Native Molecular State of Adeno-Associated Viral Vectors Revealed by Single-Molecule Sequencing

Philipp Kapranov,¹ Lingxia Chen,² Debra Dederich,¹ Biao Dong,² Jie He,¹ Kathleen E. Steinmann,¹ Andrea R. Moore,² John F. Thompson,¹ Patrice M. Milos,¹ and Weidong Xiao²

Abstract

The single-stranded genome of adeno-associated viral (AAV) vectors is one of the key factors leading to slow-rising but long-term transgene expression kinetics. Previous molecular studies have established what is now considered a textbook molecular model of AAV genomes with two copies of inverted tandem repeats at either end. In this study, we profiled hundreds of thousands of individual molecules of AAV vector DNA directly isolated from capsids, using single-molecule sequencing (SMS), which avoids any intermediary steps such as plasmid cloning. The sequence profile at 3' ends of both the regular and oversized vector did show the presence of an inverted terminal repeat (ITR), which provided direct confirmation that AAV vector packaging initiates from its 3' end. Furthermore, the vector 5'-terminus profile showed inconsistent termination for oversized vectors. Such incomplete vectors would not be expected to undergo canonical synthesis of the second strand of their genomic DNA and thus could function only via annealing of complementary strands of DNA. Furthermore, low levels of contaminating plasmid DNA were also detected. SMS may become a valuable tool during the development phase of vectors that are candidates for clinical use and for facilitating/accelerating studies on vector biology.

Introduction

ADENO-ASSOCIATED VIRUS (AAV) is a defective parvovirus with a single-stranded DNA (ssDNA) genome of approximately 4700 nucleotides. The virus contains unique 145 nucleotide *cis* elements (inverted terminal repeats, ITRs) which flank the 5' and 3' ends of the viral coding region and can assume a T-shape structure and function as the replication origin during AAV lytic infection and mediate viral integration and rescue in the latent phase of infection (Mayor *et al.*, 1969; Samulski *et al.*, 1982; Srivastava *et al.*, 1983; Muzyczka and Berns, 2002). Preservation of the ITRs alone is sufficient to confer AAV viral properties on any DNA placed between the ITRs, replacing the viral coding sequence, and thus research has focused on the development of recombinant AAV (rAAV) as a gene therapy vector (Samulski *et al.*, 1989) able to direct long-term transgene expression *in vivo* (Fisher *et al.*, 1997). Although the genetic composition of canonical AAV genomes has long been established (Srivastava *et al.*, 1983), the palindromic nature of the ITR sequence as well as its 80% GC content presents significant challenges to conventional molecular biology analysis of the viral ITR at a population level.

Nevertheless, the mechanisms leading to the persistence and stabilization of rAAV genomes remain unclear, and despite extensive research and deployment of AAV-based vectors for gene therapy applications (Manno *et al.*, 2006; Maguire *et al.*, 2008, 2009), basic questions about AAV biology, which are the key to further improvements in its application as a vector, remain. These include the following: (1) What is the encapsidation mechanism of AAV DNA? (2) Do all AAV genomes possess a similar ITR structure? (3) Can rAAV package additional nucleotide sequences beyond its normal capacity? (4) Is there a preference for early encapsidation termination?

At present, knowledge of AAV genomes has largely been derived from Sanger sequencing of DNA that was first converted into double-stranded DNA (dsDNA) form, typically obtained either by annealing of the extracted single-stranded viral DNA followed by a polymerase fill-in reaction or via isolation of the native replicative duplex (Samulski *et al.*, 1982, 1983; Srivastava *et al.*, 1983). However, this traditional plasmid-based cloning as the chosen method to recover AAV is not practical to yield sufficient sequence information to pinpoint the viral genome structure and potential heterogeneity that exists in their native ssDNA state.

¹Helicos BioSciences, Cambridge, MA 02139.

²Temple University, Philadelphia, PA 19140.

To obtain vital information related to the native ssDNA AAV genome encapsidated inside the viral capsids, care must be taken to minimize enzymatic treatments that may alter the composition of the genome. For example, the annealing of plus and minus strands of AAV along with DNA polymerase extension would allow one molecule to serve as the template for the other, obscuring the original form of the viral genome, so that the native 3' end of the ssDNA genome could not be identified. The second challenge of analyzing viral nucleic acid information is the requirement for high-throughput sequencing of substantial numbers of individual AAV viral molecules.

Therefore, a novel concept and approach is warranted in order to determine the native termini of thousands of AAV genomes in their viral capsids before any biological data can be inferred. We have taken advantage of the unique attributes of Helicos BioSciences (Cambridge, MA) single-molecule sequencing (SMS) technology and used them to analyze rAAV genomes. First, there is no cloning procedure needed, which not only makes large-scale and cost-effective analysis possible, but also eliminates potential artifacts arising from growing unstable AAV ITRs in bacteria. Second, the SMS procedures bypass the need for a DNA polymerase fill-in reaction that will obscure the 3' terminal nucleotides. Third, there is no need to ligate artificial DNA sequences to the rAAV genomes, which may have the potential drawback of favoring certain genomes. Fourth, SMS is much less sensitive to the high GC content found in the ITR of AAV (Hart *et al.*, 2010). Last, because there is no size selection involved, AAV genomes of all sizes can be detected in a simple sequencing reaction, allowing millions of molecules to be analyzed from a single rAAV preparation.

