Subgenomic satellite particle generation in recombinant AAV vectors results from DNA lesion/breakage and non-homologous end joining

Junping Zhang[1]*, Ping Guo[2]*, Xiangping Yu[2]*, Derek Pouchnik[1], Jenni Firrman[3], Hongying Wei[1], Nianli Sang[4], Dong Li[5], Roland Herzog[1], Yong Diao[2], Weidong Xiao[1]

[1]Herman B Wells Center for Pediatric Research, Indiana University, Indianapolis, IN 46202, USA;

[2]School of Biomedical Science, Huaqiao University, Quanzhou, China;

[3]United States Department of Agriculture, Agricultural Research Service, Eastern Regional Research Center, Wyndmoor, PA 19038, USA;

[4]Department of Biology, College of Arts and Sciences, Drexel University, Philadelphia, PA 19104, USA;

[5]Department of Clinical Laboratory, Shanghai Tongji Hospital, Tongji University School of Medicine, Shanghai, China.

*These authors contributed equally to this work.

Correspondence:

Weidong Xiao, PhD, Herman B Wells Center for Pediatric Research, Indiana University, 1044 W. Walnut St., R4-121, Indianapolis, IN 46202, USA.

E-mail: xiaow@iu.edu

Tel: 317-274-3155

Abstract

Recombinant AAV (rAAV) vectors have been developed for therapeutic treatment of genetic diseases. Nevertheless, current rAAV vectors administered to patients often contain non-vector related DNA contaminants. Here, we present a thorough molecular analysis of the configuration of non-standard AAV genomes generated during rAAV production. In addition to the sub-vector genomic size particles containing incomplete AAV genomes, our results found that rAAV preparations were contaminated with multiple categories of subgenomic particles with either snapback genomes or vector genomes with deletions in the mid regions. Through CRISPR and restriction enzyme-based *in vivo* and *in vitro* modeling, we identified that the main mechanism leading to the formation of non-canonical genome particles occurred through nonhomologous end joining of fragmented vector genomes caused by genome lesions or DNA breaks that were generated by the host cell/environment. The results of this study advance our understanding of AAV vectors and provide new clues on improving vector efficiency and safety profile for use in human gene therapy.

Introduction:

Recombinant adeno-associated virus (rAAV) vectors have been widely adopted as a gene delivery tool for basic research as well as a pharmaceutical drug vector for human gene therapy(*1*). The vector genome is constructed by inserting the desired expression cassette and regulatory elements between two flanking copies of inverted terminal repeats (ITR). The ITR functions as the replication origin for AAV vectors and as the packaging signal for the rAAV production process. rAAV vectors are typically produced by transfecting host cells, such as 293 cells, with plasmids coding for the vector while also supplying helper functions by delivering either a helper virus, such as adenovirus, or trans factors, such as rep78 and rep68, rep52/rep40, and VP1, VP2 and VP3. Alternatively, rAAV vectors may also be produced using a non-adenovirus helper or a non-mammalian system with baculovirus.

While rAAV vector preparation can be performed following a standard procedure, it does not result in the production of a homogenous population, even for GMP produced vectors that are used clinically. Previously identified vector related impurities include AAV particles containing plasmid backbone sequence and even host genomic sequences(*2*). Although defective interference particles are known to exist in wild type AAV populations(*3-5*), similar particles found in recombinant AAV vector preparations have never been fully characterized due to technical difficulties in obtaining the detailed vector sequences from the entire population. Previously, next generation sequencing (NGS) has been used to profile the rAAV genomic configuration and to perform transcriptomic analysis(*6*). The helicos-based sequencing platform has been used to profile the 3-‘ end of the rAAV genomes(*7*). However, all of these data have only partial genomic information

on the rAAV system. In a separate study, PacBio sequencing was used to produce more long reads and covered some special categories of rAAV genomes in the vector population (*8*). Here we systemically characterized the molecular state of rAAV vector genomes at a single virus level. In addition, through CRISPR-Cas9 based *in vivo* modeling, we identified that the host-mediated vector genome lesion/breakage as the main cause for AAV vector subgenomic particles formation.

Results:

Molecular configuration of subgenome particles in the rAAV population suggested Non-homologous end joining (NHEJ) events during AAV replication and packaging

In order to reveal the molecular state of individual AAV genomes in a population produced by the typical triple plasmid transfection method, we took advantage of the long reads and high accuracy of the PacBio Single Molecule, Real-Time (SMRT) Sequencing platform. The summary of our sequencing results from multiple vector preparations are presented in Figure 1. The highly heterogeneous rAAV population was classified within the following categories based on the our analysis of thousands of vector genomes: 1. Standard rAAV genomes which contained the complete vector sequences including transgene expression cassette and flanking AAV ITRs; 2. Snapback genomes (SBG) which had the left or right moiety of standard duplex rAAV genomes. The SBG was further classified as symmetric SBG (sSBG) or asymmetric SBG (aSBG) according to the DNA complementary state of the top and bottom strands. For sSBG, the top and bottom strands complemented each other. Unlike sSBG, DNA at the bottom strand of aSBG did not match the top strand completely and, therefore, promoted loop formation in the middle region. 3. Incomplement rAAV genomes (ICG), which had an intact 3'ITR and partial AAV genome. These were presumably formed by an aborted packaging process. 4. Genome deletion mutants (GDM), in which the middle region of the AAV genomes were deleted. 5. Secondary Derivative genomes (SDG) which were formed by using class 2-4 molecules as the template and the same mechanism to generate the next generation of subgenomic vector molecules of class 2-4.

While typical sSBG configuration may have been the product of a template switch, the existence of aSBG, GDM, and SDG could not be explained by a template switch of the DNA polymerase during AAV replication. Since there were remnant signs of multiple DNA fragments in the aSBG, GDM and SDG, we proposed that NHEJ events had occurred during the AAV replication and packaging processes.

NHEJ as the mechanism for generating subgenomic particles in an rAAV population

An NHEJ reaction requires the presence of corresponding DNA fragments. The dissection genomic configurations of subgenomic particles of AAV suggested the existence of such fragments. First we tested whether or not NHEJ events could led to the generation of snapback genomes (SBG). We transfected host cells with linear rAAV DNA fragments (Figure 2a) that were generated though restriction enzyme digestion in the presence of trans elements that complement AAV replication and packaging (Figure 2b). The parent vector plasmid pCB-EGFP-6.4K was oversized for AAV capsids. DNA recovered from vectors prepared using this oversized plasmid primarily consisted of smaller fragments, which were less than 6.4kb in size. In contrast, vectors prepared from smaller vector plasmid pCB-EGFP-3.4K, which falls within the packaging limits for the AAV capsid, mainly produced viral particles with a 3.4kb DNA genome. Interestingly, when linear fragments derived from pCB-EGFP-6.4K ranging from 0.6kb to 3.1 kb were used for transfection, the most prominent genomes recovered from the prepared vectors appeared to be resultant of intermolecular nonhomologous end joining (Figure 2). Even though intra-molecular DNA joining of the 5’end and 3’end was supposed to be more efficient, the vectors resulting from such reaction were in relatively lower yield. This was most likely

because their size was larger than the inter-molecular NHEJ products. The inter-molecular NHEJ products were confirmed to be snapback genomes (SBG, data not shown). More specifically, when vector was prepared using fCB-GFP-2.3k, the vector DNA size from AAV ITR to the breaking points was 1.8kb and 2.3kb respectively. The main vector size shown in the gel were1.8 kb, 2.3 kb, along with a faint 4.1 kb region which suggested intramolecular joining. Similar observations were obtained for vectors prepared using fCB-GFP-0.6k (0.6kb, 1.8kb), fCB-GFP-1.0k (1.0kb, 1.8kb), fCB-GFP-1.6k (1.6kb, 1.8kb), and fCB-GFP-1.8k (1.8 kb). The exception was for vectors prepared using fCB-GFP-3.1k, in which we only observed a 1.8kb genome fragment. This is likely because the 3.1kb SBG molecule was over the packaging size limit for AAV vectors.

