CELL & GENE THERAPY INSIGHTS

CHARACTERIZATION AND VALIDATION

REVIEW

Profiling AAV vector heterogeneity & contaminants using next-generation sequencing methods

Ngoc Tam Tran and Phillip WL Tai

AAV vectors continue to be the most promising gene delivery vehicle for treating rare genetic diseases through gene therapy. Understanding vector inconsistencies during the manufacturing process is vital to define batch-to-batch differences, and predicting their efficacies and safety profiles. Although AAV vectors manufactured for clinical use are rigorously tested by several analytical methods, these assays are still not able to provide comprehensive insights into a vector's composition, nor address how or why heterogeneity in vectors emerge. With the power of next-generation sequencing methods, understanding AAV vector composition and why certain designs fail to provide expected potencies can be unlocked.

Cell & Gene Therapy Insights 2023; 9(11), 1565–1583

DOI: 10.18609/cgti.2023.206

INTRODUCTION

AAVs were originally discovered in 1965 as 'virus-like' particles [1,2]. AAVs belong to a class of small, non-enveloped, dependoparvoviruses that rely on co-infection with helper viruses, such as adenovirus or herpesvirus to complete their lifecycles in the host [3]. AAV is single-stranded DNA virus that packages either the plus or minus strand of the genome at equal ratios into an icosahedral protein capsid that is approximately 20–25 nm in diameter [4,5]. The AAV genome has four known open reading frames (ORFs) that encode for the viral replication genes (*rep*), the capsid proteins (cap), the assembly-activating protein (*AAP*), and the membrane-associated accessory proteins (*MAAP*) [6]. The *rep* gene encodes for Rep40, Rep52, Rep68, and Rep78 [7]. The *cap* gene encodes three viral proteins called VP1, VP2, and VP3, which form the 60-mer capsid at approximate ratios of 1:1:10 for VP1:VP2:VP3, respectively [6,7]. The AAV family of viruses is fairly diverse. Among those that can infect humans and non-human primates, there are seven main clades (clades A-G) [8,9], which encompass AAV1/6 (clade A), AAV2 (clade B), AAV2/3-hybrid and AAV13 (clade C), AAV7 (clade D), AAV8 (clade E), AAV9 (clade F), and AAV4, AAV11, and AAV12 (clade G). AAV5 is the most distinct among the contemporary capsids, and is currently in its own class. Differences in capsid surface antigens have traditionally define viral serotypes; however, among the wildtype AAVs that have been discovered, over hundreds of naturally occurring variants have been identified based on sequence analyses [6]. Importantly, these serotypes and subvariants have different tropism profiles among several mammalian laboratory models that span an array of cell and tissue types.

The AAV genome is flanked by two inverted terminal repeats (ITRs) that are required for rescue, replication, and packaging of the genome. Similar to other parvoviruses, the ITR overcomes the end-replication problem through rolling-hairpin replication [10]. The wild-type ITR from AAV serotype 2 (AAV2) is 145 nt in length and comprises of four internal segments [11]. Its first 125 nt folds on itself to form a T-shaped hairpin with two small internal inverted repeat sequences, named the B and C arms. The stem of the T-shaped hairpin is called the A segment. The rest of the ITR, which is contiguous with the rest of the genome, forms the D sequence. The inverted nature of the ITR is essential for virus genome replication, as it serves as an origin of replication as a self-primed molecule. Embedded within the A sequence is the Rep-binding element (RBE). Together with the RBE, a sequence that is located at the tip of the cross arms, called RBE', serve to recruit Rep68/78, which nicks the terminal resolution site to separate the newly synthesized DNA strand from the template strand [12].

There are several features that make AAV ideal vehicles for gene therapy [11,13,14]. First, they cannot replicate on their own, but require factors expressed by the helper virus. These specific factors, namely those from adenovirus (E1A, E1B, E2A, E4, and viral associated RNA), can be expressed in *trans* to drive AAV replication and genome packaging. Second, AAVs confer low immunogenicity and pathogenicity. In recent years, AAV has been linked to hepatocellular carcinoma and specific cases of acute hepatitis [15–18], but the mechanisms that drive these outcomes are not fully known and are hotly debated. Third, AAV vectors can confer long-lasting transgene expression, since their genomes predominantly persist as circular double-stranded episomes in the host cell nucleus [4].

There have been multiple methods for AAV vector production described throughout the years [19–26]. However, plasmid transfection in HEK293 cells (pTx/HEK293), recombinant baculovirus infection in insect cells (rBV/Sf), and HeLa production cell lines with adenovirus are currently the three most popular production platforms for manufacturing recombinant (r)AAV for basic research, pre-clinical, and clinical use. Unfortunately, the potency of AAV vectors is inexplicably known to be impacted by the manufacturing method [14,27,28]. There are many quality control challenges in producing effective and safe vectors. Purification methods can also vary and impact the quality of AAV vectors. Vector purification chiefly involves obtaining high quality particles that are free from partial or empty particles. Many techniques have been developed for vector purification [26,29–33]; but currently, there is no single method that can completely remove empty particles from preparations. Despite well-established pipelines developed for obtaining safe and quality vectors, the final product can still contain defective vectors and contaminants [34]. Therefore, characterizing and validating AAV vectors are essential for assuring that the final product meets safety, purity, and quality standards set by the US FDA.

ANALYTICAL METHODS FOR AAV VECTOR CHARACTERIZATION & EVALUATION

Product characterization under GMP must follow guidelines required by the FDA [35]. Different analytical methods are used for characterizing and validating AAV vectors. In general, these assays evaluate a vector's identity, potency, purity, safety, and stability. [35]. These methods have been reviewed extensively [36–42]. The following metrics and the analytical methods that measure them have been industry standards for querying batch-to-batch heterogeneity.

Vector genome titration

The traditional way of quantifying viruses with infectious titers cannot be used for recombinant (r)AAVs, since the highly engineered nature of these vectors make infection a less reliable means of gauging their titers. Therefore, methods to obtain physical titers are favored. The standard means of quantifying vectors relies on the detection of vector genomes in the preparation, for which quantitative PCR (qPCR) has served as the method of choice. However, accuracy of qPCR is dependent on primer efficiencies. Since many research vectors can vary in design, primer/probe sets typically target sequences that are commonly shared, such as the polyadenylation sequence or regions proximal to the ITRs. Digital Droplet (dd)PCR has become more attractive, since the method is not as severely impacted by primer efficiencies as it is with qPCR. The only drawback of qPCR/ddPCR, as with any DNA-based detection method, is that non-encapsidated DNAs (carry-over from production) that survive endonuclease digestion during vector purification steps can be detected, leading to the overestimation of vector titers.

Particle titration

Quantification of vector DNA may not accurately reveal the abundance of vector particles in preparations, since some particles may lack vector genomes (empty capsids). Although empty capsids do not contribute to the overall transduction and potencies, they will impact how the host will respond to dosing, which is typically based on vector genome titers. Particle titers are typically quantified using ELISAs, using a monoclonal antibody that is specific to the fully assembled capsid. Antibodies are typically serotype-specific. In research settings, sodium dodecyl-sulfate polyacrylamide gel electrophoresis (SDS-PAGE), followed by silver staining or Western blotting is still used. Since silver staining does not rely on antibodies, it is typically favored for the semi-quantitative assessment of VP1, VP2, and VP3 ratios and/or capsid degradation. Although the exact ratios on the single particle scale is stochastic [6], VP ratios that deviate from 1:1:10 tend to be attributed to poor vector titers and/or associated with reduced potencies [20]. More advanced methods based on high-resolution native mass spectrometry can obtain clearer pictures of differential VP ratios in preparations [6]; but how these differences impact transduction is still unexplored.

Detection of plasmid, host cell DNA contaminants, & adventitious virus

Demonstration of vector genome purity is one means of showing that the vectors being produced are free from risks associated with the transfer of foreign DNA. Foreign DNA can encompass any material originating from the production process. This can include DNA from backbone sequences, such as antibiotic resistance genes (e.g., β-lactamase) used in the production plasmids, viral proteins originating from manufacturing schemes that use adenovirus vectors, and DNA that can originate from the packaging cell line. Importantly, detection of viral sequences not related to the production platform may signify the presence of adventitious viruses that can originate from animal serum found in cell culture media. Adventitious viruses can propagate during the manufacturing process and can elicit strong immune responses in patients, leading to adverse effects and lowered gene therapy efficacies. It should be noted that for commercial manufacturing, production schemes now typically use animal-derived component-free media, thereby limiting adventitious viruses.

