctions requires sufficient DNA sequence lengths.

Several methods to identify virus-chromosome junctions have been developed to study retrovirus integration, where generally a single proviral copy per chromosomal site is found [37,38]. The ultimate structure of the integrated long terminal repeat (LTR) is generally predictable in a way that allows an integration-specific PCR design. Linear amplification mediated (LAM)-PCR was initially designed to retrieve rare retroviral vector integration sites from small, clinical sample sizes [38]. We established a LAM-PCR with AAV primers in the "D" element of the AAV-ITR, the innermost and sole ITR region without internal inverse repetitions (Figure 1A). Unfortunately, pure AAV sequences with rearranged ITRs predominated, AAV-chromosome junctions were rare and the chromosomal DNA part often too short for unambiguous assignment to a unique genomic site. We then tested ligation-mediated (LM)-PCR that had been employed for broad surveys of lentivirus (HIV) or γ-retrovirus (MLV) integrations [39,40,41]. LM-PCR relies on a first LTR-specific primer. A linker is ligated to the first PCR strand that typically ends at the chosen restriction site within the unknown chromosomal sequence. A primer complementary to this linker ensures second strand synthesis and retrovirus-chromosome junctions are amplified by using a combination of retrovirus LTR-specific and linker-specific primer sets.

For this study a variation of LM-PCR, named linker-selection-mediated (LSM)-PCR was developed which enriched for *bona fide* AAV-chromosome fusion sequences. The genomic DNA of AAV-infected cells was cleaved with restriction enzymes that lead to sufficiently sized DNA segments to allow unambiguous genomic assignment of the chromosomal junction (Figure 1B). DNA sequences were amplified with one primer for a unique AAV-sequence, either of the p5 promoter or of the *cap* gene. The other primer binds to the linker DNA attached to the unknown chromosomal site. The structure of the linkers forces the PCR to initiate within the AAV genome, thereby suppressing amplification of chromosomal DNAs lacking integrated AAV. The use of non-cut enzymes for AAV-2 DNA helped to circumvent the problem of ligating linkers to episomal, non-integrated AAV DNA sequences. To further enrich for AAV-chromosome junctions a biotin tag was attached to the 5′-end of the linker primer. Thus, chromosome-derived PCR products could be enriched by streptavidin-mediated magnetic bead selection. This lead to PCR products selected for both, the presence of AAV and of an unknown chromosomal DNA sequence.

AAV-2 integration sites

Using LSM-PCR a total of 1700 cloned PCR fragments were screened for DNA inserts of a minimal fragment size (>500 bp) to insure unambiguous detection of AAV-chromosome junctions. Out of 350 DNA sequence runs a total of 129 unique junction sites could be assigned to the human genome. Of these, 109 fulfilled the criteria outlined in the methods for unambiguous assignment of a single chromosomal site. Junctions were retrieved with non-cut enzymes for AAV-2, PvuII or EcoRV or with DraI, which cuts once in AAV-2 DNA outside of the region covered by the PCR. In addition, 43 wildtype AAV-2 infected Hela-derived single cell clones were generated of which eight harboured AAV-chromosome junctions that fulfilled the criteria outlined in the methods.

DNA sequence analysis revealed that AAV-2 wildtype integration sites were scattered over the entire human genome. The

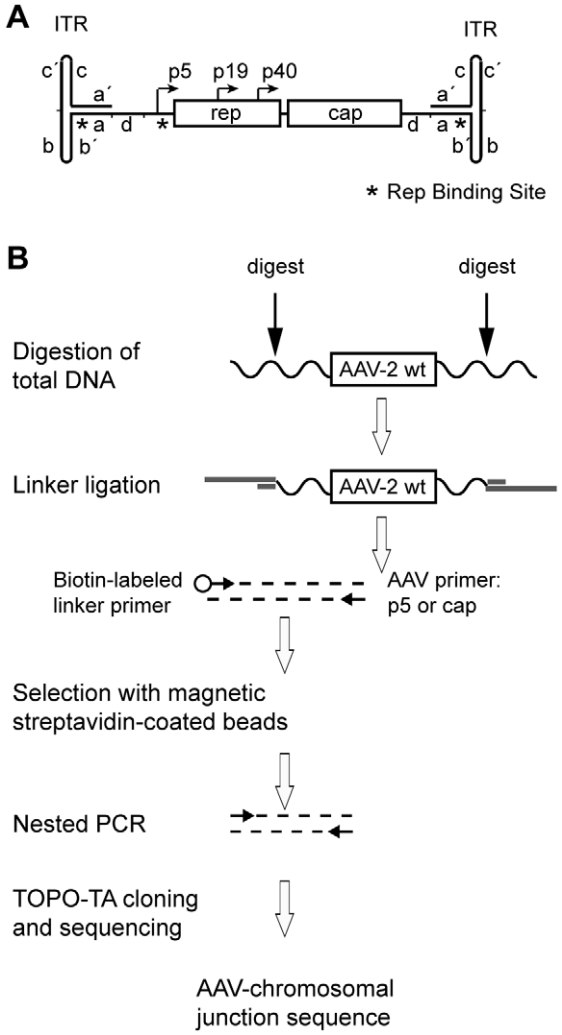

Figure 1. Linker-selection-mediated (LSM) PCR for cloning of chromosomal AAV integration sites. (A) Genome structure of AAV-2 with the *rep* and *cap* genes and their promoters flanked by inverted terminal repeats (ITRs) at either end of the ssDNA genome. The hairpin-structured ITRs contain internal repeat elements and complements thereof, represented by small letters. The positions of the Rep-binding sites (RBS) are represented by asterisks. (B) Schematic representation of the LSM-PCR strategy for amplification of AAV-chromosomal junction fragments. Wavy lines indicate chromosomal DNA. Linkers are displayed by thick, grey lines, AAV-specific primers by small, horizontal arrows. For restriction enzyme digestion indicated by vertical arrows either one of the following enzymes were used: PvuII or EcoRV (non-cut for AAV-2) or DraI (single-cut in AAV-2).
doi:10.1371/journal.ppat.1000985.g001

chromosomal distribution pattern is displayed in Figure 2A. Over one third of AAV integration sites were clustered at hotspots on chr. 19q13.42, on chr. 5p13.3 and on chr. 3p24.3 (Figure 2B–D). Infection with AAV in the absence of a helper virus leads to transient, low Rep expression. Many previous AAV integration studies used plasmid transfections of wildtype or vector AAV constructs often in combination with a high-level Rep expression construct. To evaluate whether high Rep expression influenced the target site preference of AAV, the sequence data of previously published transfection-based AAV integration sites [42] were reevaluated with the more stringent criteria outlined in the method. Of 157 DNA sequences retrieved after cotransfection of a *rep*-expression construct and an AAV vector plasmid 47 junction sequences fulfilled our criteria for unambiguous assignment of AAV to a unique chromosomal site (Table 1).

Integration hotspots

For AAV wildtype 10% of all retrieved junctions were detected at the hotspot on chr. 19q13.42 spread over a total of 33 kb around AAVS1 (Figure 2B). Only one out of twelve chr. 19q13.42-specific AAV junctions was located within the 4 kb region of AAVS1, where a consensus Rep-binding site and an adjacent *trs* site had been defined [4] The reevaluated distribution pattern of junctions generated by transfection of AAV vector- and Rep expression plasmids [42] was similar (Figure 2B). Latently AAV-infected Detroit 6 cells [43,44] were analyzed as control. Using cap-specific primers the junction was detected within AAVS1 at nucleotide position 60,319,992. A second hotspot named AAVS2 was detected on the small arm of chr. 5p13.3 within an intergenic region, where ten independent integration sites were detected within 8 kb (Figure 2C). In seven of these junctions clustered within 14 bp AAV had integrated directly into a consensus Rep binding site. The reanalyzed chromosomal integrations from AAV plasmid transfection [42] displayed a similar pattern with six integrations within 16 bp of the consensus RBS (Figure 2C). The third hotspot named AAVS3 was found on chr. 3p24.3 (Figure 2D). Out of 13 sites detected on chr. 3, three integrations were clustered in a 8 kb region where a consensus Rep binding site GAGT GAGT GAGT GAGC GAGC was detected on the complement strand (Figure 2D).

Rep-binding affinity for RBS consensus sites in AAVS1, -S2, and S3

To evaluate the binding affinities of Rep to the consensus RBS of the hotspots on chr.5 and chr. 3 compared to the RBS of chr. 19 or within the AAV genome, double-stranded oligonucleotides spanning the respective RBS regions (Figure 3) were submitted to mobility shift assays (EMSA) with increasing amounts of purified MBP-Rep78. Since it was previously shown that GAGG repeats are deficient in binding to Rep [10,45], a mutated oligo derived from the RBS of AAVS2 displaying GAGG GAGG GAGC GAGG was used as a control. As an additional control, a random oligonucleotide of similar length was used. As shown in figure 4, the RBS of AAVS3 contained five instead of four GAGY repeats and bound Rep with a two-fold higher affinity than the oligonucleotide spanning the AAVS1 RBS and trs (Figure 4B). The RBS of AAVS2 showed 76% of the Rep-binding affinity of the AAVS1 sequence (Figure 4C). In contrast, the relative binding affinity normalized to the AAVS1 sequence dropped to 13% with the mutated AAVS2 oligonucleotide, which was in the range of the random oligonucleotide (Figure 4C). These findings confirm the importance of the GAGY repeats in Rep binding. As expected, Rep78 displayed the highest affinities for oligonucleotides spanning the A-stem of the AAV-ITR or the AAV p5–promoter (Figure 4A, 4D). In summary, the newly discovered hotspots for AAV integration, AAVS2 on chr. 5 and AAVS3 on chr. 3 display RBS similarly proficient for Rep-binding as AAVS1.

