Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University Heidelberg

Presented by

Felix Bubeck

born in: Ulm (Germany)

Oral examination: 08.09.2025

Searching the hairpin in the haystack: Engineering the replication origin of AAV vectors

Referees: Prof. Dr. Robert Russell
Prof. Dr. Dirk Grimm

Abstract

Gene therapy using adeno-associated virus (AAV)-based vectors has advanced significantly in recent years, with multiple approved therapeutics now available on the market. Engineering of the capsid, transgene, and promoters has shown immense potential to improve the safety and efficacy of these drugs. The only vector component that has mostly remained in its original state is the viral replication origin, the inverted terminal repeats (ITRs). Several ITR variants have so far shown potential to improve vector efficacy and safety as well, implying that the parallel screening of large ITR libraries could improve the identification of superior variants. The reason this has never been attempted for variants of the central part of the ITR may be the strong secondary structure of the ITRs, severely impairing the cloning of novel variants. Additionally, the partial loss of the ITRs due to processing by the host's DNA repair machinery prohibits tracing of the ITR variants upon transduction. To overcome these bottlenecks, a rarely used plasmid design was harnessed in the present work to facilitate the cloning of more than 110 newly designed ITR variants. This was complemented with a novel Sanger sequencing-based method for fast and reliable ITR sequence verification in plasmid DNA, which proved effective for all conventional and most alternative ITR structures. To trace the ITR variants during production and transduction, they were tagged with a barcode sequence in the transgene of the vector genome. Intriguingly, after vector production with a pool of ITR variants, the barcoding also enabled the identification of a novel ITR repair pathway that relies on the presence of ITR sequences as repair templates *in trans*. This ITR *trans*-repair was detectable in all parts of the ITR and impedes the barcode-ITR association in pooled productions. Importantly, though, this repair could be circumvented by separate production of each ITR variant, which also facilitated the maintenance of highly diverse ITR hairpin sequences within vector genomes. Subsequently, after confirmation of the vector hairpin integrity in 90 ITR variants with high similarity to the wild-type ITR of AAV2 (wtITR2), the vectors were screened *in vitro* and *in vivo* for their effects on transduction. This revealed that the wtITR2 exhibited superior functionality *in vitro* but not *in vivo*, suggesting an immense potential for alternative ITR variants to improve transgene expression in human patients. As AAV vectors exhibit semi-random integration into the host genome, which could lead to adverse events in patients, a method was developed that enables interrogation of ITR-associated barcodes as well as the integration region. This allows the quantification of the integration propensity of different ITR variants, thereby complementing the comprehensive pipeline to screen for effects of ITR modifications on vector functionality.

Abstract

Long-term gene expression mediated by AAV vectors relies on vector genomes persisting in the cell as circular episomes or integrated into the host genome. To avoid the reliance on ITR sequences for vector genome circularization by the inefficient host cell machinery and to reduce the risks associated with host genome integration, the packaging of circular DNA genomes in AAV capsids may represent an alternative strategy to enhance the safety and efficacy of AAV-based gene therapy. Therefore, the possibility of generating circular vector genomes using a circovirus-inspired engineered AAV replication origin was examined. As hoped for, this generated circular replication intermediates using the AAV replication machinery, although packaging in AAV particles in a circular conformation was not detected.

Collectively, this work has yielded a pipeline for (i) the generation of ITR variant plasmids, (ii) the validation of the ITR sequence in the plasmids, (iii) the confirmation of ITR integrity in the vector genomes, (iv) the analysis of transduction of ITR variants by barcodes, and (v) the interrogation of the effects of ITR variants on host genome integration. Taken together, this forms a comprehensive basis for larger-scale ITR variant screens. Concomitantly, the validation of the formation of circular replication intermediates could serve as a starting point to engineer AAV vectors with circular genomes. As a whole, the results of this work could therefore facilitate and accelerate the development of next-generation AAV gene therapy vectors.

Zusammenfassung

Die Gentherapie mit Vektoren auf der Basis von Adeno-assoziierten Viren (AAV) hat in den letzten Jahren erhebliche Fortschritte gemacht, insbesondere durch mehrere bereits für den Markt zugelassene Medikamente. Die Modifikation des Kapsids sowie die Optimierung des Transgens und der Promotoren haben dabei ein immenses Potenzial zur Verbesserung der Sicherheit und Wirksamkeit dieser Medikamente gezeigt. Die einzige Vektorkomponente, die größtenteils in ihrem ursprünglichen Zustand geblieben ist, sind die viralen Replikationsursprünge, die Inverted Terminal Repeats (ITRs). Mehrere ITR-Varianten haben bisher das Potenzial gezeigt, die Wirksamkeit und Sicherheit der AAV-basierten Vektoren zu verbessern, weshalb ein paralleles Screening großer ITR-Bibliotheken die Identifizierung überlegener Varianten vereinfachen könnte. Der Grund dafür, dass dies für Varianten des zentralen Teils des ITR noch nie versucht wurde, könnte die starke Sekundärstruktur der ITRs sein, die das Klonieren neuer Varianten erschwert. Außerdem verhindert der teilweise Verlust der ITRs währed der Verarbeitung durch die DNA-Reparaturmaschinerie des Wirts die Rückverfolgung der ITR-Varianten nach der Transduktion. Um diese Engpässe zu überwinden, wurde in der vorliegenden Arbeit ein selten genutztes Plasmid-Design verwendet, das die Klonierung von mehr als 110 neuentworfenen ITR-Varianten vereinfachte. Ergänzt wurde dies durch eine neuartige, auf Sanger-Sequenzierung basierende Methode zur schnellen und zuverlässigen Verifizierung der ITR-Sequenz in Plasmid-DNA, die sich sowohl für konventionelle als auch für die meisten alternativen ITR-Strukturen als funktional erwies. Um die ITR-Varianten während der Produktion und Transduktion zu verfolgen, wurden sie mit einer Barcode-Sequenz im Transgen des Vektorgenoms ausgestattet. Bemerkenswert ist, dass das Barcoding bei der Vektorproduktion mit einem Pool von ITR-Varianten auch die Identifizierung eines neuartigen ITR-Reparaturmechanismus ermöglichte, der auf der Anwesenheit von ITR-Sequenzen *in trans* basiert, welche als Reparaturmuster dienen. Dieser ITR-*trans*-Reparaturmechanismus war in allen Teilen des ITR nachweisbar und verhindert die Barcode-ITR-Assoziation in gepoolten Produktionen. Allerdings konnte diese Reparatur durch die separate Produktion der einzelnen ITR-Varianten umgangen werden, was sogar die Erhaltung von Hairpin-Sequenzen in den Vektorgenomen ermöglichte, welche stark von AAV ITRs abweichen. Nach der Überprüfung der Vektor ITR Integrität in einer Bibliothek von 90 ITR-Varianten mit hoher Ähnlichkeit zur Wildtyp-(wt)AAV2-Sequenz wurden die Vektoren *in vitro* und *in vivo* auf ihre Auswirkungen auf die Transduktion untersucht. Dabei stellte sich heraus, dass der AAV2 wtITR (wtITR2) *in vitro*, nicht aber *in vivo* eine höhere Funktionalität

aufweist, was auf ein großes Potenzial für alternative ITR-Varianten zur Verbesserung der Transgenexpression bei Patienten schließen lässt. Da die Integration von AAV-Vektoren in das Wirtsgenom semi-randomisiert erfolgt, was zu unerwünschten Nebenwirkungen bei Patienten führen könnte, wurde eine Methode entwickelt, mit der die ITR-assoziierten Barcodes sowie die Integrationsregion gleichzeitig bestimmt werden können. Dies erlaubt eine Quantifizierung der Integrationsneigung verschiedener ITR-Varianten und vervollständigt so die umfassende Pipeline zum Screening der Auswirkungen von ITR-Modifikationen auf die Vektorfunktionalität.

Die durch AAV-Vektoren vermittelte langfristige Genexpression hängt davon ab, dass die Vektorgenome als zirkuläre Episomen in der Zelle persistieren oder in das Wirtsgenom integriert werden. Um die Abhängigkeit der Zirkularisierung des Vektorgenoms von ITR-Sequenzen durch die ineffiziente Reparaturmaschinerie der Wirtszelle zu vermeiden und gleichzeitig die mit der Integration des Wirtsgenoms verbundenen Risiken zu verringern, könnte die Verpackung zirkulärer DNA-Genome in AAV-Kapsiden eine alternative Strategie zur Verbesserung der Sicherheit und Wirksamkeit der AAV-basierten Gentherapie sein. Daher wurde die Möglichkeit untersucht, zirkuläre Vektorgenome mit Hilfe eines von Circoviren inspirierten AAV-Replikationsursprungs zu erzeugen. Wie erhofft, wurden dabei mit Hilfe der AAV-Replikationsmaschinerie zirkuläre Replikationsintermediate erzeugt, deren Verpackung in AAV-Partikel in zirkulärer Konformation jedoch nicht nachgewiesen werden konnte.

Damit hat diese Arbeit eine Pipeline für (i) die Klonierung von Plasmiden mit ITR-Varianten, (ii) die Validierung der ITR-Sequenz in den Plasmiden, (iii) die Bestätigung der ITR-Integrität in den Vektorgenomen, (iv) die Analyse der Transduktion von ITR-Varianten durch Barcodes und (v) die Untersuchung der Auswirkungen von ITR-Varianten auf die Integration in das Wirtsgenom hervorgebracht. Zusammengenommen bildet dies eine umfassende Grundlage für groß angelegte ITR-Varianten-Screens. Die Validierung der Bildung von zirkulären Replikationsintermediaten könnte als Ausgangspunkt für die Entwicklung von AAV-Vektoren mit zirkulären Genomen dienen. In ihrer Gesamtheit könnten die Ergebnisse dieser Arbeit daher die Entwicklung von AAV-Gentherapie-Vektoren der nächsten Generation erleichtern und beschleunigen.

Table of contents

List of figures

List of tables

List of abbreviations

Abbreviation	Full expression
AAP	Assembly-activating protein
AAV	Adeno-associated virus
AAVS1/2/3	AAV integration site 1/2/3
*Ae*DV	*Aedes aegypti* densovirus
ATAC-seq	Assay for transposase-accessible chromatin sequencing
ATM	Ataxia telangiectasia mutated
ATR	Ataxia telangiectasia and Rad3-related kinase
BC-seq	Barcode sequencing
CARE	*Cis*-acting replication element
cceDNA	Covalently closed-end double stranded DNA
cDNA	Complementary DNA
CFTR	Cystic fibrosis transmembrane conductance regulator
CLIC	Clathrin-independent carrier
CMV	Cytomegalovirus
DBD	DNA-binding domain
DDR	DNA damage response
DMD	Duchenne muscular dystrophy
DNA-PKCS	DNA-dependent protein kinase catalytic subunit
dsDNA	Double-stranded DNA
EMA	European Medicines Agency
FDA	U.S. Food and Drug administration
FFLuc	Firefly luciferase
G4	G-quadruplex
GEEC	Glycosylphosphatidylinositol-anchored protein-enriched early endosomal compartment
HCC	Hepatocellular carcinoma
Hek293T	Human embryonic kidney 293T
hESCs	Human embryonic stem cells
iPSCs	Induced pluripotent stem cells
ITR	Inverted terminal repeat
LINE	Long interspersed nuclear element
LB	Lysogeny Broth
MAAP	Membrane-associated accessory protein
mapq	Mapping quality
MHC-II	Major histocompatibility complex class II
miRNA	MicroRNA
MOI	Multiplicity of infection
MRN	Mre11, Rad50, and Nbs1
NGS	Next-generation sequencing
OD	Oligomerization domain
ori	Origin of replication
PCR	Polymerase chain reaction
PLA2	Phospholipase A2
polyA	Polyadenylation signal
rAAV	Recombinant AAV
RBE	Rep-binding element
RBS	Rep-binding site

Abbreviation	Full expression
RCA	Rolling circle amplification
RCR	Rolling circle replication
*R*Luc	*Renilla* luciferase
RFX	Regulatory factor X
RHR	Rolling hairpin replication
rRNA	Ribosomal RNA
S/MAR	Scaffold/matrix attachment region
scAAV	Self-complementary AAV
SINE	Short interspersed nuclear element
sITR	Single ITR
snoRNA	Small nucleolar RNA
ssAAV	ssDNA-containing AAV
ssDNA	Single-stranded DNA
T7eI	T7 endonuclease I
TB	Terrific Broth
TIIS	Type II-S restriction enzyme
TLR	Toll-like receptor
TR	Terminal repeats
trs	Terminal resolution site
UMMS	University of Massachusetts Medical School
UTR	Untranslated region
vg	Vector genomes
wt	Wild type
VP1u	VP1 unique region
wtITR x	wtITR of AAV x

1. Introduction

The following chapters will provide an overview of Adeno-associated virus (AAV) biology and the usage and challenges of AAV-based vectors for gene transfer. This will be complemented with a detailed overview over the inverted terminal repeat (ITR) sequences including their characteristics and function during AAV genome replication, packaging, and transduction as well as the potential of engineered ITR sequences.

1.1 Adeno-associated virus biology

AAV was first reported in 1965 as a contaminant in preparations of simian adenovirus by Atchison [1] and Hoggan [2], who initially identified it as a double-stranded DNA (dsDNA) virus. Soon thereafter, it was shown that the virus actually contains single-stranded DNA (ssDNA), with both positive and negative strands packaged in separate capsids at equal ratios [3-6]. Initially, simian and human adenovirus were found to induce the replication of AAV [1,2], which was later extended to vaccinia [7]-, papilloma [8]- and herpesviruses [7] as well as genotoxic stress [9]. Due to this dependency on other viruses for replication, the virus was classified into the genus *Dependoparvovirus* belonging to the *Parvoviridae* family.

1.1.1 Natural diversity of AAVs

AAVs exhibit substantial genetic variation, with multiple serotypes, variants and clades identified across different species, including birds [10] and reptiles [11,12]. These differ in their capsid proteins and thus exhibit unique tissue tropism, immune recognition, and packaging efficiency. The co-existence of AAVs in vertebrates has been traced back at least 40 million years [13], based on the presence of endogenous AAV-like sequences within genomes of different species. The evolution of AAV is thought to have originated from a common ancestor shared with other members of the *Parvoviridae* family. As noted, AAVs package positive and negative DNA strands at equal ratios, while others, especially heterotelomeric parvoviruses, package mainly the negative strand, suggesting also a parallel evolution of the replication system [3-6,14,15]. The capsid has evolved through genomic recombination and random mutations, with a total of 13 naturally occurring primate AAV serotypes (AAV1 to AAV13) described so far. Some of these serotypes are derived from contaminations in cell culture or adenovirus stocks (AAV1, 2, 6 and 12), while others have been extracted from human (AAV3 ,5 and 9) or simian (AAV4, 7, 8, 10, 11 and 13) tissue samples [16]. Phylogenetic analyses suggest that the three capsid proteins, particularly their shared external domain that is exposed on the

assembled capsid, have undergone extensive evolutionary selection to accommodate receptor binding and immune evasion functions [17]. Despite this continuous pressure to evade immune recognition, seroprevalence studies have shown that a significant proportion of the human population possesses antibodies to AAV, with rates ranging from 30% to 80% varying by age, disease, region and serotype [18-23]. Although long assumed to be of non-pathogenic nature and even being capable of ameliorating or inhibiting other viruses with oncogenic properties [24-27], AAVs have recently been implicated with cases of severe acute hepatitis in pediatric patients [28-30].

1.1.2 AAV genome organization

Despite its compact ~4.7 kb genome, AAV exhibits a remarkably intricate structure, containing overlapping genes, multiple promoters, introns, and regulatory elements that finely orchestrate its replication, packaging, and capsid assembly [31-33]. The viral genome, as shown in Figure 1, is flanked on both sides by telomeres, the ITRs. These 145 bases long sequences form hairpin structures that serve as origin of replication (ori) and encapsidation signal but are also important for genome maintenance in the host cell [34-38]. The genome contains multiple promoter regions, the two major genes *rep* (replication/non-structural) and *cap* (capsid/structural) as well as a rudimentary polyadenylation (polyA) signal [33,39-41]. The *rep* gene encodes four non-structural proteins Rep40, Rep52, Rep68 and Rep78, which mainly serve as replication proteins [42]. The expression of the large Rep proteins (Rep68, Rep78) is driven by the p5 promoter, while the small Rep proteins (Rep40, Rep52) are expressed by the p19 promoter located within the *rep* gene [32,33,42]. The shorter isoforms (Rep40 and Rep68) are formed by splicing of an intron located in the 3‘region of the *rep* gene. Notably, the presence of a single large and small Rep protein is sufficient for successful virus propagation *in vitro* [43,44]. The large Rep proteins are distinguished from the small Rep proteins by a unique N-terminal region containing a sequence-specific DNA-binding domain (DBD) and an oligomerization domain [38,45-47]. The C-terminal region contains a helicase with an ATPase domain and is present in all four Rep proteins [48].

The expression of the *cap* gene, encoding the capsid proteins VP1-3, is regulated by the p40 promoter within the *rep* gene. Alternative splicing results in two distinct RNA isoforms, depending on which splice acceptor site is used. The longer isoform arises when splicing occurs at the major splice acceptor site and encodes only VP1, which is initiated with the conventional AUG start codon [49,50]. The shorter isoform, generated by the minor splice acceptor site, leads

to the production of VP2 (initiated by an alternative ACG start codon) and VP3 (initiated by a conventional AUG start codon) [49,50]. The N-terminal extension of VP1, also called VP1 unique region (VP1u), includes a phospholipase A2 (PLA2) domain that is essential for endosomal escape [51]. Additionally, both the VP1u region and the common VP1/VP2 region feature several basic regions that serve as nuclear localization signals [52]. Based on the differential splicing and start codon usage, VP1:VP2:VP3 are expressed at a ratio of approximately 1:1:10, matching their ratio in the assembled capsid [53-55]. The capsid itself consists of 60 subunits arranged in a T=1 icosahedral structure with a diameter of around 260 Å [56,57].The *cap* gene also harbors two additional open reading frames in a +1 frameshift, encoding the assembly-activating protein (AAP) and the membrane-associated accessory protein (MAAP) as well as a hypothesized X gene (excluded from Figure 1) [58-60]. AAP promotes the assembly of the empty capsid [58] and is essential for capsid assembly of specific serotypes [61]. MAAP has been recently identified as a viral egress factor and also serves as regulator of replication [62]. Although AAV is classified as non-enveloped virus, AAVs can be embedded within exosomes, which is promoted by MAAP [59], highlighting the diversity hidden within this small genome.

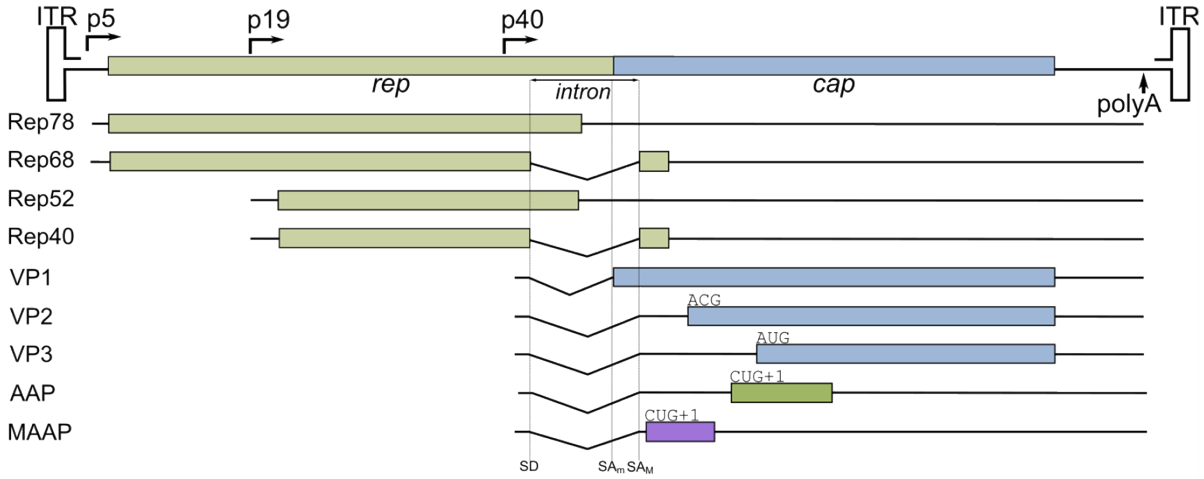


Figure 1: Genome organization of AAV. (Top) The AAV genome is flanked by ITRs and comprises two genes, *rep* and *cap*, which are regulated by three promoter regions (p5, p19 and p40) and the centrally located intron. Viral gene expression is terminated by a small poly-adenylation (polyA) site, leading to the generation of the indicated gene products (bottom) depending on the usage of the splice site. Usage of alternative start codons as well as a shift in the reading frame is indicated for various gene products. SD = splice donor, SAm = minor splice acceptor, SAM = major splice acceptor.