Materials and Methods

AAV vector production and purification

Recombinant AAV vectors were produced by the triple-transfection method, which has been described previously (Wang *et al.*, 2007). Briefly, a vector plasmid (with AAV2 inverted terminal repeats), a helper plasmid (with the AAV2 *rep* and *cap* genes), and mini-adenovirus helper plasmid (pFΔ6, with essential regions from the adenovirus genome) were cotransfected into 293 cells by calcium phosphate precipitation. AAV particles were purified by cesium chloride gradient ultracentrifugation. Vectors at a density of 1.39–1.45 g/ml³ were collected. The physical particle titers were determined by silver staining and quantitative dot-blot assay.

AAV DNA preparation

AAV vectors were digested with proteinase K at 42°C and the resulting DNA was purified with a Qiagen (Hilden, Germany) PCR purification kit. Because of the limited molarity and hairpin nature of the template DNA, sample preparation for sequencing involved a modification of the standard protocol (Powell, 2009). Reaction volumes were decreased whereas the dATP concentration was increased. For 3' end sequencing, 0.08 pmol of unsheared AAV-F9 DNA and a control reaction containing 0.074 pmol of AAV-F9 DNA and 0.25 μl of Helicos PolyA tailing control oligonucleotide (oligo) TR were denatured at 95°C for 5 min and

snap-cooled on ice to minimize reannealing of the templates. A 1.2-μl volume of terminal transferase 10× buffer (New England BioLabs [NEB], Ipswich, MA), 1.2 μl of NEB 2.5 mM CoCl₂, 10 units of NEB terminal transferase, and 1.95 μl of Helicos poly(A) (PolyA) tailing dATP were added to the denatured AAV-F9 DNA and control reaction in a final volume of 12 μl. The reactions were incubated at 37°C for 1 hr followed by 70°C for 10 min to denature the enzyme. The PolyA tail length of the oligo spike in the control reaction was monitored with an ABI 3730 DNA analyzer (Applied Biosystems, Foster City, CA). A minimal average tail length of 90 on the oligo spike ensures that the majority of the sample DNA will have a PolyA tail length of 50 or greater and will hybridize efficiently to the Helicos flow cell containing oligo-dT50 primer on the flow cell surface. It was necessary to repeat the denaturation and incubation steps, adding an additional 2 μl of Helicos PolyA tailing dATP and 20 units of NEB terminal transferase to the sample and control reactions before the average tail length on the oligo spike was greater than 90. For 3' end sequencing of AAV-F8, 0.047 pmol of unsheared AAV-F8 DNA and a control reaction containing 0.044 pmol of AAV-F8 DNA and 0.5 μl of Helicos PolyA tailing control oligo TR were denatured at 95°C for 5 min and snap-cooled on ice. A 1.5-μl volume of NEB terminal transferase 10× buffer, 1.5 μl of NEB 2.5 mM CoCl₂, 20 units of NEB terminal transferase, and 2.5 μl of Helicos PolyA tailing dATP were added to the denatured AAV-F8 DNA and control reaction in a final volume of 15 μl. The reactions were incubated at 37°C for 1 hr followed by 70°C for 10 min to denature the enzyme. It was necessary to repeat the denaturation and incubation steps two additional times, adding 2 μl of Helicos PolyA tailing dATP and 20 units of NEB terminal transferase to the sample and control reactions for the first repeat and 1 μl of dATP and 10 units of terminal transferase for the second repeat, before the average tail length on the oligo spike was greater than 90. Excess dATP was removed from both tailed sample reactions, using an EdgeBio (Gaithersburg, MD) Performa DTR gel filtration cartridge. Sample DNA was concentrated to 8 μl and denatured as described previously. The 3' ends of the sample reactions were blocked by adding 1 μl of NEB terminal transferase 10× buffer, 1 μl of NEB 2.5 mM CoCl₂, 10 units of NEB terminal transferase, and 0.25 μl of a 1:10 dilution of biotin-11-ddATP (NEL-545; PerkinElmer, Wellesley, MA) to the denatured DNA in a final volume of 10 μl. The reactions were incubated at 37°C for 1 hr followed by 70°C for 10 min to denature the enzyme.

For 5'-end sequencing, the 3' ends of the vector DNA were extended with LA Taq (TaKaRa Bio, Shiga, Otsu, Japan). AAV-F9 DNA (0.32 pmol) and 0.16 pmol of AAV-F8 DNA were extended by adding 5 μl of 10× LA PCR buffer II without Mg²⁺ (TaKaRa Bio), 5 μl of 25 mM MgCl₂, 8 μl of 2.5 mM dNTP mixture, and 2.5 units of LA Taq DNA polymerase (TaKaRa Bio) to each DNA sample in a total volume of 50 μl. The reactions were incubated at 72°C for 10 min, put on ice, and then run through an EdgeBio Performa DTR gel filtration cartridge to remove nucleotides and buffers. The extended AAV-F9 DNA (0.23 pmol) and a control reaction containing 0.12 pmol of AAV-F9 DNA and 0.25 μl of Helicos PolyA tailing control oligo TR were denatured at 95°C for 5 min and snap-cooled on ice. A 1.5-μl volume of NEB terminal transferase 10× buffer, 1.5 μl of