When these fragments were used to supply AAV production, we noticed the relative abundance differed among vectors produced. Since the Poly A containing vectors or GFP containing vectors represented different NEHJ reactions, the ratio of these two types of vectors was graphed in Figure 2C. Based on these results, it was evident that the smaller sized subgenomic particles became more dominant. This suggested that later DNA replication and packaging favor smaller genomes, which may be a major mechanism dictating the abundance of rAAV subgenomic particles.

To generate genome deletion mutant (GDM), the fragments representing 5' ITR and 3'ITR had to be linked together through NHEJ. To demonstrate this, we transfected Hek293 cells with a 5' ITR fragment carrying the CB promoter and a 3'ITR fragment carrying the GFP gene along with AAV replication and packaging helper plasmids. Interestingly, the combination of these two fragments efficiently regenerated the functional GFP expression. The infectious GFP vectors also regenerated as shown in the transduction assay. This

experiment demonstrated that the GDM molecules were produced through the same mechanism that generated the SBG virus (Figure 3).

We then created an *in vivo* model using the CRISPR-Cas9 system to mimic the breakage that occurred *in vivo.* In the vector production system, the Cas9 expression plasmid was included with the vector plasmid pCB-EGFP-3.4k. In contrast to the control without guide RNA, transfection with guide RNA produced two distinct vectors with a size of 2.3kb and 1.1kb in a native gel (Figure 4A), which corresponded to the cutting site. In the denaturing gel, the original vector was present as a 3.4kb single-stranded DNA. In contrast, the vectors produced in the presence of guide RNA appeared as single stranded DNA that were 4.6kb or 2.2kb in size. We then recovered the 2.2 kb single stranded DNA from the gel, renatured the DNA, and electrophoresed it in the native gel. The 2.2kb DNA fragment appeared as 1.1kb dsDNA, which were confirmed by restriction digestion (data not shown). These results suggested that the SBG molecules for either direction were formed in the presence of CRISPR-Cas9 induced digestion *in vivo* (Figure 4C).

DNA lesion/nicking is sufficient for generating AAV subgenomic particles

To investigate if a DNA lesion was sufficient to generate SBG molecules, CRISPR-cas9 nickase activity was introduced to the AAV production system (Figure 5). As presented in Figure 5B, the *in vivo* cutting with cas9 at various positions generated two major SBG molecules corresponding to the cutting sites. Using gRNA9 as an example, Cas9 cutting generated two vectors at 1.5kb and 1.9kb respectively. However, nicking at the top strand by gRNA9 and Cas9-H840A yielded a 1.5kb sized DNA molecule. Nicking at the bottom strand by gRNA9 and Cas9-D10A yield a 1.9kb sized vector genome. In contrast, cutting of the vector by gRNA4 and Cas9 yielded two main vectors, 0.6kb and 1.2kb in size. The

2.8kb SBG vectors that should have appeared was not observed because it exceeded the packaging capacity of AAV particle. Nicking with gRNA4 and Cas9-H840A yielded 0.6kb vector along with its dimer at 1.2kb. On the other hand, nicking at the bottom strand by gRNA4 and Cas9-D10A yielded no major bands since the theoretical 2.8kb SBG vectors are oversized for AAV capsids. Similar results were obtained from gRNA13 induced nicking or cutting. The other exception was when NHEJ product replication was overwhelmed by their relatively smaller companion fragments. In this case, the corresponding larger DNA was not greatly diminished, i.e. gRNA5 and gRNA10. This result suggested that DNA lesion/nicking was sufficient to generate DNA fragments that can lead to the creation of subgenomic particles. There was a clear strand selection, in which the nicking site and its 3' end ITR formed snapback molecules.

Cellular DNA damage in events led to subgenomic molecule formation

We further hypothesize that intracellular DNA damage events may lead to subgenomic DNA formation. As shown in Figure 6, hydrogen peroxide was added to examine the effects of this DNA damage reagent on AAV production. Corresponding to an increased concentration of hydrogen peroxide, the recovered rAAV vectors appeared as smears that were smaller in size compared to the standard AAV vectors. At 200mM hydrogen peroxide, the majority of vector DNA detected were small subgenomic DNA particles (Figure 6A). We prepared a library of the recovered DNA from these vectors and performed DNA sequencing using the PacBio platform. More than 50,000 genomes were sequenced. The majority of these sequences were not AAV vector related and appeared as short DNA fragments. Those sequences aligned to the initial vector appeared to be heavily fragmentated (Figure 6B). In addition, SBG produced from the initial vector were

recovered in the sequencing as well (Figure 6B). Some of these molecules were found to contain portions of the plasmid backbone. These results showed that global DNA damage events can ruin recombinant AAV production and lead to the production of subgenomic particles.

Discussion:

The heterogeneity in wild type AAV virus and recombinant AAV vectors has been well documented(*3, 9*). Similar to what has been observed for wtAAV, the subgenomic particles in rAAV vectors have similar molecular conformations: snapback genomes (SBG), genome deletion mutants (GDM), and incomplete genomes (IDG) (Figure 1). In addition to these three major categories, a fourth category of subgenomic particles was identified as secondary derivative genome (SDG) particles arising from damage to the SBG, GDM, and IDG forms, followed by a second round of NHEJ events. The unique molecular configuration of SDG molecules prompted us to explore NHEJ events as the main cause of subgenomic DNA particle formation. While a DNA polymerase template switch mechanism may explain the formation of sSBG(*8, 9*), the existence of GDM molecules, especially the large GDM which exceed the size of the parent AAV vector and have partial duplication of vector sequences in the junction (Figure 1), strongly favors NHEJ as the primary mechanism that leads to the formation of subgenomic particles.

The essence of the NHEJ mechanism is the ligation of various DNA fragments. Interestingly, we were able to regenerate those SBG and GDM molecules using DNA fragments derived from AAV vector genomes, either by straight *in vitro* restriction endonucleases digestion or CRISPR-cas9 *in vivo* digestion. The generation of SBG molecules was quite efficient. Often it was the dominant vector molecules produced (Figure 2, 4, 5). Furthermore, when two fragments were introduced into the AAV packaging system, the formation of GDM could be confirmed as well (Figure 3). This mechanism can also explain why AAV vectors often include host genomic DNA sequence as well materials used for AAV production.

Another key point explored in this study was how the AAV fragments originated. The nickase experiment (Figure 5) showed that simple nicking of an AAV genome was sufficient to generate corresponding SBG. Even more interesting was the identification of a stand preference for nicking. The resulting SBG contained DNA from the nicking site to its 3' ITR. This evidence indicated that the creation of such fragments was closely coupled to DNA replication.

The nicking/lesion of DNA in the rAAV genomes suggest that any host/viral factors that were associated with AAV genomes could lead to subgenomic AAV particles formation. Hydrogen peroxide is an oxidizer that can cause global DNA damage *in vivo*. Our study showed that when H_2O_2 was present at a high concentration, rAAV production was completely ruined (Figure 6) and resulted in the generation of primarily subgenomic particles. SBG and GDM can be observed in the sequencing analysis (Figure 6).