1.1.3. The AAV life cycle

The life cycle of AAV is a complex multistep process, as shown in Figure 2, beginning with viral attachment to the host cell and culminating in productive infection or latent genome persistence. For attachment, it is assumed that capsid specific binding to surface moieties, often

termed primary receptor or attachment factor, leads to the accumulation of viral particles on the cell surface [63]. These attachment factors include a variety of glycans, such as heparan sulfate proteoglycans (AAV2, 3B, 6, 13) [64-66], sialic acid (AAV1, 4, 5, 6) [67-69] and galactose (AAV9) [70]. Surface proteins may also serve as attachment factors, such as the laminin receptor (AAV2, 3, 8, 9) [71], fibroblast growth factor receptor 1 (AAV2) [72] and integrin (AAV2) [73]. Primary attachment is then followed by binding to co-receptors, such as the AAV receptor (AAVR) for a variety of serotypes (AAV1, 2, 3B, 5, 6, 8, 9) [74] or a currently unknown receptor for the AAV4 clade (AAV4, 11, 12, rh32.33) [75]. The binding specificity largely dictates the tissue tropism observed for the different AAV isolates, and synthetic capsid variants can switch both attachment factor and co-receptors [76].

AAV particles are internalized by endocytosis via diverse mechanisms. Clathrin-mediated endocytosis was described [77,78] as well as the use of Clathrin-independent carriers (CLICs) via glycosylphosphatidylinositol-anchored protein-enriched early endosomal compartments (GEECs) [79], micropinocytosis [80] and potentially also caveolin-dependent endocytosis [81]. It is suggested that the route used for endocytosis depends on the cell type, not the serotype [80], although three different uptake mechanisms were described independently in the same cell line [77,78,80].

Once the endosome has formed, AAV must escape the endosomal compartment to avoid degradation in lysosomes. Also here, different routes were described. Acidification of the endosome during maturation leads to a conformational change in the viral capsid, exposing the VP1u region and thereby the PLA2 domain [82,83]. The latter then mediates lipolytic pore formation, enabling entry of the AAV particles into the cytoplasm [84]. For the CLIC/GEEC-mediated uptake, a translocation of the AAV particles with an unfolded VP1u region to the *trans*-Golgi network and the Golgi apparatus was described [79,85]. It is hypothesized that the escape from the Golgi is then again mediated by the PLA2 domain, which is highly active at the Ca^{2+} concentration present in the Golgi [85].

Upon reaching the cytoplasm, the virus utilizes components of the cytoskeleton for transport [86], while avoiding the ubiquitin-proteasome system [87]. This is then followed by nuclear import. The basic sequences in the VP1u and VP1/2 region serve as a nuclear localization signal [52], initiating an accumulation of AAV particles at the nucleus and in the perinuclear space followed by interaction with the nuclear pore complex and transfer into the nucleoplasm [88,89]. When reaching the nuclear lumen, the ssDNA genome is ejected from the capsid via an unresolved mechanism [90]. In a latent infection in the absence of helper virus components, the

viral genome is converted to a dsDNA [89,91,92]. This is mainly achieved by second-strand synthesis, albeit annealing of two complementary AAV ssDNA genomes may also occur [91,93,94]. The AAV genomes are then converted into circular episomes, enabling stable long-term persistence [95]. Random integration into the host genome may happen both from linear ss/dsDNA and episomal AAV genome structures[96], while AAV Rep can promote the integration of the AAV genome into a genomic safe-harbor region [32,97-101]. Expression of the viral genes from integrated or episomal vector genomes leads to production of the viral Rep proteins which are then transferred back into the nucleus for viral replication. The capsid proteins of most serotypes are imported into the nucleus and assembled inside the nucleoli with the help of AAP [58], forming empty virions [102,103]. For some serotypes (AAV3, 4, 5, 9, 11, rh32.33), AAP has been described as non-essential for capsid assembly and assembly is independent of nucleolar localization[61,104].

During replicative infection in the presence of helper virus proteins (and RNAs), episome formation is reduced and the formation of replication intermediates is promoted instead [105]. Two modes of replication, rolling circle replication (RCR) and rolling hairpin replication (RHR), were described, depending on the helper virus [106]. RHR takes place in presence of adenovirus co-infection and relies on second-strand synthesis from the hairpin, nicking by Rep as well as the induction of another second-strand synthesis step [102,107-110]. The resulting DNA fragments serve as template for further replication, which is assumed to be coupled with the packaging into the assembled capsid particle [103].

The AAV genome is then transferred into the empty particle via one of the pores on the capsid surface, with the help of the small Rep helicases [111]. It is assumed that the 3’ end of the viral genome is transferred first into the particle. The fully assembled capsid is finally transported into the cytoplasm, with MAAP supporting viral secretion into the extracellular space [112] or shedding via exosomes [59].

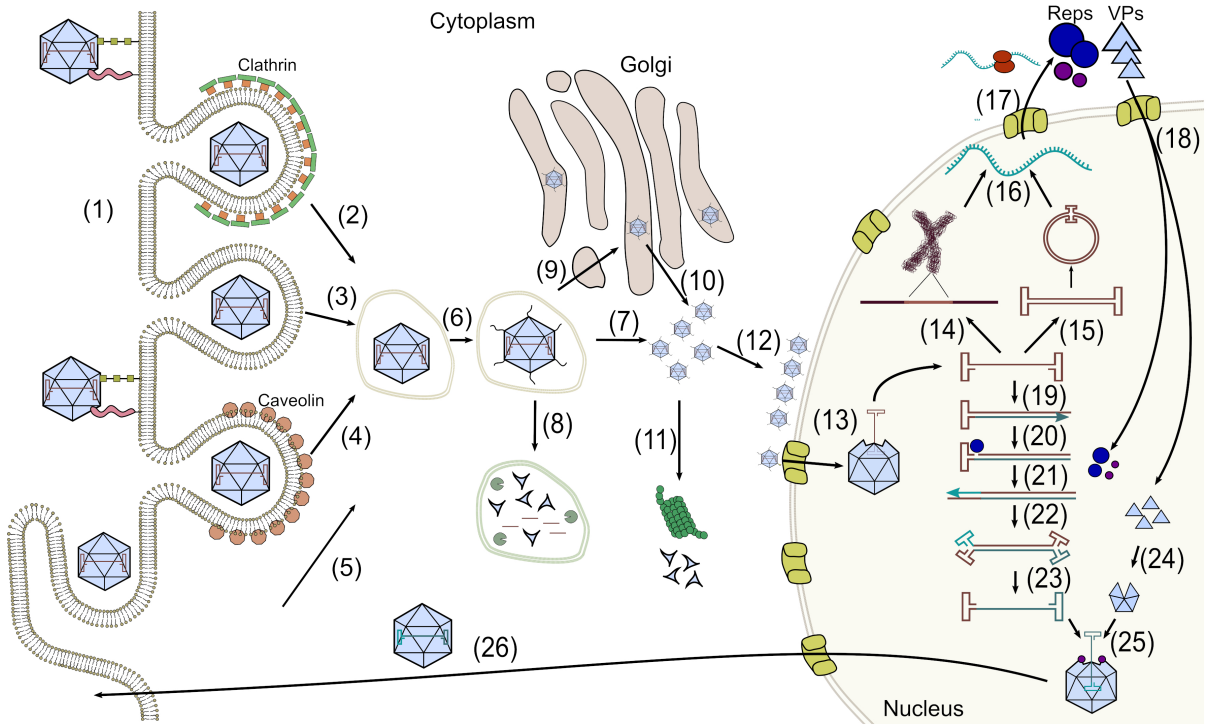


Figure 2: The AAV life cycle. (1) AAVs bind to the cell-surface glycans and receptors. Uptake can take place via (2) Clathrin-, (3) CLIC/GEEC-, (4) Caveolin- mediated endocytosis, as well as (5) macropinocytosis. (6) Endosome maturation leads to unfolding of the VP1u region from the capsid. The AAV then (7) escapes directly into the cytoplasm, (8) gets degraded in the lysosom or (9) is transported into the Golgi apparatus (10) from which it escapes into the cytoplasm. There the virus can (11) be degraded by the proteasomes or (12) accumulate at the nuclear pores before (13) being transported into the nucleus via the pores upon which viral genome is ejected from the capsid. The viral genome in the nucleus can either be (14) integrated into the host genome or (15) upon second strand synthesis be converted to a circular episome. (16) Both integrated and circular genomes allow gene expression and (17) translation in the cytoplasm generating Rep and Cap (18) which are then imported into the nucleus. During replicative infection, the RHR model proposes that (19) the second strand is synthesized, (20) enabling binding and nicking of the large Rep proteins and (21) another second strand synthesis step can take place to (22) generate two annealed complete vector genomes. Upon (23) strand separation and the (24) assembly of the capsid, (25) the genome can be pumped into the capsid and (26) exported out of the nucleus and secreted.

1.2 Significance of AAV for gene therapy

The consideration of AAVs as mostly apathogenic and the various tropisms of the diverse natural AAV variants led to major efforts to harness the versatility of AAV for gene therapy.

1.2.1 From virus to vector for gene therapy

The first and pivotal step to make AAV accessible for gene therapy was to clone a complete wild-type AAV (wtAAV) genome into a bacterial plasmid, enabling the production of infectious AAV particles in cells infected with adenovirus as helper [113,114]. Deletions in various regions of the AAV genome could be complemented by wtAAV sequences except for the ITRs,

which were found to be essential for genome replication and packaging [115,116]. The ability to partially replace viral genes with transgenes also indicated that AAV could serve as a tool for therapeutic gene delivery [116]. The discovery that a completely "gutted" vector genome supports recombinant AAV (rAAV) production proved that only the flanking ITR sequences need to be retained, while the viral genes can be supplied from a second plasmid *in trans* [36,117]. Later it could be shown that pseudotyping of rAAV enables the packaging of vector genomes with AAV2 ITRs into capsids of other natural or synthetic serotypes [118,119] and *vice versa*[120,121]. These rAAVs enable the delivery of a transgene cassette with a size of up to 5 kb DNA, with larger sizes reducing genome integrity and packaging efficiency [122,123]. Initially, vector production still relied on supplementation of the adenoviral helper genes via adenovirus co-infection of the producer cells [36,117]. Heat inactivation was used to remove adenovirus from the vector product, however, a complete removal of adenovirus could not be ensured [117]. The use of replication-defective helper virus can reduce the presence of adenovirus to non-detectable levels, but a small risk of carry-over into the product persists [124].

Ultimately, the essential helper functions encoded within the adenovirus genome were determined (E1A, E1B, E2A, E4 and VA RNA) [125-134]. These could then be supplied *in trans* without the risk of adenovirus contaminations in the final product. This led to the development of the still most commonly employed method for rAAV vector production, *i.e.*, triple-transfection of Hek293(T) cells [135-137]. The Hek239 cell line was established by transformation with sheared adenovirus 5 DNA and thus already expresses E1A and E1B, facilitating the rAAV production process [138]. Co-transfection with plasmids carrying the ITR-flanked transgene, the AAV helper genes (*rep* and *cap*) and the Adenoviral helper components E2A, E4 and VA RNA nowadays enables helper-virus free rAAV production [135-137]. Nonetheless, optimizations of this system are still ongoing to date [139,140].

A variety of strategies for rAAV production have been developed over the years, ranging from simplified transfection schemes that combine adenoviral- and AAV-helper genes into a single plasmid [121,136], to complex cell lines with stably integrated ITR-flanked transgenes, AAV and/or Adenoviral helper genes including systems with inducible gene expression [139,141-143]. Alternative production platforms lacking adenoviral genes have been established as well. These production systems include yeast [144-146], where *rep*/*cap* and the ITR-flanked transgene are transformed as plasmids, as well as Sf9 insect cells, where *rep*/*cap* and ITR-flanked transgene sequences are supplied via baculovirus infection [147]. While the yeast production platform has never been adopted due to its inadvertent formation of large amounts of

subgenomic particles [148], the Sf9-based baculovirus AAV production system is used for the production of U.S. Food and Drug administration (FDA) and European Medicines Agency (EMA) approved therapeutics (Roctavian and Hemgenix) [149].

1.2.2 AAV-based vectors for gene therapy

Before AAV gene therapy could reach the market, several hurdles needed to be overcome. The first *in vivo* use of rAAV for gene transfer was performed in rabbits, where a cystic fibrosis transmembrane conductance regulator (*CFTR*) transgene was delivered in an AAV2 capsid via a fiberoptic bronchoscope [150]. Based on the facts that *CFTR* is defective in patients with cystic fibrosis, and that *CFTR* RNA and protein was detected for 6 months after administration in the animal model [150], a phase I clinical trial was quickly initiated [151]. In this trial, no vector- related serious adverse events were observed, but the level of gene transfer was too low to achieve any measurable therapeutic effect [152,153]. Multiple factors were identified as being responsible for this outcome, including the low dose (the highest used dose was only $2x10^{12}$ vector genomes (vg) per patient [152]), as well as the used capsid serotype, which is subpar in lung tissue [65]. The first clinical trial with a rAAV that later resulted in a product approved by the EMA (Glybera) was performed using a much higher dose of up to $1x10^{12}$ vg/kg, along with immune suppression with cyclosporine A and mycophenolate mofetil [154]. This highlights two of the many challenges with AAV gene therapy, namely, efficacy and immune responses. The example of Glybera also highlights the economic challenges of AAV gene therapy, *i.e.*, high costs combined with a low demand in the case of rare diseases [155,156]. Due to these concerns, after treatment of a single patient with the approved drug, its market authorization was not renewed in 2017 [156]. Despite the high costs of AAV gene therapy, with more recent drugs reaching price tags of up to 3.5 million U.S. dollars per patient, the therapeutic benefits can be substantial, explaining the total of eight previously and currently approved EMA- or FDA-approved AAV based drugs (Table 1).

Table 1: Current and previously approved AAV gene therapy products. Shown are the names, capsid serotype (* = capsid with mutations to improve transduction and immune evasion), indication, packaged transgene, year of approval, route of administration and dose of currently or previously approved AAV gene therapy products [157,158].

Name	Capsid Serotype	Indication	Cargo	Approval	Delivery	Dose
Glybera	AAV1	Familial lipoprotein lipase deficiency (LPLD)	LPL^{S447X}	2012-2017 (EMA)	Intramuscular injection	$1x10^{12}$ vg/kg
Luxturna	AAV2	Biallelic RPE65 mutation retinal dystrophy	RPE65	2017 (FDA) 2018 (EMA)	Subretinal injection	$1.5x10^{11}$ vg/eye
Zolgensma	AAV9	Spinal muscular atrophy (SMA)	SMN1	2019 (FDA) 2020 (EMA)	Intravenous infusion	1.1×10^{14} vg/kg
Upstaza/ Kebilidi	AAV2	AADC deficiency	AADC	2022 (EMA) 2024 (FDA)	Intraputaminal infusion	0.9×10^{11} vg/ putamen
Roctavian	AAV5*	Hemophilia A	Factor VIII	2022 (EMA) 2023 (FDA)	Intravenous infusion	6×10^{13} vg/kg
Hemgenix	AAV5	Hemophilia B	Factor IX	2022 (FDA) 2023 (EMA)	Intravenous infusion	$2x10^{13}$ vg/kg
Elevidys	AAV rh74	Duchenne muscular dystrophy	µ-dystrophin	2023 (FDA)	Intravenous infusion	$1.33x10^{14}$ vg/kg (< 70 kg)
Beqvez/ Durveqtix	AAV rh74*	Hemophilia B	Factor IX	2024 (FDA) 2024 (EMA)	Intravenous infusion	$5x10^{11}$ vg/kg

Despite these major breakthroughs, the currently approved AAV-based gene therapeutics are still mostly based on natural serotype capsids requiring very high doses and are often administered along with elongated immunosuppression [159]. Patients with pre-existing antibodies against the used AAV capsid are often excluded from treatment, limiting accessibility [160]. Additionally, in recent years, AAV gene therapy has experienced setbacks. Inflammation, even after local administration in immune-privileged organs, is a common side effect, as exemplified by reports during ocular gene therapy [161]. Several vector-related deaths in different clinical trials were also observed, after patients were exposed to high doses of AAV vectors [162-165]. In one study treating X-linked myotubular myopathy, hepatotoxicity was observed in 3/17 boys at a dose of $3x10^{14}$ vg/kg, leading to two deaths, likely due to a pre-existing hepatobiliary disease [162,165]. A 27-year-old patient with Duchenne's muscular dystrophy (DMD) died after injection of $1x10^{14}$ vg/kg from acute respiratory distress syndrome and cardiac arrest, likely triggered by the innate immune response [163]. In a recent clinical trial for Rett syndrome, delivery of $3x10^{15}$ vg to the cerebrospinal fluid triggered a systemic immune reaction and a subsequent death [164]. The latter event emphasizes that even when administered in a supposedly immune-privileged organ, rAAV can trigger a systemic immune response with lethal effects.

1.2.3 Challenges of AAV-based gene therapy

Inefficient delivery or function of AAV vectors has been a continuous challenge, ranging from the first clinical trial where the used doses could not improve patient outcome, to current trials with deaths due to excessive dosing. In the following chapter, a variety of strategies to overcome the issues of efficacy, safety and limited transgene size will be described, along with the controversy on rAAV integration and its potentially oncogenic effects.

A major and predominantly pursued strategy to improve rAAV gene therapy is the development of novel capsid variants with improved specificity, efficacy as well as lower immunogenicity, harnessing methods ranging from rational design [166,167] and directed evolution [168] to *in silico* design and /machine learning [169]. Rational design includes methods such as the attachment of specific ligands to the surface of the capsid, often by chemical conjugation to non-natural amino acids [170-172] or natural amino acid residues [173,174] incorporated in the AAV capsid, or by integrating functional domains or adapter domains directly into surface exposed regions of the capsid protein [175-179]. Examples of commonly used ligands include carbohydrates [180-182], DNA aptamers [174,183] and anti- or nanobody domains [184-187]. DNA domain swapping replaces specific domains of one capsid serotype with domains found in other serotypes. While this is usually done on a very limited scale with few capsid variants, it is also possible to perform random recombination by DNA family shuffling [188,189]. There, the DNA sequences of different viral isolates are enzymatically fragmented and reassembled into a full-length capsid sequence. The incorporation of short peptide sequences within surface-exposed loops the capsid has also proved to significantly impact tropism and improve efficacy [190]. Many of these methods are not mutually exclusive and can be combined to further enhance the characteristics of the synthetic capsid [189].

The use of optimized transgene cassettes can also benefit the therapeutic efficacy of rAAV at a given dose, as exemplified by the success with an optimized version of a transgene in an approved drug. Zolgensma, one of the approved rAAV drugs, uses the human cDNA of the *SMN1* gene. However, recent research has shown that the SMN^{K186R} mutant is more resistant to adverse ubiquitination and therefore proteasomal degradation [191]. In animals, this led to a much higher therapeutic efficacy at a similar dose [192]. Also, the choice of the promoter sequence is of high importance. Zolgensma uses a synthetic constitutive promoter based on the cytomegalovirus (CMV) enhancer and the chicken-β-actin promoter [193]. While this may ensure strong expression in the target cells, the high promoter activity can increase inadvertent silencing by epigenetic modifications [194]. This is further emphasized by observations of an

elevated immune response after AAV gene therapy with ubiquitous promoters compared to tissue-specific promoters [195]. The incorporation of a variety of regulatory elements has also shown potential, such as enhancers for cell type-specific expression [196], immunomodulatory oligonucleotide sequences for reduced innate immune system activation [197], or scaffold/matrix attachment regions (S/MARs), which induce replication of the viral episomes and thereby enable stable gene expression over extended times in dividing cells [198].

The size of the packaged transgene cassette, especially when using large transgenes or the use of extensive regulatory elements, can also limit the application of rAAV gene therapy. With a maximum capacity of 5 kb [123], including transgene, promoter and polyA signal, gene replacement with the original coding sequence is impossible for some diseases. One example is DMD, where a mutation in the largest gene in the human genome, the *dystrophin* gene with a coding sequence of 11 kb, leads to a severe muscular dystrophy [199]. The now approved drug delandistrogene moxeparvovec (Elevidys) delivers a micro-*dystrophin* gene, *i.e.*, a miniaturized version containing only the key functional domains, and was shown to reduce motor function decline [200]. Alas, minigenes can have limited potency as well as unintended side effects [201]. Alternative strategies to reassemble full-length genes have been harnessed, which rely on the delivery of the transgene via two different vectors followed by reassembly of (i) the two vector genomes by recombination [202], (ii) the transcribed RNA by *trans*-splicing [203,204], or (iii) the translated protein by *trans*-splicing [205].