NEB 2.5 mM CoCl₂, 20 units of NEB terminal transferase, and 4 μ l of Helicos PolyA tailing dATP were added to the denatured AAV-F9 DNA and control reaction in a final volume of 15 μ l. Similarly, 0.12 pmol of extended AAV-F8 DNA and a control reaction containing 0.03 pmol of AAV-F9 DNA and 1 μ l of a 1:8 dilution of the Helicos PolyA tailing control oligo TR were denatured at 95°C for 5 min and snap-cooled on ice. A 1.5- μ l volume of NEB terminal transferase 10 $\times$ buffer, 1.5 μ l of NEB 2.5 mM CoCl₂, 20 units of NEB terminal transferase, and 2 μ l of Helicos PolyA tailing dATP were added to the extended AAV-F8 DNA in a final volume of 15 μ l. Half of these reagent volumes were added to the control reaction in a total volume of 7.5 μ l. The four reactions were incubated at 37°C for 1 hr followed by 70°C for 10 min to denature the enzyme. Additional incubations with more enzyme and dATP were not required to achieve an average of more than 90 tails on the control oligo spike with these DNAs. Excess dATP was removed from both tailed sample reactions, using an EdgeBio Performa DTR gel filtration cartridge. Sample AAV F9 DNA was concentrated to 15 μ l and denatured as described previously. The 3' end of the AAV F9 DNA was blocked by adding 2 μ l of NEB terminal transferase 10 $\times$ buffer, 2 μ l of NEB 2.5 mM CoCl₂, 20 units of NEB terminal transferase, and 0.5 μ l of a 1:10 dilution of biotin-11-ddATP (NEL-545; PerkinElmer) to the denatured DNA in a final volume of 20 μ l. Sample AAV-F8 was concentrated to 8 μ l. Reaction and reagent volumes were half those used for the extended AAV-F9. The reactions were incubated at 37°C for 1 hr followed by 70°C for 10 min to denature the enzyme.

SMS sequencing

After nucleic acid 3' tailing and blocking as described previously, DNA samples were hybridized to individual channels of the Helicos the oligo-dT50 flow cell; dTTP was added to fill in the corresponding PolyA tail on the individually hybridized viral DNA molecules and locked in place with the addition of the three Virtual Terminators, dATP, dCTP, and dGTP; and sequencing by synthesis proceeded for 120-nucleotide addition cycles, using the Heliscope sequencer (Harris *et al.*, 2008; Bowers *et al.*, 2009).

Data analysis

SMS reads were trimmed for leading T homopolymers and were filtered for reads with a minimal length of 8 bases after trimming with a suite of Helicos tools (available at <http://open.helicosbio.com/mwiki/index.php/Releases> and described at <http://open.helicosbio.com/HelisphereUserGuide.pdf>). Alignments were conducted with indexDP (Giladi *et al.*, 2010) software freely available at the Helicos website (<http://open.helicosbio.com/mwiki/index.php/Releases>), using the following parameters: seed size, 8; num errors, 1; weight, 6; best only; strands both; pass, 1. The aligner maximizes the aligned yield of SMS reads on the basis of the ability to align reads in which the predominant error is represented by deletions common in single-molecule sequencing. Reads were aligned to reference vector sequences with theoretical intact ITRs in FLIP and FLOP configuration at each terminal end. The alignments having a minimal normalized score (see below for description) of 4.0 were kept. The data presented

in this paper are based on aligned reads that have a minimal filtered read length of 12 bases. In Figs. 2 and 3 and Table 1, reads aligning uniquely and nonuniquely in an AAV genome were used: for reads that have more than one alignment with the same score, one such alignment was chosen randomly. In Fig. 4, all alignments (unique and all alignments for nonuniquely mapping reads) were used without randomly taking one nonunique alignment per read. The normalized score was defined as follows: $\text{Score} = (\text{no. of matches} \times 5 - \text{no. of mismatches} \times 4) / \text{length read}$. For example, in the following alignment:

Tag sequence: CCTCCGTGTTGTTCCAGCCCAGTGCTC
GCAGG

Ref. sequence: CTCCGTGTTGTTCCAGCCACAGTGCTC
GCAGG

Length of alignment block: 33

Length of tag sequence: 32

Number of matches: 31

Number of errors: 2

Score: $(31 \times 5) - (2 \times 4) = 155 - 8 = 147$

Normalized score: $147/32 = 4.59375$

Results

Strategy to map native ends of rAAV genome, using SMS

To determine the native 3' end (Fig. 1A), the vector DNA is directly extracted from purified particles without any of the additional DNA shearing often used for a standard SMS DNA sequencing protocol. This is essential to avoid creating artificial ends and to maintain the native state as much as possible for the ssDNA AAV genome. For profiling the 5' end, the T-shape structure of the 3' ITR was used to prime the terminal extension with a DNA polymerase (Snyder *et al.*, 1990a; Ward and Berns, 1991, 1995; Hong *et al.*, 1992) (Fig. 1B). The extension from the 3' ITR end would generate a new 3' end complementary to the original 5' viral end, allowing the endogenous 5'-end nucleotides to be deduced (Fig. 1B). After the DNA preparation outlined in Fig. 1A and B, the 3' ends of the native or 3'-extended AAV genomes are then directly polyadenylated, using dATP and terminal transferase. The polyadenylated DNA molecules are then hybridized to an oligo(dT) flow cell. During the normal sequencing reaction the unhybridized dAs in the tail are filled in with dTTP and each molecule is then locked in place with one of the other three Virtual Terminator nucleotides, fluorescent analogs of dATP, dCTP, and dGTP. This allows all molecules to be imaged on the flow cell surface before initiation of the sequencing by synthesis and subsequent 120-nucleotide addition cycles (Harris *et al.*, 2008). After sequencing, the quality-filtered reads with a minimal length of 12 nucleotides were aligned to the corresponding AAV reference sequence.