Unlike the DNA template switch model, which only explains the formation of largely symmetric SBG, the NHEJ mechanism seamlessly explained the formation of SBG and GDM simultaneously. Therefore, we proposed a comprehensive model of subgenomic AAV particle formation in both wild type AAV and recombinant AAV (Figure 7). When fragments with only one ITR are produced, it will undergo self-ligation or ligate to another fragment with only one ITR. In turn, this will create recombinant molecules with two ITRs. In case of molecules that are larger than the standard AAV size, it will not be packaged. The nicking of the AAV genome (DNA lesion) and breakage of AAV DNA leads to the formation of various DNA fragments. Both host factors and AAV proteins may cause nicking and breakage in the rAAV genomes. Ligation of these fragments generates both SBG and GMD molecules. Although it is also possible that such ligation will pick up any

genomes from the host cells, the majority of the products will be SBG and GMD due to their abundance in the replication center and proximity of these fragments. The subsequent DNA replication will favor SBG or GMD when they have small genomes. Alternatively, the replication dimer of rAAV has breakage points flanking the double-D ITR, and therefore, it will efficiently self-ligate and generate SBG.

For wtAAV virus, subgenomic particles are beneficial for AAV life cycle. However, the consequences of subgenomic particles in rAAV are generally harmful. As suggested in Figure 7, subgenome particles with a promoter can produce dsRNA which will be detrimental to long term gene expression. Although dsRNA was investigated in AAV vectors (*10, 11*), here we identified SBG as the true source for such dsRNA formation. In addition, the SBG containing only the promoter can potentially cause tumorigenesis events in the host cells or in human patients(*12*). From this study, it is clear that we need control the subgenomic particle formation in rAAV production. First, rAAV genomes should be optimized such that those sequences that are prone to breakage, i.e. nicking site for enzyme or with strong secondary structure, should be avoided. Second, host cells should be maintained in healthy condition to avoid DNA damage. Third, cellular factors should be controlled to reduce DNA damage. Further studies should be carried out to minimize the production of SBG in rAAV vectors and improve its safety profile.

Materials and Methods:

Cell lines and transfection

HEK293 cells and GM16095 cells (a human fibroblast cell line purchased from the Coriell Institute, Camden, NJ) were cultured in DMEM supplemented with 10% fetal bovine serum, 100 μg/mL penicillin, and 100 units/mL streptomycin (Invitrogen, Carlsbad, CA). All cells were maintained in a humidified 37°C incubator with 5% CO2. PolyJet™ DNA In Vitro Transfection Reagent (SignaGen Laboratories) was used to deliver DNA into HEK 293 cells. Cells were seeded into six-well plates or 10-cm-diameter culture dishes 18 to 24 hours prior to transfection so that the monolayer cell density reached the optimal 70~80% confluency at the time of transfection. Complete culture medium with serum was freshly added to each plate 30 minutes before transfection. Prepare PolyJet™-DNA Complex for transfection according to the ratio of 3μL PolyJet™ to 1μg DNA using serum-free DMEM to dilute DNA and PolyJet™ Reagent. This was incubated for 10~15 minutes at room temperature and then the PolyJet™/ DNA mixture was added onto the medium. The PolyJet™/DNA complex-containing medium was then removed and replaced with fresh serum-free DMEM 12~18 hours post transfection.

rAAV Infection

GM16095 cells were seeded into 12-well plates 24 hours prior to infection so that the monolayer cell density reached the optimal 70~80% confluency at the time of infection. The cells were washed with DMEM culture medium without serum twice, 3 min each time, before infection. 10μL of cell culture medium containing rAAV virions were added into the

plate and incubated at the indicated timepoints. GFP or mCherry fluorescence expression was observed using fluorescent microscopy (Leica D3000 B).

Plasmid and plasmid fragments

Plasmid pH22 was the helper plasmid used that contained the *rep* and *cap* coding sequences. Plasmid pFd6 was the miniadenovirus helper. Plasmid pCB-GFP-3.4K contained a cytomegalovirus enhancer and beta -active promoter. This plasmid was used to make vector containing the green fluorescent protein (GFP) reporter gene flanked by the AAV ITR.

Plasmid pCB-GFP-6.4K was made by cloning a 3b fragment into pCB-GFP-3.4K. The pCB-GFP-6.4K plasmid was subjected to a series of restriction digestion to produce a series of DNA fragments with different lengths, that where, fCB-GFP-0.6K, fCB-GFP-1.0K, fCB-GFP-1.6K, fCB-GFP-1.8K, fCB-GFP-2.3K and fCB-GFP-3.1K.

The gRNA target sequences in the pCB-GFP-3.4K rAAV genome were designed using Broad Institute gRNA designer tool (https://www.broadinstitute.org/rnai/public/analysis-tools/sgrna-design). The sequences for these sgRNA targets are gRNA 4: GGG AGC GGG ATC AGC CAC CG, gRNA 5: AAG CTG CGG AAT TGT ACC CG, gRNA 9: TTA GTC GAC CTC GAG CAG TG, gRNA 10: TGT TCC GGC TGT CAG CGC AG, gRNA 13: GAT CAG CGA GCT CTA GTC GA.

Virus production and purification

AAV viruses were produced using the triple plasmid transfection system in HEK 293 cells. PolyJet™ DNA In Vitro Transfection Reagent (SignaGen Laboratories) was used to deliver DNA into the HEK 293 cells. At 72 hours after transfection, medium was collected and precipitated with 40% of PEG (finial concentration 8%) overnight at 4°C. Next, it was centrifugated, resuspended and treated with DNaseI. AAV of different densities were separated using CsCl gradient ultracentrifugation. AAV of different densities were extracted and dialyzed against 5% sorbitol in Phosphate Buffered Saline (PBS, NaCl 137 mM, KCl 2.7 mM, Na2HPO4 10 mM, KH2PO4 1.8 mM, pH 7.2). Vector genome titers were determined by quantitative real-time PCR (qPCR), with vector titers expressed as vg/ml. To obtain vectors representative of all viral particles, the gradient centrifugation step was skipped. Three days after transfection, the medium was collected and precipitated into concentrated solution of rAAV particles. rAAV genomic DNA was purified and further analyzed using agarose gel electrophoresis and qPCR.

DNA agarose gel electrophoresis

rAAV genome was extracted and purified as followed: Viral vectors were treated with DNase I (1U/mL) for 30 min at 37°C, then 1 μl of 0.5 M EDTA was added (to a final concentration of 5 mM) and subsequently heated for 10 min at 75°C to cease DNase I activity. 1/2 volume of lysis buffer (Direct PCR Tail, Viagen) containing proteinase K (40 μg/mL) was added and incubated for 1 hour at 56°C and finally heated for 10 min at 95°C. One volume of phenol: chloroform: isoamyl alcohol (25:24:1) was added to the samples,

and vortexed thoroughly for approximately 20 seconds. They were then centrifuged at 4°C for 30 minutes at 16,000 × g. The upper aqueous phase was carefully removed and transferred to a fresh tube. 200 μL of 70% ethanol was added, and the tubes centrifuged at 4°C for 10 minutes at 16,000 × g. The supernatant was carefully removed and the pellet was allowed to air dry at room temperature. 20ul of TE buffer was added to dissolve DNA. DNA concentration was measured using Nanodrop. 100ng DNA was loaded on 1% of Neutralizing gel and run at 120V for 50min. The equal DNA was loaded on a 1% of Alkaline gel and run at 60V for 100min in ice-water bath. Gels were stained using 1×SYBR@ Safe DNA gel stain (Invitrogen) and a photo taken at a wavelength of 365nm using a ChemiDOC™MP Imaging System (Bio-rad).