AAV integration is another commonly observed phenomenon, which can aid long-term transgene expression but also be a safety risk. For wtAAV, a Rep-dependent site-specific integration of the wtAAV genome into a so-called genomic safe harbor region, in humans mainly the AAV integration site 1 (AAVS1) locus in chromosome 19, has initially indicated a safe usage of AAVs [100]. However, recent reports indicate a more random wtAAV integration than initially anticipated [206]. Despite initial reports of AAV-mediated suppression of tumor formation by oncogenic viruses [207-209], one report suggested wtAAV2 genome integration in proximity of cancer driver genes in humans, which has implicated AAV in the onset of hepatocellular carcinoma (HCC) [210]. But these claims are highly controversial since AAV integration may just represent a passenger mutation [211,212]. The situation regarding rAAV is similarly complex, as a variety of conflicting reports on rAAV integration and the occurrence of HCC have been published to date. The first description of HCC after rAAV exposure suggested that rAAV integration was not the driver of tumor formation [213]. Moreover, the initial analysis of rAAV integration sites, both in human cell lines and mouse liver, did not

indicate a major risk for cancer [214,215]. In contrast, a later study detected HCC formation in 30-50% of mice treated with rAAV (depending on the promoter) compared to 8% in untreated mice [216]. The rAAV integration in HCC was specific to the *Rian* locus, a locus encoding several small nucleolar RNAs (snoRNA) and microRNAs (miRNA), leading to their overexpression in the tumor. A more comprehensive study with a long-term follow-up (18-22 months) using different serotypes, transgene cassettes and doses indicated that the HCC incidence is dose-dependent and that it was differently affected by enhancer/promoter combinations [217]. In other animal models, the exposure to rAAV was not associated with HCC occurrence, as shown in cats [218], dogs [219] and primates [96]. Instead, the integration was associated with long-term gene expression, ensuring therapeutic efficacy over several years [96,219]. Only a single occurrence of HCC in clinical trials with liver-targeted AAV has so far been described [220]. A patient with hemophilia B was treated with $2x10^{13}$ vg/kg of the now approved drug Hemgenix. One year after the treatment, the patient developed HCC, without detectable clonal expansion of integrated vector genomes in the tumor tissue. It was concluded that the patient's advanced age (69 years) and a previous history of hepatitis virus infection were likely driving factors for the occurrence of HCC.

Since vector integration can be beneficial for long-term transgene expression in dividing cells [96,221], approaches have been developed to favor genomic integration[222-224]. Alternatively, to alleviate the risk associated with random integration events, approaches to artificially increase episome formation have been examined as well [225]. So far, only the intervention in pathways responding to DNA damage has been shown to affect the rate of episome formation and integration [222-225], but alternative ITR sequences could potentially achieve similar results.

1.3 AAV-ITRs – the last relic of nature in rAAV vector genomes

The ITRs are the only components in rAAV vector genomes that are also present in the wild-type virus. These sequences are crucial for vector genome replication, packaging, genome processing during transduction and thereby transgene expression, underscoring their indispensable role in rAAV function. In the following chapter, their origin, structure, and function will be explored in greater detail.

1.3.1 Parvoviral telomeres – beyond AAV ITRs

In nature, ITRs in the *Parvoviridae* family take very different shapes and forms. Some genera (*e.g.*, *Ave-, Dependo-* and *Erythroparvovirus* as well as *Ambi-* and *Iteradensovirus*) contain

genomes with two identical ITR sequences on both ends of the genome, and are thus classified as homotelomeric [226]. Homotelomeric viruses appear to be less common, but may be better researched, with AAV and the causative agent of fifth disease in children, parvovirus B19, being the most prominent members of this group [227]. Heterotelomeric viruses, with two ITRs distinct in sequence and structure, are the only known type in some genera (*e.g.*, *Amdo-* , *Boca-* and *Protoparvovirus*, *Hepan-* and *Brevidensovirus*) [226]. The main consequence of this distinction is the packaging of genomes with single strand polarity, assumed to result from inefficient nicking in one of the two distinct terminal repeats (TR) and the absence of a packaging signal [228].

The parvoviral TR sequences remain the most elusive component of the viral genomes with many unresolved sequences, *i.e.,* of AAV8-12. The known parvoviral TR sequences vary widely in composition, structure, and size. The GC content in the *Parvovirinae* subfamily is high with 60% while in the *Densovirinae* subfamily it is around 45%[226]. AAV2 ITRs have a length of 145 nt and possess a GC content of 69.7%, which is one of the highest detected in the virus family and which thus contributes to ITR stability [226]. Because the AAV ITR is very G-rich (39% of all nucleotides), G-quadruplex (G4) formation was suggested to facilitate the interaction of cellular proteins with the ITR, thereby reducing genome replication [229]. The number of G4 sequences predicted in other parvoviral TRs varies widely, with some ITRs carrying more than 200 G4 sites [226], which could indicate some unexplored regulatory function [230,231].

The secondary structure is probably the most prominent feature of the telomeric region. The simplest known ITR structures are the linear hairpin stems. These linear structures are often interrupted by short “bubbles” generated by unpaired nucleotide insertions in one strand and internal loops of unpaired nucleotides in both strands [232], such as in Parvovirus B19 (Figure 3A). In the heterotelomeric minute virus of mice (MVM), these unpaired nucleotides were found to be required for efficient viral DNA replication, generating an evolutionary advantage over ITRs without DNA “bubbles” [109].

The length of the hairpin region can also vary immensely, with some hairpins being extremely long and consisting of 740 paired nucleotides (Figure 3 B). In contrast, extremely short hairpins that seem to lack replication origin sequences in their hairpin region have been described as well [233]. One such example is the telomeric region of the *Culex pipiens* densovirus (Figure 3C) with a 64 nt hairpin (with only 42 pairing nucleotides) at the genomic termini, while the rest of the heterotelomer is single-stranded [233]. Although this theoretically suffices to prime for

DNA replication, it remains unclear whether this was a sequencing artifact or whether the hairpin is really this short [233]. In the meantime, similarly short ITR sequences have been described in other densoviruses, yet the replication mechanism for such short hairpins remains unresolved [233,234]. Most structures are relatively simple with short hairpins and small loops, but highly intricate structures can be found as well, such as in the homotelomeric Avevirus Chicken parvovirus (Figure 3D), which possesses multiple "bubbles", four hairpins as well as a large loop [235]. However, the function potentially associated with such complicated structures remains elusive.

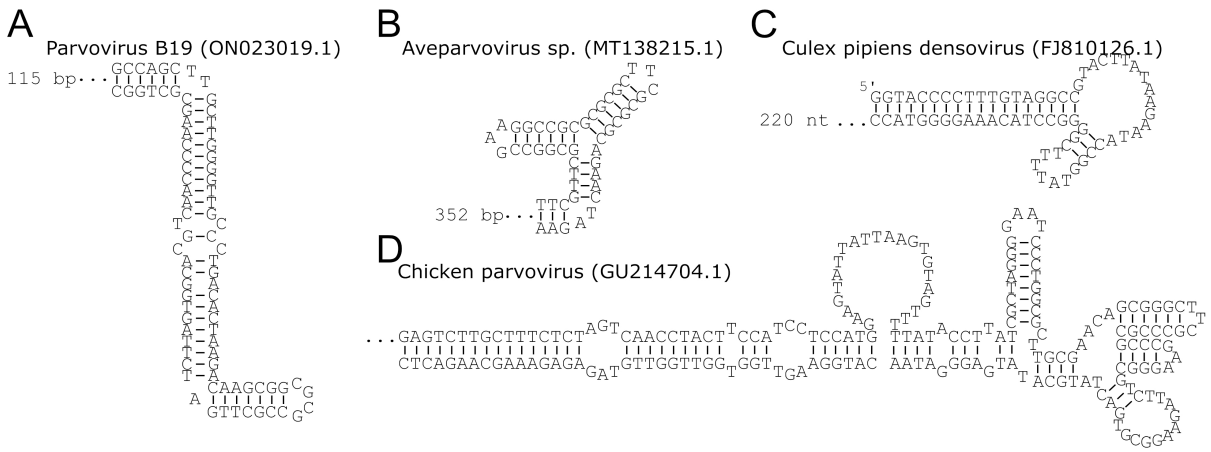


Figure 3: Diverse structures of parvoviral ITR sequences.A) Terminal part of the Parvovirus B19 ITR (Genbank: ON023019.1) hairpin with four mismatch regions (two bubbles and two bulges). **B)** Terminal region of the heterotelomeric Aveparvovirus sp. ITR (GenBank: MT138215.1). The depicted hairpin is attached to a 352 bp long stem, followed by a short region with three loops. **C)** Terminal hairpin in the *Culex pipiens* densovirus (GenBank: FJ810126.1) of only 64 nucleotides. **D)** Complete 206 nt Chicken parvovirus (GenBank: GU214704.1) ITR with complex secondary structure. Secondary structures were predicted with mfold [236] and adapted for improved visualization.

1.3.2 The elements of the AAV ITRs

In contrast to other parvoviral ITRs, the AAV ITR possesses a relatively simple secondary structure. Due to the use of AAVs in gene therapy, these ITRs are the best characterized parvoviral ITRs. For historic reasons, mainly the AAV2 ITR is used for rAAV production and thus has been studied extensively [36,117], albeit the more distantly related AAV5 ITR has been investigated as well [237]. The ITRs of AAV1 and AAV3-7 are known but have not been widely used or examined [120,121]. The AAV ITRs contain six sets of palindromic sequences (A and A', B and B', C and C'), which enable the formation of a three-way T-shaped DNA junction with two 9 nt hairpins [34], as exemplified with the AAV2 3' ITR in Figure 4A. There are inconsistencies in the literature regarding the nomenclature for the ITR region, with different names often used for the same region [34,238]. After the first sequencing the ITRs, Lusby *et al*.

proposed naming the palindromic repeats A-B-C-D-B'-A'-E and the two terminal hairpins C and D [239], however, this nomenclature was later simplified to A-B-C-A'-D [33]. The AAV2 ITR has a length of 145 nt, of which the first 125 nt form the hairpin sequence, leaving the last 20 nt unpaired, which is commonly referred to as D-sequence [34]. Another aspect where terminology is inconsistent is the configuration of the ITR [34,240], thus the more commonly used terminology from Lusby *et al*. will be used here [34,241,242]. The ITR exists in two configurations due to inversion of the hairpins during replication, resulting from hairpin transfer following the model proposed for the replication of eukaryotic telomeres [243]. The two configurations are termed "flip" and "flop", with a flip ITR being defined as ITR with the B-B' hairpin arm closest to the 3' end of the viral genome and the flop ITR with the C-C' hairpin closest to the 3' end [34,244]. Within the A- and D- sequences, the RBE and the terminal resolution site (trs) are embedded [245-247]. During AAV genome replication, the helicase activity of the Rep proteins leads to a restructuring of the ITR region resulting in the formation of an additional stem-loop structure around the trs (Figure 4B), although this restructuring may not be essential for processing [93,248]. Then the RBE is bound by the N-terminal domain of the large Rep proteins and the DNA is nicked at the trs leading to a covalent attachment of the Rep protein to the free 5' end via a phosphotyrosine linkage generated in a transesterification reaction [107,245-247,249,250]. The nicking activity of Rep can be enhanced by an interaction with the RBE' region, located at the tip of the hairpin opposite of the trs [247].

The consensus ITR generated from the ITR sequences of AAV1-4 and 6-7 (Figure 4C) [238], shows a maintained T-shaped hairpin structure, despite differences in the sequence. This partially enables an interchangeable use of Rep proteins, from different serotypes [121,251]. Specifically, the ITRs of AAV1-4 and 6 can be used with AAV1-4 and 6 Rep for production, due to the high sequence similarity in the RBE and trs hairpin [121]. The wtITR5, with only 58% homology to the wtITR2 [252], can only mediate genome replication in presence of AAV5 Rep [121]. This incompatibility is based on the more divergent RBE and trs sequences as well as a distinct structure with a 11 nt spacer between the RBE and trs hairpin, compared to the single-nucleotide spacer in the other known ITRs [252].

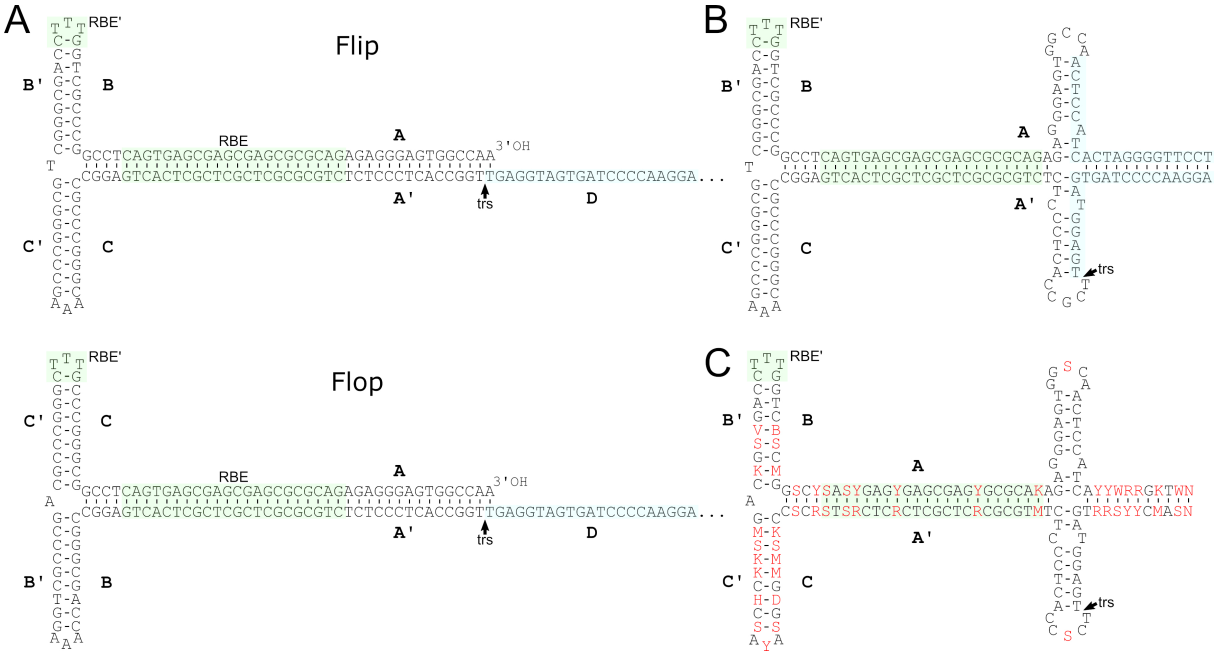


Figure 4: The AAV ITR structure and its elements. A) AAV2 3' ITR with T-shape formed by the repetitive palindromic occurrence of sequences in a D-A'-C-C'-B-B'-A pattern in the flip configuration and D-A'-B-B'-C-C'-A in flop configuration. Green nucleotides represent the RBE and RBE', and blue nucleotides representing the D-sequence. **B)** A Rep-induced conformational change in the wtITR2 leads to formation of a stem-loop with the trs located near the tip of the hairpin. **C)** Consensus ITR sequence generated from the ITR sequences of the serotypes AAV1-4 and 6-7, adapted from Earley *et al.*, 2020 [238] with red nucleotides shown in IUPAC code (Y = C or T, R = A or G, S = G or C, W = A or T, K = G or T, M = A or C, B = G or T or C, V = G or C or A, and N = any nucleotide).

1.3.3 ITR sequences in vector plasmids for AAV vector productions

As already indicated in the chapter describing the major elements of the ITR, the AAV ITRs serve as ori and also aid in the packaging of the genome. It is thus very surprising that efforts to engineer and improve AAV ITRs have remained scarce since the original cloning of the AAV2 wild-type genome into plasmids [113,114] and the discovery that removal of the viral genes enables the packaging of transgene cassettes [35,36,117]. Initially, vector plasmids contained full length ITRs adjacent to a GC-polynucleotide, severely impacting plasmid stability [239]. To improve plasmid handling, the GC-repeats were removed and the A-sequence of the ITR was truncated reducing the length of the hairpin to 95 nt (Figure 5A) [35]. This plasmid, termed pSub201 by Samulski *et al.* in 1987 [35] and its derivatives such as pSSV9, are still used as the basis for most current AAV vector plasmids. Interestingly, this plasmid also still contains 46 nt of the ITR adjacent region from the viral genome, which functions in rAAV as liver-specific enhancer with a binding site for the hepatic transcription factor HNF1-α [253]. Despite the hairpin length reduction to only 95 nt, a recent study found that the

ITR integrity in AAV vector plasmids remains a major issue [254]. This was emphasized by the observation that 69 out of 123 vector plasmids obtained from the plasmid repository Addgene contained unintended deletions in the ITRs [254], which could severely impact vector productivity and function [255]. Nevertheless, a rarely used plasmid setup uses a single extended ITR sequence of 165 nt in the vector plasmid, which was shown sufficient for rAAV production [93,94,256].

Several alternative vector plasmids and production strategies have been designed to overcome limitations that rely on the conventional plasmid design. To overcome the secondary strand synthesis as limiting factor during transduction of ssDNA AAV (ssAAV) vectors [91,92], a vector plasmid has been developed which enables the packaging of dsDNA vector genomes in the capsid [257,258]. To this end, one of the ITRs is maintained intact while the other contains a mutation that prevents replication initiation, usually a deletion of the trs region (Δtrs) [257,258] (Figure 5B). The rAAV DNA replication, initiated at the intact ITR, progresses to the Δtrs ITR, where the hairpin shape can induce a template switching process that redirects synthesis along the previously synthesized strand back towards the intact ITR. This generates a dsDNA or self-complementary (sc)AAV vector genome (Figure 5C). While scAAV show a 5- to 140-fold higher transduction efficiency [257,258], the major trade-off is the significant reduction of the packaging capacity to about 2.3 kb, *i.e.*, half of the wild-type genome, as well as an increased innate immune response [259]. Despite these limitations, scAAV have already reached the clinic, with Zolgensma being the first approved drug to employ a scAAV vector genome [260].

While the error-prone replication of rAAV can generate alternative vector genomes, it can also lead to packaging of unwanted sequences. Contaminating sequences stemming from plasmid backbones are considered a serious safety issue, as these prokaryotic sequences can persist *in vivo* after rAAV therapy [261]. For scAAV productions, the inefficient replication detouring at the Δtrs ITR hairpin can lead to contaminating prokaryotic sequences in more than 20% of the packaged genomes [262]. To address this issue, different production strategies have been developed, such as the use of minicircle vector plasmids. There, the prokaryotic plasmid backbone is flanked with recombination sites, allowing the plasmid to be recombined *in vitro* and the prokaryotic sequences to be selectively degraded [262] (Figure 5D). This can substantially reduce the packaging of plasmid backbone-derived sequences and improve the transduction, especially for scAAV preparations [262]. A similar strategy harnesses “doggybone” or covalently closed-end double-stranded DNA (cceDNA) templates. There, the ITR-flanked vector genome sequence is enveloped by phage protelomerase recognition sites (Figure 5E) [263]. Upon

processing by the protelomerase, which cleaves and covalently closes the DNA ends, two linear cceDNA fragments are obtained. Following the removal of the cceDNA containing the bacterial backbone, the linear cceDNA vector genome fragment with the two ITRs is then used for vector production. For scAAV generation with cceDNA, the Δtrs ITR can be replaced by a protelomerase recognition site. The hairpin formed after processing with protelomerase very efficiently generates scAAV genomes without contaminating plasmid backbone sequences. These engineering efforts highlight the diversity of vector plasmids and strategies that are already commonly used for rAAV production.