The direct method for detecting DNA contaminants is via qPCR/ddPCR using primer/probes that target specific sequences. For example, to detect plasmid contaminants, primer/probes targeting the antibiotic resistance gene can be used; for targeting host-cell contaminants, 18S ribosomal RNA or Alu targets is routinely used [43]; and for adventitious viruses, a panel primer/probes that target a range of viral DNAs are employed [44]. However, PCR-based methods are inherently problematic, since low abundance contaminants can be hard to detect, even under exponential amplification. In addition, only known target sequences can be queried, limiting the detection of host-cell DNA and adventitious viruses.

Full versus empty capsids

A common assessment of vector quality has been the detection of empty capsids in preparations. Since the percentage of empty capsids in final preparations can range widely from 50–90% (depending on the purification method), they are large determinants of vector potency. Transmission electron microscopy is a classical way to visually observe and count the ratio of full-to-empty capsids [45]. However, it cannot reveal information on partial or oversized vectors (e.g., truncated genomes or genomes that exceed the design length). Analytical ultracentrifugation (AUC) can yield sedimentation velocities of particles and relies on the density profiles of empty and full capsids [46]. AUC can reveal species that can deviate from the main empty and full capsid peaks, which can typically point towards the presence of partial or oversized packaged genomes. Unfortunately, AUC cannot further characterize the genomes of these non-unit length species. Direct quantification of DNA and capsid proteins can be measured by optical density using A260/A280 [47]. This method also cannot describe vector genome heterogeneity for preparations. Charge detection mass spectrometry can quantify capsid content by measuring the mass-to-charge ratios unique to empty and full capsids [48]. AUC and charge detection mass spectrometry methods require high amounts of material, have long turnaround times, and require technical training and knowhow. Mass photometry is a fast and label-free orthogonal technique that was developed recently [49]. This technique can be used for multiple serotypes [49], and can also work with low amounts of sample [49]. Unfortunately, all analytical methods mentioned above still lack the capacity to characterize the genomes of non-unit length species nor describe vector genome heterogeneity.

Other recently developed analytical methods and advanced orthogonal approaches, such as size-exclusion chromatography with UV and multiangle light scattering can provide insights into vector genome heterogeneity [38,40,50]. However, these methods do not have the ability to disclose the structure or sequences of truncated or oversized forms, chimeric genomes, and the composition of DNA contaminants. These shortcomings inspired the development of a new class of methods that employ next-generation sequencing (NGS) technology. These bioinformatics-reliant methods have opened the door for gaining insights into AAV biology and vectorology, and have revealed the types of structures that can be packaged into capsids that are impossible with other methodologies.

The remainder of this review will discuss the challenge of sequencing AAV, and the use of NGS in vector characterization and evaluation. We will also address some of the shortcomings related to these NGS-based methods.

THE CHALLENGE OF SEQUENCING AAV VECTORS

The wild-type AAV2 sequence was the first AAV genome cloned into plasmids [51], enabling genetic studies [52,53]. The ITRs of AAV2 were first sequenced in 1980 by the Maxam-Gilbert method [54]. Since then, sequencing the full AAV vector genome has been notoriously challenging. This problem has been mainly due to the complexity of the ITRs [55]. In addition, the ITR is GC-rich (70%), which makes standard methods like Sanger sequencing, difficult. Substitution of dGTPs with 7-deaza-dGTP during amplification of the ITRs can help to overcome sequencing issues related to GC content [56]. However, such methods are less than ideal. Until recent times, sequencing AAV vectors sans ITRs has been the staple in the AAV gene therapy field. Nevertheless, ITRs are critical for replication, rescue, and packaging; thus, further understanding of these crucial viral elements substantiates the need to develop robust means to sequence AAV vectors with their ITRs.

NGS-BASED METHODS FOR VECTOR CHARACTERIZATION & EVALUATION

Although many sophisticated methods have been used for assessing the AAV vector product, including those mentioned above, they are unable to provide comprehensive insight into the genome compositions of truncated vectors and DNA contaminants (e.g., plasmid backbone DNA, *rep-cap* genes, and adenovirus helper genes) [27,57–59]. Replication-competent AAVs are also another form of contaminant that can alter the safety of gene therapy vectors [60]. Profiling of packaged content in a population of diverse, and low-abundance species, remains challenging with standard methods like qPCR/ddPCR. NGS has been used widely in many disciplines, but has only recently gained use for characterizing and evaluating AAV vectors. NGS-based methods have the ability to reveal the contents of vectors at the level of the DNA sequence, and can identify contaminants that cannot be captured by standard molecular biology methods. Furthermore, it can detect/identify multiple contaminants in a single experiment, as opposed to using multiple molecular assays. Since NGS can achieve read depths of millions to hundreds of millions of sequences, rare DNA species can be semi-quantitatively profiled, and with certain platforms (discussed below), can be quantitatively assessed [57,61].

Since there are no standardized methods to sequence AAV vectors using NGS, investigators usually look for the most efficient way that is best fit for their research goals. Employment of NGS approaches can depend on different factors, such as budget, time sensitivity, accuracy of the results, and technical proficiency.

Short-read sequencing technology & next-generation sequencing-based methods

Illumina is well-known for its popular short-read sequencing technology [62–64]. It is based on a sequencing by synthesis approach that employs cyclic reversible termination [62,65,66]. Currently, Illumina is still the most popular NGS solution [65,67–69]. It has several advantages, including its established technology, high level of cross-platform compatibility, high accuracy, and a wide range of instruments that span low-throughput to high-throughput options [65,67]. However, Illumina has some drawbacks, including its short-read lengths, high instrument costs, some poor coverage across GC-rich regions, and a tendency towards substitution errors [65,70].

Unfortunately, short-read sequencing has poor coverage at the ITR regions, and fails to capture the full and intact AAV vector genome [27,58]. Nevertheless, they are best for detecting single-nucleotide variants (SNVs) and insertions and deletions (indels), because of their ability to achieve high sequencing depths; and for detecting low-abundance contaminants.

Single-stranded DNA virus sequencing

Single-stranded DNA virus sequencing (SSV-Seq) was the first NGS-based method developed for characterizing AAV vector genomes and residual DNA [57,61], and was developed to address the shortcomings of qPCR. The SSV-Seq method is based on Illumina's short-read sequencing technology. The major steps in SSV-Seq protocol are as follows. First, the preparation is treated with DNase to digest non-encapsidated DNA. Second, DNA extraction is performed, followed by second-strand synthesis with random hexamers to convert ssAAV to double-stranded genomes. Subsequently, the dsDNA template is sonicated into small fragments for NGS library preparation. Next, libraries are sequenced with Illumina HiSeq. Finally, the sequencing data are analyzed by using ContaVect bioinformatics tool. Figure 1 summarizes the SSV-Seq protocol [57]. A PCR-free version of the method called SSV-Seq 2.0 was also developed for optimizing vector genomes with a high percentage of GC and homopolymers [71]. Although SSV-Seq is successful at characterizing AAV vector genomes including residual DNA, the major drawback of this method is its inability to interrogate full and intact vector genomes. Another limitation is the amount of purified rAAV preparation required for input (2×10^{11} vector genomes of purified rAAV). Furthermore, SSV-Seq cannot provide optimal coverage of the ITRs, since Illumina's short-read sequencing requires amplification of the target using polymerases that have low processivity across the ITRs either at the library preparation stage with PCR, or on the flow cell during bridge amplification steps.

Fast-Seq

The development of Fast-Seq was inspired by the limitations of the traditional Sanger method [72], which requires slow and labor-intensive manual evaluation of sequencing reads. Importantly, the Sanger method it is unable quantitate single nucleotide polymorphisms (SNPs) and indels as a result of low sequencing depths. Fast-Seq relies on a Tn5-based library generation that is compatible with single-strand (ss)AAV genomes [72]. The Fast-Seq approach is an end-to-end method for the extraction, purification, sequencing, and data analyses of packaged vector genomes [72] (Figure 1). Fast-Seq's reliance on fragmentation and simultaneous adapter ligation using Tn5 transposase is inexpensive and relatively easy compared to sonication followed by adapter ligation. In addition, it requires less input DNA, which makes it well-suited for inexpensive and lower throughput instruments, such as MiSeq and iSeq. Furthermore, Fast-Seq provides opensource code with a prebuilt customizable Docker container on GitHub for data analysis. However, Fast-Seq also requires double-stranded genome conversion and it can miss single-stranded genomes that fail to convert. Because it is based on short-read sequencing, it inherits all the limitations described above for SSV-Seq. In addition, Fast-Seq was designed primarily for analyzing variants such as SNPs and indels, but not for analyzing contaminants.