Native genome profile of regular-sized vector

The first rAAV vector we attempted to analyze was AAV-F9, a typical, canonical rAAV vector containing a 4.3-kb human factor IX cDNA expression cassette that can be accommodated by a 20- to 25-nm capsid. By following the 3' termini profiling strategy outlined in Fig. 1A, we

FIG. 1. Strategies to profile the 5' and 3' termini of rAAV vectors, using single-molecule sequencing (SMS). **(A)** Purified viral DNA is directly tailed at the 3' end and subjected to SMS. Any viral genome with a 3'-OH end will be sequenced. **(B)** The 5' profiling uses an additional extension step to extend the 3' ends of the folded 3' ITR to the 5' end of the viral genome. Thus, the 5' ends of vectors that have intact 3' ITRs will be profiled.

obtained the sequence read distribution of the 3' end for this vector, based on profiling of ~576K genomes (Fig. 2A). Interestingly, significant AAV-F9 alignments were in the AAV ITR region. As a unique feature of the SMS method employed in this study, the start position of the SMS reads will always mark the 3' ends of the sequenced DNA molecules. Therefore, the sequence strategy of Fig. 1A only captures the 3' termini of the individual DNA molecules representing the native AAV genomes. Importantly, the sequences of 5' and 3' ITRs are identical and thus cannot be differentiated by the short read SMS technology. Nonetheless, the distribution profile of 3' AAV termini allows us to conclude that all encapsulated AAV genomes have their 3' ends in an ITR.

The existence of an ITR at the 3' termini allowed us to analyze the 5' termini of AAV vectors, using the strategy illustrated in Fig. 1B. To determine the 5' termini, vector DNA was subjected to one round of a fill-in reaction, using TaKaRa LA Taq at 65°C before SMS analysis. This allows extension of the 3' end of the dsDNA through to the end of the single-stranded DNA. In the case of AAV-F9, the overall 5' profile was similar to the 3' profile, based on analysis of ~14.5M individual genomes (Fig. 2B). The overall symmetrical distribution pattern of the 5' and 3' ITR termini suggests that the 5' termini are similar to the 3' termini. A notable difference was the presence of a significant signal outside of the ITR regions indicative of truncated viral genomes (Fig. 2B; and see Discussion).

FIG. 2. DNA termini distribution of a regular-size vector, AAV-F9. **(A)** 3' end profiling. **(B)** 5' end profiling. The x axis represents the vector reference nucleotide position. The y axis represents counts of starts of SMS reads mapping to each position. The first and the last 145 nucleotides are ITR sequences (shaded). The upper half of each graph represents the distribution of termini corresponding to the virus with DNA of plus-strand polarity. The lower half of each graph represents the distribution of termini corresponding to the virus with DNA of minus-strand polarity. The data have been normalized to 1 million total aligned reads for comparison purposes. Regions corresponding to defective interfering (DI) particles are identified. Color images available online at www.liebertonline.com/hum

Native genome profile of oversized vector

We carried out the same experimental sequencing with an oversized AAV vector that cannot be packaged normally. AAV-F8 is a nontypical rAAV vector containing a 5.8-kb human factor VIII cDNA expression cassette and thus is ~ 1.1 kb over the normal 4.7-kb capacity of an AAV capsid. If AAV initiates packaging from the 5' end, it should, because of the oversized insert, terminate before reaching the 3' ITR and thus have 3' ends distal to the true 3' ITR. Therefore,

after SMS analysis we would expect that most of the reads would not map to an ITR but, rather, are found elsewhere in the sequence of the vector genome. If, on the other hand, the packaging initiates from the 3' end (in the 3' ITR), the incomplete packaging will generate truncated 5' ends that will not be detected by the native 3' SMS and the 3' ends of the vector genomes will always map to an ITR.

The results of the overall AAV-F8 3' end profiling are shown in Fig. 3A. The sequence data clearly show no signal over the general background level at any region in the body

FIG. 3. DNA termini distribution of an oversized vector AAV-F8. **(A)** 3' end profiling. **(B)** 5' end profiling. The x axis is the vector reference nucleotide position. The y axis represents counts of starts of SMS reads mapping to each position. The first and the last 145 nucleotides are ITR sequences (shaded). The upper half of each graph represents the distribution of termini corresponding to the virus with DNA of plus-strand polarity. The lower half of each graph represents the distribution of termini corresponding to the virus with DNA of minus-strand polarity. The data have been normalized to 1 million total aligned reads for comparison purposes. Regions corresponding to DI particles are identified. Color images available online at www.liebertonline.com/hum

of the vector beyond the ITRs. Therefore, through the analysis of DNA representing $\sim 684\text{K}$ independent AAV-F8 genomes, we confirmed that AAV packaging initiates from the 3' end, in the 3' ITR.