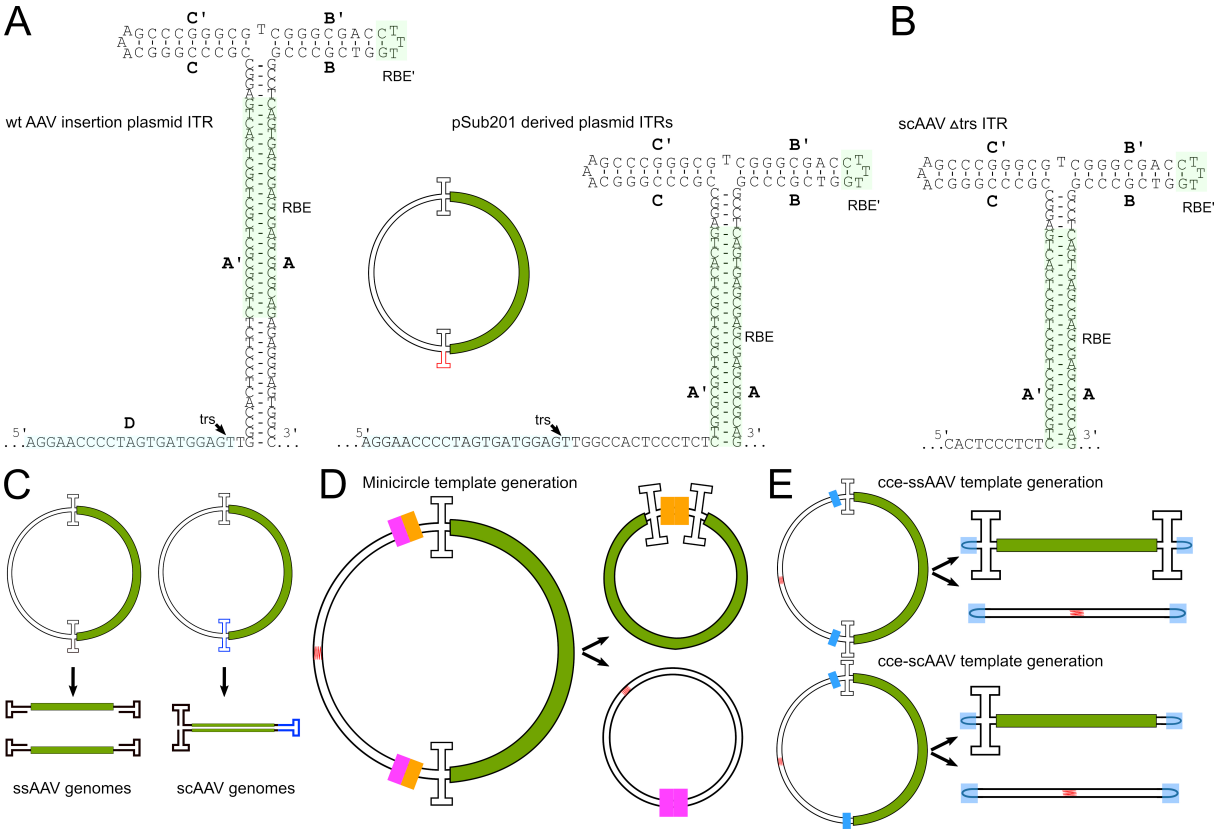


Figure 5: Common plasmid setups for AAV production. A) 3‘ITR sequence of the full length AAV genome directly inserted into a plasmid (left) and the pSub201/pSSV9 truncation of the A-sequence (right). The plasmid construct (center) with transgene region (green) and hairpin sequences is given as well. The 3’ hairpin sequence is highlighted in red. **B)** 3‘ITR sequence after truncation of the D-sequence and trs for scAAV generation. **C)** Schematic comparison of vector plasmids for ssAAV (left) and scAAV (right) vector genome generation. Top: production plasmid with ITRs. Bottom: vector genomes generated using the indicated plasmids. The transgene region is indicated in green and the Δtrs ITR in dark blue. **D)** Minicircle AAV production plasmid with recombination sites indicated in pink/orange and a single cutter cleavage site in the plasmid backbone indicated in red. Upon recombination, the plasmid backbone minicircle is degraded by restriction enzyme and exonuclease digest, and the ITR-containing minicircle can be used for vector production. **E)** cceDNA ssAAV vector production template (top) and cceDNA scAAV vector production template (bottom). The protelomerase binding domain is indicated in blue. Upon processing, the protelomerase covalently closes the generated DNA strands. To remove the cceDNA fragment with the bacterial backbone, a restriction enzyme cleavage site (red) is used for initial digest before exonuclease-mediated degradation.

1.3.4 ITRs role in rAAV genome plasmid resolution, replication and encapsidation

It has been recently shown that AAV genomes can replicate using two different mechanisms depending on the helper virus [106], namely, either via the well-described RHR mechanism in presence of adenovirus or a RCR mechanism in presence of herpesvirus. The RCR mechanism, which has also been suggested for the replication of other parvoviruses [264-266], is far less understood. The hypothesized mechanism for RCR postulates the formation of a circular episome with an intact ITR, from which a genome concatemer is replicated and then fragmented into unit-length AAV genomes by an unresolved mechanism [106].

Most research has focused on the RHR mechanism, as the adenoviral helper genes are most commonly used for vector production. The RHR model during super-infection with AAV and adenovirus was described in chapter 1.1.3 (see Figure 2), but resolution of vector genomes from plasmid templates is more complex and assumed to mimic the replication of integrated AAV genomes [262,267]. There are several additional steps required to generate replication intermediates resembling two annealed vector genomes from vector genomes (Figure 6A). The replication intermediates can finally serve as templates to initiate packaging into the pre-assembled capsids, with genome packaging assumed to be coupled with replication [103,268]. Accordingly, upon re-initiation of strand-displacement synthesis at one of the free 3' end termini, the displaced strand with the 5' attached Rep may guide the DNA to the capsid. There, translocation into the capsid via the pore at the five-fold symmetry axis is mediated by the helicase/ATPase domain of the small Rep proteins [111]. The current model assumes that during encapsidation, the 3' end of the genome is pumped into the capsid first [111], while the D-sequence of the ITR is presumed to serve as packaging signal [93,94,110,269].

While plasmid ITR mutation are frequent, due to the strong secondary structure and replication errors during plasmid propagation, they do not always prohibit vector production [254,270]. During early studies on AAV replication, it was discovered that the mutation or loss of one of the two ITRs could be corrected by the intact ITR [270]. The proposed model for this *cis*-repair mechanism includes the formation of a pan-handle DNA structure, enabling an unknown repair machinery to use the intact ITR as a repair template (Figure 6B).

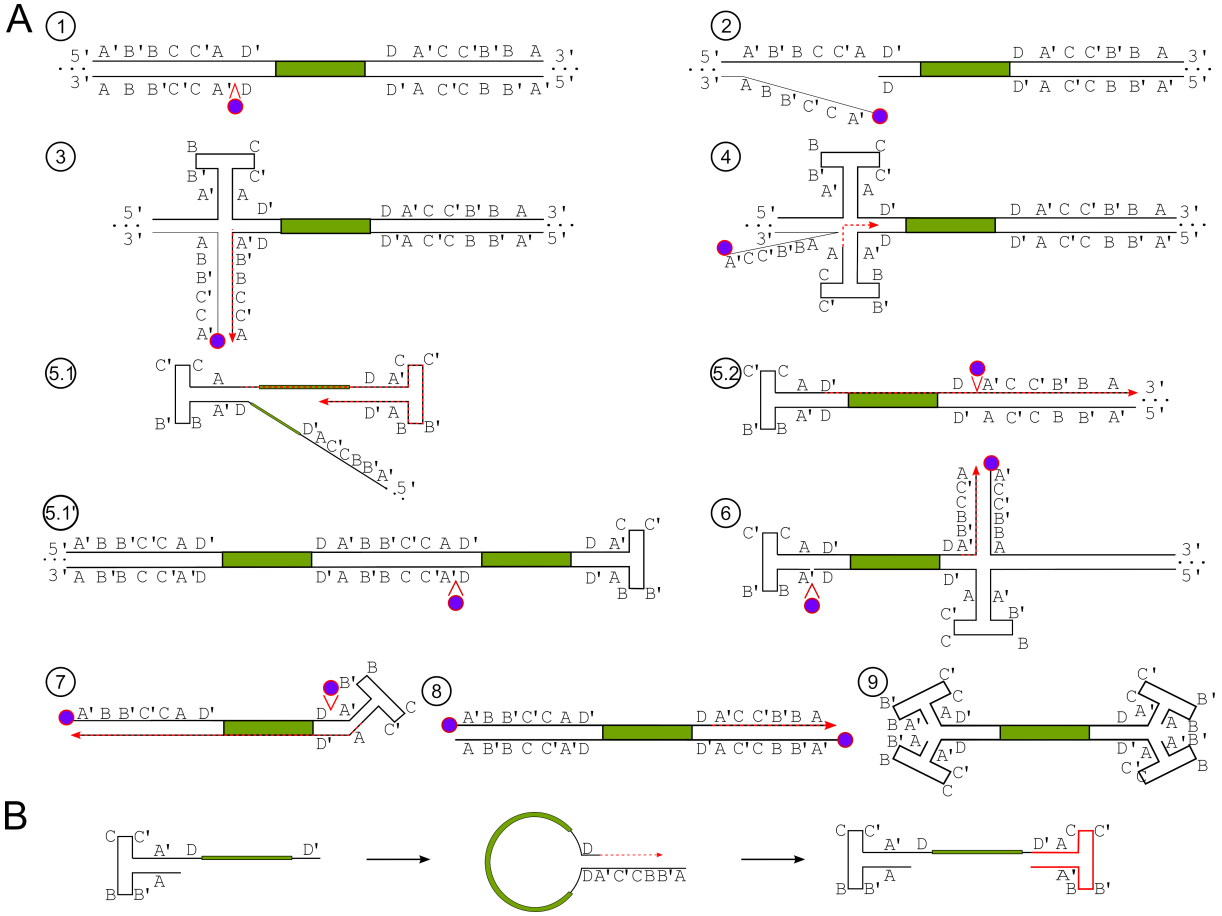


Figure 6: Proposed models for AAV genome resolution from plasmids by replication and *cis*-repair A) (1-2) Replication is initiated by binding and nicking of the large Rep proteins (purple circle with red outline) at the trs in the plasmid. (3) After formation of the hairpin in the complementary strand, the unfolded ITR sequence can serve as template for second-strand synthesis from the newly formed 3' end (red color dotted line). (4) The newly synthesized ITR sequence can now fold back and the ITR with its free 3' terminus serving as primer for second-strand synthesis. (5.1) The strand-displacing synthesis can then either be detoured at the ITR or (5.2) continue across the ITR. In the first case, (5.1') the strand-displacing second-strand synthesis continues back into the vector genome generating a concatemeric intermediate, which resembles the intermediate that is formed by read-through into the plasmid backbone (5.2) and can thus be processed similarly. (6) There, Rep binds in the ITR sequence reconstituted by the second-strand synthesis, and the unfolded ITR can again serve as template for synthesis. (7) After nicking by Rep at the terminal hairpin and refolding of the newly synthesized hairpin, the strand displacement synthesis resolves the vector genome from the plasmid backbone sequences. (8-9) After another resolution step by Rep, both ITRs are fully reconstituted and represent two annealed complete genomes. **B)** Proposed model for repair of damaged ITRs via a *cis*-repair mechanism. After formation of a pan-handle shape, the damaged or deleted ITR can be corrected via second-strand synthesis (red color dotted line), reconstituting the full AAV genome.

1.3.5 Additional functions of AAV ITRs

Besides their role in replication and encapsidation, ITRs play a significant role in a variety of processes during transduction, such as genome processing, transgene expression, immune response, and persistence via episome formation or integration, as described in the following.

1.3.5.1 DDR-mediated genome processing and toxicity

The induction of a DNA damage response (DDR) is essential for the replication of different parvoviruses [271-273] and mainly facilitated by the non-structural proteins. Similarly, in AAV, a DDR-dependent cell cycle arrest in S phase upon infection is mediated by the AAV Rep proteins [274,275]. Research on the effects of the DDR upon rAAV transduction has been initiated after the discovery that artificial induction of the DDR can improve transduction[276] and processing of rAAV genomes into stable episomes [277]. A variety of DDR-associated proteins, mainly associated with DNA double-strand break repair pathways[278], have since been linked with the processing of the AAV/rAAV genome during transduction, with several of them directly influenced by the ITRs.

Several DDR-associated proteins have been implicated in processing AAV/rAAV genomes, with an ITR interaction only hypothesized. An ataxia telangiectasia and Rad3-related kinase (ATR)-dependent DDR was initially described to be elicited after exposure of cells to UV-inactivated wtAAV [279]. Initially suggested to rely on the ITRs structure, resembling stalled replication forks [279-281], later an involvement of the p5 promoter sequence and its cryptic RBE was assumed [282]. Recently, ATR was implicated in reduced transgene expression from rAAV without p5 promoter [278], but the mechanism remains to be resolved. Other DDR proteins involved in AAV/rAAV genome processing may be linked with ITRs due to their natural DNA processing function. The homology-directed repair associated protein Rad52 increases rAAV transduction and promotes random integration [277], potentially due to its inhibition of DNA end resection [283]. Ku86, which initiates non-homologous end joining, was shown to reduce functional AAV transduction [277]. Although the interaction of Rad52 and Ku86 with the rAAV genome was never mapped to the ITR [277], an interaction may be assumed based on their DNA-termini binding characteristics [284,285].

For other DDR proteins, the involvement of the ITRs has been more clearly demonstrated. FKBP52, a co-chaperone in the steroid receptor complex [286] is not a typical DDR protein, but it may be involved in the response to ionizing radiation-induced DNA damage [287]. In its phosphorylated form FKBP52 can interact directly with the ITRs and block DNA second-strand synthesis initiated at the free 3' end [288,289]. In its unphosphorylated form FKBP52 may aid the nuclear transport of AAV particles thereby promoting transduction [288]. The Mre11, Rad50, and Nbs1 (MRN) complex has been shown to directly recognize AAV ITRs and accumulate at AAV replication centers [290]. Initially assumed to inhibit replication and rAAV transduction by preventing second strand synthesis [89,290], an inhibitory effect during scAAV

transduction led to the hypothesis of a direct binding of Mre11 to the ITR as origin of the inhibition [291]. MRN could be shown to inhibit transduction of vector genomes with wtITR2, wtITR5 and a chimeric ITR [291,292], suggesting a sequence- and structure-independent silencing mechanism. The Ataxia telangiectasia mutated (ATM) kinase is usually activated by a functional MRN complex after DNA double-strand breaks [293]. However, the inhibitory effect of ATM on rAAV transduction was MRN-independent [291]. The mechanism was initially identified in patient-derived fibroblasts with mutated ATM, where rAAV transduction was found to be elevated compared to fibroblasts from healthy donors [294]. ATM-deficient cells also exhibited reduced host genomic integration levels but only little impact on vector circularization [223]. The silencing effect was found to be dependent on the T-shaped hairpin of the AAV ITRs, with linear U-shaped hairpins being unaffected by ATM-mediated silencing [292,295]. DNA-dependent protein kinase catalytic subunit (DNA-PKcs) and Artemis deficiency, such as in severe-combined-immunodeficient (SCID) mice, have also been implicated with reduced transduction, circularization and integration of both ssAAV and scAAV [223,296,297]. In normal mouse tissue, DNA-PKcs and Artemis have since been recognized as main mediators of vector recombination into episomes by recognizing and processing the ITRs [298]. Their absence in SCID mice leads to the accumulation and persistence of linear double-stranded AAV genomes [296,297], except for liver tissue where a slower alternative pathway facilitates the processing of the ITRs [299]. Intriguingly, the effects of pharmacological inhibitors of DDR proteins on rAAV transduction often vary between cell culture models and often contradict published data [278,300], potentially indicating cell-type specific processing pathways. This can be exemplified by the well-established reliance on DNA-PKcs for efficient transduction [223,298], while the inhibition of DNA-PKcs in primary human airway epithelia (pHAE) appeared to boost rAAV transduction [300]

The involvement of DDR proteins in AAV genome processing is undisputed, but the extent of the DDR-signaling upon rAAV transduction is controversial. In most cells no increase in DDR signaling occurs during transduction with rAAV [282], but in cells with a certain degree of stemness a fatal DDR upon rAAV transduction was ascribed to the unique structure of the AAV ITRs [280,281,301,302].

In cells lacking active p53, rAAV and UV-inactivated wtAAV transduction was shown to induce apoptosis via an ATM-mediated DDR [280,302,303]. By micro-injection of single-stranded DNA with ITR structures, it was shown that the p53-deficient cells sense the abnormal DNA structures of the ITR and undergo G2 arrest and apoptosis [280]. This effect was attributed to the

structural similarity of ITRs to stalled replication forks, triggering an ATR-dependent DDR leading to cell cycle arrest [281]. In p53-deficient cells, this cell cycle arrest cannot be maintained, thus the cells progress into mitotic catastrophe [281]. Cells that express p53 can also undergo apoptosis upon rAAV exposure, such as human embryonic stem cells (hESCs) and induced pluripotent stem cells (iPSCs) where p53 is a major regulator of apoptosis [304,305]. After delivery of the viral single-stranded DNA via rAAV, a p53-dependent apoptosis is initiated, which could be traced back to the ITR sequences by microinjection of a 39 nucleotide A-A' ITR oligonucleotide [301]. Later, six p53 binding motifs were predicted in the ITR sequence whose mutation in the ITRs of AAV2 prevented p53-dependent cell death [306]. A similar toxic effect was observed *in vivo*, where injection of rAAV1 into the dentate gyrus of mice caused ablation of neural progenitor cells and immature dentate granule cells [307]. The use of a different capsid serotype prevented cell death, likely due to the different tropism [307]. The toxicity in neural progenitor cells could also observed *in vitro* upon electroporation of AAV2 ITR sequences, especially in dividing cells [307], but the exact mechanism was not further investigated.

Thus, ITRs are important for the processing of the rAAV genome by DDR-related proteins, with varying effects in different cell types, opening the possibility of cell-type-specific responses to modified ITRs as was previously hypothesized [298]. However, unintended DDR activation by the ITRs could also lead to severe side effects.

1.3.5.2 ITRs and the integration of wtAAV and rAAV

In chapter 1.2.3, the controversy of the tumorigenic potential of AAV and rAAV integration was described, including the fact that in several studies with AAV and rAAV, the formation of HCC was observed in mice [210,216,217]. In contrast, in dogs [219], primates [220] and humans [220], rAAV therapy has so far not been associated with the direct causation of HCC, but cannot fully be excluded [210,220]. This chapter will focus more on the molecular mechanisms and the potential role of ITRs that lead to the integration of AAV and rAAV.

Shortly after the discovery of AAVs, latently infected cells were described from which infectious AAV could be recovered after co-infection with adenovirus even after extended passaging [308,309]. These AAV genomes were thus speculated to be integrated within the host genome [308]. Subsequently, a hotspot for site-specific integration of wtAAV in the long arm of human chromosome 19 was described, later termed AAVS1 [98-101]. This targeted integration relies on the large Rep proteins, which can bind and recognize a Rep binding site (RBS) within the AAVS1 locus [310,311]. Integration in the AAVS1 locus is not fixed to a specific location, but

occurs within 4 kb from the RBS and is often associated with sequence rearrangements [100,101,312-316]. Initially, the ITRs were assumed as the only viral element besides Rep that is required for targeted wtAAV integration [314]. Later, using a mutated ITR, it was shown that the ITR has only a minor effect on integration [310] and that the integration is mediated by Rep protein interaction with a cryptic RBE present in the p5 promoter [317]. Together with the previous observation that ITRs are lacking in the actively integrated sequences [314], the currently accepted model for Rep-mediated AAV integration does not rely on the ITRs, but rather on the cryptic RBE within the p5 promoter [318]. Still, other publication question the importance of the cryptic RBE for integration, as site-specific integration of rAAV was not significantly affected *in vivo* by the inclusion of the p5 promoter sequence in rAAV genomes when Rep was supplied *in trans* via a wtAAV [319]. The notion of efficient and targeted integration of wtAAV into the AAVS1 locus also has also come under scrutiny. Initially, site-specific integration in AAVS1 was thought to make up nearly 50% of all integration events in cell lines [101]. The use of non-selective methods to detect genomic integration of wtAAV in cell lines [320,321], non-human primate and human samples [206] suggest a maximum of 10% integration events targeted to the AAVS1 locus, lower levels at alternative preferred integration sites (AAVS2 and AAVS3) and most integrations randomly distributed across the genome.

In early studies with rAAV, where Rep is not expressed, the integration was shown to be random and the rate of integration was thought to be minor [322,323], nevertheless potential adverse effects were discussed [322]. Notably, rAAV with homology arms were used for gene targeting with site-specific integration efficiency of nearly 1% [324], with the ITRs potentially reducing random integration [325]. Recent studies have shown that random rAAV integration is a very common event and may be responsible for long-term expression of the delivered transgene [96,215,221], although circular episomal AAVs remain the predominant form even several years after vector administration [326]. During random integration of rAAVs the genomes are processed by DDR proteins [225], generating short truncations in the ITRs and vector genome [327], but most of the vector genomes remain intact and functional [328]. The integration sites appear randomly distributed in the host genome, albeit with a preference for open chromatin [326,329], repetitive sequences such as the ribosomal RNA (rRNA) repeats [206,215] and in the vicinity of DNA palindromes with a hairpin longer than 20 bp [330]. Thus, it was hypothesized that integration may preferentially occur in naturally occurring DNA strand breaks, as repeat and hairpin regions represent are more instable [331]. This was supported by data from immature T-cells, where genomic integration was highly enriched in the T-cell receptor locus, due to the

naturally occurring DNA strand breaks during maturation[332]. Additionally, in artificially induced DNA strand breaks, generated using endonuclease I-SceI [333], zinc finger nuclease [334] and CRISPR/Cas9 [335-337], rAAV integration was highly enriched. Further observations indicated that integration is mediated by a non-homologous end-joining mechanism [337,338], and by modulation of DDR pathways rAAV genomic integration rates can be increased [223,224] or decreased [223,225]. A recent publication monitoring AAV integration in mouse liver over one year found that most rAAV integration events are already observable within the first week [339], suggesting that unprocessed AAV genomes are responsible for integration events, in line with the mostly monomeric nature of integrated genomes [327]. Nonetheless, processed concatenated genomes were also shown to integrate [96,224]. Altogether this suggests that synthetic ITR structures with alternative recognition and processing by DDR proteins could alter the rate of genomic integration.