H_2O_2 Treatment

HEK 293 cells were seeded in twenty 15-cm dishes and incubated for 18 hours. The old culture medium was replaced with free FBS DMEM containing a final concentration of 0μM, 50μM, 100μM and 200μM H2O2 at 60 min prior to transfection. Three plasmids, pH22, pFΔ6 and pssAAV-CB-GFP-4.7K, were transfected into HEK 293 cells using PolyJet™ DNA In Vitro Transfection Reagent (SignaGen Laboratories). 72 hours after transfection, medium was collected, precipitated with 40% of PEG (finial concentration 8%), and purified by Cscl gradient method. rAAV DNA was extracted and purified using the phenol: chloroform: isoamyl alcohol (25:24:1) method. 500ng of rAAV DNA was subjected to sequence by PacBio SMRT platform, meanwhile, 30ng of DNA was loaded on 1% agarose gel and run at 120V for 50min.

Quantitative Real Time PCR (qPCR) Assay

Viral vectors (1 × 1010 vg, 1ul) in solution containing DNase I (1U/mL) were incubated for 30 min at 37°C, add 1 μl of 0.5 M EDTA (to a final concentration of 5 mM) and subsequently heated for 10 min at 75°C to cease DNase I activity. Control samples each received lysis buffer (Direct PCR Tail, Viagen) containing proteinase K (40 μg/mL), and were incubated for 1 hour at 56°C and finally heated for 10 min at 95°C. The samples intended for thermal treatment were directly heated following heat inactivation of DNase I treatment at the indicated temperatures. The copy numbers of viral genomes subsequently released were quantified by real-time PCR and expressed in vg/ml. The primers include GFP forward: AGTCCGCCCTGAGCAAAGAC and GFP reverse: CTCGTCCATGCCGAGAGTGA; polyA forward: GTGCCTTCCTTGACCCTGGA and polyA reverse: CACCTACTCAGACAATGCGATGC.

AAV Genome Sequencing and Data Analysis

For long-read PacBio SMRT sequencing, AAV samples were prepared according to SMRTbell™ procedures. DNA was extracted and purified by AMPure PB Beads and then repaired by SMRTbell™ Damage Repair kit. The adaptor ligation reaction was performed, and then ExoIII and ExoVII were added to remove failed ligation products. AMPure PB were performed three times.

SMRT subread filtering and the high-quality circular consensus sequences corresponding to the rAAV library were generated using SMRT analysis portal (minimum accuracy of 0.99 and minimum of 3 CCS passes) and considered for further analysis. Filtered reads

were mapped to the rAAV genome using the Minimap2 (https://github.com/PacificBiosciences/pbmm2) and processed alignments to demonstrate configuration categories of molecules in the rAAV population.

Figure legend

Figure 1. Molecular configuration of DNA genomes in rAAV vectors. AAV genomes were sequenced using the PacBio platform and compared to the reference sequences on the top. Besides the standard sized AAV vector genomes, four typical categories of subgenomic rAAV genomes were found in the rAAV vectors: i). Symmetric snapback genomes (sSBG) and asymmetric snapback genomes (aSBG); ii). Genome deletion mutants (GDM); iii). Incomplete genomes (ICG); iv). Secondary derivative genomes (SDG).

Figure 2. Intermolecular NHEJ is a mechanism leading to the formation of SBG molecules. The parent plasmid 6.4kb pCB-GFP-6.4k was linearized with varying restriction enzymes to obtain linear fragments as shown in A). The plasmid backbone is depicted with a dotted line. HEK 293 cells with rAAV packaging helper functions were transfected with DNA fragments in (A), plasmid pCB-GFP-6.4k, or pCB-GFP-3.4k. B) The resulting rAAV vectors in the media were harvested and the DNA in the vectors were extracted and analyzed for genome status using a 1% agarose gel. For simplicity, fragments such as fCB-GFP-0.6K were referred to as 0.6k on top of the gel in B. Red arrows indicate key fragments. The vectors recovered were quantified by qPCR using primers specific for poly A or GFP. The ratio of vectors containing poly A or GFP are shown in C. In pCB-GFP-3.4K, the vector size is 3.4kb. For DNA fragments, the size of 5'ITR-GFP is 1.8kb and the size of the poly A to 3'ITR are indicated as the last three letters in the name.

Figure 3. Intermolecular NHEJ is a mechanism leading to AAV genome deletion mutants (GDM). A) Hek293 cells were transfected with ITR fragments containing the CB promoter or GFP gene alone or combined with supplemental helper plasmids for rAAV replication

and packaging. The positive control was a 2.3kb intact pCB-GFP-3.4k plasmid. At 3 days post-transfection, the GFP expression was monitored by fluorescence microscopy (mid panel). The harvested vectors were used to transduce GM16095 cells, and the GFP expression was monitored at 24 hours post-infection (bottom panel). B). The vector DNA recovered from panel A was electrophoresed in 1% agarose gel and AAV genomes were detected by Southern blot using an ITR specific probe. Δ indicate key fragments 0.9kb, 1.4kb and 2.3 kb.

Figure 4. Intra-host cell vector DNA breakage is the mechanism for AAV subgenomic particle formation. Hek293 cells were transfected with AAV plasmid pCB-GFP-3.4k along with Cas9 expressing plasmids with or without corresponding guide RNA. A) The resulting vector DNA was electrophoresed in native agarose gel. B) The resulting vector DNA was electrophoresed in denaturing agarose gel. C) The denatured fragments of B (indicated as ① ②) were collected and renatured and were electrophoresed again in the native gel.

Figure 5. Intra-host cells vector DNA Lesion is sufficient for SBG formation. A) Illustration of gRNA sites in pCB-EGFP-3.4k for Cas9 nicking or digestion. Hek293 cells were transfected with plasmid pCB-EGFP-3.4k for vector production in the presence of cas9 or cas9 mutants (H840A or D10A) and corresponding guide RNA. B) The resulting vector DNA was electrophoresed in native agarose gel with EB staining. C) Cas9-double cut, D: D10A-nicking H: H840A-nicking. H stands for Cas9-H840A nicking. D stands for Cas9-D10A nicking. C stands for Cas9 cutting. The table summarized the potential DNA sizes that can be generated by nicking or cutting. The actual observed bands are summarized in the brackets. - indicates “not observed”.

Figure 6. DNA damaging conditions in the host cells promoted subgenomic particle formation. Hydrogen peroxide at varying concentrations was added to the rAAV production system after transfection. A) The resulting rAAV vectors were purified by CsCl gradient, and the vector DNA analyzed by gel analysis. B) Partially recovered vector genomes were sequenced and aligned to the reference sequence. The coverage is marked by blue lines. Exemplary DNA configurations of AAV subgenomic particles are illustrated at the bottom.

Figure 7. A model of subgenomic particle formation. The key point is that varying DNA fragments with only one ITR were generated from the lesion/breakage on monomer or dimer of replication form of AAV genomes. NHEJ then rejoin these fragments and the resulting products restore two ITRs in a molecule, which can be replicated and packaged in an AAV capsid. This mechanism readily led to the generation of SBG, GDM, and various forms that were not illustrated in the figure.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (NIH) of United States (HL142019, HL114152 and HL130871)

References:

1. R. J. Samulski, N. Muzyczka, AAV-Mediated Gene Therapy for Research and Therapeutic Purposes. *Annu Rev Virol* **1**, 427-451 (2014).
2. J. F. Wright, Product-Related Impurities in Clinical-Grade Recombinant AAV Vectors: Characterization and Risk Assessment. *Biomedicines* **2**, 80-97 (2014).
3. C. A. Laughlin, M. W. Myers, D. L. Risin, B. J. Carter, Defective-interfering particles of the human parvovirus adeno-associated virus. *Virology* **94**, 162-174 (1979).
4. L. M. de la Maza, B. J. Carter, Heavy and light particles of adeno-associated virus. *J Virol* **33**, 1129-1137 (1980).
5. L. M. de la Maza, B. J. Carter, Molecular structure of adeno-associated virus variant DNA. *J Biol Chem* **255**, 3194-3203 (1980).
6. E. Lecomte *et al.*, Advanced Characterization of DNA Molecules in rAAV Vector Preparations by Single-stranded Virus Next-generation Sequencing. *Mol Ther Nucleic Acids* **4**, e260 (2015).
7. P. Kapranov *et al.*, Native molecular state of adeno-associated viral vectors revealed by single-molecule sequencing. *Hum Gene Ther* **23**, 46-55 (2012).
8. J. Xie *et al.*, Short DNA Hairpins Compromise Recombinant Adeno-Associated Virus Genome Homogeneity. *Mol Ther* **25**, 1363-1374 (2017).
9. P. W. L. Tai *et al.*, Adeno-associated Virus Genome Population Sequencing Achieves Full Vector Genome Resolution and Reveals Human-Vector Chimeras. *Mol Ther Methods Clin Dev* **9**, 130-141 (2018).
10. W. Shao *et al.*, Double-stranded RNA innate immune response activation from long-term adeno-associated virus vector transduction. *JCI Insight* **3**, (2018).