Genomic features

To evaluate whether AAV-2 wildtype prefers specific motifs or genomic features for chromosomal integration the detected chromosomal junctions were compared to integration sites described for infection of human cells with a rep-deleted AAV-2 based vector [46]. The published DNA sequence files were reanalyzed using the criteria as outlined in the methods. This led to 450 junctions that could be included as an AAV vector-specific

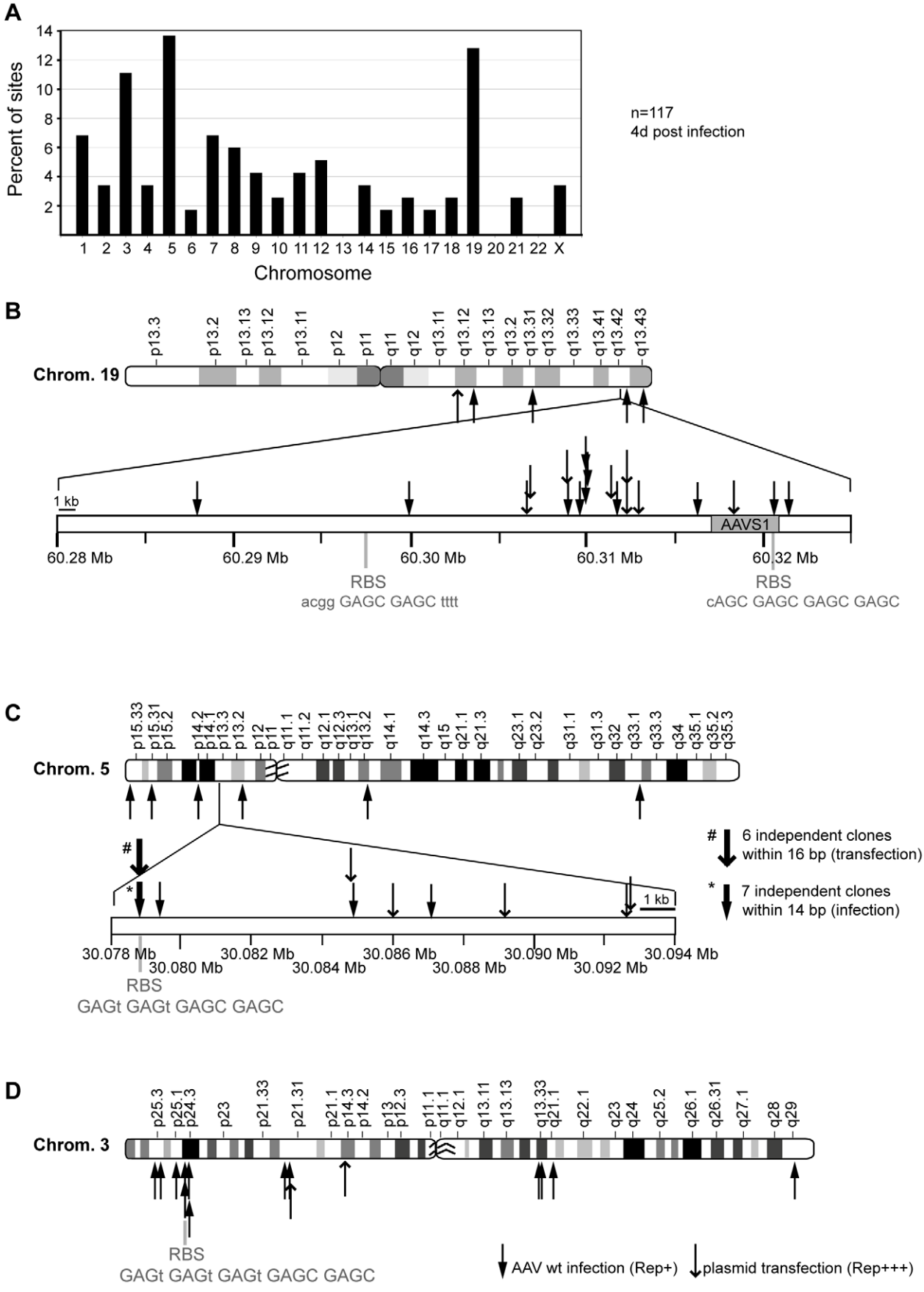
A
Percent of sites
Chromosome
n=117
4d post infection
B
Chrom. 19
1 kb
AAVS1
60.28 Mb
60.29 Mb
60.30 Mb
60.31 Mb
60.32 Mb
RBS
acgg GAGC GAGC tttt
RBS
cAGC GAGC GAGC GAGC
C
Chrom. 5
6 independent clones within 16 bp (transfection)
* 7 independent clones within 14 bp (infection)
1 kb
30.078 Mb
30.080 Mb
30.082 Mb
30.084 Mb
30.086 Mb
30.088 Mb
30.090 Mb
30.092 Mb
30.094 Mb
RBS
GAGt GAGt GAGC GAGC
D
Chrom. 3
RBS
GAGt GAGt GAGt GAGC GAGC
AAV wt infection (Rep+)
plasmid transfection (Rep+++)

Figure 2. Chromosomal distribution of AAV-2 integration sites. DNA of AAV-2 wildtype infected Hela cells was analyzed for viral integration by the LSM-PCR-method. DNA sequence data from 117 cloned AAV-chromosome junctions were assigned to unique loci. (A) Distributions of junctions on individual chromosomes are shown as percentage of the total 117. (B–D) AAV integration sites drawn to scale for chromosome 19 (B), 5 (C), and 3 (D). Shown are the chromosome ideograms and enlarged bands of hotspots found on chr. 19q13.42 (AAVS1) and chr. 5p13.3 (AAVS2). Solid arrows represent sites detected upon wildtype AAV-2 infection. Open arrows represent junctions stemming from cotransfection of AAV vector- and Rep-expression plasmids.
doi:10.1371/journal.ppat.1000985.g002

data set (Table 1). The preference for integration next to selected genomic features was analyzed for rep-positive AAV wildtype and for rep-deficient AAV vectors (Table 2). The data showed that the integration frequency of AAV wildtype in genes was higher than expected by chance (Table 2). The frequency was comparable to that of rep-deficient AAV vectors, thus confirming the findings by Miller et al. [46].

Chromatin state at AAV integration sites

To analyze the effect of epigenetic modifications on AAV integration the association of integration sites with histone modifications as markers for open or closed chromatin were assessed by chromatin immunoprecipitation sequencing (ChIP-Seq) analysis as outlined in the methods. Trimethylated lysine 27 of histone 3 (H3K27me3) is correlated with gene repression (closed chromatin) [47], while methylation of lysine 4 in H3K4me3 and H3K4me1 is indicative of promoter or enhancer regions (open chromatin) [48]. As shown in table 2 the association of AAV wildtype with open chromatin regions is significantly higher than expected from random controls. Conversely, the respective association with closed chromatin is significantly reduced. In summary, AAV wildtype prefers integration into open chromatin whereas closed chromatin was avoided.

Bioinformatic analysis of the AAV integration sites

A series of publications have shown that fused combinations of two to four GAGC motifs bind to Rep78/68 of AAV-2 [4,49,50,51,52,53]. Moreover, *in vitro* ternary complex formation of Rep68 with the AAV-2 ITR and AAVS1 of chr. 19q13.42 [10] led to the concept of Rep acting as an adapter that targets AAV to the human genome. Although only AAV-2 has been analyzed for chromosomal integration so far, all known AAV serotypes displayed various combinations of GAGC and/or GAGT motifs in the ITR and the p5 promoter. An alignment of these AAV elements to the integration hotspots AAVS1, AAVS2 and AAVS3 is displayed in Figure 3.

Based on these data we hypothesized that AAV-2 wildtype, due to the presence of Rep, prefers integration at chromosomal sites in closer proximity to consensus Rep binding sites than would be expected from control sites. The hypothesis was tested with the three sets of junctions derived from: 1. Infection with AAV-2 wildtype, 2. Cotransfection of plasmids coding for an AAV vector and a constitutive Rep-expression cassette, and 3. Infection with Rep-deficient AAV vectors (Table 1). The distances between any one integration site and its nearest Rep-binding site were determined in the human genome and compared to similarly determined distances of individual control sites to the nearest Rep-binding sites. Calculations were repeated using various combinations of RBS as displayed in Figure 5.

The choice of randomly generated genomic control sites was considered optimal for comparative analysis of the three sets of data. Yet, a concern was the choice of restriction endonucleases for the identification of the wildtype AAV-2 integration sites by LSM-PCR. To control a bias introduced by a conceivable non-random genomic distribution of the restriction sites, the average distance of PvuII, EcoRV, or DraI-generated restriction sites to putative Rep-binding sites was compared to the average distances of random sites to Rep-binding sites. PvuII restriction sites were found to be closer to Rep-binding sites than random control sites (Figure S1). This was assumedly due to the high G+C content of the PvuII recognition sequence and of the consensus Rep-binding sites. Both EcoRV and DraI sites were found further apart from Rep-binding sites in accordance with their high A+T content (Figure S1). To circumvent any bias arising from the use of PvuII, the data set for AAV wildtype infection was calculated against the data set of random control sites as well as against the data sets for the restriction site–related controls. Since not more than two thirds of sites were generated with PvuII, the PvuII-related control sites would at most underestimate the association to Rep-binding sites and was therefore used as the most stringent control set. In addition all calculations were also performed with the set of random controls leading to similar findings (Figure S2).

Table 1. Summary of data sets analysed in this study.