1.3.5.3 Promoter activity of ITRs

The ITR-adjacent region of AAV2 was shown to function as enhancer of transgene expression in liver [253], while involvement of the entire ITR in transgene expression was also suggested [340]. Then, ITRs of AAV2 [341] of other serotypes [238,342] were found to serve as promoters, with the promoter activity traced back to the A- and D-sequences [341]. For wtITR2, the presence of transcription factor binding sites within the D-sequence of the ITR was found responsible for the promoter activity, owing to binding of Regulatory factor X (RFX)1 and RFX3 [343]. In a study comparing the promoter activity of ITRs from different serotypes, the ITRs of AAV2 and AAV3 showed the strongest promoter activity in cell culture while the ITR of AAV1 was most active in murine liver [238]. There, RFX1/3 interaction may be responsible for promoter function of the AAV1 and AAV2 ITR due to similar D-sequences, while the transcription factor interacting with the AAV3 ITR remains unknown [343]. However, it should be noted that some of the detected expression could also stem from host promoters after integration into the genome. Nevertheless, altering the intrinsic promoter or enhancer activity of ITRs by engineering may be an elegant approach to improve control of the transgene expression, potentially even in a tissue-specific manner.

1.3.5.4 Interplay of the ITRs and the immune response

The adaptive immune response is a major limiting factor of AAV gene therapy, including pre-existing immunity as well as cytotoxic T-cell responses to the capsid [344] or the delivered

transgene [345]. Limited research has directly implicated the ITRs in immune response modulation, but there are indications that ITRs may both increase and decrease the immune response. A major part of the innate immune response is the recognition of viral elements in endosomal compartments and the cytoplasm by Toll-like receptors (TLRs) [346]. The packaged rAAV genome being non-methylated [347] and the ITRs being rich in CpG, making them an ideal substrate for TLR9 recognition [348]. This, in turn, makes the ITRs a highly immunogenic component of the AAV genome, although they could never be proven responsible for the induction of the innate immune response upon rAAV exposure. Another immunogenic effect is derived from the aforementioned promoter activity from within the ITR sequences (see section 1.3.5.3). Initiation of transcription by the ITRs on the opposite strand of the transgene can lead to the formation of dsRNA products. A delayed dsRNA-dependent innate immune response was observed in a chimeric mouse model with human hepatocytes [349]. There, the immune response triggered by the dsRNA sensor MDA5 was found responsible for the elimination of the transduced liver cells.

On the other hand, the ITR's promoter function may also lead to the suppression of the immune response. The D-sequence of the ITRs interacts with the RFX transcription factors thereby mediating the ITR's promoter function [343], but also competing with the host's natural promoter regions for RFX, *i.e.*, the genes encoding the major histocompatibility complex class II (MHC-II) [350]. This competition reduces expression of MHC-II genes, which could dampen the adaptive immune response. Notably, though, the effect on MHC gene expression was only observed *in vitro* but could not yet be verified *in vivo*.

1.3.5.5 miRNA sequences encoded within the ITR region

Viruses, especially DNA viruses, often encode miRNAs, which are used to modify cellular processes in the host cells or by exosome packaging also in bystander cells [351,352]. One publication has examined and identified miRNA expression from AAV genomes, with hotspots found near the p40 promoter and within the ITRs [353]. The bidirectional activity of the p5 promoter can drive not only the expression of Rep [354], but also of the miRNAs encoded within the ITRs. Several of these miRNAs were shown to be processed by the host miRNA processing machinery, but a function in the host cell or an effect on the helper virus could not be identified [353]. Expression of the ITR-encoded miRNAs could, in theory, still occur from rAAV granted the promoter used for transgene expression also exhibits bidirectional activity. However, the partial loss of the ITR sequences during episome formation and integration [355]

would likely further dampen any effect. Despite the likely absence of effects from ITR-encoded miRNA during rAAV transduction, a potential disruption or the inclusion of novel miRNA sequences may need to be considered when engineering the ITR sequences in rAAVs.

1.3.6 Engineered ITR sequences – current state and prospects

While capsid engineering has quickly evolved, and a plethora of methods are now used to identify new capsids with superior tropism (see chapter 1.2.3), the ITRs have remained mostly the same, with all approved drugs relying on the wtITR2 sequences. The outlier is Zolgensma with its scAAV genome that also relies on wtITR2 but carries the Δtrs mutation in one of the ITRs [121,260]. ITRs from other serotypes have been employed for scientific research [251], but their use has not been widely adopted. This is even more relevant considering that the use of alternative wtITRs or even synthetic ITRs could yield major advantages and could thereby provide additional levers to improve rAAV vector functionality.

Several factors have contributed to the scarcity of attempts to engineer ITR sequences. These include the observation that six different wtITR sequences from different AAV serotypes did not lead to any detectable differences in the expression of a delivered transgene *in vivo* [120], implying that the standard wtITR2 is sufficient. Additionally, achieving high vector titers during vector production already represents a major challenge [356], but altered ITR sequences and structures are generally considered to reduce productivity rates [241,306]. The third and probably most important limitation is the complex secondary structure of the ITRs, making cloning procedures complicated and time-consuming. The ITR structure can be split in half for cloning and the fragment can be reassembled in a multi-fragment ligation [357]. However, with conventional vector plasmids containing two ITR sequences, this is an inefficient and cost-intensive cloning procedure, especially when generating multiple ITR variants. Subsequently, the sequence of the ITR variant needs to be verified, which has been previously achieved by radiolabeled chain-terminator sequencing [358] or by using conventional Sanger sequencing after linearization of the ITR hairpin with a sequence-specific restriction enzyme [357]. However, these approaches are either time-consuming or restricting the sequence diversity during ITR engineering. Recently, the emergence of commercially available Sanger sequencing services for ITR sequences in plasmids [240] and Nanopore technology-based whole-plasmid sequencing [254] has facilitated the sequencing of plasmids with ITRs.

Despite these challenges, rAAV vectors with ITR variants have been generated and examined. Initially this research was focused on understanding the function of the ITRs. The first ITRs

with non-natural sequences were generated during examination of the ITR *cis*-repair mechanism, where partial deletions of one of the two ITRs were shown to be repaired by the intact ITR present on the opposite end [270]. Later, the use of ITR mutants was mostly focused on deciphering the sequence requirements for AAV replication [358] and integration [310]. Also, the importance of the D-sequence, participating in trs hairpin formation (see Figure 4B), was determined during vector production with partial or complete deletions within ITR sequences [110,269,359]. During evaluation of the requirements for Rep functionality, especially the binding and nicking of ITR sequences during replication and integration [310], a variety of different ITR mutants were examined as well. However, these assays were mostly performed *in vitro* with purified Rep proteins and ITR-like DNA oligonucleotides [246,247].

Only later synthetic ITRs were developed to serve a specific purpose relevant for gene therapy and used to produce rAAVs. To alter the function of rAAV during transduction, it is sufficient to rely solely on wtITR sequences in a synthetic configuration, *i.e.,* equipping a single rAAV genome with ITRs from two replication incompatible serotypes. During vector production, the incompatibility of AAV2 Rep with the wtITR5 and, *vice versa*, AAV5 Rep with the wtITR2 [121] allowed the incorporation of two distinct ITRs at both ends of the genome, granted production takes place in presence of Reps from both serotypes [37,360]. These rAAV are less prone to forming monomeric episomes upon transduction [37,360], instead promoting the assembly into di- or multimeric episomes [37]. This facilitates the reassembly of an oversized split-transgene delivered via two different vectors by conventional RNA splicing [37,360], without relying on a *trans*-splicing mechanism [360]. Chimeric ITRs of different wtITRs were developed to prevent vector remobilization, the theoretical process where a superinfection with wtAAV and adenovirus leads to generation of infectious rAAV particles in the patient's body. There, a chimeric AAV2/AAV5 ITR was identified that replicates only in presence of a chimeric AAV2/AAV5 Rep proteins, thereby reducing the risk of vector remobilization [357].

The ITR's D-sequence is the most easily altered sequence of the ITR due to the lack of a secondary structure in the typical vector plasmid with two ITR sequences (see Figure 5A). This accessibility, combined with the transcription factor binding and immune evasion as well as packaging signal function [110,248,269,289,350], makes the region relevant for engineering. The removal of the D-sequence in one of the ITRs was shown to induce the packaging of DNA strands of a single polarity without affecting transduction, also indicating a function of this region as packaging signal [93,248]. The substitution of a potential transcriptional repressor binding site in the D-sequence of one ITR with an inert DNA sequence maintained single

polarity genome packaging but increased transgene expression *in vitro* and *in vivo* [248]. The only randomization-based ITR screen, comparable to the long-established randomization of capsid sequences, was limited to D-sequence variation [361]. To facilitate library generation, the screen was performed using scAAV, thus only the functional full-length ITR needed to be engineered to contain the randomized D-sequence library. This screen suggested that the D-sequence is of low importance for replication and packaging [361]. Notably, this finding and the use of the scAAV context were criticized, as the plasmid resolution, replication and packaging of scAAVs may differ from traditional ssAAVs [362].

ITR engineering within the central dual hairpin region has been even less frequent, with several publications only examining the functionality of a single synthetic ITR variant. A deletion of the B-B' and C-C' hairpin was shown to improve transduction both *in vitro* and *in vivo* [292]. The reduced ATM-mediated inhibition towards such U-shaped hairpins, compared to the T-shaped wtAAV ITR was previously known (see section 1.3.5.1) [295], but this had not been demonstrated using rAAV vectors with altered ITRs. Using this ITR variant for transgene delivery substantially increased expression *in vitro*, especially at lower multiplicities of infection (MOIs), and *in vivo* by up to six-fold [292]. This came at the expense of a ten-fold reduction in productivity, thereby limiting applicability since AAV manufacturing remains a major challenge [356]. The same deletion variant was also included in a recent analysis of common random ITR plasmid mutants (ΔB-B', ΔC-C', ΔB-B'-C-C', ΔD and a partial ΔA-A' ITR) where the reduced productivity was confirmed but an improved transduction was not observed [255].

Since the innate immune response antagonizes successful gene therapy, it was reasoned that the disruption of CpG dinucleotide motifs within the ITR sequences could reduce recognition by TLR9 and thus reduce the immune response (see section 1.3.5.4) [241]. A dual ITR vector plasmid was generated in which one nucleotide of all CG motifs was replaced by either A or T , reducing rAAV productivity by three-fold compared to conventional wtITR2 [241]. Vector transduction was not affected by this CpG depletion and an effect on the immune response was not examined [241]. Another recently described synthetic variant of wtITR2 lacks potential p53 binding sites [306], which have been implicated in cell death of hESCs and iPSCs upon exposure to rAAVs [304,305] (see section 1.3.5.1). The authors describe an increased productivity compared to wtITR2 as well as the absence of p53 activation in hESCs, albeit there the transgene was barely expressed [306]. Thus, this ITR variant could, after further improvements, facilitate the safe use of rAAV for gene therapy even in stem or progenitor cells.

In conclusion, the field of ITR engineering lags far behind the field of capsid engineering, with only a single ITR randomization screen limited to D-sequence [361]. Undoubtedly, engineered ITRs can provide capabilities that cannot be achieved with engineered capsid mutants, such as the described reduction of toxicity in hESC [306], the prevention of rAAV remobilization as well as altered levels of episome formation [357] and potentially host genome integration [223-225]. Thus, new approaches to enable the screening of large ITR variant libraries could greatly benefit the development of safer and more efficient rAAV gene therapy vectors.

1.4 Aims of this work

A main focus of AAV engineering has traditionally been the optimization of the capsid structure or of the transgene expression cassette including promoters and regulatory elements. In recombinant AAV vectors, the ITRs, the replication origin, are the last remnant of the original virus. This thesis work shifts the spotlight onto these critical yet often overlooked sequences, aiming to explore and engineer the ITRs to better understand and potentially enhance their function.

The main aim of this work was to create a platform that facilitates the screening of ITR variants. Recently, the use of alternative ITR sequences with improved characteristics has been described, including ITR variants mediating increased levels of transgene expression[292], reduced toxicity[306] and a potential reduction of immune recognition[241]. However, large screens are impeded by cloning difficulties due to the hairpin structure and the loss of the altered ITR during transduction after episome formation or integration. To overcome these issues, I aimed to establish (i) a plasmid design and cloning strategy that facilitates the generation of ITR variants in larger numbers, (ii) a Sanger-sequencing based approach to facilitate ITR sequence confirmation in a plasmid context, (iii) a barcoding strategy that enables the parallel interrogation of multiple ITR variants and their effects on vector production as well as their function during transduction, despite the partially loss of the ITRs, and (iv) confirm the sequence of the altered ITRs at the termini of the rAAV genomes. All this was performed with the final goal of creating and screening a comprehensive ITR variant library. After successful completion of the first three steps, during the ITR sequence confirmation in the vector genomes, a novel ITR repair mechanism was discovered that suggests ITRs can serve as repair templates to alter the sequence of other ITRs. This ITR *trans*-repair made the large ITR variant library temporarily unattainable, thus I further explored the extent of the ITR *trans*-repair mechanism and strategies to circumvent it. After confirming the ITR integrity of two libraries

with a total of 90 different ITR mutations with high similarity to the wtITR2, their function was analyzed *in vitro* and *in vivo*. There, the impact of the ITRs on production, transgene expression, and episome formation was examined. Additionally, the samples from the *in vivo* screen were used to develop a novel method that enabled the simultaneous interrogation of both vector barcode and integration region had to be developed.

In a second and smaller project, the possibility was examined to alter the AAV replication origin with the aim to facilitate the generation and packaging of circular genomes. The delivery of circular genomes via AAV capsids could circumvent the circularization step required for episome formation with conventional AAV genomes, thus potentially rendering such vectors safer and more efficient. To this end, I devised a novel plasmid design that has the potential to form circular genomes and then traced the genomes during replication and after packaging in AAV particles.

2. Material and Methods

2.1 Materials

2.1.1 Devices

Table 2: Devices used in this thesis.

Application	Device	Provider
Pipetting		
Pipetting	Accu-jet pro	Brand (Wertheim am Main, Germany)
Pipetting	Pipetboy acu 2	Integra (Zizers, Switzerland)
Pipetting	Research plus (2.5 µl, 10 µl, 100 µl, 200 µl, 1000 µl)	Eppendorf (Hamburg, Germany)
Pipetting	Research plus multichannel (12-well; 10 µl, 100 µl, and 300 µl)	Eppendorf (Hamburg, Germany)
Centrifugation		
Centrifugation	Allegra X-12; Allegra 25R	Beckman Coulter GmbH (Brea, USA)
Centrifugation	Allegra 25R	Beckman Coulter GmbH (Brea, USA)
Centrifugation	Avanti J-26XP	Beckman Coulter GmbH (Brea, USA)
Centrifugation	Microcentrifuge 5415R, 5427R, 5425	Eppendorf (Hamburg, Germany
Centrifugation	Optima L-90K	Beckman Coulter GmbH (Brea, USA)
Centrifugation	JA-10 rotor	Beckman Coulter GmbH (Brea, USA)
Centrifugation	Fixed angle rotor type 70.1 Ti	Beckman Coulter GmbH (Brea, USA)
Gel electrophoresis		
Agarose gel running system	EasyPhor	Biozym Scientific (Hessisch Oldendorf, Germany)
Agarose gel power device	E385 power supply	Consort (Turnhout, Belgium)
Agarose gel device	Mupid-One electrophoresis system	Nippon Genetics (Düren, Germany)
Agarose gel imaging	Gel Doc XR	Bio-Rad (Hercules, USA)
Agarose gel imaging	Azure 300 Imager	Azure Biosystems (Dublin, USA)
Droplet digital (dd)PCR		
Droplet generation	QX200 Droplet Generator	Bio-Rad (Hercules, USA)
Plate sealing	PX1 PCR Plate Sealer	Bio-Rad (Hercules, USA)
PCR	C1000 Touch Thermal Cycler	Bio-Rad (Hercules, USA)
Droplet analysis	QX200 Droplet Reader	Bio-Rad (Hercules, USA)
Cell/Bacteria culture		
Automated cell counter	Countess	Thermo Fisher Scientific (Waltham, USA)
Bacteria incubator	Heraeus Function Line Incubator	Thermo Fisher Scientific (Waltham, USA)
Bacteria incubator (shaking)	Multitron	INFORS HT (Basel, Switzerland)
Bacteria incubator + shaking	Ecotron	INFORS HT (Basel, Switzerland)
Cell counting	Neubauer counting chamber	Brand (Wertheim am Main, Germany)
Cell culture hood	Herasafe KS12	Thermo Fisher Scientific (Waltham, USA)
Incubator	Heracell 150 CO_2 incubator	Thermo Fisher Scientific (Waltham, USA)
Miscellaneous		
DNA purification	QIAvac 24 Plus	Qiagen (Hilden, Germany)

Application	Device	Provider
DNA purification	MagnaBot 96 Magnetic Separation Device	Promega (Madison, USA)
DNA quality assessment	2100 Bioanalyzer	Agilent Technologies (Santa Clara, USA)
DNA/RNA quantification	NanoDrop 2000 Spectrophotometer	Thermo Fisher Scientific (Waltham, USA)
Incubation and mixing	Mixing Block MB 102	Bioer Technology (Hangzhou, China)
Iodixanol density gradient tube sealing	Tube Sealer 342428	Beckman Coulter GmbH (Brea, USA)
Luminescence measurement	GloMax Navigator Microplate Luminometer	Promega (Madison, USA)
Mixing	Vornado Mini Vortex Mixer	Benchmark Scientific (Sayreville, USA)
Mixing	Heathrow HS120214 Vortexer	Heathrow Scientific (Vernon Hills, USA)
N_2 bucket	Nalgene Dewar-NB4	Thermo Fisher Scientific (Waltham, USA)
Nanopore	MinION Mk1b	Oxford Nanopore Technologies (Oxford, UK)
PCR cycler	Mastercycler nexus GSX1/GSX2e	Eppendorf (Hamburg, Germany)
PCR hood	Captair Bio Smart PCR-Hood	Erlab (Val-de-Reuil, France)
qRT-PCR cycler	Corbett Rotor-Gene 6000	Qiagen (Hilden, Germany)
Scale	Kern EG 2200-2NM	Kern & Sohn (Balingen, Germany)
Storage	Forma 900 Series (-80°C)	Thermo Fisher Scientific (Waltham, USA)
Storage	Liebherr Comfort (-20°C)	Liebherr (Bulle, Switzerland)
Storage	Liebherr Comfort (4°C)	Liebherr (Bulle, Switzerland)
Tissue lysis	TissueLyser LT	Qiagen (Hilden, Germany)
Ultrasound bath	Sonorex Digitec DT31H	Bandelin (Berlin, Germany)
Water bath	Model TW20 Water Bath	Julabo (Seelbach, Germany)

2.1.2 Software

Table 3: Software employed in this work.