11. C. Stutika *et al.*, A Comprehensive RNA Sequencing Analysis of the Adeno-Associated Virus (AAV) Type 2 Transcriptome Reveals Novel AAV Transcripts, Splice Variants, and Derived Proteins. *J Virol* **90**, 1278-1289 (2016).

12. L. E. Rosas *et al.*, Patterns of scAAV vector insertion associated with oncogenic events in a mouse model for genotoxicity. *Mol Ther* **20**, 2098-2110 (2012).

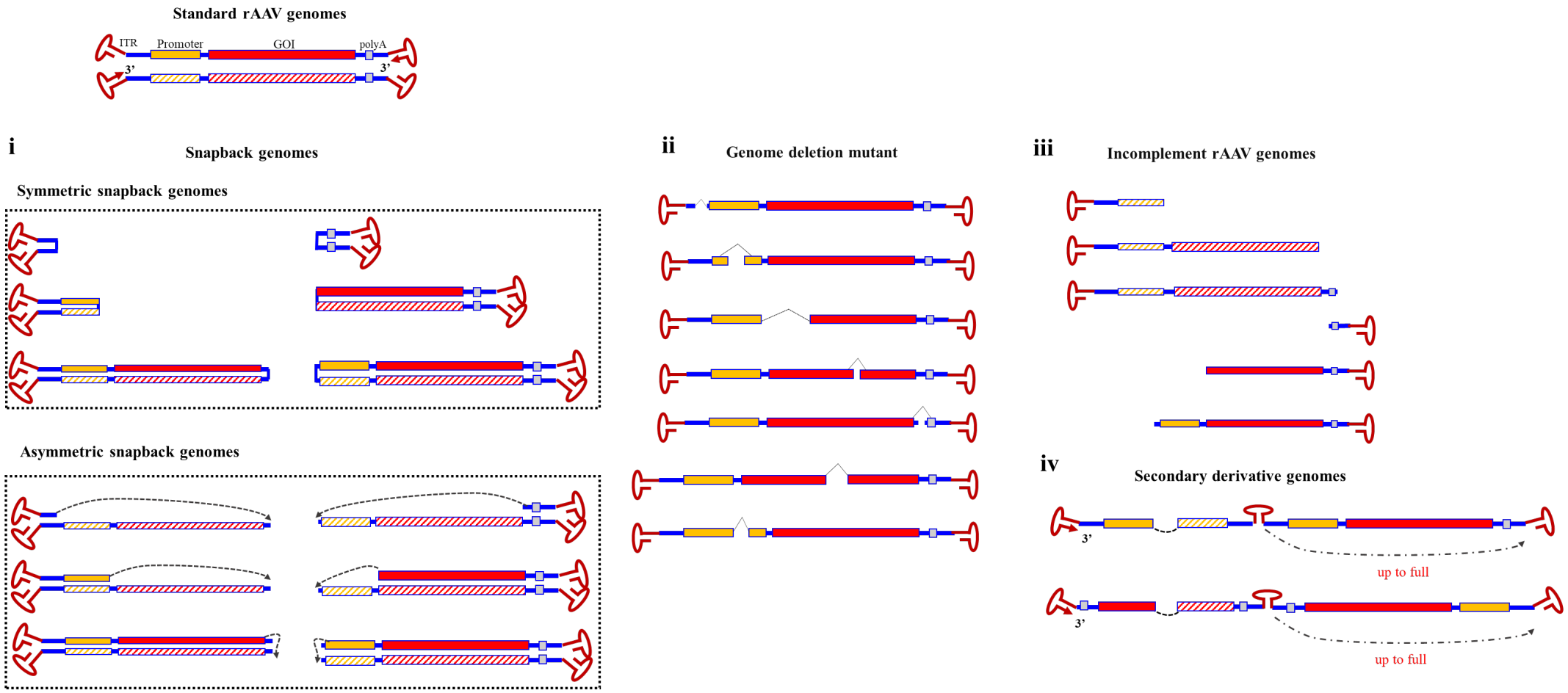

Standard rAAV genomes
ITR
Promoter
GOI
polyA
3'
3'
i
Snapback genomes
Symmetric snapback genomes
Asymmetric snapback genomes
ii
Genome deletion mutant
iii
Incomplement rAAV genomes
iv
Secondary derivative genomes
3'
up to full
3'
up to full