Author	Reference	Source of integration sites	Number of junctions	Aim of study
Drew et al.	J Gen Virol, 2007	Cotransfection of *neo*-expressing AAV vector- and rep expression plasmids in HeLa cells. Selection of G418-resistent cell clones	47 (157)	Characterization of Rep78-dependent AAV-2 vector integration sites
Miller et al.	J Virol, 2005	AAV-2 vector infection of human diploid fibroblasts, no cell selection, analysis between 14–40 days p.i.	450 (1172)	Integration site pattern of AAV-2 vector integration (no rep)
Hüser et al.	This study	AAV-2 wt infection of HeLa cells, analysis at 4 days p.i.	109 [a]	Integration site pattern of wildtype AAV-2 integration
Hüser et al.	This study	AAV-2 wt infection of HeLa cells, expansion of single cell clones, no selection, analysis at 3–4 weeks p.i.	8 [a]	Integration site pattern of wildtype AAV-2 integration

Numbers in brackets represent the total numbers of junctions published in the given reference.
[a]Junctions derived from AAV-2 wildtype infected cells four days p.i. and from those after clonal expansion were combined for statistical analyses.
doi:10.1371/journal.ppat.1000985.t001

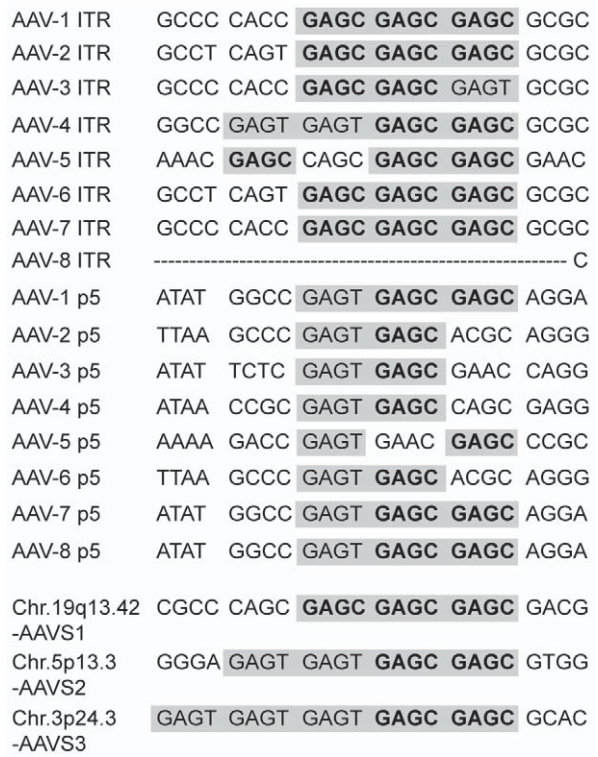

Figure 3. Sequence alignment of Rep-binding sites. RBS elements present in the ITR and the p5 promoter of all known AAV serotypes are aligned and related to consensus RBS sites present at chromosomal integration hotspots. Rep binding element GAGC is displayed in bold letters. Both GAGC and GAGT elements are highlighted in grey.
doi:10.1371/journal.ppat.1000985.g003

The bioinformatic calculations with GAGC GAGC as a minimal Rep-binding site strikingly confirmed our hypothesis that integration of wildtype AAV takes place close to Rep-binding sites with very high significance ($p < 0.0001$). A comparable effect was seen with the data set for AAV vectors in the presence of Rep ($p < 0.001$). Most importantly, the set of integration sites for AAV vectors in the absence of Rep did not show any difference of integration site preference compared to random control sites (Figure 5A). With a frequency of 15,707 sites per human genome the Rep binding motif GAGC GAGC occurs sufficiently frequent to lead to a mean distance of around 50 kb to the next AAV integration site in the presence of Rep. In the absence of Rep the mean distance to AAV (vector) integration sites rises to around 130 kb (Figure 5A). To ensure that the presence of repetitive DNA in the random controls did not lead to a bias in the analysis, an independent control calculation was performed for AAV wt data using AAV vector infection data as background. The high significance level was maintained (data not shown). The significance of the Rep-associated preferential integration near GAGC GAGC sequences was further underlined by the results of similar calculations for the putative Rep-binding motif GAGT GAGC, where no such association was found. Only in the presence of presumably large amounts of Rep (AAV vector transfection, Rep+++) a small effect was seen (Figure 5B). Obviously the GAGT GAGC motif is not sufficient to attract Rep and the AAV genome for integration. When an additional GAGC repeat is added (GAGY GAGC GAGC) the integration preferences of AAV wildtype and Rep-expressing AAV vectors shifted to closer proximity to Rep-binding sites ($p < 0.0001$). This is especially surprising since only 616 sites per human genome are found for GAGY GAGC GAGC (Figure 5C). To allow more potential Rep-binding site permutations, calculations were repeated with the consensus GAGC GAGC GAGC with one or two random mismatches. This led to a significantly decreased mean distance to AAV junctions in spite of the fact that up to 100-fold more genomic hits were found for the motifs (Figure 5D; E). A single nucleotide exchange in the GAGY GAGC GAGC motif (Figure 5F, GAGY GAGC GAGA) on the other hand led to a complete loss of association to AAV integration sites. This is surprising in view of the reported *in vitro* binding of Rep to this motif [45] and supports the assumption that the C at the 3′ end of the Rep binding motif is relevant for Rep-binding *in vivo*. Motifs GCCC GAGT GAGC and GAGT GAGC ACGC are part of the RBS in the viral p5 promoter. The individual motifs are found at very low frequency ($n = 85$, or $n = 82$, respectively) in the human genome, so that either no RBS was found in the same contig or the distance to the next RBS was more than several thousands kb. For these reasons we did not proceed with calculations for these motifs. To further exclude the possibility that the calculated associations with Rep binding sites were predominantly based on sequences assigned to the hotspots AAVS1 and AAVS2, the significance of the associations was re-evaluated with data sets omitted for the hotspot sequences (Table 3). The robustness of the data becomes evident by the fact that the highly significant association of AAV junctions to motifs GAGC GAGC and GAGY GAGC GAGC is maintained. In summary, AAV prefers integration sites in the vicinity of consensus Rep-binding elements, most prominently on chr. 19q13.42 (AAVS1), chr. 5p13.3 (AAVS2), and chr. 3p24.3 (AAVS3). But even in the absence of hotspots AAV still shows a highly significant integration preference for Rep-binding motifs at numerous additional sites in the human genome.

Discussion

This study represents the first genome-wide survey of wildtype AAV-2 integration in the human genome combined with a thorough bioinformatic analysis of the surrounding genome. We show here that wildtype AAV-2 infection leads to preferential integration in the vicinity of consensus Rep-binding sites (RBS) at defined hotspots as well as at numerous additional genomic sites. In contrast, AAV-2 vectors in the absence of Rep-expression integrate without discernable preference for consensus Rep-binding sites.

Hotspots of AAV integration

At the hotspot on chr. 19q13.42, up to 10% of all AAV junctions were scattered over a region of 33 kb, mostly in centromeric direction with regard to the previously defined core 4 kb AAVS1 site. AAV vectors in the absence of Rep expression do not show any preference for chr. 19q13.42 [46]. The here identified, novel hotspot AAVS2 on chr. 5p13.3 displayed roughly 8% of all junctions retrieved from wildtype AAV-2 infection and 23% of those retrieved from cotransfection of AAV vectors in the presence of Rep distributed over a region of 14 kb. A cluster of 13 independent junctions was found within 14 bp of the AAVS2 RBS that was shown to be similarly proficient in binding to Rep *in vitro* as is the RBS of AAVS1 (Figure 4). The high *in vivo* integration numbers may in part be due to the choice of HeLa as target cells. These are hypertriploid with up to 12 copies of the p-arm of chr. 5 [54]. The extra gain of integrations within the described 8 kb region is however unique for the AAVS2 site and not accompanied by a parallel increase of integrations at additional sites on the overrepresented p-arm of chr. 5, where 201 additional