Viral genome sequencing

Viral genome sequencing (VGS) was developed with the aim to overcome the double strand synthesis requirement from the other

NGS-based methods [73]. The VGS method is based on the assumption that rAAV DNA extracts are primarily double-stranded species due to the natural base pairing of complementary plus and minus strands [73]. VGS also utilizes a tagmentation-based library construction approach (Figure 1), and was designed to profile the rAAV genome, as well as detecting the presence of contaminants [73]. Because VGS bypasses the double-strand synthesis step, VGS can save time and costs related to sample preparation. In addition, VGS provides Python scripts for validating serotype and Cre-independent DNA recombination events in rAAVs. However, VGS may miss some single-stranded genomes, because its design is based on the assumption that double-stranded configurations are naturally formed from annealing of plus and minus stranded genomes after DNA extraction. VGS also inherits the limitations associated with Illumina short-read sequencing.

Long-read sequencing technologies & NGS-based methods

For many years, the major limitation with using NGS to sequence AAV vectors has been the need to rely on reconstruction of the genome from small read fragments. Although the approach can be useful in determining SNPs and indels, it fails to reveal the structures of the genomes. AUC analyses and gel electrophoresis can reveal the heterogeneity in vector preparations; however, cannot provide sequence information. An NGS approach that can produce reads that capture targets spanning the entirety of the vector genome would be ideal.

In 2009, the first single-molecule sequencing technology was developed and commercialized by Helicos BioSciences [64]. This approach permitted single-molecule representation of AAV for the first time [74]. However, single-molecule sequencing could not achieve complete coverage of the AAV genome. Fortunately, two sequencing technologies were maturing.

Pacific Biosciences and AAV genome population sequencing

Pacific Biosciences (PacBio) is well-known for its long-read sequencing technology called single molecule, real-time (SMRT) sequencing [64]. This technology has the advantages of achieving long-read lengths (approximately 10–20 kb) and shorter instrument execution times [64,70]. However, its accurate base calling is dependent on the consensus reads of multiple passes across a target template. Therefore, the longer the read fragment, the lower the base calling accuracy. The technology also has high operational costs [70]. Coupled with AAV genome population sequencing (AAV-GPseq) [58], SMRT sequencing can accurately profile genomes that are in double-stranded configurations. Double-stranded genomes can be achieved by annealing plus and minus stranded genomes by heat-treating and slow cooling the rAAV genomes (thermal annealing) [59]. The steps for preparing samples for AAV-GPseq is summarized in Figure 1.

Due to the advantage of covering long sequences in a single read, AAV-GPseq has opened the door for gaining insights into the composition of vector genomes, as well as other packaged elements in the vector product that would be elusive with other analytical methods. The significant feature of AAV-GPseq is its ability to capture the intact vector genome from ITR-to-ITR without the need for bioinformatic re-construction from short reads. AAV-GPseq also requires a significant amount of purified vector genomes for input (1×10^{11}–1×10^{12} vector genomes). Because AAV-GPseq requires the ligation of the SMRT bell adapters to double-stranded genomes, this method can also miss the single-stranded genomes that fail to anneal into double-stranded targets. Due to the lower sequencing depths achieved by SMRT sequencing, the accuracy of SNV and indels is lower than can be achieved with Illumina short-read sequencing. In contrast, SMRT sequencing can capture full and intact AAV vector genomes, and can cover the ITR regions, since the phi29-derived polymerase has strand-displacement activity, and sequencing in real time efficiently unwinds the ITR structure. High base calling is achieved through multiple passes of adaptered genomes. This overcomes the inherent error of single passes, yielding base calling errors that are approximately 1%. Nevertheless, due to the nature of its flow cell design, its sequencing depth is relatively low (approximately 5–8 million reads can typically be obtained on a Sequel II). Another shortcoming for SMRT sequencing is its bias towards smaller DNA targets. Typically, SMRT reads need to be normalized to a spike in standard ladder such as BstEII-digested lambda phage DNA or calibrated on fragmented bacterial DNA in order to assess relative abundances [58,75].

Oxford Nanopore & ssDNA sequencing

Another well-known long-read sequencing technology is nanopore sequencing. Oxford Nanopore technology can produce the longest read lengths (approximately 2 Mb) [64,67,76]. The MinION instrument is also small and portable, and can be operated with a laptop computer. Running samples using nanopore is quick, relatively easy, and has lower operational costs [70,77]. However, it has high error rates [64,70]. Nevertheless, nanopore sequencing has high processivity through ITRs and it can directly sequence AAV vectors without amplification [78].

The inspiration for developing ssDNA sequencing was to overcome the need to convert single-stranded DNA (ssDNA) into double-stranded templates, which is a prerequisite for existing NGS-based methods. This conversion can again cause bias [78]. Since nanopore uses a transposase that was found to have residual activity on ssDNA, ssDNA sequencing with nanopore bypasses the need for double-strand conversion of the AAV genomes [78]. However, the efficiency of sequencing double-strand templates was still shown to be much higher than ssDNA. Furthermore, similar to the AAV-GPseq method, AAV genomes can be converted to double-stranded templates and sequenced directly as an intact molecule from ITR to ITR [77]. ssDNA sequencing can also detect contaminants and it can reveal the molecular state of vector genomes [78]. Nanopore sequencing has similar capabilities to SMRT sequencing as a long-read technology, but since DNA strands are only covered through a single pass of the DNA, the accuracy of base calling at each position is low [77]. An illustration of the ssDNA sequencing workflow is shown in Figure 1.

TYPES OF NON-UNIT LENGTH GENOMES FOUND AMONG AAV VECTORS

With the development of NGS methods to profile rAAV, the diversity of non-unit vectors has been revealed. In addition, some of the mechanisms by which they arise are being slowly solved.

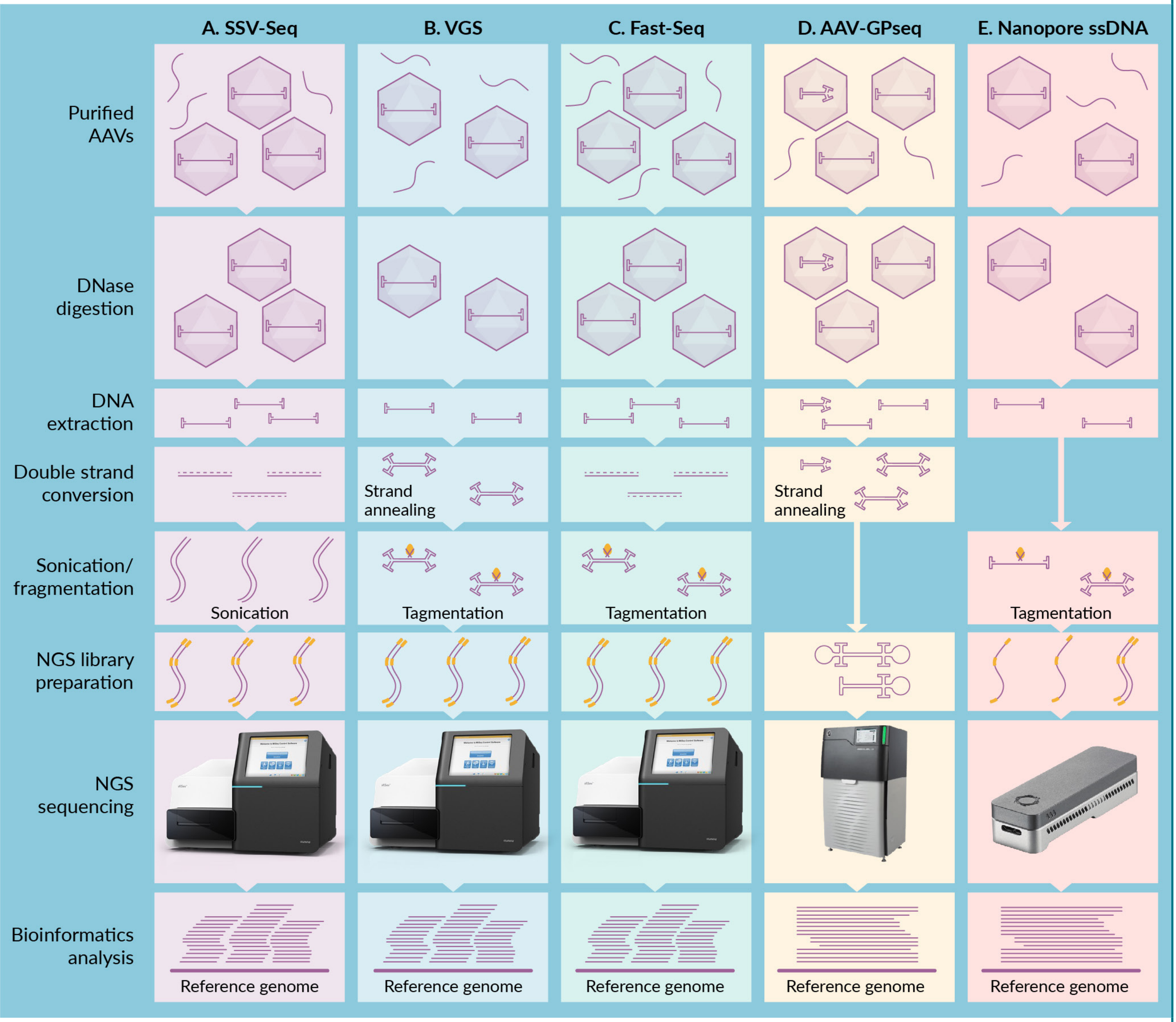
Furthermore, vectors produced by different platforms can have varying degrees of heterogeneity. For example, AAV-GPseq has revealed a diversity of vector genomes including completed genomes, truncated genomes, chimeric genomes, and oversized genomes [27,58,59] (Figure 2). The following section will review commonly identified non-unit length genomes.