Name	Provider	Source
Geneious V7.1.7.	Biomatters (Auckland, New Zealand)	
IGV (v2.18.4).	IGV Team (UC San Diego & Broad Institute, USA)	
ImageJ V1.53k	Wayne Rasband (NIH)	
Inkscape (V0.92)	Inkscape Community	
Microsoft Office 365	Microsoft (Redmond, USA)	
Minimap2 (v2.24)	Li H	https://doi.org/10.1093/bioinformatics/bty191
MinKNOW (v22.05.5)	Oxford Nanopore Technologies (Oxford, UK)	
GraphPad Prism5 (v5.03)	GraphPad Software Inc (Boston, USA)	
QuantaSoft (1.7.4.09179	Bio-Rad (Hercules, USA)	
R (v4.3.3)	R Core Team	
RotorGene 6000 Series Software 1.7	Qiagen	

Name	Provider	Source
Rstudio (2024.09.1)	Posit PBC (Boston, USA)	
samtools	Li H	doi: 10.1093/gigascience/giab008
R packages		
annotatr (v1.28.0)	Cavalcante RG & Sartor MA	doi: 10.1093/bioinformatics/btx183.
bamsignals (v1.34.0)	Mammana A & Helmuth J	doi:10.18129/B9.bioc.bamsignals
BSgenome.Mmusculus.UCSC.mm10 (v3.20)	The Bioconductor Dev Team	10.18129/B9.bioc.BSgenome.Mmusculus.UCSC.mm10
dplyr (v2.5.0)	Wickham H et al.	https://dplyr.tidyverse.org
GenomicAlignments (v1.38.2)	Lawrence M et al.	doi:10.1371/journal.pcbi.1003118
GenomicRanges (v1.43.1)	Lawrence M et al.	doi:10.1371/journal.pcbi.1003118
ggplot2 (v3.5.0).	Wickham H et al.	https://ggplot2.tidyverse.org
ggpubr (v0.6.0)	Kassambara A	https://rpkgs.datanovia.com/ggpubr/
regioneR (v1.34.0)	Gel B et al.	doi:10.1093/bioinformatics/btv562
Rsamtools (v2.18.0)	Morgan M et al.	doi:10.18129/B9.bioc.Rsamtools,
ShortRead (v1.60.0)	Morgan M et al.	doi:10.1093/bioinformatics/btp450
stringr (v1.5.1)	Wickham H	https://stringr.tidyverse.org
txDb.Mmusculus.UCSC.mm10.knownGene(v3.20)	Carlson M	10.18129/B9.bioc.TxDb.Mmusculus.UCSC.mm10.knownGene
Galaxy web platform tools		https://usegalaxy.eu/
BBMap (v39.08)		
Je-clip (v1.2.1)		
Trim (v0.0.2)		

2.1.3 Consumables

Table 4: Consumable material used for the experiments

Consumables	Specification	Provider
Amicon Ultra-15 Centrifugal Filter Unit	MWCO 100000	Merck (Darmstadt, Germany)
Bacteria culture tubes	Tube 13 ml, 100x16 mm	Sarstedt (Nümbrecht, Germany)
Cell culture dishes	150 x 20 mm	Sarstedt (Nümbrecht, Germany)
Cell culture flasks	75 cm2; 175 cm2	Greiner Bio-One (Kremsmünster, Austria)
Cell lifter	3008	Corning (New York, USA)
Countess cell counting chamber slides		Thermo Fisher Scientific (Waltham, USA)
ddPCR plates	96-well, semi-skirted	Bio-Rad (Hercules, USA)
DG8 Cartridges for Droplet Generator		Bio-Rad (Hercules, USA)
DG8 Gaskets for Droplet Generator		Bio-Rad (Hercules, USA)
DNA LoBind Tubes	1.5 mL	Eppendorf (Hamburg, Germany)
Erlenmeyer flasks	250 ml; 500 ml; 1000 ml	DWK Life Sciences (Wertheim, Germany)
Glass bottles	50 ml; 100 ml; 250 ml; 500 ml; 1000 ml	Thermo Fisher Scientific (Waltham, USA)
Glass test tubes	160x15 mm	Thermo Fisher Scientific (Waltham, USA)
Microcentrifuge tubes	1.5 ml; 2 ml	Sarstedt (Nümbrecht, Germany)
Microplate 96-well	F-bottom, white	Greiner Bio-One (Kremsmünster, Austria)
Nanopore	Spot-ON Flow Cell, R9 Version	Oxford Nanopore Technologies (Oxford, UK)

Consumables	Specification	Provider
N-well tissue culture plates	6 well; 96 well	Greiner Bio-One (Kremsmünster, Austria)
Pasteur capillary pipettes	230 mm	Corning (New York, USA)
PCR plate	96 well, semi-skirted	Biozym Scientific (Hessisch Oldendorf, Germany)
PCR strip tubes	0.2 mL	Biozym Scientific (Hessisch Oldendorf, Germany)
Petri dishes	94 x 16 mm	Greiner Bio-One (Kremsmünster, Austria)
Pierceable foil heat seal		Bio-Rad (Hercules, USA)
Reaction tubes	1.5 ml, 2 ml	DWK Life Sciences (Wertheim, Germany)
Re-seal polyallomer centrifuge tubes	16 × 76 mm	Seton Scientific (Petaluma, USA)
Scalpel blades	figure 23, Carbon steel	Heinz Herenz (Hamburg, Germany)
Serological pipettes	5 ml; 10 ml, 25 ml; 50 ml	Greiner Bio-One (Kremsmünster, Austria)
Stainless steel beads	3 mm	Qiagen (Hilden, Germany)
Syringes	Luer-Lok (3 mL; 5 ml)	BD (Franklin Lakes, USA)
TipOne - pipet tips	10/20 µl; 200 µl; 1000 µl	Starlab (Hamburg, Germany)
TipOne - pipet tips filtered	10/20 µl; 200 µl; 1000 µl	Starlab (Hamburg, Germany)
Tubes (Falcon)	15 mL / 50 mL	Corning (New York, USA)

2.1.4 Kits

Table 5: Commercial kits employed in this thesis work

Application	Kits	Provider
Bead-based DNA purification	ProNex Size-Selective Purification System	Promega (Madison, USA)
Bioanalyzer	DNA 1000 Kit	Agilent Technologies (Santa Clara, USA)
cDNA synthesis	SuperScript IV VILO Master Mix with ezDNase Enzyme	Invitrogen/Thermo Fisher Scientific (Waltham, USA)
cDNA synthesis	Tetro cDNA Synthesis Kit	Meridian Bioscience (Cincinnati, USA)
ddPCR primer/probe kit	ddPCR GEX HEX Assay Rpp30, Mmu (#10031255)	Bio-Rad (Hercules, USA)
ddPCR	ddPCR Supermix for Probes (No dUTP)	Bio-Rad (Hercules, USA)
DNA and RNA extraction (cells and tissues)	AllPrep Mini Kit	Qiagen (Hilden, Germany)
DNA extraction (agarose gels)	QIAquick Gel Extraction Kit	Qiagen (Hilden, Germany)
DNA purification	QIAquick PCR Purification Kit	Qiagen (Hilden, Germany)
DNA purification	QIAquick Nucleotide Removal Kit	Qiagen (Hilden, Germany)
DNA purification/concentration	DNA Clean & Concentrator-5	Biozym Scientific (Hessisch Oldendorf, Germany)
Dual luciferase assay	Dual-Luciferase Reporter Assay System	Promega (Madison, USA)
NGS library preparation	Nextera XT Index Kit v2	Illumina (San Diego, USA)
On-column DNA digest for RNA extraction	RNase-Free DNase Set	Qiagen (Hilden, Germany

Application	Kits	Provider
Plasmid purification	QIAprep Spin Miniprep Kit	Qiagen (Hilden, Germany)
Plasmid purification	QIAGEN Plasmid Plus Midi Kit	Qiagen (Hilden, Germany)
Plasmid purification	PureYield Plasmid Midiprep System	Promega (Madison, USA)
Plasmid purification	NucleoBond PC 500 Maxi Kit	Macherey-Nagel (Düren, Germany)
qPCR (probe-based)	Sensimix II Probe Kit	Bioline (London, UK)
Qubit DNA quantification	Qubit dsDNA HS Assay Kit	Thermo Fisher Scientific (Waltham, USA)
Nanopore	Rapid Barcoding Kit SQK-RBK004	Oxford Nanopore Technologies (Oxford, UK)
Nanopore	Flow Cell Wash Kit	Oxford Nanopore Technologies (Oxford, UK)
adapted ITR-seq	NEBNext UltraShear	NEB (Ipswich, USA)
adapted ITR-seq	NEBNext End Repair Module	NEB (Ipswich, USA)
adapted ITR-seq	NEBNext dA-Tailing Module	NEB (Ipswich, USA)
adapted ITR-seq	Quick Ligation Kit	NEB (Ipswich, USA)

2.1.5 Reagents

Table 6: Reagents used for experiments in this thesis.

Reagent	Provider
0.25% Trypsin / EDTA	Gibco/Thermo Fisher Scientific (Waltham, USA)
1 kb Plus DNA ladder	Thermo Fisher Scientific (Waltham, USA)
1× Dulbecco's phosphate buffered saline (PBS)	Gibco/Thermo Fisher Scientific (Waltham, USA
2.5% Trypsin / EDTA	Gibco/Thermo Fisher Scientific (Waltham, USA)
Acetic acid	VWR chemicals (Radnor, USA)
Ampicillin	Roth (Karlsruhe, Germany)
Aqua B. Braun	B. Braun Avitum Saxonia GmbH (Melsungen, Germany)
Bacto agar	BD (Franklin Lakes, USA)
Bacto tryptone	BD (Franklin Lakes, USA)
Bacto yeast extract	BD (Franklin Lakes, USA)
Calcium chloride ($CaCl_2$)	Sigma-Aldrich (St. Louis, USA)
Carbenicillin	Sigma-Aldrich (St. Louis, USA)
CutSmart/NEB1/NEB2/NEB2.1/NEB3.1 buffer	NEB (Ipswich, USA)
ddPCR droplet generation oil	Bio-Rad (Hercules, USA)
ddPCR droplet reader oil	Bio-Rad (Hercules, USA)
Deoxynucleotide (dNTP) Solution Mix (10 mM of each)	NEB (Ipswich, USA)
Diluent B	NEB (Ipswich, USA)
DMEM, high glucose, GlutaMAX (61965026)	Gibco/Thermo Fisher Scientific (Waltham, USA)
DNaseI Reaction Buffer	NEB (Ipswich, USA)
EDTA	GRÜSSING GmbH (Filsum, Germany)
Ethanol absolute	Merck (Darmstadt, Germany)
Ethidium bromide 1%	Roth (Karlsruhe, Germany)
Fetal bovine serum (FBS)	Capricorn Scientific (Ebsdorfergrund, Germany)
Gelred Nucleic Acid Gel Stain	Biotium (Fremont, USA)
GeneRuler DNA Ladder Mix	Thermo Fisher Scientific (Waltham, USA)
Glycerol	VWR chemicals (Radnor, USA)
Isopropanol	Merck (Darmstadt, Germany)
Kanamycin	Sigma-Aldrich (St. Louis, USA)
Magnesium chloride (MgCl2)	Applichem (Darmstadt, Germany)

Reagent	Provider
Nuclease-free H_2O	Qiagen (Hilden, Germany)
OptiPrep (Iodixanol)	Progen (Heidelberg, Germany)
Penicillin / Streptomycin (P/S)	Gibco/Thermo Fisher Scientific (Waltham, USA)
Phenol red	Merck (Darmstadt, Germany)
Polyethyleneimine (PEI MAX)	Polysciences Europe GmbH (Eppelheim, Germany)
Potassium acetate (KAc)	Honeywell (Seelze, Germany)
Potassium chloride (KCl)	GRÜSSING GmbH (Filsum, Germany)
Sodium chloride (NaCl)	GRÜSSING GmbH (Filsum, Germany)
Sodium Dodecylsulfate (SDS)	Serva (Heidelberg, Germany)
Sodium hydroxide (NaOH) 2 M	Merck (Darmstadt, Germany)
T4 DNA Ligase Buffer	NEB (Ipswich, USA)
TE Buffer	Thermo Fisher Scientific (Waltham, USA)
Terrific Broth	Roth (Karlsruhe, Germany)
TrickTrack DNA Loading dye (6x)	Thermo Fisher Scientific (Waltham, USA)
TRIS	Roth (Karlsruhe, Germany)
Tris-Cl	
TRIS-HCl	Roth (Karlsruhe, Germany)
Trypan Blue Solution, 0.4%	Thermo Fisher Scientific (Waltham, USA)
β-mercaptoethanol	Sigma Aldrich

2.1.6 Enzymes

Table 7: Enzymes employed in this thesis work.

Enzymes	Provider
AccI	NEB (Ipswich, USA)
AgeI	NEB (Ipswich, USA)
Antarctic phosphatase	NEB (Ipswich, USA)
EMPROVE Benzonase	Merck (Darmstadt, Germany)
BglII	NEB (Ipswich, USA)
BsaI-HF	NEB (Ipswich, USA)
BsmbI-v2	NEB (Ipswich, USA)
BsrGI	NEB (Ipswich, USA)
DNase I (RNase-free)	NEB (Ipswich, USA)
EcorI-HF	NEB (Ipswich, USA)
Esp3I	NEB (Ipswich, USA)
HindIII-HF	NEB (Ipswich, USA)
KAPA HiFi HotStart ReadyMix PCR Kit	Roche (Basel, Switzerland)
NaeI	NEB (Ipswich, USA)
NheI	NEB (Ipswich, USA)
NotI-HF	NEB (Ipswich, USA)
OneTaq Quick-Load 2× Master Mix with Standard Buffer	NEB (Ipswich, USA)
PacI	NEB (Ipswich, USA)
Phusion Flash High-Fidelity PCR Master Mix	Thermo Fisher Scientific (Waltham, USA)
Proteinase K	Roche (Basel, Switzerland)
PvuI	NEB (Ipswich, USA)
Q5 High-Fidelity 2X Master Mix	NEB (Ipswich, USA)
SacII	NEB (Ipswich, USA)
SalI	NEB (Ipswich, USA)
ScaI	NEB (Ipswich, USA)
SpeI	NEB (Ipswich, USA)
T4 DNA Ligase	NEB (Ipswich, USA)

Enzymes	Provider
T4 Polynucleotide Kinase	NEB (Ipswich, USA)
T7 endonuclease I	NEB (Ipswich, USA)
XbaI	NEB (Ipswich, USA)
XmaI	NEB (Ipswich, USA)

2.1.7 Buffer

Table 8: Utilized buffers and their ingredients.

Buffer	Ingredients
15% iodixanol	75.00% PBS-MK-NaCl, 25.00% OptiPrep
25% iodixanol	58.19% PBS-MK, 41.56% OptiPrep, 0.25% Phenol red stock
40% iodixanol	66.67% OptiPrep, 33.33% PBS-MK
60% iodixanol	99.75% OptiPrep, 0.25% Phenol red stock
Benzonase buffer	150 mM NaCl, 50 mM TRIS-HCl (pH 8.0), 2 mM $MgCl_2$
Hirt lysis buffer (+SDS)	10 mM Tris-Cl, 10 mM EDTA, 0.6% SDS (SDS added short before use)
LB agar	1.5% Bacto agar, 1.0% NaCl, 1% Bacto tryptone, 0.5% Bacto yeast extract
LB medium	1.0% Bacto tryptone, 1.0% NaCl, 0.5% Bacto yeast extract
P1 resuspension buffer	50 mM Tris-HCl (pH 8.0), 10 mM EDTA, 100 µg/mL RNase A
P2 lysis buffer	200 mM NaOH, 1% SDS
P3 neutralization buffer	2.8 M KAc (pH 5.1)
PBS-MK	PBS (1×), 2.5 mM KCl, 1 mM $MgCl_2$
PBS-MK-NaCl	1 M NaCl in PBS-MK
Phenol red stock	Nuclease-free H_2O, 0.5% Phenol red
TAE buffer	2 M TRIS, 1 M acetic acid, 50 mM EDTA
TB-medium	1.2% Casein, 2.4% Yeast extract, 1.254 K_2HPO_4, 0.231% KH_2PO_4, 0.4% glycerol

2.1.8 Cells and cell lines

Table 9: Prokaryotic and eukaryotic cells used in this thesis work

Prokaryotic / Eukaryotic cells	Provider/Source
One Shot TOP10	Invitrogen
Hek293T	Human (embryonic kidney)
HuH7	Human (hepatoma)

2.1.9 Services

Table 10: Commercial services employed for this thesis work.

Services	Provider
AAV-ITR sequencing	Azenta/Genewiz (Chelmsford, USA)
Amplicon-EZ NGS	Azenta/Genewiz (Chelmsford, USA)
NGS (MiSeq & NextSeq)	EMBL (Heidelberg, Germany)
Sanger sequencing (Light Run/Light Run Express)	Eurofins Scientific (Luxemburg, Luxemburg)

2.1.10 Oligonucleotides

All DNA oligonucleotides were purchased from Integrated DNA Technologies (IDT; San Jose, USA) or Merck (Darmstadt, Germany). Probes with the indicated fluorophores and quenchers were purchased from IDT (San Jose, USA). The following Table 11 provides the used oligonucleotide sequences, except the ones used for the cloning of the barcode and hairpin sequences which are described in section 2.2.3.

Table 11: DNA oligonucleotides used for this thesis project. Oligonucleotides are given in IUPAC code. Oligonucleotides described in cloning sections indicating the design of the oligonucleotides are not included in this table, this covers all barcode nucleotides as well as most ITR hairpin nulceotides. Chemical modifications of primers include phosphothioate bonds between nucleotides (*) and 5‘phosphorylation ([phos]).

Name	Sequence	Application
CMV_probe	FAM-AGTCATCGCTATTACCATGG-BHQ1	qPCR/ddPCR
CMV_fw	TGCCCAGTACATGACCTTATGG	qPCR/ddPCR
CMV_rv	GAAATCCCCGTGAGTCAAACC	qPCR/ddPCR
muRPP30_c9_probe	/5HEX/ATCTAGGCT/ZEN/TGCTGTTTGGGCTCT/3IABkFQ/	qPCR/ddPCR
muRPP30_c9_fw	TGTCCAGTGCTGCAGAAAG	qPCR/ddPCR
muRPP30_c9_rv	GCCCAAACAGCAGTCCTAA	qPCR/ddPCR
mCherry_probe	/56-FAM/CCAACTTGA/ZEN/TGTTGACGTTGTAGGCG/3IABkFQ/	qPCR/ddPCR
mCherry_fw	GAGGCTGAAGCTGAAGGAC	qPCR/ddPCR
mCherry_rv	GATGGTGTAGTCCTCGTTGTG	qPCR/ddPCR
85FB	CGCCTACAACGTCAACATC	plasmid/T7eI ITR-seq
156FB	CAAGTCTGGAAAGAAATGC	plasmid/T7eI ITR-seq
357FB	CGATTTTTGTGATGCTCGTCAG	plasmid/T7eI ITR-seq
358FB	CCATTATAAGCTGCAATAAAC	plasmid/T7eI ITR-seq
606FB	GTAATGCAGAAGAAAA*C*C	RCA
607FB	ACCTTGTAGATGAACT*C*G	RCA
608FB	CATGGACGAGCTGTACA*A*G	RCA
609FB	TTCGTACTGTTCCACGA*T*G	RCA
268FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATCGTGGAACAGTACGAA	NGS (1st. Gen sITR)
269FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATAAACGATCCTTTATTGCTAG	NGS (1st. Gen sITR)
270FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNATCGTGGAACAGTACGAA	NGS (1st. Gen sITR)
271FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNATAAACGATCCTTTATTGCTAG	NGS (1st. Gen sITR)
272FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNATCGTGGAACAGTACGAA	NGS (1st. Gen sITR)
273FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNATAAACGATCCTTTATTGCTAG	NGS (1st. Gen sITR)
274FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNATCGTGGAACAGTACGAA	NGS (1st. Gen sITR)
275FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNATAAACGATCCTTTATTGCTAG	NGS (1st. Gen sITR)

Name	Sequence	Application
276FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNATCGTGGAACAGTACGAA	NGS (1st. Gen sITR)
277FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNNNATAAACGATCCTTTATTGCTAG	NGS (1st. Gen sITR)
278FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNATCGTGGAACAGTACGAA	NGS (1st. Gen sITR)
279FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNNNNATAAACGATCCTTTATTGCTAG	NGS (1st. Gen sITR)
280FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNNATCGTGGAACAGTACGAA	NGS (1st. Gen sITR)
281FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNNNNNATAAACGATCCTTTATTGCTAG	NGS (1st. Gen sITR)
282FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNNNATCGTGGAACAGTACGAA	NGS (1st. Gen sITR)
283FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNNNNNNATAAACGATCCTTTATTGCTAG	NGS (1st. Gen sITR)
448FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GCCTTCACTAGTATGCGACTG	NGS (2nd. Gen sITR)
449FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNCCTTCACTAGTATGCGACTG	NGS (2nd. Gen sITR)
450FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNCCTTCACTAGTATGCGACTG	NGS (2nd. Gen sITR)
451FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNCCTTCACTAGTATGCGACTG	NGS (2nd. Gen sITR)
452FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNCCTTCACTAGTATGCGACTG	NGS (2nd. Gen sITR)
453FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNCCTTCACTAGTATGCGACTG	NGS (2nd. Gen sITR)
454FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNNCCTTCACTAGTATGCGACTG	NGS (2nd. Gen sITR)
455FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNNNCCTTCACTAGTATGCGACTG	NGS (2nd. Gen sITR)
456FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GCACCTACAAGGCCAAGAAG	NGS (2nd. Gen sITR)
667FB	[phos]CCTGTGCGATTCC*T*C	adapted ITR-seq (Y-adapter)
668FB	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN NWNNWNNTCCGAATCGCACAGG*T	adapted ITR-seq (Y-adapter - no dephasing)
679FB	GTTCGCAGCTTCGTACATGNNWNNWNNTCCGA ATCGCACAGG*T	adapted ITR-seq (Y-adapter - dephasing)
669FB	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTG AGCTGTACAAGTAAGTCGACGTAC	adapted ITR-seq (no dephasing)
670FB	ACACTCTTTCCCTACACGACG	adapted ITR-seq (no dephasing)
680FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GGCTGTACAAGTAAGTCGACG	adapted ITR-seq (dephasing)
681FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNGCTGTACAAGTAAGTCGACG	adapted ITR-seq (dephasing)
682FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNGCTGTACAAGTAAGTCGACG	adapted ITR-seq (dephasing)
683FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNGCTGTACAAGTAAGTCGACG	adapted ITR-seq (dephasing)
684FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNGCTGTACAAGTAAGTCGACG	adapted ITR-seq (dephasing)