Figure 1

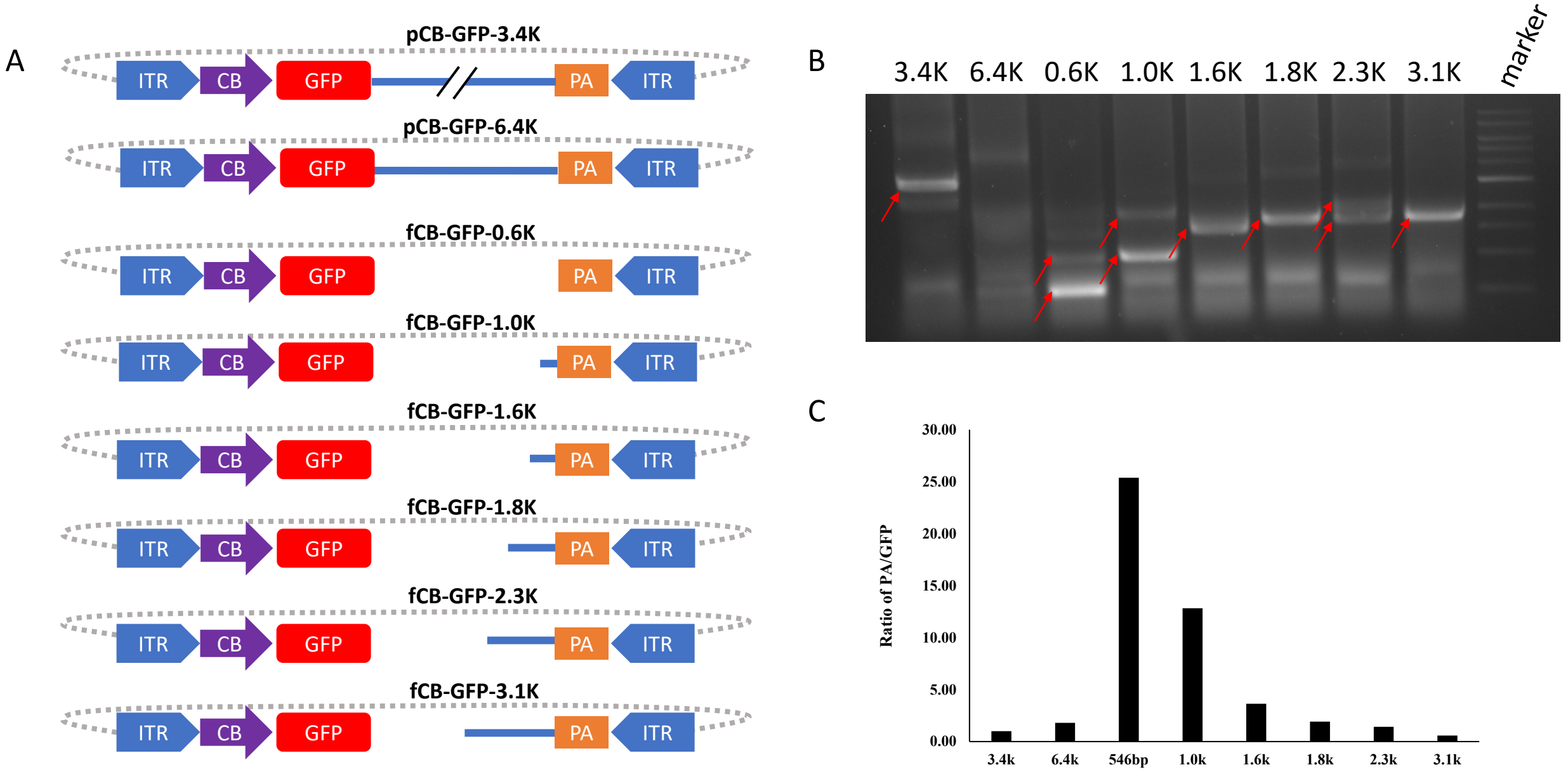

A
pCB-GFP-3.4K
ITR
CB
GFP
PA
ITR
pCB-GFP-6.4K
fCB-GFP-0.6K
fCB-GFP-1.0K
fCB-GFP-1.6K
fCB-GFP-1.8K
fCB-GFP-2.3K
fCB-GFP-3.1K
B
3.4K
6.4K
0.6K
1.0K
1.6K
1.8K
2.3K
3.1K
marker
C
Ratio of PA/GFP
30.00
25.00
20.00
15.00
10.00
5.00
0.00
3.4k
6.4k
546bp
1.0k
1.6k
1.8k
2.3k
3.1k

Figure 2

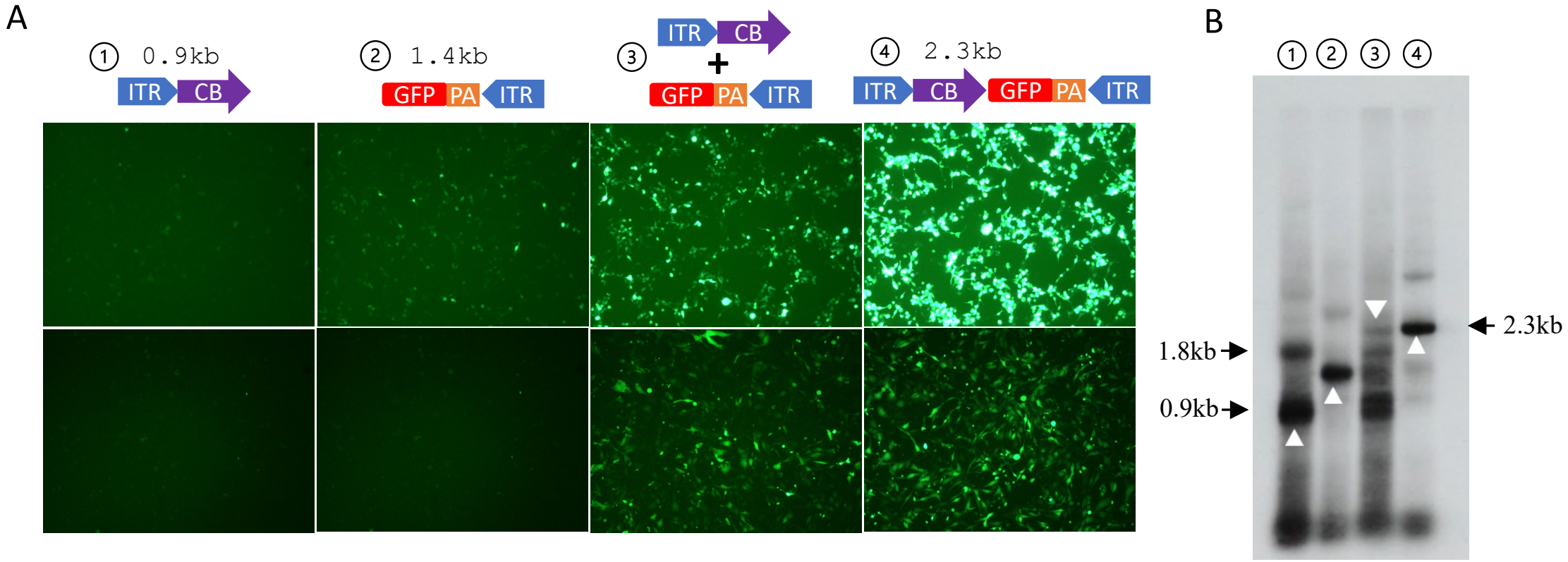

A
① 0.9kb
ITR
CB
② 1.4kb
GFP
PA
ITR
③
ITR
CB
+
GFP
PA
ITR
④ 2.3kb
ITR
CB
GFP
PA
ITR
B
① ② ③ ④
1.8kb
0.9kb
2.3kb