Truncated genomes

Truncated AAV genomes were first described with wtAAVs as a hallmark of defective interfering particles [79,80]. Previous studies on the incorporation of short hairpin (sh)RNA or short hairpin-like structures into vector constructs showed that they can lead to truncated vector genomes that have self-complementary configurations. These types of genomes are also commonly called snapback genomes [79,81]. The mechanisms that underpin these events are hypothesized to be due to polymerase redirection or template-switching during viral

FIGURE 1

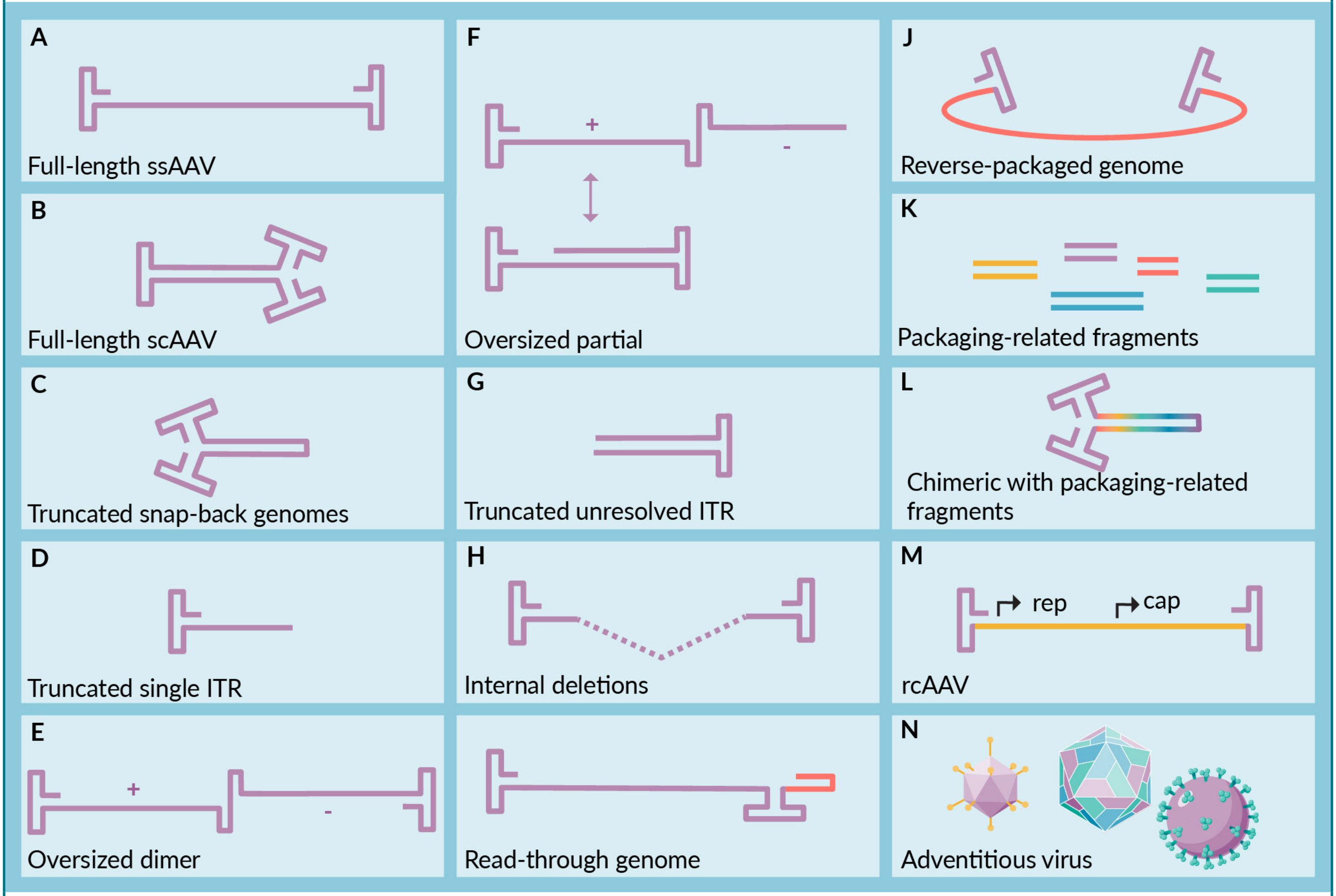
Workflow illustrations for NGS-based methods.



(A) SSV-Seq protocol [57]. The purified particles are digested to remove non-encapsidated DNAs. Next, viral DNAs are extracted and then subjected to second-strand conversion. The double-stranded genomes are then sonicated into fragments for Illumina library preparation. The libraries are then sequenced on an Illumina instrument. Finally, the sequencing data are analyzed by using ContaVect bioinformatic tool. (B) VGS workflow [73]. Purified particles are digested with DNase. Next, viral DNAs are extracted and then subjected to library preparation with tagmentation. Libraries are then sequenced with Illumina MiSeq. Finally, the sequencing data are analyzed with Geneious software and custom Python scripts. (C) Fast-Seq workflow [72]. The purified particles are digested with nuclease treatment. Viral DNAs are then extracted from purified particles followed by second-strand conversion. Double-stranded genomes are fragmented and immediately adaptered by tagmentation with Tn5 transposase. Libraries are sequenced by MiSeq, iSeq, MiniSeq, NextSeq, etc. Lastly, the sequencing data are analyzed with opensource code and a prebuilt customizable Docker container on GitHub. (D) AAV-GPseq workflow [59]. Purified particles are digested by using DNase I treatment. Following digestion, viral DNAs are extracted by using phenol/chloroform. Vector genomes then go through second-strand conversion with heat treatment and cool annealing. Next, vector genomes are prepared for sequencing with SMRT sequencing. Lastly, the sequencing data are analyzed by using custom bioinformatics pipelines. (E) Nanopore ssDNA sequencing workflow [78]. Purified particles are digested with Benzonase nuclease, followed by vector DNA extraction. Next, vector genomes go through nanopore library preparation, which includes adaptering by tagmentation and then sequencing. The sequencing data are then analyzed by using custom bioinformatics pipelines. (NGS: next-generation sequencing; SMRT: single molecule, real-time; SSV-Seq: single-stranded DNA virus sequencing; VGS: viral genome sequencing.)

FIGURE 2

Diagrams of different types of non-unit length genomes and DNA contaminants.



(A) Full-length single-strand (ss)AAV. (B) Full-length self-complementary (sc)AAV. (C) Truncated or snapback genomes (SBG). (D) Truncated genomes with a single ITR. (E) Oversized dimer genome with an unresolved ITR. (F) Oversized partial dimer with an unresolved ITR. (G) Truncated genome with a single unresolved ITR. (H) Truncated genome with deletion between the two ITRs. (I) Read-through genome carrying backbone DNA in red. (K) Reverse-packaged genome carrying backbone DNA in red. (K) DNA fragments from packaging components (ITR-free) (L) Chimeric genome carrying non-vector DNA from multiple fragments following recombination. (M) Replication-competent (rc)AAV with intact *rep* and *cap* ORFs recombined with flanking ITRs. (N) Representation of adventitious viruses. ITR: inverted terminal repeats; ORF: open reading frame.

genome replication [82]. Another means of snapback formation is due to DNA damage [83]. Additionally, truncated genomes are also present in wild-type AAV genomes (Figure 2C) [81]. Truncated genomes with a single ITR can also be observed, but these are rarely identified by NGS. Also directed-repeats are also predicted to cause internal deletions in AAV vectors, but these are also not well represented in sequencing data (Figure 2H).