Name	Sequence	Application
685FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNGCTGTACAAGTAAGTCGACG	adapted ITR-seq (dephasing)
686FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNNGCTGTACAAGTAAGTCGACG	adapted ITR-seq (dephasing)
687FB	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNNNGCTGTACAAGTAAGTCGACG	adapted ITR-seq (dephasing)
688FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GGTTCGCAGCTTCGTACATG	adapted ITR-seq (dephasing)
689FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNGTTCGCAGCTTCGTACATG	adapted ITR-seq (dephasing)
690FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNGTTCGCAGCTTCGTACATG	adapted ITR-seq (dephasing)
691FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNNGTTCGCAGCTTCGTACATG	adapted ITR-seq (dephasing)
692FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNNNGTTCGCAGCTTCGTACATG	adapted ITR-seq (dephasing)
693FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNNNNGTTCGCAGCTTCGTACATG	adapted ITR-seq (dephasing)
694FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNNNNNGTTCGCAGCTTCGTACATG	adapted ITR-seq (dephasing)
695FB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GNNNNNNNGTTCGCAGCTTCGTACATG	adapted ITR-seq (dephasing)
38FB	GCCTCTAGATTAGAAAAACTCATCGAGCATC	cloning (base construct sITR)
39FB	TACGGAATTCGTTTTTCCATAGGCTCCG	cloning (base construct sITR)
40FB	TAAACGAATTCCGTCGAGGTTACATAACTTACG	cloning (base construct sITR)
41FB	GGCGACCGGTGGATCCGGATC	cloning (base construct sITR)
1FB	TTCGTCTCTGCCTCAGTGAGCGAGCGA	cloning (1st generation sITR)
284FB	TCGACGTACTGAGACCAATAGGTCTCTCTAGCA GTCGCATA	cloning (2nd generation sITR)
285FB	CTAGTATGCGACTGCTAGAGAGACCTATTGGTC TCAGTACG	cloning (2nd generation sITR)
93FB	TTCGTCTCTGCCTCAGTGAGCGAGCGAGCGCGC AGAGAGGGAGTG	cloning (2nd generation sITR)
176FB	GTACAAGTAAGTCGACCATGTACTAGAGCAGG TAGGAACCTGCGGCACTAGTGAAGGAG	cloning (2nd generation sITR)
177FB	CTAGCTCCTTCACTAGTGCCGCAGGTTCCTACC TGCTCTAGTACATGGTCGACTTACTT	cloning (2nd generation sITR)
289FB	GAGCGCGCAGAGAGGGAGTGGCCAACTCCATC ACTAGGGGTTCCTGCG	cloning (2nd generation sITR)
523FB	TCGTCTCTGGGAGTGGCCAACTCCATCACTAGG GGTTCCTGAATTCCGTCGAGGTTAC	cloning (2xRBE*)
477FB	TCGTCTCTGGGAGTGGCCAACTCCATCACTAGG GGTTCCTGGCGTTTTTCCATAGGCTCC	cloning (2xRBE*)
478FB	TCGTCTCTTCCCTCTCTGCGCGCTCGCTCGCTCA CTGAGGCTGCGCGGAACCCCTATTTG	cloning (2xRBE*)
663FB	TCGTCTCTTAGTGGCGTTTTTCCATAGGCTCCG	cloning (Luc-switching reporter 1)

Name	Sequence	Application
653FB	TCGTCTCTAGTTATGTAACCTCGACGGAATTCAG	cloning (Luc-switching reporter 1)
654FB	TCGTCTCTAACTCGACTCACTATAGGCTAGCCAC	cloning (Luc-switching reporter 1)
655FB	TCGTCTCTTCCCTCTCTGCGCGCTCGCTCGCTCACTGAGGCCCAACAGTACCGGAATGCC	cloning (Luc-switching reporter 1)
656FB	TCGTCTCTGGGAGTGGCCAACTCCATCACTAGGGGTTCCTGTTGGTAAAGCCACCATGGC	cloning (Luc-switching reporter 1)
657FB	TCGTCTCTACTAGTAATTGTTGTTGTTAACTTGTTTATTGCAGC	cloning (Luc-switching reporter 1)
671FB	TTCGTCTCTAGGTAAGTAGGCATTCCGGTACTGTTGG	cloning (Luc-switching reporter 2)
672FB	TTCGTCTCTGGCCTGAGGGAGGGAAAGGTGGCTTTACCAACAGG	cloning (Luc-switching reporter 2)
673FB	TTCGTCTCTGGCCGATGCTAAGAACATTAAGAAG	cloning (Luc-switching reporter 2)
674FB	TTCGTCTCTAATGTAATAGATCGTTAGTAGCTAGCCTATAGTGAGTCGAGTTATG	cloning (Luc-switching reporter 2)
675FB	TTCGTCTCTCATTTTTTCTTTACAGGATAAGATCAAGGCCATCGTC	cloning (Luc-switching reporter 2)
676FB	TTCGTCTCTGAAGCTTTTAAGCGGGTCGCTG	cloning (Luc-switching reporter 2)
677FB	TTCGTCTCTCTTCACCATGGCTTCCAAGGTG	cloning (Luc-switching reporter 2)
678FB	TTCGTCTCTACCTGGTGCTCGTAGGAGTAGTG	cloning (Luc-switching reporter 2)
51FB	CCATTCTCTATAAGTTTAAACCGCTGATCAGCCT	cloning (CMV-Rep)
52FB	CCTCTGGCGAGCCTGTCTGCGTAGTTGATCGAAGC	cloning (CMV-Rep)
16FB	TTTTGCTAGCATGCCGGGGTTTTACGAGAT	cloning (CMV-Rep)
17FB	TTTTTCTAGATTATTGTTCAAAGATGCAGTCATCCA	cloning (CMV-Rep)
304FB	TTCGTCTCTGCCACAGTCTGTAAGCCACTGTGCTGGATATCTGC	cloning (CMV-Rep)
305FB	TTCGTCTCTTGGCCTCGTGCTAGCCTGTCTGCGTAGTTGATCGAAGC	cloning (CMV-Rep)
427FB	TTCGTCTCTCTAGCCACCATGCCGGGGTTTTACGAGATTGTG	cloning (CMV-Rep)
428FB	TTCGTCTCTAATTCTTACAGACTGTGGCCTCGTGCTAGCCTGTCTG	cloning (CMV-Rep)
659FB	TCGTCTCTTTCTTGCTCCCCAAAACCCAG	cloning (CMV-Rep Y156F)
660FB	TCGTCTCTAGAAATTGGGGATGTAGCACTCATC	cloning (CMV-Rep Y156F)
661FB	TCGTCTCTCACACCAACATCGCGGAGG	cloning (CMV-Rep K340H)
196FB	TTCGTCTCTGCCTAGTGAGCGAGCGAGCGCGCAGAGAGGGAGTG	cloning (VS2)
201FB	TTCGTCTCTGCCTGCAGTGAGCGAGCGAGCGCGCAGAGAGGGAGTG	cloning (VS3)
290FB	GAGCGCGCAGAGAGGGAGTGGCCAACTCCATCACTAGAGGTTCCTGCG	cloning (VS4)

Name	Sequence	Application
93FB	TTCGTCTCTGCCTCAGTGAGCGAGCGAGCGCGC AGAGAGGGAGTG	cloning (VS4)
98FB	AGGCCGGGCGACCAAAGGTCGCCCGACGCCCG GGCTTTGCCCGGGCG	cloning (VS2-4)
99FB	AGGCCGCCCGGGCAAAGCCCGGGCGTCGGGCG ACCTTTGGTCGCCCG	cloning (VS2-4)

2.1.11 Plasmids

Table 12: List of plasmids cloned and used in this work.

Plasmid ID	Name	Origin
Addgene ID #61591	pX601-AAV-CMV::NLS-SaCas9-NLS-3xHA-bGHpA;U6::BsaI-sgRNA	Feng Zhang (Addgene plasmid #61591)
Addgene ID #113077	pEvolvR-enCas9-PolI3M-TBD	John Dueber & David Schaffer (Addgene plasmid #113077)
pSiCheck2	pSiCheck2	Promega (Madison, USA)
pBlueScript	PBlueScript II SK(+)	Agilent (Santa Clara, USA)
#0183	WHc2 (SpeI)	Grimm Lab
#0714	pSSV9-pSiCheck	Grimm Lab
#1111	AdH	Grimm Lab
#1558	AdH Rep (pDGΔVP)	Grimm Lab
#1729	WH-Rep2-Cap1WT	Grimm Lab
#1736	WH-Rep2-Cap2WT	Grimm Lab
#1743	WH-Rep2-Cap3bWT	Grimm Lab
#1750	WH-Rep2-Cap4WT	Grimm Lab
#1757	WH-Rep2-Cap5WT	Grimm Lab
#1764	WH-Rep2-Cap6WT	Grimm Lab
#1771	WH-Rep2-Cap7WT	Grimm Lab
#1778	WH-Rep2-Cap8WT	Grimm Lab
#2887	CMV-mCherry/U6-ccdB	Grimm Lab
ITR variant screening		
FB14	FB14 SBC: ITR2_Kan_ori_ CMV_mCherry_pA_ITR2 (ori +amp)	This thesis
FB36	FB36 sITR – INS no BC	This thesis
FB64	FB64 sITR-INS (AACAAGT)	This thesis
FB68	FB68 sITR-INS (AGAATAC)	This thesis
FB69	FB69 sITR-INS (AGAGTTG)	This thesis
FB72	FB72 sITR-INS (ATAACGC)	This thesis
FB77	FB77 sITR-INS (CCTTTAT)	This thesis
FB78	FB78 sITR-INS (CGTGGAA)	This thesis
FB80	FB80 sITR-INS (GAAGTGC)	This thesis
FB82	FB82 sITR-INS (GATTTAG)	This thesis
FB83	FB83 sITR-INS (GCATGGT)	This thesis
FB84	FB84 sITR-INS (GCTTAAC)	This thesis
FB85	FB85 sITR-INS (GGGAAAC)	This thesis
FB86	FB86 sITR-INS (GGGATTG)	This thesis
FB87	FB87 sITR-INS (GGTGCTA)	This thesis
FB90	FB90 sITR-INS (TAAGAGG)	This thesis
FB98	FB98 sITR-INS (AACTTAC)	This thesis
FB99	FB99 sITR-INS (ACCGACA)	This thesis
FB111	FB111 sITR-INS (CTAGGTC)	This thesis
FB51	FB51 sITR AptCV_BC (AGAATAC)	This thesis
FB52	FB52 sITR AptMG_BC(AGAGTTG)	This thesis

Plasmid ID	Name	Origin
FB53	FB53 sITR wtAAV2 _BC(ATAACGC)	This thesis
FB54	FB54 sITR 2loops_BC(CCTTTAT)	This thesis
FB55	FB55 sITR 15 TA Hairpin_BC(CGTGGAA)	This thesis
FB56	FB56 sITR 17 AAV2-loop-indent_BC(GAAGTGC)	This thesis
FB57	FB57 sITR 19 AAV2-AT-loop_BC(GATTTAG)	This thesis
FB58	FB58 sITR 20 AAV2-loop_BC(GCATGGT)	This thesis
FB59	FB59 sITR 21 AAV2triple_BC(GCTTAAC)	This thesis
FB60	FB60 sITR22 wtAAV1_BC(GGGAAAC)	This thesis
FB61	FB61 sITR27 PorcParvo_BC(TAAGAGG)	This thesis
FB134	FB134 sITR wtAAV3b_BC(CTAGGTC)	This thesis
FB135	FB135 sITR wtAAV4_BC(AACTTAC)	This thesis
FB125	FB125 sITR wtAAV5_BC(GGGATTG)	This thesis
FB136	FB136 sITR wtAAV6_BC(ACCGACA)	This thesis
FB137	FB137 sITR wtAAV7_BC(CACTGAA)	This thesis
FB126	FB126 sITR AeDV_BC(GGTGCTA)	This thesis
FB185	FB185 ITR INS (Ins-ITR: BsmBI; INS-BC:BsaI)	This thesis
FB270	FB270 sITR-INS (AGACTCGTTGTATAT)	This thesis
FB271	FB271 sITR-INS (TAGAGATTTAAACCG)	This thesis
FB272	FB272 sITR-INS (CGTGACAGCGGATGG)	This thesis
FB273	FB273 sITR-INS (TGGGCGGTCAGGGTC)	This thesis
FB274	FB274 sITR-INS (TTGCCGTCCTTCGAG)	This thesis
FB275	FB275 sITR-INS (TTCAGCGGACGGGCC)	This thesis
FB276	FB276 sITR-INS (GTCAGTCCGCTCTTT)	This thesis
FB277	FB277 sITR-INS (TTAAGATCCTGGTCG)	This thesis
FB278	FB278 sITR-INS (TCAACATGGGCAACG)	This thesis
FB279	FB279 sITR-INS (CTTGATCGACGCCCA)	This thesis
FB280	FB280 sITR-INS (TACGCTATTCAATCT)	This thesis
FB281	FB281 sITR-INS (GTGCTTCTGGCGGAT)	This thesis
FB282	FB282 sITR-INS (CGGCTGTCGGTCGCC)	This thesis
FB283	FB283 sITR-INS (ATCGTACGTTACTGA)	This thesis
FB284	FB284 sITR-INS (GATTCGAAAGCATAG)	This thesis
FB285	FB285 sITR-INS (CGTATCGGGTCCGGA)	This thesis
FB286	FB286 sITR-INS (TGGTTGGGTTTGTGG)	This thesis
FB287	FB287 sITR-INS (TCGTTGTAACGGTAC)	This thesis
FB289	FB289 sITR-INS (GACCACTAGAAGGGC)	This thesis
FB290	FB290 sITR-INS (CTGCATGGCGGAGTT)	This thesis
FB291	FB291 sITR-INS (TCAACGATTGTCTGG)	This thesis
FB292	FB292 sITR-INS (ACGTCGCACCGTTTG)	This thesis
FB293	FB293 sITR-INS (TAACGTTGGGTTGCC)	This thesis
FB294	FB294 sITR-INS (CAGGCTTAACGCGGG)	This thesis
FB295	FB295 sITR-INS (ACCATAGCGCCACGA)	This thesis
FB296	FB296 sITR-INS (GTCCCGACTAGGACT)	This thesis
FB297	FB297 sITR-INS (GTCTTGATTGCTTCG)	This thesis
FB298	FB298 sITR-INS (ATTTGGCACAGGATG)	This thesis
FB299	FB299 sITR-INS (GGCCACCGTGTGTGA)	This thesis
FB300	FB300 sITR-INS (ATGAGCAGCGAATGA)	This thesis
FB301	FB301 sITR-INS (ATGTTTAACGGCATA)	This thesis
FB302	FB302 sITR-INS (TTGGACTCACAGATG)	This thesis
FB303	FB303 sITR-INS (AAGGTGACCTAGTGT)	This thesis
FB304	FB304 sITR-INS (CCCTCATGAGGTCCG)	This thesis
FB305	FB305 sITR-INS (ATGACAATGTGCAGG)	This thesis
FB306	FB306 sITR-INS (GCGAGGTCGTTAGTT)	This thesis
FB307	FB307 sITR-INS (TAAGACTGTTCCGGG)	This thesis
FB308	FB308 sITR-INS (GTTTGTAATCTCTAC)	This thesis

Plasmid ID	Name	Origin
FB309	FB309 sITR-INS (GTTAACGCGGCCATT)	This thesis
FB310	FB310 sITR-INS (AGCGGCGTTTATCGT)	This thesis
FB311	FB311 sITR-INS (TTGGTATGTGTCAAT)	This thesis
FB313	FB313 sITR-INS (GAGCGTAATTGTGAG)	This thesis
FB315	FB315 sITR-INS (GTGACATGCAGGTAG)	This thesis
FB316	FB316 sITR-INS (ACGATCGTACGTCTT)	This thesis
FB317	FB317 sITR-INS (GTTCAGGTCAGGTCT)	This thesis
FB364	FB364 sITR-INS (TAAGGAGGGCTGTAG)	This thesis
FB365	FB365 sITR-INS (TATCAAGCTAACGTT)	This thesis
FB366	FB366 sITR-INS (GCTCTGGATGTAGTA)	This thesis
FB367	FB367 sITR-INS (TAGATGTGGCGGACA)	This thesis
FB368	FB368 sITR-INS (GTCAACATCGTTACA)	This thesis
FB369	FB369 sITR-INS (GGGCCCTAGCGCGTG)	This thesis
FB370	FB370 sITR-INS (GATAGGCTGGTCCAA)	This thesis
FB371	FB371 sITR-INS (TATTTGTGTCGTTCC)	This thesis
FB372	FB372 sITR-INS (AGTTAGGGCGCTGCG)	This thesis
FB373	FB373 sITR-INS (GCGGAACATAGGCGG)	This thesis
FB374	FB374 sITR-INS (GCCCTTCAGTCAGCT)	This thesis
FB375	FB375 sITR-INS (CGGTCGCGTGACGTG)	This thesis
FB377	FB377 sITR-INS (CGAGTCGTATGTGGC)	This thesis
FB382	FB382 sITR-INS (TCCACGGAGGCTGCG)	This thesis
FB386	FB386 sITR-INS (TGGTTTACAAATTAT)	This thesis
FB387	FB387 sITR-INS (TGTCCGGAAAGGACA)	This thesis
FB325	FB325 WT-BC(AGACTCGTTGTATAT)	This thesis
FB393	FB393 V1-BC(TAGAGATTTAAACCG)	This thesis
FB394	FB394 V2-BC(CGTGACAGCGGATGG)	This thesis
FB395	FB395 V3-BC(TGGGCGGTCAGGGTC)	This thesis
FB396	FB396 V4-BC(TTGCCGTCCTTCGAG)	This thesis
FB397	FB397 V5-BC(TTCAGCGGACGGGCC)	This thesis
FB398	FB398 V6-BC(GTCAGTCCGCTCTTT)	This thesis
FB399	FB399 V7-BC(TTAAGATCCTGGTCG)	This thesis
FB400	FB400 V8-BC(TCAACATGGGCAACG)	This thesis
FB401	FB401 V9-BC(CTTGATCGACGCCCA)	This thesis
FB402	FB402 V10-BC(TACGCTATTCAATCT)	This thesis
FB403	FB403 V11-BC(GTGCTTCTGGCGGAT)	This thesis
FB404	FB404 V12-BC(CGGCTGTCGGTCGCC)	This thesis
FB405	FB405 V13-BC(ATCGTACGTTACTGA)	This thesis
FB406	FB406 V14-BC(GATTCGAAAGCATAG)	This thesis
FB407	FB407 V15-BC(CGTATCGGGTCCGGA)	This thesis
FB408	FB408 V16-BC(TGGTTGGGTTTGTGG)	This thesis
FB409	FB409 V17-BC(TCGTTGTAACGGTAC)	This thesis
FB410	FB410 V18-BC(TGGTTTACAAATTAT)	This thesis
FB411	FB411 V19-BC(GACCACTAGAAGGGC)	This thesis
FB412	FB412 V20-BC(CTGCATGGCGGAGTT)	This thesis
FB413	FB413 V21-BC(TCAACGATTGTCTGG)	This thesis
FB414	FB414 V22-BC(ACGTCGCACCGTTTG)	This thesis
FB415	FB415 V23-BC(TAACGTTGGGTTGCC)	This thesis
FB416	FB416 V24-BC(CAGGCTTAACGCGGG)	This thesis
FB417	FB417 V25-BC(ACCATAGCGCCACGA)	This thesis
FB418	FB418 V26-BC(GTCCCGACTAGGACT)	This thesis
FB419	FB419 V27-BC(GTCTTGATTGCTTCG)	This thesis
FB420	FB420 V28-BC(ATTTGGCACAGGATG)	This thesis
FB421	FB421 V29-BC(GGCCACCGTGTGTGA)	This thesis
FB422	FB422 V30-BC(ATGAGCAGCGAATGA)	This thesis