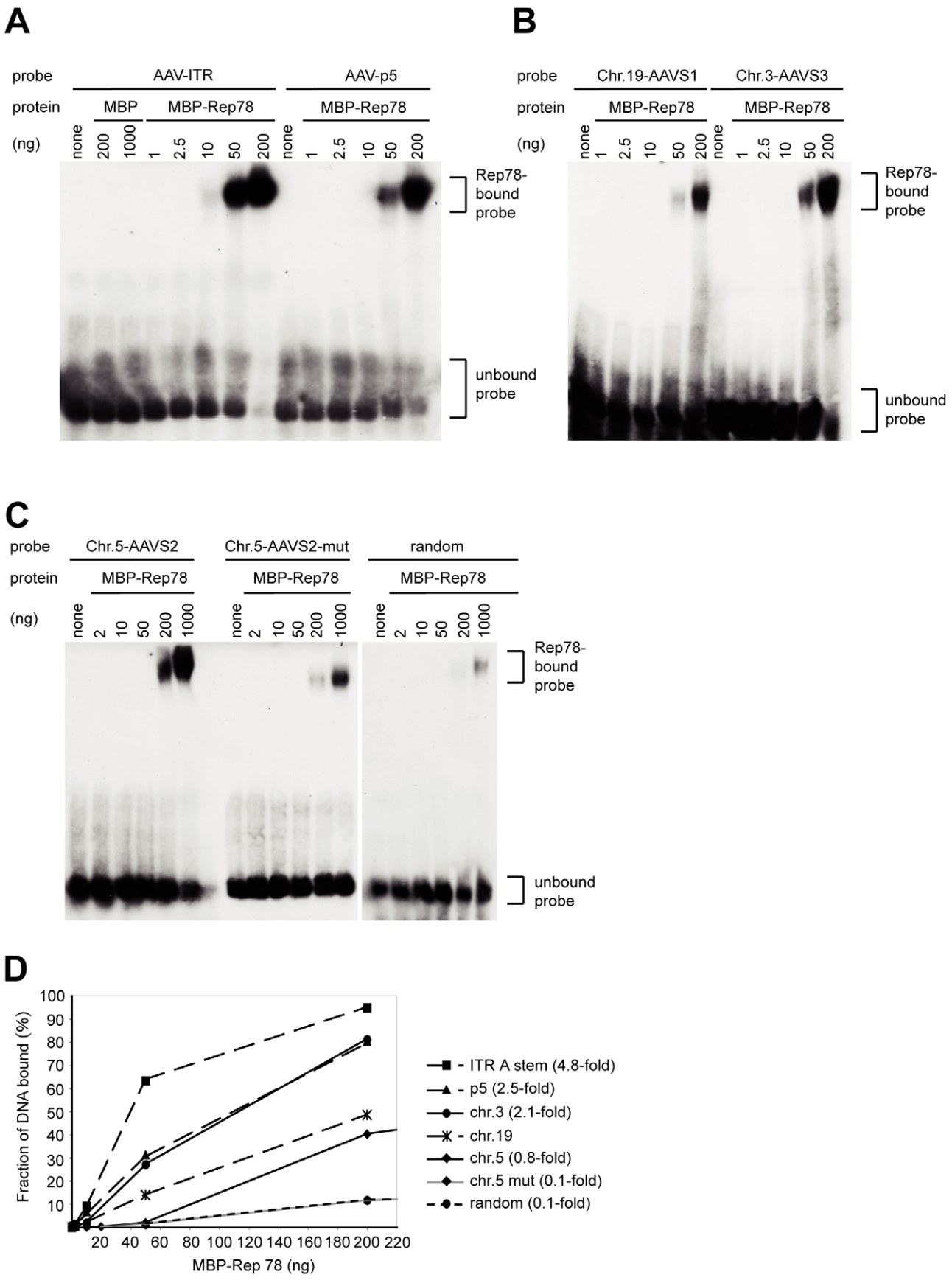
A
probe
AAV-ITR
AAV-p5
protein
MBP
MBP-Rep78
MBP-Rep78
(ng)
none
200
1000
1
2.5
10
50
200
none
1
2.5
10
50
200
Rep78-bound probe
unbound probe
B
probe
Chr.19-AAVS1
Chr.3-AAVS3
protein
MBP-Rep78
MBP-Rep78
(ng)
none
1
2.5
10
50
200
none
1
2.5
10
50
200
Rep78-bound probe
unbound probe
C
probe
Chr.5-AAVS2
Chr.5-AAVS2-mut
random
protein
MBP-Rep78
MBP-Rep78
MBP-Rep78
(ng)
none
2
10
50
200
1000
none
2
10
50
200
1000
none
2
10
50
200
1000
Rep78-bound probe
unbound probe
D
Fraction of DNA bound (%)
100
90
80
70
60
50
40
30
20
10
20
40
60
80
100
120
140
160
180
200
220
MBP-Rep 78 (ng)
ITR A stem (4.8-fold)
p5 (2.5-fold)
chr.3 (2.1-fold)
chr.19
chr.5 (0.8-fold)
chr.5 mut (0.1-fold)
random (0.1-fold)

Figure 4. Binding of MBP-Rep78 to Rep-binding sites of AAV-2 and of chromosomal integration hotspots. (A) to (C) Electrophoretic mobility shift assays (EMSA) were performed with ^{32}P-labeled double-stranded oligonucleotides in the presence of increasing amounts of affinity-purified MBP-Rep78 as indicated above the autoradiograms. (D) Quantitative determination of the bound fractions of the different RBS and control oligonucleotide probes as a function of the amount of MBP-Rep78 protein in the binding reaction. EMSA gels shown in (A) to (C) were subjected to phosphorimager analysis to determine the relative amount of unbound and bound ^{32}P-labeled oligonucleotides. The relative binding affinity was calculated as follows: The highest amount of Rep used in this assay (1000 ng) bound 31% of the random oligonucleotide. The amount of Rep that bound the same fraction of the other oligonucleotides was determined and normalized to the binding of the chr. 19 (AAVS1) oligonucleotide.
doi:10.1371/journal.ppat.1000985.g004

GAGC GAGC repeats and three additional GAGY GAGC GAGC repeats were counted. The only fourfold tetranucleotide repeat on the chr.5 p-arm is found in AAVS2 (GAGT GAGT GAGC GAGC; Figure 2C). In addition, junctions of rep-deficient AAV vector were reported to be underrepresented on chr. 5 [46].

A major difference between the hotspots on chr. 5 and chr. 19 concerns the presence of genes. The junctions identified on chr. 19 span the region of the transcribed gene for protein phosphatase 1, regulatory subunit 12C (*PPP1R12C*). The 8 kb AAVS2 sequence identified on chr. 5p13.3 represents an intergenic region to the best of current knowledge. It is well known that Rep expression leads to extensive rearrangements of AAVS1 [18,55,56]. Apparently, *PPP1R12C* is essential, since the majority of latently infected cell lines display gene duplications [57] and simultaneous AAV integrations in both alleles have never been reported. A currently unresolved question concerns the presence of a terminal resolution site (*trs*) next to the RBS of AAVS2 and AAVS3. In AAVS1 the spatial configuration of RBS and *trs* resembles that of the AAV-ITR. The *trs* element lies next to the RBS and serves as a nicking site for Rep [4]. In AAVS2 and AAVS3 the nearest perfect *trs* elements (5′-GTTGG-3′) are 400 and 500 bp away from the RBS, which represents the mean statistical occurrence for this motif. Unfortunately, the consensus nucleotide requirements for a functional *trs* element are not defined well enough to conduct a meaningful bioinformatic search. Therefore, the presence of nicking sites next to the RBS in AAVS2 or AAVS3 remains open at present.

Target site choice for AAV integration

Besides the identified integration hotspots numerous additional chromosomal junction sites were found for integrated wildtype AAV-2, scattered over the human genome. From the bioinformatic calculations it appeared that the perfect tetranucleotide repeat GAGC GAGC represented the minimal requirement for Rep-dependent targeted integration, and GAGY GAGC GAGC represents the optimized *in vivo* target sequence for wildtype AAV-2. Hotspots AAVS1, AAVS2, and AAVS3 display this core sequence fused to additional imperfect GAGY repeats. Other AAV serotypes display RBS sequences with similar numbers of GAGC and/or GAGT repeats, extended by additional imperfect repeats. AAV5 Rep co-crystallised with the hairpin-structured AAV5-ITR revealed that five Rep monomers bind to five consensus tetranucleotide repeats of the RBS, each of which was contacted by two Rep monomers from opposite faces of the DNA [58]. AAV2-Rep78/68 was shown to simultaneously bind to the RBS of the AAV-2 ITR and to that of AAVS1 [10]. Although it is currently unknown whether other AAV serotypes integrate at all, this is highly likely in view of the ability of both AAV-2 Rep and the relatively distant AAV-5 Rep to multimerize and simultaneously bind to clustered GAGY repeats.

In the initial descriptions of AAVS1, site-specific nicking of the *trs* by Rep bound to the adjacent RBS was viewed as preferred entry site for AAV recombination [4]. Meanwhile the majority of AAV integrations on chr. 19q13.42 were found many kb away from the RBS-*trs* combination, and neither AAVS2 or AAVS3 display obvious *trs* homologues next to the RBS. Therefore alternative explanations for RBS-dependent AAV integration should be considered. The potential use of preexisting chromosomal breakage sites recalls a mechanism already proposed for the integration of rep-deficient AAV vectors [34,59]. Alternative integration concepts include the use of imperfect *trs* elements for nicking as shown *in vitro* [4,60,61], or the ability of Rep78 to induce DNA damage *in vivo* by single-strand nicking of cellular

Table 2. Genomic features and chromatin state associated with AAV-2 wildtype- and AAV vector-derived integration sites.

Genomic feature	AAV wt infection % of sites (n = 117)	AAV vector infection % of sites (n = 450)	Random control % of sites	p-Value wt[a]	p-Value vector [a]
RefSec genes	50.4	51.8	40.1	<0.05	<0.01
RefSec genes tss+/−2 kb	6.8	5.6	2.9	<0.05	<0.01
Ens genes	51.1	51.3	40.7	<0.05	<0.0001
GenScan genes	83.8	74.0	69.5	<0.001	<0.05
Known genes	57.3	55.8	45.2	<0.01	<0.0001
Known genes exons	3.4	3.6	2.7	n.s.	n.s.
CpG	0.9	1.6	0.7	n.s.	<0.05
CpG+/−2 kb	7.7	6.4	4.1	n.s.	<0.05
Histone modification					
H3K4me1	13.7	n.d.	7.5	<0.05	n.d.
H3K4me3	6.8	n.d.	2.3	<0.01	n.d.
H3K27me3	28.2	n.d.	42.2	<0.01	n.d.