Oversized genomes

Oversized genomes are those that go beyond the single-unit length of the ITR flanked construct. This type of genome can be a result of abnormal packaging of vector genomes in production or those that are produced from transgene cassettes that are shorter than approximately 5 kb, and are packaged with unresolved ITRs. These types of genomes have been observed with particularly high frequencies with vectors produced by the rBV/Sf system [27], and if designed to exactly half the packaging limit of AAV will form dimers and self-complementary AAV vectors (Figure 2E). Truncated genomes have also been detected with AAV-GPseq from unresolved ITRs. This type of truncated genome was found to be predominantly produced by rBV/Sf production system, and can possibly arise from partial oversized genomes that have undergone cleavage during library preparation steps (Figure 2F,G) [27].

TYPES OF VECTOR CONTAMINANTS

As mentioned above, overcoming the limitations of qPCR in characterizing contaminants in vector product was among the motivations for the developing NGS-based methods to sequence rAAV preparations. As a result, long-read sequencing in particular has helped to reveal a diversity of packaged genomes that can wind up in manufactured rAAV. This section will review the different types of contaminants that are known to exist.

Vector backbone DNA

The most dominant DNA contaminant comes directly from the vector backbone itself. The vector backbone refers to the construct that houses the ITR-flanked vector genome. For platforms utilizing plasmid transfections into producer cell lines, the auxiliary elements within the vector plasmid, also referred to as the *cis* plasmid, would be considered the backbone. In rBV/Sf systems, the recombinant baculovirus vector would be considered the backbone. SSV-Seq has shown that backbone contaminants can range from 0.84–5.97% with different purification techniques [57]. Identification of these types of contaminants are critical. For example, in plasmid-based platforms, antibiotic-resistance genes such as KanR and AmpR may be transferred to patients, potentially increasing risks related to the spread of antibiotic resistance in microbes or hypersensitivity to antibiotics in some patients [84].

Backbone contaminants can originate from read-through genomes. These genomes are packaging events that are characterized by the encapsidation of DNAs that extend beyond the ITR and into the backbone sequence (Figure 2I). Backbone DNA can also be packaged via reverse-packaging events, whereby genomes are packaged from ITR-to-ITR but encompass product that exclusively spans the backbone (Figure 2J).

Helper DNA

Helper gene contaminants encompass those originating from the helper plasmid. This type of contaminant is less common than vector plasmid contaminants. The common targets for observing contaminant include viral associated RNA, *E2* genes, *E4orf6* gene.

rep-cap DNA

rep-cap DNA contaminants are related to the AAV *rep* and *cap* genes. These genes are required for replication and packaging of the vector genome. These contaminants are also less common, but can be problematic for preparations, as they can hint at the presence of replication-competent (rc)AAV (see below). Expression of Cap can also lead to immune responses in the target tissue, resulting in the loss of transduced tissues.

Host-cell DNA

Host-cell (hc)DNA contaminants are infrequent, but can be problematic. Packaging of promoter sequences or full open reading frames can be transferred to the patient with unknown consequences. However, previous studies have shown that host-cell contaminants are higher in mammalian cell production platforms than with insect cell platforms [34]. It has been hypothesized that regions that bare motifs with sequence similarity to RBEs and are in open regions of chromatin are preferentially packaged.

Chimeric genomes

Typically, contaminants described above are packaged into AAV capsids because they may contain sequence similarity to the RBE

[85], or are packaged passively as fragmented DNA (Figure 2K). However, AAV-GPseq has revealed the presence of chimeric genomes [58], which are contaminants that are contiguous with ITRs, and result from recombination events (Figure 2L). Chimeric genomes would therefore be actively packaged into capsids via the packaging signal within the ITRs. To detect these species, it is important to employ A-tail adaptering methods, whereby the AAV genome is end processed to carry an A-tail, and the SMRTbell adapter is T-tailed. This eliminates false-positives from fragment-to-fragment ligation.

Replication-competent AAV

The formation of replication-competent (rc)AAV is a result of recombination events between the ITR in vector plasmid with the *rep* and *cap* genes during vector production (Figure 2M). These recombination events that generate intact, replicative, and potentially infectious virus-like virions are thought to occur randomly and without sequence specificity [60]. This type of contaminant is very rare and must be detected following amplification in cells in the presence of a helper virus. SMRT sequencing and AAV-GPseq uncovered a diversity of recombination events that provide insights into how rcAAVs can emerge [60].

Adventitious viruses/pathogens

As described above, the detection of adventitious virus by qPCR/ddPCR is complicated by the fact that one has to have foreknowledge of the viral contaminant. With NGS, reads that fail to map to the provided user-defined references can be used to megablast to viral genomes in order to detect any potential viruses in the rAAV preparation (Figure 2N). This approach was taken to validate the purity of plasmid DNA used to generate the vectors used in the first-in-human IND trial for Tay-Sachs disease gene therapy [86].

ITR HETEROGENEITY: TRUNCATIONS, MUTATIONS, & DELETIONS

The wild-type AAV2 ITR is widely used in most vector constructs. SMRT sequencing has also permitted the interrogation of ITR heterogeneity [87], allowing for a more comprehensive understanding of the ITR composition in plasmid DNA and in the vector product. Uncovering the ITR composition is crucial for validating vector design, as well as improving vector quality. Furthermore, there is a correlation between ITR configuration and vector heterogeneity [27]. ITR truncations can occur in different vector production systems [27]. The ITR structures are inherently unstable in the bacteria used in plasmid production and during baculovirus replication [88,89]. The truncation can vary and they can bear several configurations [27]. Deletions can also occur in any region of the ITR. As a result, ITRs can lack the B arm, C arm, or both B and C arms. Trident-shaped ITRs can also result from errors in ITR replication [27]. There is a strong correlation between mutations in ITRs with unresolved AAV genomes, which can lead to higher degrees of heterogeneity [27]. Intriguingly, the phenomenon of ITR repair, which presumably occurs through the copying of the opposing intact ITR, was verified by AAV-GPseq [59].

USE OF NGS IN AAV POST-ENTRY EVENTS

The use of NGS platforms to interrogate the composition of AAV vectors has led to a better understanding for vector integrity, heterogeneity, and risk. However, many of these new concepts have yet to be linked to any functional knowledge related to the potency and safety of vectors. Another significant role that NGS has played in the field of AAV biology and vectorology has been its support in investigating AAV integration. Wildtype (wt)AAV has long been known for its ability to integrate into the human genome following infection. Classical studies have shown that wtAAV can integrate into several genomic locations including the well-known AAVS1 site on human chromosome 19q13.42 [90–93]. In the early years of AAV integration analysis, molecular methods such as PCR and Southern blot were used to detecting integration events [92,94,95]. Later investigations become more comprehensive, as a result of advancements in NGS approaches [90,96–100]. The integration of AAV2 in host cell genome has been studied extensively, as a result of the concerns for hepatocellular carcinoma found in AAV-positive patients [15,101–104]. In these studies, advanced molecular method and notably high-throughput sequencing have been used for detecting and analyzing AAV integration sites. Recent efforts to understand this potential link continue to reveal aspects of AAV biology that were previously unknown [15,103–106]. Most recently, a comprehensive analyses of human and non-human primates tissues using target-enrichment sequencing and NGS have shown that wildtype AAV and recombinant AAV show preferential integration of respective viral and vector genomes into and near gene bodies of highly expressed genes [107,108]. Such studies have shed more light into the biology and consequences of AAV in gene therapy applications.

Although AAV vectors are considered safe, studies in rodents have shown that AAV vector integration can lead to oncogenesis [97]. At present, there is no evidence that AAV vector integration can cause oncogenesis in humans; although, the FDA now recommends long-term follow-ups after AAV administration. Furthermore, integration analyses in mouse or non-human primates (or other relevant large animal models) are required as part of pre-clinical evaluations of vector safety. For example, a recent long-term study in dogs treated for hemophilia A identified clonal expansion of transduced liver cells [96].