We then performed the extension reaction to analyze the 5' termini of the $\sim 29\text{M}$ individual genomes of the oversized vector. The results are shown in Fig. 3B. The notable observation is that the pattern of signal is represented by a single peak on each strand—in the 3' ITR. Moreover, this peak has been shifted completely from the position of 122 nucleotides to 142 nucleotides in the ITR, indicative of the complete extension reaction. The absence of the peak in the 5' ITR is consistent with the packaging starting at the 3' ITR and terminating after packaging the allowed $\sim 4.7\text{kb}$ of the vector genome before reaching the 5' ITR. The notable difference compared with the 3' ITR mapping for this vector is the increase in the distribution of 5' ends across the whole AAV genome with some higher density in the first 500 nucleotides, similar to the case for the AAV-F9 vector (Fig. 3B). This is all consistent with the presence of a large frac-

tion of truncated genomes in the preparations of both vectors (see Discussion).

Details of inverted terminal repeats in AAV vectors

It has been documented that an ITR is the only *cis* element required for AAV packaging and thus is absolutely critical for the viral infection process (Muzyczka and Berns, 2002). We thus looked further into the integrity of the AAV ITR. The ITR can occur in two alternative configurations—FLIP and FLOP—as shown in Fig. 4C and D. The two configurations arise from AAV replication, which can be predicted by a Cavalier-Smith model (Muzyczka and Berns, 2002). Interestingly, the distribution of reads before or after the extension reaction in the ITR region indicated that both the regular vector (AAV-F9) and oversized vector (AAV-F8) followed a reproducible pattern specific to each vector (Fig. 4A and B). For AAV-F9, the highest peaks are at the end of the ITR, which suggests that the main population of AAV particles have intact ITRs at both the 5' and 3'

FIG. 4. DNA termini distribution in the ITR regions. (A and B) 3' and 5' profiling of AAV ITRs in AAV-F9. (C and D) 3' and 5' profiling of AAV ITRs in AAV-F8. The y axis represents the distribution percentage based on all alignments (see Materials and Methods) falling into the ITR regions. The x axis represents the ITR reference nucleotide in either the FLIP or FLOP configuration, which are shown. Symbols on the x axis: *, a and a' sequences; periods (.), b and b' sequences; ^^, AAA or TTT loop; colons (:), c and c' sequences; |, d region.

termini. At the nucleotide level, the actual peak is not at the first nucleotide of the ITR. Rather, it was at the first C of the last four bases of the virus: ...CCAA-3'. This initiation or termination appears exactly as predicted based on the resolution of the SMS technology. SMS reads typically start at the second nucleotide position after a non-A base because of the A-fill and nucleotide lock required for imaging (Lipson *et al.*, 2009). However, the 3' termini distribution for both vectors shows that in addition to a "perfect" initiation from the very end of the 3' ITR, there are a considerable number of encapsidation events initiating from other positions within the ITR. This suggests that viral packaging mechanisms are relatively sloppy as at most 34.5–65% of the AAV-F9 and AAV-F8 vectors, respectively, whose 3' ends are within the 3' ITR, have 3' ends within 10 bases of the annotated 3' end.

Although data in Fig. 3B clearly suggest that ITRs are absent from the 5' termini of AAV-F8, there are still sequencing reads aligned to the 3' ITR region after extension reaction (Fig. 4C and D). These reads are derived from the 3' ITRs that could not be extended, mostly from the molecules with truncated 3' ITRs. Molecules of viral genomes with intact ITRs that form the proper T-shape configuration that can prime the extension reaction (Fig. 1B) were extended and reads marking their extended 3' ends no longer contribute to the ITR regions shown in Fig. 4C and D. The molecules with truncated 3' ITRs that can no longer serve as the primers formed a special "finger print" that did not change significantly compared with the original 3' profiling. Interestingly, the profile changes illustrated the consistency of both SMS technology and the AAV 3' distribution profile. In addition, for the first time, the profile difference between the FLIP and FLOP configurations shown in Fig. 4C and D suggests that these two configurations are not equal in the packaging process.

Single-molecule sequencing may be used as a vector quality control tool

In the AAV termini mapping experiments, we observed SMS reads that correspond to the plasmid backbone flanking the AAV vector sequences. An example comparison of alignments to various regions for AAV-F9 vector plasmids used for production is presented in Table 1; the fraction of

these alignments is minor compared with the reads that map to the ITRs (1:1000). There is not much difference in the fraction of reads mapping to non-ITR portions of the AAV vector sequence or the plasmid backbone sequence. We concluded that these reads were derived from residual DNA fragments created by extensive nuclease treatment. Because SMS can detect such fragments, it may be a useful tool as a quality control assay.

Discussion

Detection of various forms of vector DNA genomes in their native states without modifications can provide valuable clues about the basic mechanisms of vector biology related to vector production and performance *in vivo*. Until more recently no direct methods were available to achieve this goal. Current molecular biology methods require converting vector genomes into forms that can be amplified and studied by the canonical tool kit of molecular biology, such as cloning in plasmids followed by sequencing selected clones, Southern blot analyses, PCR amplification, and so on. For example, an ssDNA virus had to be converted to a dsDNA form before sequencing, or an RNA virus had first to be converted into a cDNA form. Such manipulations can hide or distort the native forms of the vector genome and thus influence our understanding of the basic vector biogenesis processes. Here we present a study to uncover the basic mechanisms of AAV vector biology via mapping the native states of the vector genomes.