Figure 3

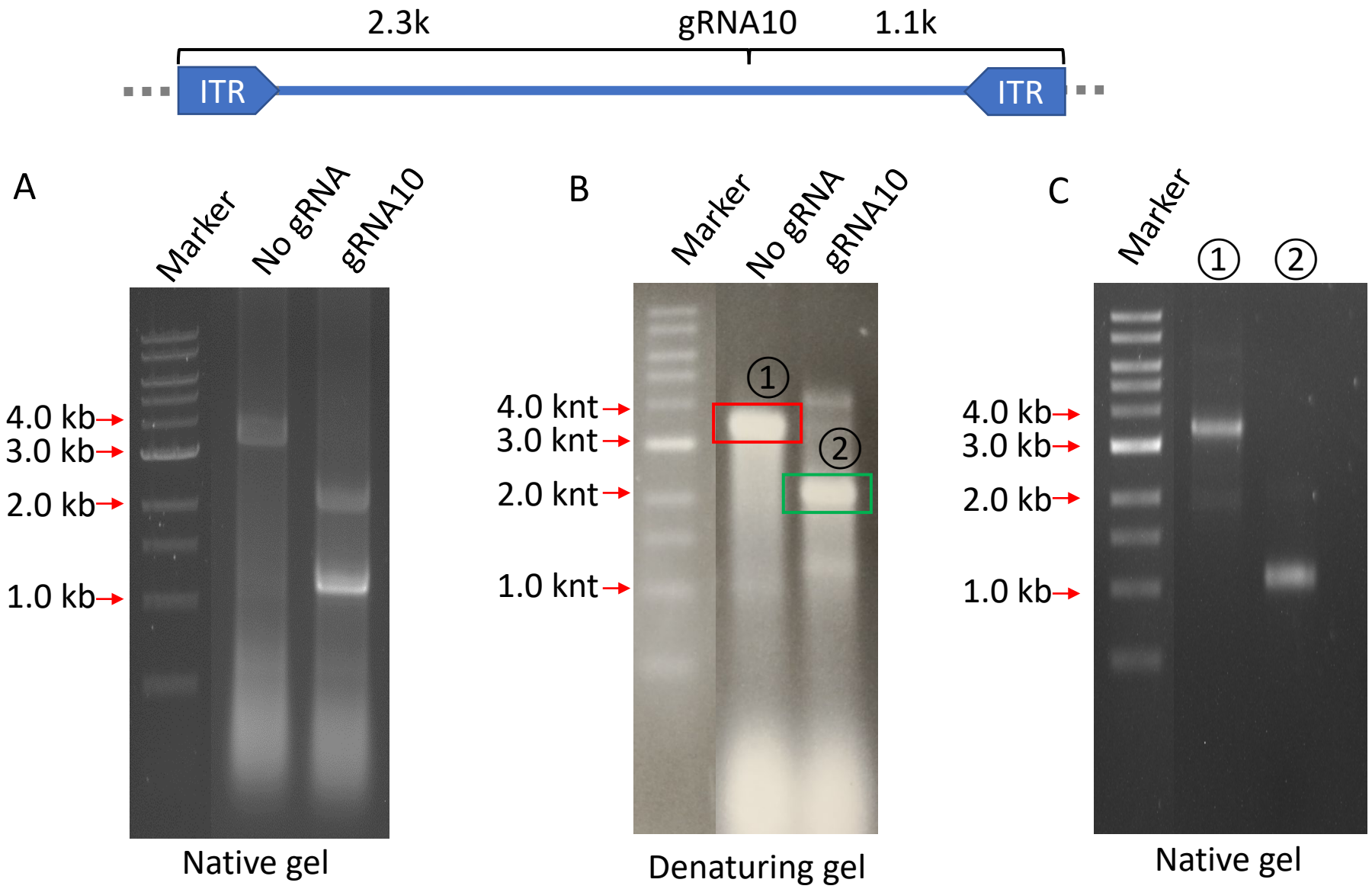
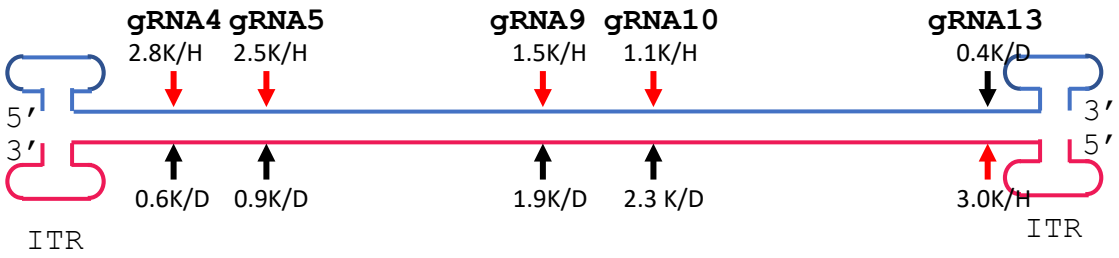

2.3k
gRNA10
1.1k
ITR
ITR
A
Marker
No gRNA
gRNA10
4.0 kb
3.0 kb
2.0 kb
1.0 kb
Native gel
B
Marker
No gRNA
gRNA10
4.0 knt
3.0 knt
2.0 knt
1.0 knt
1
2
Denaturing gel
C
Marker
1
2
4.0 kb
3.0 kb
2.0 kb
1.0 kb
Native gel

Figure 4

a



c

Guide RNA	Vector size expected (monomer/dimer)	
gRNA4	0.6K/1.2k(C,D,-)	2.8K/n.a (-,-,-)
gRNA5	0.9K/1.8k(C,D,-)	2.5K/n.a (-,-,H)
gRNA9	1.9K/n.a (C,D,-)	1.5K/n.a (C,-,H)
gRNA10	2.3K/n.a (-,D,-)	1.1K/2.2K(C,-,H)
gRNA13	3.0K/n.a (-,-,-)	0.4K/0.8k(C,D,-)