LIMITATIONS & FUTURE DEVELOPMENT OF NGS-BASED METHODS

Unfortunately, there are no standards or universal means of manufacturing AAV. There are also no standardized and universal methods for assessing vector quality control. Recent guidelines from the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), section Q5A(R2) EWG, indicate that NGS is appropriate for viral safety evaluation of biotechnology products derived from cell lines of human or animal origin. Specifically, this pertains to the detection of adventitious viruses. Due to the sensitivity of assay and the breadth of virus detection, NGS can also be used for replacing cell-based infectivity assays [109]. Current NGS-based protocols used to profile AAV have inherit all of the advantages and disadvantages of their adopted NGS sequencing technologies. Therefore, comprehensive interrogation of AAV requires the adoption of both short and long-read sequencing technologies. Long-read sequencing technologies can provide full coverage at the ITR regions to allow a more complete characterizing and understanding of the ITRs in plasmid as well as in vector product. On the other hand, short-read sequencing is more accurate for detecting indels and SNVs in vector genomes. Short-read sequencing can also allow for the detection of very low abundance contaminants. An approach that can encompass the best of both worlds, can achieve reliable quantification of heterogenous populations without biases, requires less genomic input, and can be easily adopted, remains an aspiration for the field. Until such a technique is developed, a combination of long- and short-read techniques may be the most ideal approach for obtaining a complete genomic profiling of AAV-based gene therapy vectors.

REFERENCES

1. Atchison RW, Casto BC, Hammon WM. Adenovirus-associated defective virus particles. *Science* 1965; 149(3685), 754–756.

2. Hoggan MD, Blacklow NR, Rowe WP. Studies of small DNA viruses found in various adenovirus preparations: physical, biological, and immunological characteristics. *Proc. Natl. Acad. Sci. USA* 1966; 55(6), 1467–1474.

3. Carter BJ. Adeno-associated virus and the development of adeno-associated virus vectors: a historical perspective. *Mol. Ther.* 2004; 10(6), 981–989.

4. Wang D, Tai PWL, Gao G. Adeno-associated virus vector as a platform for gene therapy delivery. *Nat. Rev. Drug Discov.* 2019; 18(5), 358–378.

5. Gigout L, Rebollo P, Clement N, *et al.* Altering AAV tropism with mosaic viral capsids. *Mol. Ther.* 2005; 11(6), 856–865.

6. Wörner TP, Bennett A, Habka S, *et al.* Adeno-associated virus capsid assembly is divergent and stochastic. *Nat. Commun.* 2021; 12(1), 1642.

7. Naso MF, Tomkowicz B, Perry WL 3rd, Strohl WR. Adeno-associated virus (AAV) as a vector for gene therapy. *BioDrugs* 2017; 31(4), 317–334.

8. Gao G, Vandenberghe LH, Alvira MR, *et al.* Clades of adeno-associated viruses are widely disseminated in human tissues. *J. Virol.* 2004; 78(12), 6381–6388.

9. Mietzsch M, Jose A, Chipman P, *et al.* Completion of the AAV structural atlas: serotype capsid structures reveals clade-specific features. *Viruses* 2021; 13(1), 101.

10. Cotmore SF, Tattersall P. The autonomously replicating parvoviruses of vertebrates. *Adv. Virus Res.* 1987; 33, 91–174.

11. Berns KI. The unusual properties of the aav inverted terminal repeat. *Hum. Gene Ther.* 2020; 31(9–10), 518–523.

12. Wilmott P, Lisowski L, Alexander IE, Logan GJ. A user's guide to the inverted terminal repeats of adeno-associated virus. *Hum. Gene Ther. Methods* 2019; 30(6), 206–213.

13. Qu Y, Liu Y, Noor AF, Tran J, Li R. Characteristics and advantages of adeno-associated virus vector-mediated gene therapy for neurodegenerative diseases. *Neural Regen. Res.* 2019; 14(6), 931–938.

14. Clément N, Grieger JC. Manufacturing of recombinant adeno-associated viral vectors for clinical trials. *Mol. Ther. Methods Clin. Dev.* 2016; 3, 16002.

15. Nault JC, Datta S, Imbeaud S, *et al.* Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat. Genet.* 2015; 47(10), 1187–1193.

16. Morfopoulou S, Buddle S, Torres Montaguth OE, *et al.* Genomic investigations of unexplained acute hepatitis in children. *Nature* 2023; 617(7961), 564–573.

17. Servellita V, Sotomayor Gonzalez A, Lamson DM, *et al.* Adeno-associated virus type 2 in US children with acute severe hepatitis. *Nature* 2023; 617(7961), 574–580.

18. Ho A, Orton R, Tayler R, *et al.* Adeno-associated virus 2 infection in children with non-A-E hepatitis. *Nature* 2023; 617(7961), 555–563.

19. Kondratov O, Marsic D, Crosson SM, *et al.* Direct head-to-head evaluation of recombinant adeno-associated viral vectors manufactured in human versus insect cells. *Mol. Ther.* 2017; 25(12), 2661–2675.

20. Mietzsch M, Hering H, Hammer EM, Agbandje-McKenna M, Zolotukhin S, Heilbronn R. OneBac 2.0: Sf9 cell lines for production of AAV1, AAV2, and AAV8 vectors with minimal encapsidation of foreign DNA. *Hum. Gene Ther. Methods* 2017; 28(1), 15–22.

21. Clément N, Knop DR, Byrne BJ. Large-scale adeno-associated viral vector production using a herpesvirus-based system enables manufacturing for clinical studies. *Hum. Gene Ther.* 2009; 20(8), 796–806.

22. Flotte TR, Trapnell BC, Humphries M, *et al.* Phase 2 clinical trial of a recombinant adeno-associated viral vector expressing α1-antitrypsin: interim results. *Hum. Gene Ther.* 2011; 22(10), 1239–1247.

23. Thomas DL, Wang L, Niamke J, *et al.* Scalable recombinant adeno-associated virus production using recombinant herpes simplex virus type 1 coinfection of suspension-adapted mammalian cells. *Hum. Gene Ther.* 2009; 20(8), 861–870.

24. Gao GP, Qu G, Faust LZ, *et al.* High-titer adeno-associated viral vectors from a Rep/Cap cell line and hybrid shuttle virus. *Hum. Gene Ther.* 1998; 9(16), 2353–2362.

25. Clark KR, Voulgaropoulou F, Fraley DM, Johnson PR. Cell lines for the production of recombinant adeno-associated virus. *Hum. Gene Ther.* 1995; 6(10), 1329–1341.

26. Grieger JC, Soltys SM, Samulski RJ. Production of recombinant adeno-associated virus vectors using suspension HEK293 cells and continuous harvest of vector from the culture media for GMP FIX and FLT1 clinical vector. *Mol. Ther.* 2016; 24(2), 287–297.

27. Tran NT, Lecomte E, Saleun S, *et al.* Human and insect cell-produced recombinant adeno-associated viruses show differences in genome heterogeneity. *Hum. Gene Ther.* 2022; 33(7–8), 371–388.

28. Giles A, Lock M, Chen SJ, *et al.* Significant differences in capsid properties and potency between adeno-associated virus vectors produced in Sf9 and HEK293 cells. *Hum. Gene Ther.* 2023; 34(19–20), 1003–1021.

29. Qu G, Bahr-Davidson J, Prado J, *et al.* Separation of adeno-associated virus type 2 empty particles from genome containing vectors by anion-exchange column chromatography. *J. Virol. Methods* 2007; 140(1–2), 183–192.

30. Lock M, Alvira MR, Wilson JM. Analysis of particle content of recombinant adeno-associated virus serotype 8 vectors by ion-exchange chromatography. *Hum. Gene Ther. Methods* 2012; 23(1), 56–64.

31. Nass SA, Mattingly MA, Woodcock DA, *et al.* Universal method for the purification of recombinant AAV vectors of differing serotypes. *Mol. Ther. Methods Clin. Dev.* 2017; 9, 33–46.

32. Qu W, Wang M, Wu Y, Xu R. Scalable downstream strategies for purification of recombinant adeno- associated virus vectors in light of the properties. *Curr. Pharm. Biotechnol.* 2015; 16(8), 684–695.

33. Gao G, Sena-Esteves M. Introducing genes into mammalian cells: viral vectors. In: *Molecular Cloning: A Laboratory Manual* (Editors: Gao G, Sena-Esteves M). 2012; 2, 1209–1313. Cold Spring Harbor Laboratory Press

34. Penaud-Budloo M, François A, Clément N, Ayuso E. Pharmacology of recombinant adeno-associated virus production. *Mol. Ther. Methods Clin. Dev.* 2018; 8, 166–180.