[a]P-values were determined in comparison to random controls, n.d. = not determined, p>0.05 were not considered statistically significant (n.s.).
doi:10.1371/journal.ppat.1000985.t002

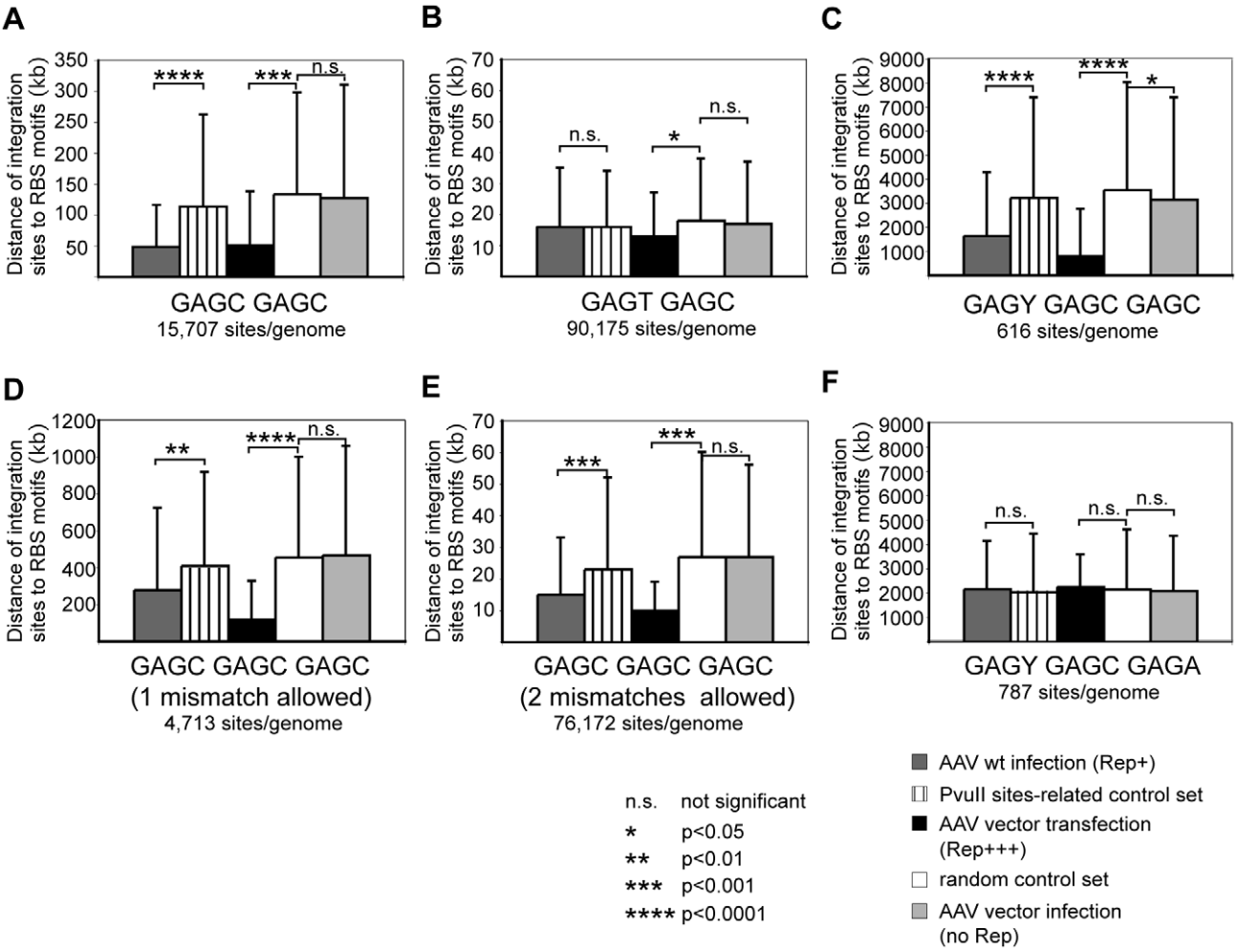

Figure 5. Statistical analysis of distances of integration sites from human Rep-binding motifs. AAV integration sites of different data sets were analyzed for the proximity to the next Rep binding sites. Calculations were performed with the following putative Rep binding sites: GAGC GAGC (A); GAGT GAGC (B); GAGY GAGC GAGC (C); GAGC GAGC GAGC, one mismatch (D), or two mismatches allowed (E), and GAGY GAGC GAGA (F). Average distances of integration sites from putative Rep binding sites are displayed as mean +/− S.D. The levels of significance are marked by asterisks. P-values >0.05 were not considered statistically significant. P-values <0.01 were considered highly significant. For the analysis of wildtype AAV integration data a PvuII-related control set was used in addition to the random control set.
doi:10.1371/journal.ppat.1000985.g005

chromatin [19]. It is conceivable that the introduction of single-strand nicks occurs anywhere in accessible chromatin, even if the nicking site is hundreds or thousands of bp apart from the RBS on an extended DNA strand. HMGB1, an ubiquitous architectural protein that serves as key component of the chromatin remodelling complex may be of help [62]. Its long-known *in vivo* interaction with Rep [63] may help remodel the chromatin to make it accessible for nicking by Rep. Rep was also shown to contact other key players of the nucleosome remodelling complex as components of the transcription- or DNA replication machinery

Table 3. Neighbourhood analysis of wildtype AAV-2 integration sites and RBS motifs found outside of the hotspots on chr. 19q13.42 (AAVS1) and on chr. 5p13.3 (AAVS2).

RBS motif	all integration sites (n = 117)	no chr.19q13.42 (n = 106)	no chr.19q13.42 no chr.5p13.3 (n = 96)
GAGC GAGC	<0.0001	<0.0001	<0.001
GAGY GAGC GAGC	<0.0001	<0.001	<0.01
GAGC GAGC GAGC (1 mismatch allowed)	<0.01	<0.05	not significant [a]
GAGC GAGC GAGC (2 mismatches allowed)	<0.001	<0.01	<0.05

Displayed are p-values calculated with the PvuII control set.
[a]p-values >0.05 were not considered statistically significant.
doi:10.1371/journal.ppat.1000985.t003

[64,65,66]. Any of these mechanisms can be exploited to open the chromatin for AAV integration. In summary, Rep with its combined DNA-binding and endonuclease activity appears to serve as a relatively imprecise targeting tool for AAV integration preferably in open chromatin regions in the reach of consensus Rep-binding sites prevalent in the human genome.

Implications for Rep-dependent targeting of AAV vector integration

The early finding that Rep would mediate site-specific AAV integration on chr. 19q13.42 had immediate implications for gene therapy. A variety of concepts were devised to incorporate Rep as an adapter to target AAV-ITR flanked transgenes to a specific site [26,27,28,57,67]. In the majority of cases appropriate cell selection or PCR for AAVS1 led to cells displaying the desired integration. The reported high frequencies of integration into AAVS1 are difficult to reconcile with our findings, unless the level of Rep expression is considered to have an impact on target site choice. Upon AAV infection Rep is only moderately expressed due to autoregulation of the AAV p5 promoter. Rep-dependent AAV vector transductions typically use strong heterologous promoters that lead to high and sustained Rep expression levels. Increasing Rep levels may increase the overall probability for integration anywhere in the genome, including at hotspots. Under these conditions AAVS1-specific integration will be detected more readily. This appears however to come at the price of genomic rearrangements in reach of alternative Rep-binding sites. Therefore, it is plausible that in the absence of any selection AAV integration into AAVS1 is typically unstable and difficult to detect.

In summary, Rep expression increases the probability for integration next to one of several genomic hotspots. However, the net genotoxic effect is unpredictable both with respect to the integrity of the AAV integration locus itself and with respect to the numerous additional sites where Rep binds and initiates chromosomal damage. Therefore, the current concept of a relatively precise site-specific targeting of AAV should be extended to a concept of a relative preference for accessible chromatin regions in the neighbourhood of any of the numerous consensus Rep-binding sites. More recent approaches for site-specific vector targeting try to exploit DNA sequence-specific zinc-finger nucleases to target a genomic sequence of wish [68]. Although zinc-finger nucleases are not free from off-target genotoxicity, at least the genomic targeting site for the transgene can be more precisely defined, a goal that appears to be inherently unachievable using Rep as an adapter molecule.

Materials and Methods

Cells

Detroit 6 cells harbouring latent AAV-2 genomes and HeLa cells were grown in Dulbecco's modified Eagles's medium (Gibco) supplemented with 10% fetal calf serum, penicillin (100 U/ml), and streptomycin (100 µg/ml).

AAV infection

Viral stocks of wildtype AAV-2 with infectious titers of 5×10^9 i.u./ml were prepared on HeLa cells as described before [16]. For the analysis of AAV integration sites 1.7×10^6 HeLa cells were seeded overnight on 10 cm diameter dishes and infected with AAV-2 at a MOI of 500. Cells were harvested at 96 hours post infection (p.i.) for the extraction of genomic DNA. The period of cell growth after infection was minimized to reduce the chances of selection of particular integration sites during cell proliferation. Alternatively, AAV-infected HeLa cells were seeded to microtiter plates at a dilution of 60 cells per plate and grown up as single-cell clones without drug selection.

Plasmids

Plasmid pTAV2-0 covers the AAV-2 wildtype genome (GenBank accession number AF043303), pRVK the 4 kb fragment of the AAVS1 locus on chromosome 19 (GenBank accession number S51329), and pAAVS1-TR covers an AAV-ITR/AAVS1 junction [16]. Plasmid pMBP-Rep78 encoding Rep78 fused to maltose-binding protein (MBP) was described before [69].

Production and purification of MBP-Rep78 fusion protein

MBP-Rep78 encoding Rep78 fused to maltose-binding protein was expressed und purified essentially as described [69]. Briefly, E.coli strain BL21 transformed with pMBP-Rep78 was grown at 30°C to an $OD_{600\ nm}$ of 0.6 to 0.8. Production of MBP-Rep78 was induced with 0.3 mM IPTG for 3 h at 30°C. Cells were harvested by centrifugation and lysed by sonication for 2 min (30% duty cycle) in lysis buffer of 50 mM phosphate pH 7.8, 300 mM NaCl, 1% (v/v) Triton X-100, 0.1 mM PMSF. Cell debris was removed by centrifugation at 6500×g for 20 min at 4°C. The supernatant was adsorbed to amylose resin (New England Biolabs) in a batch process and the resin was washed extensively (5 washes with about 100 volumes of the resin) with lysis buffer. The adsorbed proteins were eluted with lysis buffer containing 10 mM maltose and analyzed for purity by SDS-polyacrylamide gel electrophoresis.