b

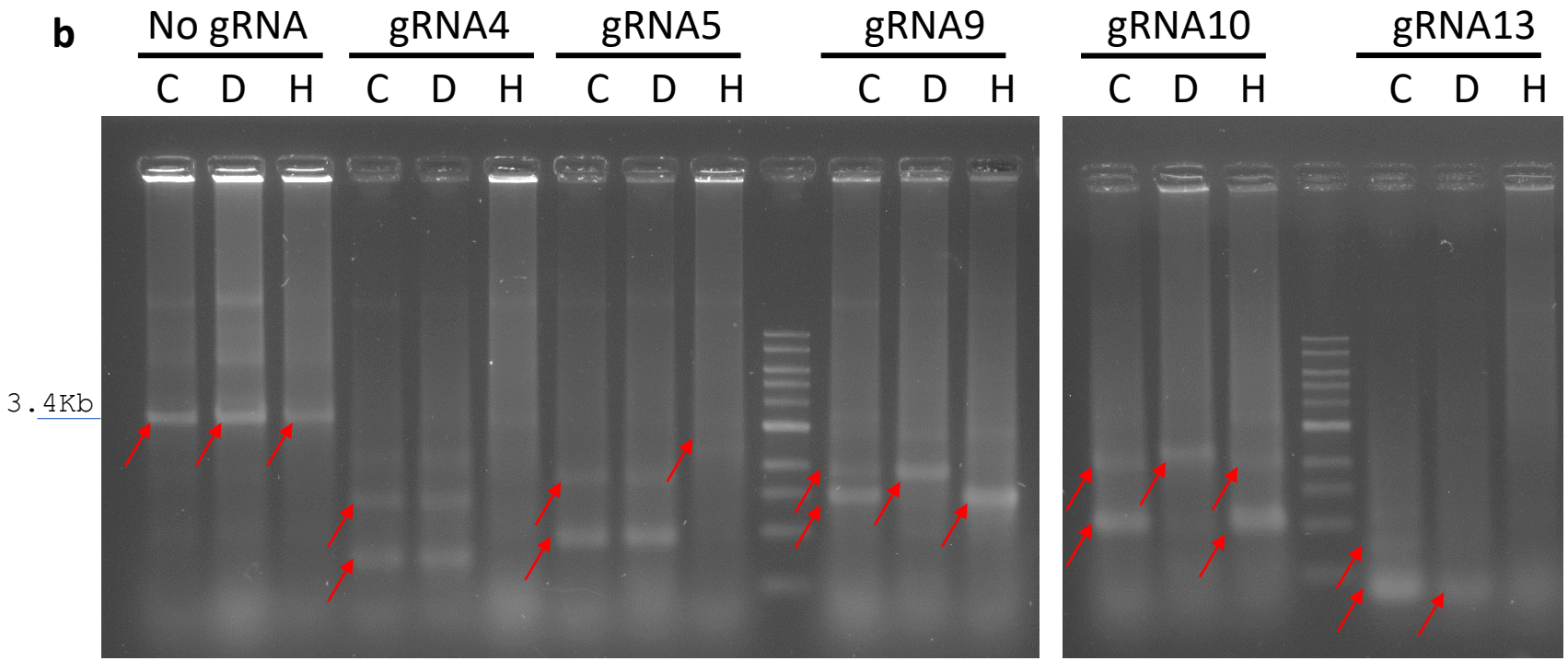


Figure 5

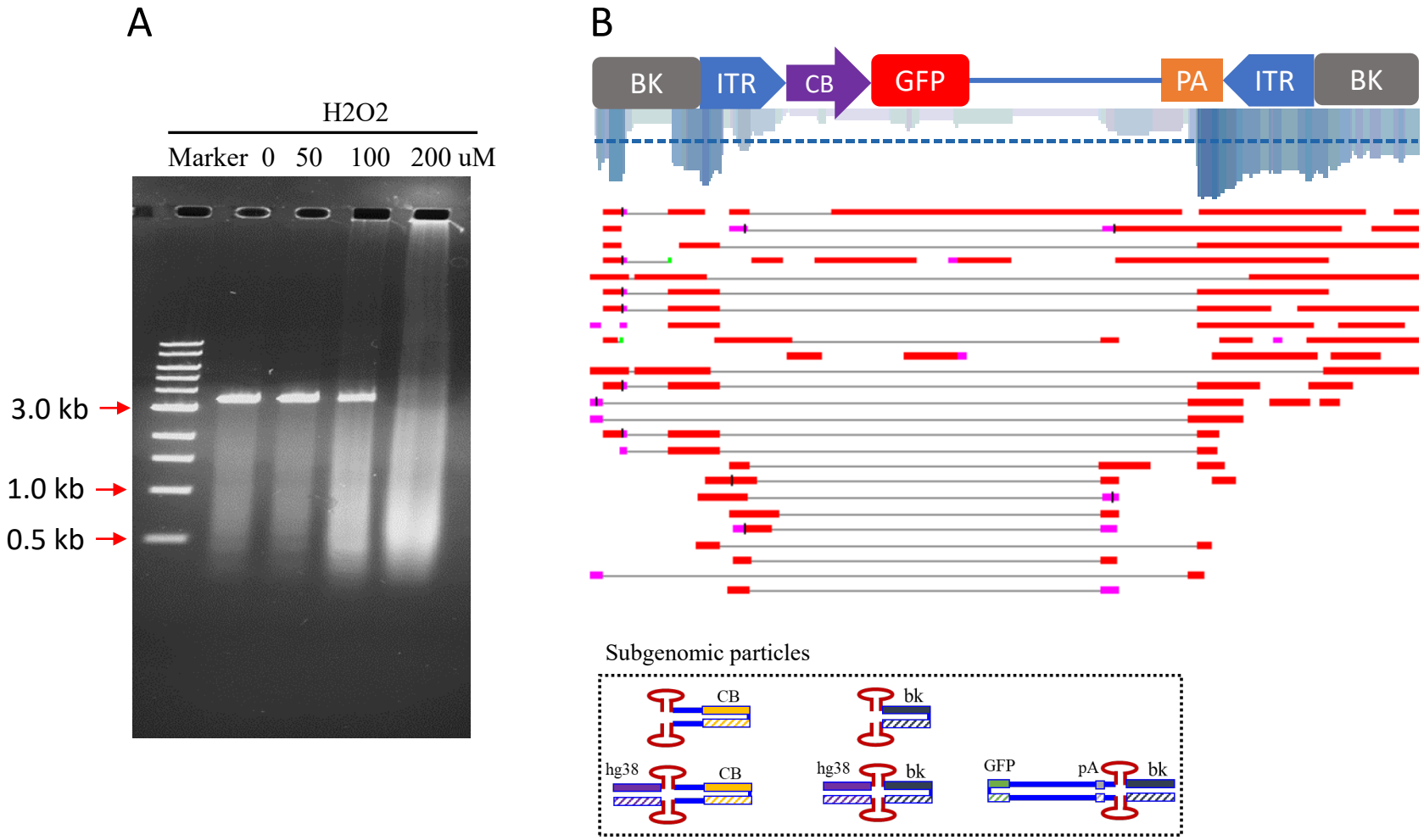

A
H2O2
Marker 0 50 100 200 uM
3.0 kb
1.0 kb
0.5 kb
B
BK
ITR
CB
GFP
PA
ITR
BK
Subgenomic particles
CB
bk
hg38
CB
hg38
bk
GFP
pA
bk

Figure 6

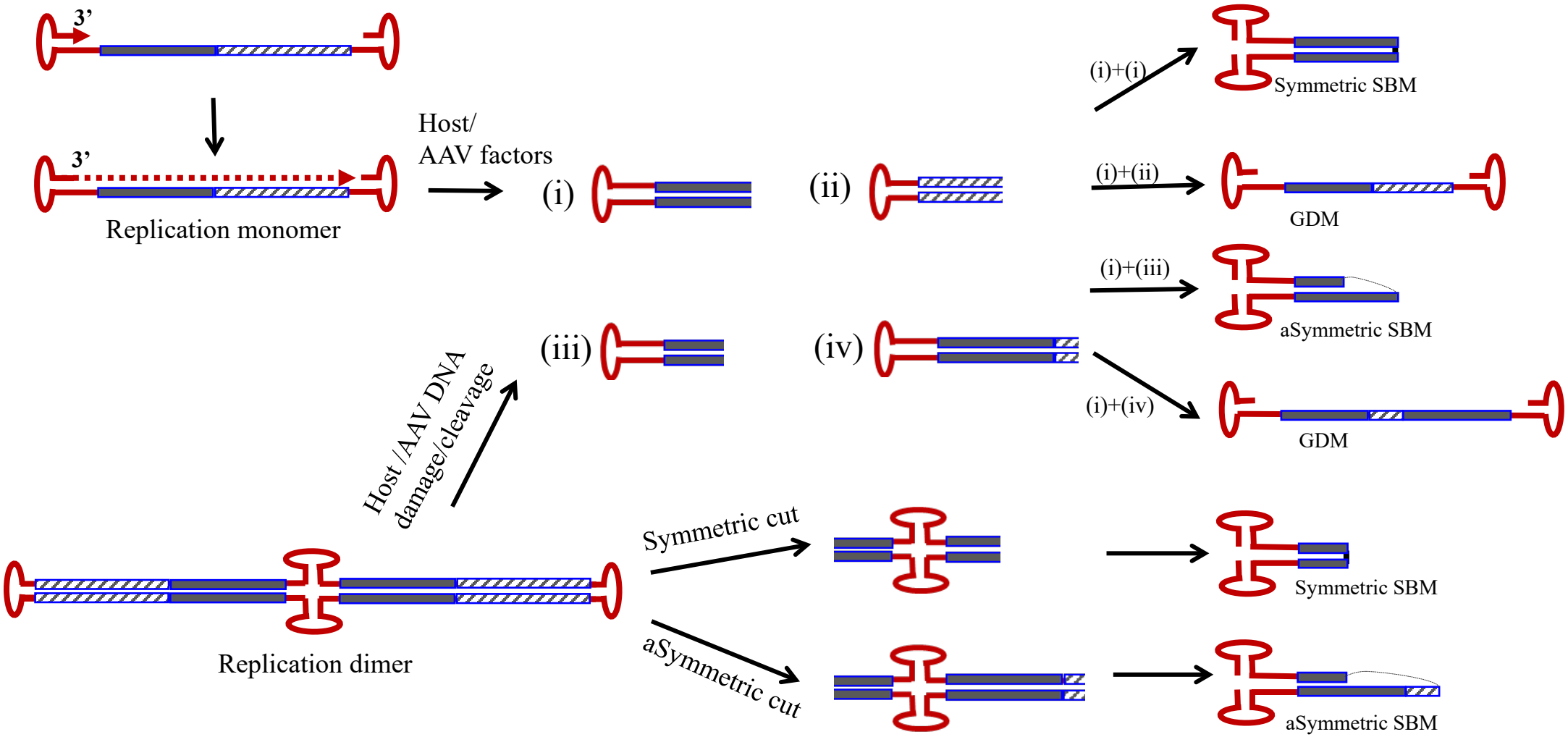

3'
3'
Replication monomer
Host/
AAV factors
(i)
(ii)
(iii)
(iv)
(i)+(i)
Symmetric SBM
(i)+(ii)
GDM
(i)+(iii)
aSymmetric SBM
(i)+(iv)
GDM
Host /AAV DNA
damage/cleavage
Replication dimer
Symmetric cut
aSymmetric cut
Symmetric SBM
aSymmetric SBM

Figure 7