35. King D, Schwartz C, Pincus S, Forsberg N. Viral vector characterization: a look at analytical tools. *Cell Culture Dish* Oct 10, 2018; https://cellculturedish.com/viral-vector-characterization-analytical-tools/. (Accessed Aug 2023)

36. Gimpel AL, Katsikis G, Sha S, *et al.* Analytical methods for process and product characterization of recombinant adeno-associated virus-based gene therapies. *Mol. Ther. Methods Clin. Dev.* 2021; 20, 740–754.

37. Dobnik D, Kogovšek P, Jakomin T, *et al.* Accurate quantification and characterization of adeno-associated viral vectors. *Front. Microbiol.* 2019; 10, 1570.

38. McIntosh NL, Berguig GY, Karim OA, *et al.* Comprehensive characterization and quantification of adeno associated vectors by size exclusion chromatography and multi angle light scattering. *Sci. Rep.* 2021; 11(1), 3012.

39. Cole L, Fernandes D, Hussain MT, Kaszuba M, Stenson J, Markova N. Characterization of recombinant adeno-associated viruses (rAAVs) for gene therapy using orthogonal techniques. *Pharmaceutics* 2021; 13(4), 586.

40. Wagner C, Innthaler B, Lemmerer M, Pletzenauer R, Birner-Gruenberger R. Biophysical characterization of adeno-associated virus vectors using ion-exchange chromatography coupled to light scattering detectors. *Int. J. Mol. Sci.* 2022; 23(21), 12715.

41. Prantner A, Maar D. Genome concentration, characterization, and integrity analysis of recombinant adeno-associated viral vectors using droplet digital PCR. *PLoS One.* 2023; 18(1):e0280242.

42. Dorange F, LeBec C. Analytical approaches to characterize AAV vector production & purification Advances and challenges. *Cell & Gene Therapy Insights* 2018; 4(2), 119–129.

43. Higashiyama K, Yuan Y, Hashiba N, Masumi-Koizumi K, Yusa K, Uchida K. Quantitation of residual host cell dna in recombinant adeno-associated virus using droplet digital polymerase chain reaction. *Hum. Gene Ther.* 2023; 34(11–12), 578–585.

44. Barone PW, Wiebe ME, Leung JC, *et al.* Viral contamination in biologic manufacture and implications for emerging therapies. *Nat. Biotechnol.* 2020; 38(5), 563–572.

45. Werle AK, Powers TW, Zobel JF, *et al.* Comparison of analytical techniques to quantitate the capsid content of adeno-associated viral vectors. *Mol. Ther. Methods Clin. Dev.* 2021; 23, 254–262.

46. Burnham B, Nass S, Kong E, *et al.* Analytical ultracentrifugation as an approach to characterize recombinant adeno-associated viral vectors. *Hum. Gene Ther. Methods* 2015; 26(6), 228–242.

47. Schnödt M, Büning H. Improving the quality of adeno-associated viral vector preparations: the challenge of product-related impurities. *Hum. Gene Ther. Methods* 2017; 28(3), 101–108.

48. Pierson EE, Keifer DZ, Asokan A, Jarrold MF. Resolving adeno-associated viral particle diversity with charge detection mass spectrometry. *Anal. Chem.* 2016; 88(13), 6718–6725.

49. Wagner C, Fuchsberger FF, Innthaler B, Lemmerer M, Birner-Gruenberger R. Quantification of empty, partially filled and full adeno-associated virus vectors using mass photometry. *Int. J. Mol. Sci.* 2023; 24(13), 11033.

50. Ryan JP, Kostelic MM, Hsieh CC, *et al.* Characterizing adeno-associated virus capsids with both denaturing and intact analysis methods. *J. Am. Soc. Mass Spectrom.* 2023; 34(12), 2811–2821.

51. Srivastava A, Lusby EW, Berns KI. Nucleotide sequence and organization of the adeno-associated virus 2 genome. *J. Virol.* 1983; 45(2), 555–564.

52. Samulski RJ, Berns KI, Tan M, Muzyczka N. Cloning of adeno-associated virus into pBR322: rescue of intact virus from the recombinant plasmid in human cells. *Proc. Natl. Acad. Sci. USA* 1982; 79(6), 2077–2081.

53. Laughlin CA, Tratschin JD, Coon H, Carter BJ. Cloning of infectious adeno-associated virus genomes in bacterial plasmids. *Gene* 1983; 23(1), 65–73.

54. Lusby E, Fife KH, Berns KI. Nucleotide sequence of the inverted terminal repetition in adeno-associated virus DNA. *J. Virol.* 1980; 34(2), 402–409.

55. Petri K, Fronza R, Gabriel R, *et al.* Comparative next-generation sequencing of adeno-associated virus inverted terminal repeats. *Biotechniques* 2014; 56(5), 269–273.

56. Mroske C, Rivera H, Ul-Hasan T, Chatterjee S, Wong KK. A capillary electrophoresis sequencing method for the identification of mutations in the inverted terminal repeats of adeno-associated virus. *Hum. Gene Ther. Methods* 2012; 23(2), 128–136.

57. Lecomte E, Tournaire B, Cogné B, *et al.* Advanced characterization of DNA molecules in raav vector preparations by single-stranded virus next-generation sequencing. *Mol. Ther. Nucleic Acids* 2015; 4(10):e260.

58. Tai PWL, Xie J, Fong K, *et al.* Adeno-associated virus genome population sequencing achieves full vector genome resolution and reveals human-vector chimeras. *Mol. Ther. Methods Clin. Dev.* 2018; 9, 130–141.

59. Tran NT, Heiner C, Weber K, *et al.* AAV-genome population sequencing of vectors packaging crispr components reveals design-influenced heterogeneity. *Mol. Ther. Methods Clin. Dev.* 2020; 18, 639–651.

60. Yip M, Chen J, Zhi Y, *et al.* Querying recombination junctions of replication-competent adeno-associated viruses in gene therapy vector preparations with single molecule, real-time sequencing. *Viruses* 2023; 15(6), 1228.

61. Lecomte E, Leger A, Penaud-Budloo M, Ayuso E. Single-stranded DNA virus sequencing (SSV-Seq) for characterization of residual DNA and AAV vector genomes. *Methods Mol. Biol.* 2019; 1950, 85–106.

62. Pervez MT, Hasnain MJU, Abbas SH, Moustafa MF, Aslam N, Shah SSM. A comprehensive review of performance of next-generation sequencing platforms. *Biomed. Res. Int.* 2022; 2022, 3457806.

63. History of sequencing by synthesis. Illumina 2023; https://www.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html. (Accessed Aug 2023)

64. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: history and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* 2019; 18, 9–19.

65. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 2016; 17(6), 333–351.

66. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics* 2016; 107(1), 1–8.

67. Kanzi AM, San JE, Chimukangara B, *et al.* Next generation sequencing and bioinformatics analysis of family genetic inheritance. *Front. Genet.* 2020; 11, 544162.

68. Segerman B. The most frequently used sequencing technologies and assembly methods in different time segments of the bacterial surveillance and RefSeq genome databases. *Front. Cell Infect. Microbiol.* 2020; 10, 527102.

69. Cantu M, Morrison MA, Gagan J. Standardized comparison of different DNA sequencing platforms. *Clin. Chem.* 2022; 68(7), 872–876.

70. Garrido-Cardenas JA, Garcia-Maroto F, Alvarez-Bermejo JA, Manzano-Agugliaro F. DNA sequencing sensors: an overview. *Sensors (Basel)* 2017; 17(3), 588.

71. Lecomte E, Saleun S, Bolteau M, *et al.* The SSV-Seq 2.0 PCR-free method improves the sequencing of adeno-associated viral vector genomes containing GC-rich regions and homopolymers. *Biotechnol. J.* 2021; 16(1):e2000016.

72. Maynard LH, Smith O, Tilmans NP, *et al.* Fast-Seq: a simple method for rapid and inexpensive validation of packaged single-stranded adeno-associated viral genomes in academic settings. *Hum. Gene Ther. Methods* 2019; 30(6), 195–205.

73. Guerin K, Rego M, Bourges D, *et al.* A novel next-generation sequencing and analysis platform to assess the identity of recombinant adeno-associated viral preparations from viral DNA extracts. *Hum. Gene Ther.* 2020; 31(11–12), 664–678.

74. Kapranov P, Chen L, Dederich D, *et al.* Native molecular state of adeno-associated viral vectors revealed by single-molecule sequencing. *Hum. Gene Ther.* 2012; 23(1), 46–55.

75. Soares LMM, Hanscom T, Selby DE, *et al.* DNA read count calibration for single-molecule, long-read sequencing. *Sci. Rep.* 2022; 12(1), 17257.

76. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 2020; 21(1), 30.