Electophoretic mobility shift assays (EMSA)

Binding of MPB-Rep78 fusion protein to ^{32}P- labeled double-stranded oligonucleotide probes was detected by altered mobility of the probes in nondenaturating polyacrylamide gels essentially as described previously [70]. Briefly, oligonucleotides of 46–49 nt length were end-labeled with T4 polynucleotide kinase and annealed. EMSA reactions were performed for 20 min at 20°C as follows: 0.015 pmol of labeled DNA substrate was incubated with the indicated amounts of MBP or MBP-Rep78 in a binding buffer containing 25 mM HEPES-KOH (pH 7.8), 10 mM $MgCl_2$, 40 mM NaCl, 1 mM DTT, 2% glycerol, 12.5 µg/ml BSA, 0,01% Nonidet P40 and 5 µg/ml salmon sperm DNA. The following oligonucleotides were used:

AAV-ITR (nucleotide position 85–133): GCCTCAGTGAGCGAGCGAGCGCGCAGAGAGGGAGTGGCCAACTCCATCA;

AAV-ITR complementary strand: TGATGGAGTTGGCCACTCCCTCTCTGCGCGCTCGCTCGCTCACTGAGGC

Chr. 19q13.42 (AAVS1): TGGCGGCGGTTGGGGCTCGGCGCTCGCTCGCTCGCTGGGCGGGCGGGC

Chr19 (AAVS1) complementary strand: GCCCGCCCGCCCAGCGAGCGAGCGAGCGCCGAGCCCCAACCGCCGCCA

Chr. 5p13.3 (AAVS2): AGCTGGACCCCACGCTCGCTCACTCACTCTCCCCTCACCGCTTTGT

Chr. 5 (AAVS2) complementary strand: ACAAAGCGGTGAGGGGAGAGTGAGTGAGCGAGCGTGGGGTCCAGCT

Chr. 3p24.3 (AAVS3) GCTTCCCAAGGGGAATGAATGTGCGCTCGCTCACTCACTCACTCCTCAC

Chr.3 (AAVS3) complementary strand: GTGAGGAGTGAGTGAGTGAGCGAGCGCACATTCATTCCCCTTGGGAAGC

Chr. 5MUT (AAVS2 mutated): AGCTGGACCCCACC**C**CTCGCTC**C**CTC**C**CTCTCCCCTCACCGCTTTGT

Chr.5MUT (AAVS2 mutated), complementary strand: ACAAAGCGGTGAGGGGAGAG**G**GAG**G**GAGCGAG**G**GTGGGGTCCAGCT

AAV p5 (nucleotide position 245–292): TCACGCTGGGTATTTAAGCCCGAGTGAGCACGCAGGGTCTCCATTTTG

AAV p5 complementary strand: CAAAATGGAGACCCTGCGTGCTCACTCGGGCTTAAATACCCAGCGTGA

random control: CAGAGCAGCAGCACAGACGCTAGCAGATCTCCTGCGACCGGAGATGTG

random control, complementary strand: CACATCTCCGGTCGCAGGAGATCTGCTAGCGTCTGTGCTGCTGCTCTG

Preparation of genomic DNA

Total genomic DNA was extracted by SDS/proteinase K digestion followed by repeated phenol/chloroform extractions and ethanol precipitation, as described before [71]. High molecular weight DNA (2 µg) was digested with restriction enzymes that lead to a mean genomic fragment size of around 4 kb and produce blunt-ends ready for linker/adapter ligation. Non-cut enzymes for AAV-2 DNA were preferred, PvuII, EcoRV. Additional junctions were retrieved with DraI (one cut in AAV-2 wildtype DNA). Digested genomic DNA was purified by repeated phenol-chloroform extractions and precipitated with ethanol.

Linker-Selection-Mediated (LSM) PCR

A linker-based strategy described in [39,40] and outlined in more detail in the manual of the GenomeWalker kit (Clontech) was modified as outlined in Figure 1B. The following oligos were used for linker construction: "Linkerlong" (5′GTA ATA CGA CTC ACT ATA CGG CAC GCG TGG TCG ACG GCC CGG GCT GGT 3′) and "linkershort" (5′ACC AGC CC 3′modifikation: 2′,3′-dideoxyC). Equal amounts of "linkerlong" and phosphorylated "linkershort" (100pmol each) were annealed and ligated to restriction enzyme-digested genomic DNA.

PCR-primers: The linker-primers were "P linker outside" with biotin attached to its 5′ end (5′-GTA ATA CGA CTC ACT ATA CGG C; $T_m = 58.4°C$) and "P linker nested" (5′-ACT ATA CGG CAC GCG TGG T; $T_m = 58.8°C$). Two AAV-2-specific primer sets were used. The first primer set covered the AAV p5 promoter: "AAV2p5" (5′-TCA AAA TGG AGA CCC TGC GTG CTC A; $T_m = 64.6°C$, AAV-2, nt 293–269), primer "AAV2p5 nested" (5′-TAA ATA CCC AGC GTG ACC ACA TGG TG; $T_m = 64.8°C$, AAV-2, nt 260–235). The other primer set is located in the *cap* gene region, as described before [2]: "CAPgsp1" (5′-GTC TGT TAA TGT GGA CTT TAC TGT GGA CAC; $T_m = 65.4°C$, AAV-2, nt 4320–4349) and "CAPgsp2" 5′-GTG TAT TCA GAG CCT CGC CCC AT; $T_m = 64.2°C$, AAV-2 nt 4357–4379).

The PCR reaction contained 0.2 mM dNTPs, linker primer and AAV specific primer (0.25 µM, each), 2.5 U proofreading hot-start polymerase (Herculase) in reaction buffer, as provided by the supplier (Stratagene). Of the preceding linker-ligation reaction 1–5 µl was added to a final volume of 50 µl. PCR conditions were as follows: 3 min at 98°C, followed by 10 cycles of 40 sec at 98°C, 30 sec at 65°C, and 4 min at 72°C, followed by 20 cycles of 40 sec at 98°C, 30 sec at 65°C, and 4 min + 10 sec per cycle at 72°C, terminated by an extension period of 10 min at 72°C. Biotin-labelled PCR products were further enriched on streptavidin-labelled Dynabeads M-280, as outlined by the supplier (Invitrogen). Subsequent nested PCR used conditions identical to the first round but with pairs of the nested PCR primers, as outlined above. Finally, to add overhangs of multiple As, PCR products were incubated with 1 U Taq polymerase (New England Biolabs).

Analysis of LSM PCR products

Products of LSM-PCR reactions were separated on agarose gels. To ensure sufficient chromosomal fragment lengths, PCR bands of a calculated minimal length (>500 bp) were excised and purified by the QIAEX II Gel extraction kit (Qiagen, Hilden, Germany). TOPO-TA cloning was performed as described [72]. Colonies were PCR-screened with the M13 forward (-20) and reverse primer pair (0.4 µM, each) with 0.2 mM dNTP, 2 U Taq polymerase (New England Biolabs) at the following conditions: 10 min at 94°C, followed by 30 cycles of 30 sec each at 94°C, 52°C, and 72°C, followed by 10 min at 72°C. Column-purified PCR products were submitted to DNA sequencing using the primer provided by the TOPO-TA cloning kit. DNA sequences were run on a CEQ2000 genetic analysis system (Beckman) using the CEQ Dye Terminator Cycle Sequencing Quick start kit (Beckman) and the run method LFR-a. Cycling conditions were as follows: 1 min at 96°C, followed by 30 cycles 20 sec at 96°C, 20 sec at 50°C and 4 min at 60°C.

Integration site determination

The genomic positions of AAV integration sites in the human genome (assembly from March 2006, hg18) were determined using the BLAT tool from the UCSC Genome Browser web site (http://genome.ucsc.edu/cgi-bin/hgBlat) [73]. A match was defined as a BLAT search result fulfilling all of the following criteria:

1. A human chromosome-derived part of the DNA sequence is at least 100 bp in length and of 98% or higher homology to the database.
2. A shorter chromosomal match is acceptable if it displays a minimum of 25 bp of a contiguous DNA sequence match.
3. A part of the sequence allows assignment of AAV.
4. In the case of unassigned base pairs between the AAV and the human part of the sequence, this sequence is no longer than 20 bp.
5. Sequences matching to multiple chromosomal regions (i.e. repeat regions) were discarded in view of the inability to unambiguously assign the surrounding genome for subsequent bioinformatic analysis (see below).
6. Duplicate AAV-chromosomal fusion sequences (identical viral and identical human DNA sequences) were counted only once.

In addition to the LSM-PCR derived sequences, the original DNA sequence files of 157 chromosomal junctions [42] kindly provided by Dr. G.W. Both, North Ride, Australia were reanalyzed applying the above inclusion criteria. This led to 47 DNA sequences suitable for our analysis (Table 1). In their study, HeLa cells had been cotransfected with plasmids for constitutive RSV-promoter-driven Rep78 expression and for recombinant AAV vectors expressing a SV40-promoter-driven neomycin gene [42]. Furthermore, 1100 DNA sequences from a published analysis of *rep*-deficient AAV vector integration sites in diploid human cells [46] were reanalyzed. Since the PCR methods employed in our study and in the one by Drew et al. [42] cannot detect the matching left and right junction sites generated by one AAV integration event, only one chromosomal junction was analyzed per rescued provirus. The original DNA sequence files (DU711025.1 to DU709854.1) of Miller et al. [46] were downloaded from the Genome Survey Sequences (GSS) Database of NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucgss) and reanalyzed using the UCSC March 2006 human genome build. The analysis led to a total of 450 junction sequences that fulfilled all of the above inclusion criteria for bioinformatic comparisons. For the subsequent data analysis we implemented software in C++ using the software library SeqAn [74] and several Python scripts.