Direct analysis of the native ends of AAV vectors allowed us to conclude that AAV packaging was initiated from the 3' end (Fig. 5). SMS results also allowed us to investigate the AAV termination preference. The observed diversity of locations of SMS reads marking 5' ends throughout the viral genome allowed us to conclude that the termination process is more relaxed and could thus generate less consistently defined 5' ends. Thus, this leads to the formation of particles that have smaller genomes, referred to as defective interfering (DI) particles (Hauswirth and Berns, 1977, 1979; Laughlin *et al.*, 1979). The 5' end pattern of the AAV-F8 vector provides additional insights concerning AAV vector genomes. Figure 3B suggested the existence of large amounts of DI particles that contain only the ITR sequence: extensions of 3' ends of such particles were found to

TABLE 1. COMPARISON OF FIRST NUCLEOTIDE DISTRIBUTION RELATED TO VARIOUS POSITIONS IN AAV-F9 VECTOR PLASMIDS

Final 10 nucleotides of ITR		Representative internal viral region		Representative plasmid backbone	
Nucleotide ^a	No. of reads ^b	Nucleotide ^a	No. of reads ^b	Nucleotide ^a	No. of reads ^b
4318	465	2104	11	7684	11
4319	127	2105	12	7685	12
4320	1,181	2106	26	7686	11
4321	2,050	2107	18	7687	12
4322	3,933	2108	25	7688	6
4323	13,350	2109	18	7689	18
4324	25,273	2110	21	7690	15
4325	5,553	2111	9	7691	17
4326	5,546	2112	22	7692	14
4327	166	2113	17	7693	24

^aNucleotide position in the reference sequence.

^bNumber of SMS reads whose 5' ends map to this position.

FIG. 5. A model for rAAV packaging. The 3' end of AAV ITR is inserted into the preformed capsid first. *a*, early termination signal (ETS) leading to DI particles; *b*, mature termination signal (MTS), which is the 5' ITR; *c*, maximal capacity of the vector; Rep78 and Rep52, viral proteins involved in rAAV packaging; *, terminal resolution site of the ITR.

terminate at the 5' end of an ITR: the 122- to 142-nucleotide shift mentioned above. It represents a new class of DI particles besides the similar DI particles observed with the AAV-F9 vector. Such premature termination had been observed previously (King *et al.*, 2001), yet these data were apparently ignored and no further follow-up studies have been reported.

It is clear that DI particles found in the AAV-F8 preparation are different from those found in the AAV-F9 preparation. This is probably due to the oversized nature of AAV-F8, which leads to more erratic termination. Mechanistically, DI particles with just the ITRs can be explained readily by the known biochemical activity of AAV replication proteins (Snyder *et al.*, 1990b; Kotin *et al.*, 1992; Muzyczka, 1992). If AAV Rep nicks at the terminal resolution site within the ITR, in the same manner as it does for the mature ITR at the 5' ITR, this would lead to premature termination of AAV packaging (Fig. 5). We speculate that the encapsidation machinery probably nicks at the initiating 3' ITR more frequently when the 5' ITR is not within the regular packaging limit, which is the case with the oversized AAV-F8 vector, which is larger than 4700 nucleotides. This is consistent with the notion that the ITR is supposedly the binding site for cellular and viral proteins involved in the packaging machinery (Snyder *et al.*, 1990b). Because DI particle formation is not a rare event, DI particles less than half the size of the genome can account for as much as 50% of all 5' ends detected (Figs. 2B and 3B). Such defective vector genomes would have a dramatic and negative effect in vectors produced for gene therapy, resulting in significantly lower rAAV production yields as well as reducing the transgene expression level or efficiency. This result is also consistent with our observation using a rescue assay (Dong *et al.*, 2010), in which we showed the majority of recovered molecules are small in size. Furthermore, we did not observe a dominant termination site/region 4.7 kb distal to the 3' ITR from the AAV-F8 vector (Fig. 3B), consistent with otherwise random termination.

Such an imperfect genome profile revealed by SMS would exert a significant impact on the mechanism of rAAV vector transgene expression. Transgene expression derived from an ssDNA AAV vector requires duplex DNA presumably arising from either of the two highly disputed mechanisms in the field of gene therapy: (1) annealing of AAV genomes of opposite polarities (Nakai *et al.*, 2000); or

(2) synthesis of the complementary strand, using the T shape-structured ITR (Ferrari *et al.*, 1996; Fisher *et al.*, 1996). ITRs that cannot form the correct ITR shape would not be able to self-prime as required for second-strand DNA synthesis. This would explain physically why AAV would be inherently inefficient for transgene expression. Rather, those genomes would require the annealing mechanism to make them competent for transgene expression. As noted, the profile of 3' ends within each ITR was unique to each vector. The reason for this is currently unknown, but it suggests that sequences outside the ITR influence 3' end initiation of packaging.