77. Namkung S, Tran NT, Manokaran S, *et al.* Direct ITR-to-ITR nanopore sequencing of AAV vector genomes. *Hum. Gene Ther.* 2022; 33(21–22), 1187–1196.

78. Radukic MT, Brandt D, Haak M, Müller KM, Kalinowski J. Nanopore sequencing of native adeno-associated virus (AAV) single-stranded DNA using a transposase-based rapid protocol. *NAR Genom. Bioinform.* 2020; 2(4):lqaa074.

79. Hauswirth WW, Berns KI. Adeno-associated virus DNA replication: non-unit-length molecules. *Virology.* 1979; 93(1), 57–68.

80. Laughlin CA, Myers MW, Risin DL, Carter BJ. Defective-interfering particles of the human parvovirus adeno-associated virus. *Virology.* 1979; 94(1), 162–174.

81. Zhang J, Yu X, Guo P, *et al.* Satellite subgenomic particles are key regulators of adeno-associated virus life cycle. *Viruses* 2021; 13(6), 1185.

82. Xie J, Mao Q, Tai PWL, *et al.* Short DNA hairpins compromise recombinant adeno-associated virus genome homogeneity. *Mol. Ther.* 2017; 25(6), 1363–1374.

83. Zhang J, Guo P, Yu X, *et al.* Subgenomic particles in rAAV vectors result from DNA lesion/break and non-homologous end joining of vector genomes. *Mol. Ther. Nucleic Acids* 2022; 29, 852–861.

84. Vandermeulen G, Marie C, Scherman D, Préat V. New generation of plasmid backbones devoid of antibiotic resistance marker for gene therapy trials. *Mol. Ther.* 2011; 19(11), 1942–1949.

85. Wright JF. Product-related impurities in clinical-grade recombinant AAV vectors: characterization and risk assessment. *Biomedicines.* 2014; 2(1), 80–97.

86. Flotte TR, Cataltepe O, Puri A, *et al.* AAV gene therapy for Tay-Sachs disease. *Nat. Med.* 2022; 28(2), 251–259.

87. Tran NT, Lecomte E, Saleun S, *et al.* Human and insect cell-produced recombinant adeno-associated viruses show differences in genome heterogeneity. *Hum. Gene Ther.* 2022; 33(7–8), 371–388.

88. Jacob A, Brun L, Jiménez Gil P, *et al.* Homologous recombination offers advantages over transposition-based systems to generate recombinant baculovirus for adeno-associated viral vector production. *Biotechnol. J.* 2021; 16(1):e2000014.

89. Pijlman GP, van den Born E, Martens DE, Vlak JM. Autographa californica baculoviruses with large genomic deletions are rapidly generated in infected insect cells. *Virology.* 2001; 283(1), 132–138.

90. Hüser D, Gogol-Döring A, Chen W, Heilbronn R. Adeno-associated virus type 2 wild-type and vector-mediated genomic integration profiles of human diploid fibroblasts analyzed by third-generation PacBio DNA sequencing. *J. Virol.* 2014; 88(19), 11253–11263.

91. Kotin RM, Siniscalco M, Samulski RJ, *et al.* Site-specific integration by adeno-associated virus. *Proc. Natl. Acad. Sci. USA* 1990; 87(6), 2211–2215.

92. Samulski RJ, Zhu X, Xiao X, *et al.* Targeted integration of adeno-associated virus (AAV) into human chromosome 19. *EMBO J.* 1991; 10(12), 3941–3950.

93. Deyle DR, Russell DW. Adeno-associated virus vector integration. *Curr. Opin. Mol. Ther.* 2009; 11(4), 442–447.

94. Hamilton H, Gomos J, Berns KI, Falck-Pedersen E. Adeno-associated virus site-specific integration and AAVS1 disruption. *J. Virol.* 2004; 78(15), 7874–7882.

95. Henckaerts E, Dutheil N, Zeltner N, *et al.* Site-specific integration of adeno-associated virus involves partial duplication of the target locus. *Proc. Natl. Acad. Sci. USA* 2009; 106(18), 7571–7576.

96. Nguyen GN, Everett JK, Kafle S, *et al.* A long-term study of AAV gene therapy in dogs with hemophilia A identifies clonal expansions of transduced liver cells. *Nat. Biotechnol.* 2021; 39(1), 47–55.

97. Dalwadi DA, Calabria A, Tiyaboonchai A, *et al.* AAV integration in human hepatocytes. *Mol. Ther.* 2021; 29(10), 2898–2909.

98. Gil-Farina I, Fronza R, Kaeppel C, *et al.* Recombinant AAV integration is not associated with hepatic genotoxicity in nonhuman primates and patients. *Mol. Ther.* 2016; 24(6), 1100–1105.

99. Hanlon KS, Kleinstiver BP, Garcia SP, *et al.* High levels of AAV vector integration into CRISPR-induced DNA breaks. *Nat. Commun.* 2019; 10(1), 4439.

100. Li H, Malani N, Hamilton SR, *et al.* Assessing the potential for AAV vector genotoxicity in a murine model. *Blood* 2011; 117(12), 3311–3319.

101. Berns KI, Byrne BJ, Flotte TR, *et al.* Adeno-associated virus type 2 and hepatocellular carcinoma? *Hum. Gene Ther.* 2015; 26(12), 779–781.

102. Russell DW, Grompe M. Adeno-associated virus finds its disease. *Nat. Genet.* 2015; 47(10), 1104–1105.

103. La Bella T, Imbeaud S, Peneau C, *et al.* Adeno-associated virus in the liver: natural history and consequences in tumour development. *Gut* 2020; 69(4), 737–747.

104. Schäffer AA, Dominguez DA, Chapman LM, *et al.* Integration of adeno-associated virus (AAV) into the genomes of most Thai and Mongolian liver cancer patients does not induce oncogenesis. *BMC Genomics* 2021; 22(1), 814.

105. Janovitz T, Klein IA, Oliveira T, *et al.* High-throughput sequencing reveals principles of adeno-associated virus serotype 2 integration. *J. Virol.* 2013; 87(15), 8559–8568.

106. Meumann N, Schmithals C, Elenschneider L, *et al.* Hepatocellular carcinoma is a natural target for adeno-associated virus (AAV) 2 vectors. *Cancers (Basel)* 2022; 14(2), 427.

107. Greig JA, Martins KM, Breton C, *et al.* Integrated vector genomes may contribute to long-term expression in primate liver after AAV administration. *Nat. Biotechnol.* Epub ahead of print. doi: 10.1038/s41587–023–01974–7.

108. Martins KM, Breton C, Zheng Q, *et al.* Prevalent and disseminated recombinant and wild-type adeno-associated virus integration in macaques and humans. *Hum. Gene Ther.* 2023; 34(21–22), 1081–1094.

109. ICH, Q5A(R2). EWG viral safety evaluation of biotechnology products derived from cell lines of human or animal origin, Dec 2023.

AFFILIATIONS

Ngoc Tam Tran
Horae Gene Therapy Center,
UMass Chan Medical School,
Worcester,
MA 01605, USA
and
Department of Microbiology
and Physiological Systems,
UMass Chan Medical School,
Worcester,
MA 01605, USA

Phillip WL Tai
Horae Gene Therapy Center,
UMass Chan Medical School,
Worcester,
MA 01605, USA
and
Department of Microbiology
and Physiological Systems,
UMass Chan Medical School,
Worcester,
MA 01605, USA
and
Li Weibo Institute of Rare Diseases
Research,
UMass Chan Medical School,
Worcester,
MA 01605, USA

AUTHORSHIP & CONFLICT OF INTEREST

Contributions: The named authors take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Acknowledgements: None.

Disclosure and potential conflicts of interest: Tai PWL received support for this manuscript from The Bassick Family Foundation Award and the BRIDGE Fund Award (UMass Chan Medical School). Tai PWL had a leadership/fiduciary role in association with the American Society of Gene and Cell Therapy.

Funding declaration: Tai PWL received grants/contracts from Spirovant Sciences, AAVnerGene, Apic Bio, Uniqure/CSL, Cambridge Life Sciences, and Sparing Vision. Tai PWL has received consulting fees from Solid Biosciences Inc., Gordian Biotechnology, BioMarin Pharmaceutical Inc., Shape Therapeutics, FormBio, Biogen, Tavistock, and REGENXBIO Inc.

ARTICLE & COPYRIGHT INFORMATION

Article source: Invited; externally peer reviewed.

Submitted for peer review: Nov 1, 2023; **Revised manuscript received:** Jan 10, 2024; **Publication date:** Jan 31, 2024.