Determination of distances of integration sites to putative Rep binding sites

For different Rep binding motifs, we computed the average distance of virus integration sites to the closest occurrences of Rep binding motifs within the genome. We supposed that the observed integration events were independent from each other and the sample size was high enough for assuming the distance to be normally distributed. To assess whether these distances differ significantly from expectation, several background models were generated:

(1) For the background model "random", we assumed that the probabilities for the observation of virus integrations were equally distributed among all conceivable positions in the genome. A program was implemented that computed the exact mean and standard deviation of this background distribution.

(2) Since the integration site analysis required a suitable restriction enzyme site in the neighbourhood of the integrated virus three background-models for the restriction enzymes DraI, EcoRV or PvuII were generated. These models served as a corrective tool for an eventual bias of a non-uniform distribution of the respective restriction enzyme sites in the genome. For each AAV integration site observed, the distance to the closest restriction site was determined individually. Then, 1000 control sites per integration site were generated that displayed the same distance to randomly chosen restriction sites.

The generation of both, the data analysis and the background model was confined to those genomic contigs that contained at least one Rep binding motif, since otherwise the distance to the "closest Rep binding motif" would not be defined. A given set of AAV integration sites was considered to be significantly closer to Rep binding motifs than expected by chance, if the significance was calculated for *all* relevant background models. Data sets of AAV vectors were analyzed with the "random" background model. We applied a Z-test for determining statistical significances for the distances of integration sites to Rep binding sites. For comparing integration sites from AAV wildtype infection sites against those from rep-deficient AAV vector infection we applied the Student's t-test.

Presence of genomic features

AAV integration sites were examined for the occurrence of various genomic features using tables available in the UCSC database. For the determination of significant divergences from expectations, we compared the actual integration sites with a set of 100,000 randomly chosen control sites in the human genome using a two-tailed binomial test.

Analysis of chromatin state

Chromatin immunoprecipitation sequencing (ChIP-Seq) data were used to define the state of histone modifications in genomic regions of AAV integration. H3K27me3 domains determined by Cuddapah et al. were used as markers for closed chromatin (http://www.wip.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSM325898) [75]. Regions enriched for H3K4 methylation (open chromatin) were determined as follows: The raw ChIP-Seq reads by Robertson et al. [76] (http://www.bcgsc.ca/data/histone-modification) were mapped to the human genome using Bowtie [77], and peaks were called using MACS [78]. H3K4me1/3 domains are then defined as 5 kb windows around the centers of the peaks.

Supporting Information

Figure S1 Distribution of restriction sites in relation to Rep binding sites. (A) Cleavage sites of restriction enzymes used to digest genomic DNA of wildtype AAV-2-infected Hela cells and the numbers of occurrences per human genome are shown. (B) The mean distances of restriction enzyme cleavage sites to Rep binding sites were compared to those of random control sites to Rep Binding Sites. Calculations are displayed for consensus RBS that yielded significant proximity of integration sites to RBS as displayed in Figure 5. P-values were < 0.00002 for all motifs.
Found at: doi:10.1371/journal.ppat.1000985.s001 (0.13 MB TIF)

Figure S2 Bioinformatic analysis of AAV-2 wildtype integration sites. Distances of integration sites from Rep Binding Sites were calculated and the z-score was assessed in relation to the following control sites: (A) GAGC GAGC; (B) GAGT GAGC; (C) GAGY GAGC GAGC; (D) GAGC GAGC GAGC one mismatch allowed; (E) GAGC GAGC GAGC two mismatches allowed. In order to analyze only the Rep-binding sites outside of integration hotspot regions, sites within the hotspots of chr. 19 (AAVS1) and/or chr. 5 (AAVS2) were omitted in separate calculations.
Found at: doi:10.1371/journal.ppat.1000985.s002 (0.25 MB TIF)

Acknowledgments

We are grateful to G. Both CSIRO, North Ride, Australia for providing the original DNA sequence files of his study and to M. and A. Fromm, Institute of Clinical Physiology, Charité, Berlin for their friendly and knowledgeable help with Phosphorimager analysis.

Author Contributions

Conceived and designed the experiments: DH RH. Performed the experiments: DH TL KW EMH. Analyzed the data: DH AGD TL SW TC KR RH. Contributed reagents/materials/analysis tools: SW. Wrote the paper: DH AGD KR RH.

References

1. Muzyczka N, Berns KI (2001) Parvoviridae: The viruses and their replication. In: Knipe DM, Howley PM, eds. Fields Virology. Philadelphia: Lippincott. pp 2327–2359.
2. Schnepp BC, Jensen RL, Chen CL, Johnson PR, Clark KR (2005) Characterization of adeno-associated virus genomes isolated from human tissues. J Virol 79: 14793–14803.
3. Kotin RM, Siniscalco M, Samulski RJ, Zhu XD, Hunter L, et al. (1990) Site-specific integration by adeno-associated virus. Proc Natl Acad Sci U S A 87: 2211–2215.
4. Linden RM, Winocour E, Berns KI (1996) The recombination signals for adeno-associated virus site-specific integration. Proc Natl Acad Sci U S A 93: 7966–7972.
5. Snyder RO, Im D-S, Ni T, Xiao X, Samulski RJ, et al. (1993) Features of the adeno-associated virus origin involved in substrate recognition by the viral Rep protein. J Virol 67: 6096–6104.
6. Im D-S, Muzyczka N (1990) The AAV origin-binding protein Rep68 is an ATP-dependent site-specific endonuclease with helicase activity. Cell 61: 447–457.
7. Philpott NJ, Gomos J, Berns KI, Falck-Pedersen E (2002) A p5 integration efficiency element mediates Rep-dependent integration into AAVS1 at chromosome 19. Proc Natl Acad Sci U S A 99: 12381–12385.
8. Samulski RJ, Zhu X, Xiao X, Brook JD, Housman DE, et al. (1991) Targeted integration of adeno-associated virus (AAV) into human chromosome 19 [published erratum appears in EMBO J 1992 Mar;11(3):1228]. EMBO J 10: 3941–3950.
9. Kotin RM, Linden RM, Berns KI (1992) Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. Embo J 11: 5071–5078.
10. Weitzman MD, Kyöstiö SRM, Kotin RM, Owens RA (1994) Adeno-associated virus (AAV) Rep proteins mediate complex formation between AAV DNA and its integration site in human DNA. Proc Natl Acad Sci U S A 91: 5808–5812.