Although the data from Figs. 2B and 3B suggest that the entire ITR to ITR sequences fitting the AAV capsid cannot be detected when AAV genomes are much larger than the regular capacity of AAV capsids, it supports a hypothesis that gene expression for large AAV genomes comes from complementation of incomplete genomes (DI) (Dong *et al.*, 2010; Lai *et al.*, 2010; Wu *et al.*, 2010). Whether such vectors should be used clinically deserves further deliberation.

SMS is a sensitive technology that may be extended as a quality control tool. We consistently detected residual plasmid contamination from the vector preparation in the SMS results. Although SMS cost is currently high, it is projected to be lower when this technology becomes a routine assay in the future. It would be a good tool with which to verify vector genome composition and contaminant levels before a vector can be used clinically.

It is worth noting that the overall reads postextension are much more than those of the preextended sequencing reaction. We speculate that the reason for this is that the native 3' ends would be in a recessive configuration whereas the 3' ends produced by DNA polymerase extension would be in the blunt-end configuration. It is known that the latter would be a better substrate for the terminal transferase and thus would explain the difference. In this respect, it is important to emphasize that we always use counts normalized to the total number of mapped reads per experiment—by doing so we remove the experiment-specific differences, such as the total number of viral DNAs loaded, DNA purity, tailing efficiency, and so on. Another issue is that there is no “direct” way to map the profile of the 5' ends. So far, all information on 5' end positions has been obtained from enzymatic extension of the viral DNA. Because the DNA polymerase will stop at the true 5' end, the longest extension product would represent the native 5' end. This assumption has guided us for the previous 20 years of study using traditional approaches. At the individual molecule level, it is not possible to assert that every read is error free. However, the entirety of the individual molecule data should be able to tell us where the true 5' ends are, based on the accumulation of reads at these locations. Our results suggest that this approach is working: for example, the 5' results from the short factor IX vector showed the largely intact 5' ITR whereas the intact 5' ITR is not shown in the large factor VIII vectors. Because we used DNA polymerase to extend the 3' end, there is also concern about the reinitiation of extension. Because reinitiation would be a rare event under the conditions we used for extension, it should have minimal impact on the overall results.

In summary, we have extended the SMS technology to study AAV genomes at the single-molecule DNA level. The

original molecular states of hundreds of thousands and millions of individual AAVs were accurately assessed, which would not have been possible without SMS. The fine details of the AAV ITR were mapped at both the 5' and 3' ends. Our results confirm the hypothesis that AAV packaging initiates from the 3' end. The 5' end mapping revealed premature termination as a mechanism that leads to inefficient rAAV packaging. It is commonly thought that different AAV vectors are packaged with different efficiency. One can envision that further studies using SMS can be carried out to map immature termination in the AAV packaging process and thus allow further improvement of AAV vector preparation for human gene therapy. Also, given the fact that SMS technology can sequence RNA directly (Ozsolak *et al.*, 2009, 2010), one can also apply this approach to vectors based on RNA viruses such as retroviral or lentiviral vectors.

Acknowledgments

W.X. is supported by National Institutes of Health grants R01HL080789 and R01HL084381. A.R.M. is supported by NIH training grant in Thrombosis (T32 HL077777).

Author Disclosure Statement

P.K., D.D., J.H., K.S., J.F.T., and P.M.M. are or were employees of Helicos BioSciences Corporation throughout the duration of this research.