11. Meneses P, Berns KI, Winocour E (2000) DNA sequence motifs which direct adeno-associated virus site-specific integration in a model system. J Virol 74: 6213–6216.
12. Yang CC, Xiao X, Zhu X, Ansardi DC, Epstein ND, et al. (1997) 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.
13. Tsunoda H, Hayakawa T, Sakuragawa N, Koyama H (2000) Site-specific integration of adeno-associated virus-based plasmid vectors in lipofected HeLa cells. Virology 268: 391–401.
14. Palombo F, Monciotti A, Recchia A, Cortese R, Ciliberto G, et al. (1998) Site-specific integration in mammalian cells mediated by a new hybrid baculovirus-adeno-associated virus vector. J Virol 72: 5025–5034.
15. Pieroni L, Fipaldini C, Monciotti A, Cimini D, Sgura A, et al. (1998) Targeted integration of adeno-associated virus-derived plasmids in transfected human cells. Virology 249: 249–259.
16. Hüser D, Weger S, Heilbronn R (2002) Kinetics and frequency of adeno-associated virus site-specific integration into human chromosome 19 monitored by quantitative real-time PCR. J Virol 76: 7554–7559.
17. Hüser D, Heilbronn R (2003) Adeno-associated virus integrates site-specifically into human chromosome 19 in either orientation and with equal kinetics and frequency. J Gen Virol 84: 133–137.
18. McCarty DM, Young SM, Jr., Samulski RJ (2004) Integration of adeno-associated virus (AAV) and recombinant AAV vectors. Annu Rev Genet 38: 819–845.
19. Berthet C, Raj K, Saudan P, Beard P (2005) How adeno-associated virus Rep78 protein arrests cells completely in S phase. Proc Natl Acad Sci U S A 102: 13634–13639.
20. Schmidt M, Afione S, Kotin RM (2000) Adeno-associated virus type 2 Rep78 induces apoptosis through caspase activation independently of p53. J Virol 74: 9441–9450.
21. Di Pasquale G, Chiorini JA (2003) PKA/PrKX activity is a modulator of AAV/adenovirus interaction. Embo J 22: 1716–1724.
22. Heilbronn R, Bürkle A, Stephan S, zur Hausen H (1990) The adeno-associated virus *rep* gene suppresses herpes simplex virus-induced DNA-amplification. J Virol 64: 3012–3018.
23. Cortes ML, Oehmig A, Saydam O, Sanford JD, Perry KF, et al. (2008) Targeted integration of functional human ATM cDNA into genome mediated by HSV/AAV hybrid amplicon vector. Mol Ther 16: 81–88.
24. Zhang C, Cortez NG, Berns KI (2007) Characterization of a bipartite recombinant adeno-associated viral vector for site-specific integration. Hum Gene Ther 18: 787–797.
25. Wang H, Lieber A (2006) A helper-dependent capsid-modified adenovirus vector expressing adeno-associated virus rep78 mediates site-specific integration of a 27-kilobase transgene cassette. J Virol 80: 11699–11709.
26. Howden SE, Voullaire L, Wardan H, Williamson R, Vadolas J (2008) Site-specific, Rep-mediated integration of the intact beta-globin locus in the human erythroleukaemic cell line K562. Gene Ther 15: 1372–1383.
27. Recchia A, Parks RJ, Lamartina S, Toniatti C, Pieroni L, et al. (1999) Site-specific integration mediated by a hybrid adenovirus/adeno- associated virus vector. Proc Natl Acad Sci U S A 96: 2615–2620.
28. Recchia A, Perani L, Sartori D, Olgiati C, Mavilio F (2004) Site-specific integration of functional transgenes into the human genome by adeno/AAV hybrid vectors. Mol Ther 10: 660–670.
29. Wonderling RS, Owens RA (1997) Binding sites for adeno-associated virus Rep proteins within the human genome. J Virol 71: 2528–2534.
30. Schnepp BC, Jensen RL, Clark KR, Johnson PR (2009) Infectious molecular clones of adeno-associated virus isolated directly from human tissues. J Virol 83: 1456–1464.
31. Penaud-Budloo M, Le Guiner C, Nowrouzi A, Toromanoff A, Cherel Y, et al. (2008) Adeno-associated virus vector genomes persist as episomal chromatin in primate muscle. J Virol 82: 7875–7885.
32. Nakai H, Iwaki Y, Kay MA, Couto LB (1999) Isolation of recombinant adeno-associated virus vector-cellular DNA junctions from mouse liver. J Virol 73: 5438–5447.
33. Vincent-Lacaze N, Snyder RO, Gluzman R, Bohl D, Lagarde C, et al. (1999) Structure of adeno-associated virus vector DNA following transduction of the skeletal muscle. J Virol 73: 1949–1955.
34. Miller DG, Rutledge EA, Russell DW (2002) Chromosomal effects of adeno-associated virus vector integration. Nat Genet 30: 147–148.
35. Dyall J, Szabo P, Berns KI (1999) Adeno-associated virus (AAV) site-specific integration: formation of AAV-AAVS1 junctions in an in vitro system. Proc Natl Acad Sci U S A 96: 12849–12854.
36. Rizzuto G, Gorgoni B, Cappelletti M, Lazzaro D, Gloaguen I, et al. (1999) Development of animal models for adeno-associated virus site-specific integration. J Virol 73: 2517–2526.
37. Bushman F, Lewinski M, Ciuffi A, Barr S, Leipzig J, et al. (2005) Genome-wide analysis of retroviral DNA integration. Nat Rev Microbiol 3: 848–858.
38. Schmidt M, Schwarzwaelder K, Bartholomae C, Zaoui K, Ball C, et al. (2007) High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat Methods 4: 1051–1057.
39. Wu X, Li Y, Crise B, Burgess SM (2003) Transcription start regions in the human genome are favored targets for MLV integration. Science 300: 1749–1751.
40. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, et al. (2002) HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110: 521–529.
41. Meekings KN, Leipzig J, Bushman FD, Taylor GP, Bangham CR (2008) HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP. PLoS Pathog 4: e1000027.
42. Drew HR, Lockett LJ, Both GW (2007) Increased complexity of wild-type adeno-associated virus-chromosomal junctions as determined by analysis of unselected cellular genomes. J Gen Virol 88: 1722–1732.
43. Cheung AKM, Hoggan MD, Hauswirth WW, Berns KI (1980) Integration of the adeno-associated virus genome into cellular DNA in latently infected Human Detroit 6 cells. J Virol 33: 739–748.
44. Kotin RM, Berns KI (1989) Organization of adeno-associated virus DNA in latently infected Detroit 6 cells. Virology 170: 460–467.
45. Chiorini JA, Yang L, Safer B, Kotin RM (1995) Determination of adeno-associated virus Rep68 and Rep78 binging sites by random sequence oligonucleotide selection. J Virol 69: 7334–7338.
46. Miller DG, Trobridge GD, Petek LM, Jacobs MA, Kaul R, et al. (2005) Large-scale analysis of adeno-associated virus vector integration sites in normal human cells. J Virol 79: 11434–11442.
47. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.
48. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459: 108–112.
49. Chiorini JA, Wiener SM, Owens RA, Kyöstiö SRM, Kotin RM, et al. (1994) Sequence requirements for stable binding and function of Rep68 on the adeno-associated virus type 2 inverted terminal repeats. J Virol 68: 7448–7457.
50. Im D-S, Muzyczka N (1989) Factors that bind to adeno-associated virus terminal repeats. J Virol 63: 3095–3104.
51. McCarty DM, Ryan JH, Zolotukhin S, Zhou X, Muzyczka N (1994) Interaction of the adeno-associated virus Rep protein with a sequence within the A palindrome of the viral terminal repeat. J Virol 68: 4998–5006.
52. Ryan JH, Zolotukhin S, Muzyczka N (1996) Sequence requirements for binding of Rep68 to the adeno-associated virus terminal repeats. J Virol 70: 1542–1553.
53. Owens RA, Weitzman MD, Kyöstiö SRM, Carter BJ (1993) Identification of a DNA-binding domain in the amino terminus of adeno-associated virus rep proteins. J Virol 67: 997–1005.
54. Macville M, Schrock E, Padilla-Nash H, Keck C, Ghadimi BM, et al. (1999) Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. Cancer Res 59: 141–150.
55. Young SM, Jr., Samulski RJ (2001) Adeno-associated virus (AAV) site-specific recombination does not require a Rep-dependent origin of replication within the AAV terminal repeat. Proc Natl Acad Sci U S A 98: 13525–13530.
56. Hamilton H, Gomos J, Berns KI, Falck-Pedersen E (2004) Adeno-associated virus site-specific integration and AAVS1 disruption. J Virol 78: 7874–7882.
57. Henckaerts E, Dutheil N, Zeltner N, Kattman S, Kohlbrenner E, et al. (2009) Site-specific integration of adeno-associated virus involves partial duplication of the target locus. Proc Natl Acad Sci U S A 106: 7571–7576.
58. Hickman AB, Ronning DR, Perez ZN, Kotin RM, Dyda F (2004) The nuclease domain of adeno-associated virus rep coordinates replication initiation using two distinct DNA recognition interfaces. Mol Cell 13: 403–414.
59. Russell DW (2003) AAV loves an active genome. Nat Genet 34: 241–242.
60. Brister JR, Muzyczka N (1999) Rep-mediated nicking of the adeno-associated virus origin requires two biochemical activities, DNA helicase activity and transesterification. J Virol 73: 9325–9336.
61. Jang MY, Yarborough OH, 3rd, Conyers GB, McPhie P, Owens RA (2005) Stable secondary structure near the nicking site for adeno-associated virus type 2 Rep proteins on human chromosome 19. J Virol 79: 3544–3556.
62. Bianchi ME, Agresti A (2005) HMG proteins: dynamic players in gene regulation and differentiation. Curr Opin Genet Dev 15: 496–506.
63. Costello E, Saudan P, Winocour E, Pizer L, Beard P (1997) High mobility group chromosomal protein 1 binds to the adeno-associated virus replication protein (Rep) and promotes Rep-mediated site-specific cleavage of DNA, ATPase activity and transcriptional repression. Embo J 16: 5943–5954.
64. Hermonat PL, Santin AD, Batchu RB, Zhan D (1998) The adeno-associated virus Rep78 major regulatory protein binds the cellular TATA-binding protein in vitro and in vivo. Virology 245: 120–127.
65. Weger S, Wendland M, Kleinschmidt J, Heilbronn R (1999) The adeno-associated virus type 2 regulatory proteins Rep78/Rep68 interact with the transcriptional coactivator PC4. J Virol 73: 260–269.
66. Nash K, Chen W, Salganik M, Muzyczka N (2009) Identification of cellular proteins that interact with the adeno-associated virus rep protein. J Virol 83: 454–469.
67. Goncalves MA, van Nierop GP, Tijssen MR, Lefesvre P, Knaan-Shanzer S, et al. (2005) Transfer of the full-length dystrophin-coding sequence into muscle cells by a dual high-capacity hybrid viral vector with site-specific integration ability. J Virol 79: 3146–3162.
68. Cathomen T, Joung JK (2008) Zinc-finger nucleases: the next generation emerges. Mol Ther 16: 1200–1207.
69. Chiorini JA, Weitzman MD, Owens RA, Urcelay E, Safer B, et al. (1994) Biologically active rep proteins of adeno-associated virus type 2 produces as fusion proteins in *Escherichia coli*. J Virol 68: 797–804.

70. Cathomen T, Collete D, Weitzman MD (2000) A chimeric protein containing the N terminus of the adeno-associated virus rep protein recognizes its target site in an In vivo assay. J Virol 74: 2372–2382.
71. Heilbronn R, zur Hausen H (1989) A subset of herpes simplex replication genes induces DNA amplification within the host cell genome. J Virol 63: 3683–3692.
72. Hüser D, Weger S, Heilbronn R (2003) Packaging of human chromosome 19-specific adeno-associated virus (AAV) integration sites in AAV virions during AAV wild-type and recombinant AAV vector production. J Virol 77: 4881–4887.
73. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12: 656–664.
74. Döring A, Weese D, Rausch T, Reinert K (2008) SeqAn An efficient, generic C++ library for sequence analysis. BMC Bioinformatics 9: 11.
75. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, et al. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res 19: 24–32.
76. Robertson AG, Bilenky M, Tam A, Zhao Y, Zeng T, et al. (2008) Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. Genome Res 18: 1906–1917.
77. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
78. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.