References

- Bowers, J., Mitchell, J., Beer, E., *et al.* (2009). Virtual terminator nucleotides for next-generation DNA sequencing. *Nat. Methods* 6, 593–595.
- Dong, B., Nakai, H., and Xiao, W. (2010). Characterization of genome integrity for oversized recombinant AAV vector. *Mol. Ther.* 18, 87–92.
- Ferrari, F.K., Samulski, T., Shenk, T., and Samulski, R.J. (1996). Second-strand synthesis is a rate-limiting step for efficient transduction by recombinant adeno-associated virus vectors. *J. Virol.* 70, 3227–3234.
- Fisher, K.J., Gao, G.P., Weitzman, M.D., *et al.* (1996). Transduction with recombinant adeno-associated virus for gene therapy is limited by leading-strand synthesis. *J. Virol.* 70, 520–532.
- Fisher, K.J., Jooss, K., Alston, J., *et al.* (1997). Recombinant adeno-associated virus for muscle directed gene therapy. *Nat. Med.* 3, 306–312.
- Giladi, E., Healy, J., Myers, G., *et al.* (2010). Error tolerant indexing and alignment of short reads with covering template families. *J. Comput. Biol.* 17, 1397–1411.
- Harris, T.D., Buzby, P.R., Babcock, H., *et al.* (2008). Single-molecule DNA sequencing of a viral genome. *Science* 320, 106–109.
- Hart, C., Lipson, D., Ozsolak, F., *et al.* (2010). Single-molecule sequencing: Sequence methods to enable accurate quantitation. *Methods Enzymol.* 472, 407–430.
- Hauswirth, W.W., and Berns, K.I. (1977). Origin and termination of adeno-associated virus DNA replication. *Virology* 78, 488–499.
- Hauswirth, W.W., and Berns, K.I. (1979). Adeno-associated virus DNA replication: Nonunit-length molecules. *Virology* 93, 57–68.
- Hong, G., Ward, P., and Berns, K.I. (1992). *In vitro* replication of adeno-associated virus DNA. *Proc. Natl. Acad. Sci. U.S.A.* 89, 4673–4677.
- King, J.A., Dubielzig, R., Grimm, D., and Kleinschmidt, J.A. (2001). DNA helicase-mediated packaging of adeno-associated virus type 2 genomes into preformed capsids. *EMBO J.* 20, 3282–3291.
- Kotin, R.M., Linden, R.M., and Berns, K.I. (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.
- Lai, Y., Yue, Y., and Duan, D. (2010). Evidence for the failure of adeno-associated virus serotype 5 to package a viral genome ≥ 8.2 kb. *Mol. Ther.* 18, 75–79.
- Laughlin, C.A., Myers, M.W., Risin, D.L., and Carter, B.J. (1979). Defective-interfering particles of the human parvovirus adeno-associated virus. *Virology* 94, 162–174.
- Lipson, D., Raz, T., Kieu, A., *et al.* (2009). Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.* 27, 652–658.
- Maguire, A.M., Simonelli, F., Pierce, E.A., *et al.* (2008). Safety and efficacy of gene transfer for Leber's congenital amaurosis. *N. Engl. J. Med.* 358, 2240–2248.
- Maguire, A.M., High, K.A., Auricchio, A., *et al.* (2009). Age-dependent effects of RPE65 gene therapy for Leber's congenital amaurosis: A phase 1 dose-escalation trial. *Lancet* 374, 1597–1605.
- Manno, C.S., Pierce, G.F., Arruda, V.R., *et al.* (2006). Successful transduction of liver in hemophilia by AAV-Factor IX and limitations imposed by the host immune response. *Nat. Med.* 12, 342–347.
- Mayor, H.D., Torikai, K., Melnick, J.L., and Mandel, M. (1969). Plus and minus single-stranded DNA separately encapsidated in adeno-associated satellite virions. *Science* 166, 1280–1282.
- Muzyczka, N. (1992). Use of adeno-associated virus as a general transduction vector for mammalian cells. *Curr. Top. Microbiol. Immunol.* 158, 97–129.
- Muzyczka, N., and Berns, K.I. (2002). Parvoviridae: The viruses and their replication. In *Fields Virology*. D.M. Knipe, P.M. Howley, D.E. Griffin, *et al.*, eds. (Lippincott Williams & Wilkins, Philadelphia).
- Nakai, H., Storm, T.A., and Kay, M.A. (2000). Recruitment of single-stranded recombinant adeno-associated virus vector genomes and intermolecular recombination are responsible for stable transduction of liver *in vivo*. *J. Virol.* 74, 9451–9463.
- Ozsolak, F., Platt, A.R., Jones, D.R., *et al.* (2009). Direct RNA sequencing. *Nature* 461, 814–818.
- Ozsolak, F., Kapranov, P., Foissac, S., *et al.* (2010). Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143, 1018–1029.
- Powell, J.S. (2009). Recombinant factor VIII in the management of hemophilia A: Current use and future promise. *Ther. Clin. Risk Manag.* 5, 391–402.
- Samulski, R.J., Berns, K.I., Tan, M., and Muzyczka, N. (1982). Cloning of adeno-associated virus into pBR322: Rescue of intact virus from the recombinant plasmid in human cells. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2077–2081.
- Samulski, R.J., Srivastava, A., Berns, K.I., and Muzyczka, N. (1983). Rescue of adeno-associated virus from recombinant plasmids: Gene correction within the terminal repeats of AAV. *Cell* 33, 135–143.
- Samulski, R.J., Chang, L.S., and Shenk, T. (1989). Helper-free stocks of recombinant adeno-associated viruses: Normal integration does not require viral gene expression. *J. Virol.* 63, 3822–3828.
- Snyder, R.O., Im, D.S., and Muzyczka, N. (1990a). Evidence for covalent attachment of the adeno-associated virus (AAV) Rep

- protein to the ends of the AAV genome. *J. Virol.* 64, 6204–6213.
- Snyder, R.O., Samulski, R.J., and Muzyczka, N. (1990b). *In vitro* resolution of covalently joined AAV chromosome ends. *Cell* 60, 105–113.
- Srivastava, A., Lusby, E.W., and Berns, K.I. (1983). Nucleotide sequence and organization of the adeno-associated virus 2 genome. *J. Virol.* 45, 555–564.
- Wang, J., Xie, J., Lu, H., *et al.* (2007). Existence of transient functional double-stranded DNA intermediates during recombinant AAV transduction. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13104–13109.
- Ward, P., and Berns, K.I. (1991). *In vitro* rescue of an integrated hybrid adeno-associated virus/simian virus 40 genome. *J. Mol. Biol.* 218, 791–804.
- Ward, P., and Berns, K.I. (1995). Minimum origin requirements for linear duplex AAV DNA replication *in vitro*. *Virology* 209, 692–695.
- Wu, Z., Yang, H., and Colosi, P. (2010). Effect of genome size on AAV vector packaging. *Mol. Ther.* 18, 80–86.

Address correspondence to:

Dr. Patrice M. Milos
Helicos BioSciences
Building 700, One Kendall Square
Cambridge, MA 02139

E-mail: pmilos@helicosbio.com

Dr. Weidong Xiao
Temple University
3400 N. Broad Street, OMS-531
Philadelphia, PA 19140

E-mail: wxiao@temple.edu

Received for publication August 25, 2011;
accepted after revision August 29, 2011.

Published online: August 29, 2011.