Plasmid ID	Name	Origin
FB423	FB423 V31-BC(ATGTTTAACGGCATA)	This thesis
FB424	FB424 V32-BC(TTGGACTCACAGATG)	This thesis
FB425	FB425 V33-BC(AAGGTGACCTAGTGT)	This thesis
FB426	FB426 V34-BC(CCCTCATGAGGTCCG)	This thesis
FB427	FB427 V35-BC(ATGACAATGTGCAGG)	This thesis
FB428	FB428 V36-BC(GCGAGGTCGTTAGTT)	This thesis
FB429	FB429 V37-BC(TAAGACTGTTCCGGG)	This thesis
FB430	FB430 V38-BC(GTTTGTAATCTCTAC)	This thesis
FB431	FB431 V39-BC(GTTAACGCGGCCATT)	This thesis
FB432	FB432 V40-BC(AGCGGCGTTTATCGT)	This thesis
FB433	FB433 V41-BC(TTGGTATGTGTCAAT)	This thesis
FB434	FB434 V42-BC(TGTCCGGAAAGGACA)	This thesis
FB435	FB435 V43-BC(GAGCGTAATTGTGAG)	This thesis
FB436	FB436 V44-BC(GTGACATGCAGGTAG)	This thesis
FB437	FB437 V45-BC(ACGATCGTACGTCTT)	This thesis
FB438	FB438 V46-BC(GTTCAGGTCAGGTCT)	This thesis
FB439	FB439 V47-BC(TAAGGAGGGCTGTAG)	This thesis
FB440	FB440 V48-BC(TATCAAGCTAACGTT)	This thesis
FB441	FB441 V49-BC(GCTCTGGATGTAGTA)	This thesis
FB442	FB442 V50-BC(TAGATGTGGCGGACA)	This thesis
FB443	FB443 V51-BC(GTCAACATCGTTACA)	This thesis
FB444	FB444 V52-BC(GGGCCCTAGCGCGTG)	This thesis
FB445	FB445 V53-BC(GATAGGCTGGTCCAA)	This thesis
FB446	FB446 V54-BC(TATTTGTGTCGTTCC)	This thesis
FB454	FB454 V55-BC(AGTTAGGGCGCTGCG)	This thesis
FB455	FB455 V56-BC(GCGGAACATAGGCGG)	This thesis
FB456	FB456 V57-BC(GCCCTTCAGTCAGCT)	This thesis
FB457	FB457 V58-BC(CGGTCGCGTGACGTG)	This thesis
FB458	FB458 V59-BC(CGAGTCGTATGTGGC)	This thesis
FB459	FB459 V60-BC(TAAGACTGTTCCGGG)	This thesis
FB460	FB460 V61-BC(TAGAGATTTAAACCG)	This thesis
FB461	FB461 V62-BC(CGTGACAGCGGATGG)	This thesis
FB462	FB462 V63-BC(TGGGCGGTCAGGGTC)	This thesis
FB463	FB463 V64-BC(TTGCCGTCCTTCGAG)	This thesis
FB464	FB464 V65-BC(TTCAGCGGACGGGCC)	This thesis
FB465	FB465 V66-BC(GTCAGTCCGCTCTTT)	This thesis
FB466	FB466 V67-BC(TTAAGATCCTGGTCG)	This thesis
FB467	FB467 V68-BC(TCAACATGGGCAACG)	This thesis
FB468	FB468 V69-BC(CTTGATCGACGCCCA)	This thesis
FB469	FB469 V70-BC(TACGCTATTCAATCT)	This thesis
FB470	FB470 V71-BC(GTGCTTCTGGCGGAT)	This thesis
FB471	FB471 V72-BC(TATTTGTGTCGTTCC)	This thesis
FB472	FB472 V73-BC(ATCGTACGTTACTGA)	This thesis
FB473	FB473 V74-BC(GATTCGAAAGCATAG)	This thesis
FB474	FB474 V75-BC(CGTATCGGGTCCGGA)	This thesis
FB475	FB475 V76-BC(TGGTTGGGTTTGTGG)	This thesis
FB476	FB476 V77-BC(TCGTTGTAACGGTAC)	This thesis
FB477	FB477 V78-BC(TAACGTTGGGTTGCC)	This thesis
FB478	FB478 V79-BC(GACCACTAGAAGGGC)	This thesis
FB479	FB479 V80-BC(CTGCATGGCGGAGTT)	This thesis
FB480	FB480 V81-BC(TCAACGATTGTCTGG)	This thesis
FB481	FB481 V82-BC(ACGTCGCACCGTTTG)	This thesis
FB482	FB482 V83-BC(GTTTGTAATCTCTAC)	This thesis
FB483	FB483 V84-BC(CAGGCTTAACGCGGG)	This thesis

Plasmid ID	Name	Origin
FB484	FB484 V85-BC(ACCATAGCGCCACGA)	This thesis
FB485	FB485 V86-BC(GTCCCGACTAGGACT)	This thesis
FB486	FB486 V87-BC(GTCTTGATTGCTTCG)	This thesis
FB487	FB487 V88-BC(ATTTGGCACAGGATG)	This thesis
FB488	FB488 V89-BC(GGCCACCGTGTGTGA)	This thesis
FB489	FB489 V90-BC(ATGAGCAGCGAATGA)	This thesis
FB539	FB539 Y1-BC(TTCAGCGGACGGGCC)	This thesis
FB540	FB540 Y2 -BC(TCAACATGGGCAACG)	This thesis
FB541	FB541 Y3-BC(TACGCTATTCAATCT)	This thesis
FB542	FB542 Y4-BC(CGGCTGTCGGTCGCC)	This thesis
FB543	FB543 Y5-BC(ATCGTACGTTACTGA)	This thesis
FB544	FB544 Y6-BC(GATTCGAAAGCATAG)	This thesis
FB545	FB545 Y7-BC(CGTATCGGGTCCGGA)	This thesis
FB546	FB546 Y8-BC(TGGTTGGGTTTGTGG)	This thesis
FB547	FB547 Y9-BC(GACCACTAGAAGGGC)	This thesis
FB548	FB548 Y10-BC(ACGTCGCACCGTTTG)	This thesis
FB549	FB549 Y11-BC(CAGGCTTAACGCGGG)	This thesis
FB550	FB550 Y12-BC(ATTTGGCACAGGATG)	This thesis
FB551	FB551 Y13-BC(GGCCACCGTGTGTGA)	This thesis
FB552	FB552 Y14-BC(ATGAGCAGCGAATGA)	This thesis
FB553	FB553 Y15-BC(ATGTTTAACGGCATA)	This thesis
FB554	FB554 Y16-BC(TCCACGGAGGCTGCG)	This thesis
FB555	FB555 Y17-BC(TAGAGATTTAAACCG)	This thesis
FB562	FB562 VS2-BC(TAGAGATTTAAACCG)	This thesis
FB563	FB563 VS3-BC(GTCAGTCCGCTCTTT)	This thesis
FB564	FB564 VS4-BC(CTTGATCGACGCCCA)	This thesis
Circular AAV project		
FB503	FB503 AAV circ V2 (2xRBE*)	This thesis
FB535	FB535 2xRBE* AAV LuciferaseSwitch 1	This thesis
FB538	FB538 2xRBE AAV LuciferaseSwitch 2.0	This thesis
FB4	FB4 pcDNA Rep78	This thesis
FB18	FB18 pcDNA Rep68 (no Kozak)	This thesis
FB319	FB319 pcDNA Rep68	This thesis
FB536	FB536 pcDNA Rep68 (Y156F)	This thesis
FB537	FB537 pcDNA Rep68 (K340H)	This thesis

2.2 Methods

2.2.1 Molecular Biology Methods

2.2.1.1 In Silico Cloning

Plasmid maps, cloning strategy and primer design as well as the analysis of Sanger sequencing results were performed with Geneious V7.1.7. Primer annealing temperatures were determined using the NEB Tm calculator (https://tmcalculator.neb.com). Potential primer secondary structures as well as secondary structures of the ITR sequences were predicted using the mFold application for DNA on the UNAFold Web Server (http://www.unafold.org/) [236].

2.2.1.2 Polymerase chain reaction

Polymerase chain reaction (PCR) for amplification of DNA templates was performed using the Q5 High-Fidelity 2x Master Mix (NEB) or Phusion Flash High-Fidelity PCR Master Mix (Thermo Fisher Scientific) according to the manufacturer's protocol. In brief, 500 nM of forward and reverse primers were used, with the exception of PCR primers with a binding site within the ITR region in plasmids. There, a single primer was used with a final concentration of 1 µM. For plasmid DNA templates 1 ng was used as template, for genomic DNA typically 150 ng were used. Primer annealing temperatures were used as suggested by the manufacturer or adapted if necessary. Elongation times were adjusted to the template size, as indicated in Table 13 showing the cycling conditions with the used polymerases.

Table 13: PCR cycling conditions for Q5 and PhusionFlash Polymerase

Step	Temperature [°C]	Time [min:sec] – Q5	Time [min:sec] – Phusion	Cycles
Initial denaturation	98	0:30	0:30	x1
Denaturation	98	0:10	0:10	
Annealing	58-72	0:15	0:15	x30
Elongation	72	30 sec/kb	30 sec/kb	
Final Elongation	72	2:00	2:00	x1
Hold	4-10	∞	∞	

For verification of some multi-fragment cloning steps, colony PCRs were performed. For this purpose, single bacterial colonies from agar plates were diluted in 50 µl H_2O. 2 µl of this dilution served as template for the PCR reaction with the OneTaq Quick-Load 2x Master mix (NEB) with 200 nM of forward and reverse primers in a total reaction volume of 15 µl according to the manufacturer's recommendation. The initial denaturation step at 94°C (30 sec) was followed by 30 cycles of denaturation at 94°C (15 sec) annealing at 45-68°C (15 sec) followed by the elongation step at 68°C (1 min/kb). After a final elongation step at 68°C for 5 min, the sample was held at 10°C.

For direct purification of the PCR product the QIAquick PCR purification Kit or the QIAquick Nucleotide Removal Kit (both Qiagen) were used according to the manufacturer's instructions.

2.2.1.3 Agarose gel electrophoresis

DNA fragments were separated by agarose gel electrophoresis. The gels were prepared by dissolving 1% agarose (Biozym) in 1x TAE buffer. To enable DNA visualization under UV light irradiation, the DNA was mixed with 6x TriTrack DNA loading dye (ThermoScientific)

supplemented with 1% 10.000x GelRed Nucleic Acid Gel stain (Biozol). As GelRed in the loading can lead to overestimation of the DNA band height, alternatively the 10.000x GelRed Nucleic Acid Gel stain (Biozol) was added directly to the dissolved agarose to reach a 1x solution. The third option was to use 0.5 µg/ml ethidium bromide added before gel hardening. Electrophoresis was performed at 100 to 130 V. The DNA was then detected with UV light and compared the size of the DNA was determined by comparison with the GeneRuler DNA Ladder Mix or 1kb Plus DNA ladder (both ThermoScientific) mixed with GelRed supplemented TriTrack DNA loading dye or GelRed-free TriTrack DNA loading dye when using GelRed or ethidium bromide in the gel. Imaging was performed using the GelDoc XR system (Bio-Rad) or the Azure 300 (AzureBiosystems).

For purification of DNA fragments excised from agarose gels, the QIAquick Gel Extraction Kit (Qiagen) was employed, following the manufacturer's instruction.

2.2.1.4 Annealing of DNA oligonucleotides

Annealing of oligonucleotides for subsequent use in cloning or NGS pipelines was performed by mixing 5 µl of each oligo (100 µM) with 5 µl NEB buffer 2 (NEB) and 35 µl H_2O. The mixture was placed in a thermocycler and heated to 95°C for 5 minutes followed by a cooldown to 25°C with a slow ramping speed of 0.1°C /sec.

2.2.1.5 De-/Phosphorylation of PCR products

PCR product phosphorylation for blunt end cloning was performed using the T4 Polynucleotide Kinase (PNK; NEB). For this purpose, 500 ng of purified PCR products were phosphorylated with 1 µl T4 PNK, 2 µl T4 PNK reaction buffer and adjusted with H_2O to a final volume of 20 µl with H_2O. Reactions were incubated at 37°C for 30 min and heat in activated at 65°C for 20 min. For dephosphorylation 500 ng of the purified PCR product were mixed with 1 µl of Antarctic phosphatase (NEB) with 2 µ(a) of Antarctic phosphatase reaction buffer (10x) in a total volume of 20 µl, and incubated at 37°C for 30 min. For both phosphorylation and dephosphorylation the fragments were then purified using the QIAquick PCR purification kit (Qiagen).

2.2.1.6 Preparation of chemically competent cells

Chemocompetent cells were prepared with One Shot Top10 *E. coli* stock (Invitrogen). Cells were thawed on ice and 5 µl were used to inoculate 50 ml of Lysogeny Broth (LB) medium

and incubated in a shaking incubator for 16 h at 37°C at 180 rpm. Then, 10 ml of the culture was used to inoculate 400 ml LB-medium. The OD600 was measured and once reaching a value between 0.5-0.6, cells were spun down for 20 min at 3570 g at 4°C. Supernatant was discarded and the cell pellets were resuspended in 200 ml pre-cooled 100 mM $CaCl_2$ solution and incubated on ice for 30 min. Then the cells were centrifuged again for 20 min at 3570 g at 4°C, the supernatant discarded, and the cell pellet dissolved in 80 ml pre-cooled 100 mM CaCl2 supplemented with 10% glycerol. 100 µl aliquots were generated, snap-frozen in liquid N_2 and stored at -80°C until usage for transformation.

2.2.1.7 Ligation and transformation

Ligations of different pre-digested amplicons and/or annealed oligonucleotides were typically performed at a 5:1 (insert: backbone) ratio and a total DNA content of 100-200 ng. For the reaction, 2 µl T4 Ligase Buffer and 1 µl T4 Ligase (both NEB) were added to the DNA and the volume was adjusted to 20 µl with H_2O. Ligations were incubated at room temperature or 21°C in a thermocycler for 30 min before inactivation of the ligase at 70°C for 10 min. Blunt end ligations were incubated at 12°C for 16 hours.

For transformation of chemically competent *E. coli* Top10 the cells were thawed on ice. Then, the ligation reaction was added to the cells at a volume not exceeding 10% of the cell volume. For transformation of purified plasmid DNA (retransformation) a maximum of 10 ng DNA were added to the cells. Transformations were performed at volumes of 25 µl (for increased throughput), 50 µl or 100 µl. At the two higher volumes, the samples were kept on ice for 10-30 minutes followed by a heat shock at 42°C for 45 seconds in a heating block and a subsequent incubation on ice for another 5 minutes. For the low volume transformation, the samples were kept in a thermocycler at 4°C for 25 min followed by a 42°C heat shock for 30 sec and another incubation step at 4°C for 5 minutes. The program was run with maximum ramp speed.

For selection on ampicillin or carbenicillin, the cells were immediately plated after transformation. For selection on kanamycin or a combination of kanamycin and ampicillin, cells were recovered for 45 min at 37°C at 180 rpm in 100, 500 or 1000 µl of antibiotic-free LB medium, depending on the scale of the transformation. Cells were then spun down; excess supernatant was removed and resuspended in 50 µl remnant volume for plating.

2.2.1.8 Golden Gate Assembly

Cloning via Golden Gate assembly was used for cloning of DNA fragments with two terminal Type IIS restriction enzyme cleavage sites, the attachment of annealed oligonucleotides to a whole plasmid PCR amplicon with two terminal Type IIS cleavage sites as well as for the insertion of annealed oligonucleotides into an acceptor plasmid DNA with two such cleavage sites. For this purpose, the DNA fragments or annealed oligonucleotides were mixed with an insert: vector ratio of 10:0, then 2 µl T4 ligase buffer, 1 µl T4 ligase as well as the type IIS restriction enzyme, *e.g.*, BsaI or Esp3I (all NEB) and adjusted to a total volume of 20 µl. The mixture was incubated in a thermocycler for 30 cycles at 37°C for 5 min and 16°C for min followed by an inactivation step at 80°C for 10 min. The product was subsequently used for transformation like a ligation product.

2.2.1.9 Bacterial culture conditions and plasmid preparation

Bacteria were cultured in LB medium supplemented with ampicillin (50 µg/ml), carbenicillin (100 µg/ml) or kanamycin (50µg/ml) or a combination of both ampicillin and kanamycin depending on the resistance genes present on the plasmid. For preparation of agar plates, the LB medium was supplemented with 1.5% (w/v) agar and the antibiotic was used at the same final concentrations as for liquid medium. Liquid cultures for the propagation of plasmids with large ITR sequences were performed in Terrific Broth (TB) medium supplemented with kanamycin (50 µg/ml). Bacteria on agar plates and in liquid cultures were grown for 16 hours at 37°C. To reduce recombination events when carrying plasmids with large ITR structures or sequences prone to recombination, the temperature was reduced to 33°C and the incubation elongated to 20-24 hours.

For small scale plasmid preparations bacteria were grown in a volume of 4 ml LB medium or 6 ml TB medium. Plasmids were then isolated using the QIAprep Spin Miniprep Kit (Qiagen). For mid-scale plasmid preparation (Midi-preps) the volume was set to 50 ml for the QIAGEN Plasmid Plus Midi Kit (Qiagen), 100-150 ml for using the PureYield Plasmid Midiprep System (Promega). Plasmid DNA concentration before commercial ITR-sequencing was performed using the DNA Clean & Concentrator-5 Kit (Biozym). For large-scale plasmid preparation (maxi-Preps) a culture volume of 400 ml was used, and plasmids were isolated using the NucleoBond PC 500 kit (Macherey-Nagel). Liquid cultures were incubated at 37°C for 16 h. When propagating plasmids with sequences that could facilitate recombination or large ITR

structures, the temperature was reduced to 33°C and the incubation time was elongated to 20-24 h.

Small scale plasmid preparations that were not intended for subsequent transfection or were only used as intermediate cloning step were not purified using commercial silica columns but instead via a protocol relying on self-made buffers and isopropanol precipitation. For this purpose, bacteria from liquid cultures were pelleted at 5000 g for 3 min, resuspended in 300 µl P1 (Resuspension Buffer). After 3 min incubation time at room temperature, 300 µl of P2 (Lysis Buffer) were added and the sample was incubated for another 5 min. Subsequently, 300 µl of P3 (Neutralization Buffer) were added, the sample mixed gently, and centrifuged at 21000 g for 10 min. 800 µl of the supernatant were mixed with 600 µl isopropanol and centrifuged again at 21000g for 10 min. The supernatant was completely discarded and pelleted DNA was washed with 500 µl 70% ethanol, centrifuged at 21000 g for 5 min. After discarding the supernatant, the pellet was air dried for 10-20 min and resuspended in 50 µl H2O.

All Kits were used according to the manufacturer's instruction. After extraction, the DNA concentrations were determined using NanoDrop 2000 Spectrophotometer (ThermoScientific).

2.2.1.10 Sanger Sequencing

2.2.1.10.1 Conventional Sanger Sequencing

Sanger sequencing was performed for all plasmids using the Eurofins Genomics Sanger sequencing service. The sequencing reaction was prepared with a total volume of 10 µl, containing 2.5 µM of an appropriate sequencing primer and 100-500 ng DNA template (volume adjusted with H_2O).

Sequencing of plasmids with ITR was initially done after digest with suitable restriction cleavage sites inside the hairpin region, *e.g.*, EcoRI-HF for the single ITR base construct. Therefore, 500 ng plasmid DNA, 1 µl of the restriction enzyme and 2 µl of the associated buffer were combined, adjusted to 20 µl, with H_2O, and incubated for 30 min at 37°C and then inactivated when possible. This restriction enzyme digest resolved the hairpin, and the linearized plasmid DNA was sequenced using the aforementioned sequencing reaction composition using 7.5 µl of the restriction digest as plasmid input. Sequencing results were aligned to the plasmid reference using Geneious V7.1.7.

2.2.1.10.2 Commercial Sanger sequencing for ITR interrogation

The Genewiz AAV-ITR Sequencing service was used as commercial sequencing service for ITR sequence verification. For each sequencing reaction, 5-10 µl of ITR plasmid at concentrations between 200 ng/µl and 300 ng/µl were submitted along with 30 µl of 5 µM primer with appropriate binding sites 150-350 bp upstream of the ITR sequence. Sequencing results were aligned to a plasmid reference sequence using Geneious V7.1.7.

2.2.1.10.3 T7eI-ITR-sequencing

The ITR-sequencing by conventional Sanger sequencing upon pre-digest by T7 endonuclease 1 (T7e1) used different input volumes depending on the length and number of ITR sequences in the plasmid. For plasmids with two ITR sequences of 129 bp (of which 91 nt are forming the hairpin secondary structure) typically 2000 ng were used as input for the pre-digest, or the indicated amount. For plasmids with a single ITR, where all nucleotides participate in secondary structure formation, different ITR sequence lengths were examined. There, the input for the T7eI pre-digest ranged from 250 to 750 ng. The T7e1 pre-digest was performed with 2 µl NEB2 buffer, 1 µl T7e1 (both NEB) and the volume was adjusted to 20 µl with H_2O. The samples were then incubated for 30 min at 37°C in a thermocycler followed by a heat inactivation step at 95°C for 5 min.

For conventional Sanger sequencing, the Eurofins Genomics Sanger sequencing service was used. The sequencing reaction was prepared in a volume of 10 µl with 7.5 µl of the T7e1 digest and 2.5 µM of an appropriate sequencing primer. The sequencing primer binding sites were 200-300 bp upstream of the ITR sequence. Sequencing results were aligned to the plasmid reference containing the intended ITR hairpin mutant variant or the wtITR2 using Geneious V7.1.7.

2.2.1.11 Extraction of DNA and or RNA from cells and tissues

Parallel isolation of DNA and RNA from tissues and cell culture samples was performed using the AllPrep DNA/RNA Mini Kits (Qiagen) according to the manufacturer's instruction with the following adaptations to the protocol.

For cell culture samples, cell lysis was performed in the wells with 600 µl RLT buffer supplemented with 1% β-mercaptoethanol. To reduce DNA fragmentation during sample homogenization, the samples were only homogenized by snap-freezing in liquid N_2 and vortexing. 200 µg of Proteinase K (Roche) were added to the sample and incubated for 15 min

at 55°C. To avoid column blockage, the lysate was then centrifuged at 13000 g for 3 min and the supernatant was used for further processing on the columns.

For extracted murine tissues, a small piece (10-25 mg) was sliced from the tissue sample, mixed with 600 µl RLT buffer supplemented with 1% β-Mercaptoethanol. One 5 mm stainless steel bead (Qiagen) was added to the sample and the tissue was homogenized with the TissueLyser LT (Qiagen) for at least 45 sec at 50 Hz. The lysate was then digested with 200 µg Proteinase K (Roche) for 15 min at 55°C centrifuged for 3 min at 13000 g and the supernatant was used for further processing.

For the RNA purification after loading the samples as indicated by the manufacturer's protocol, the columns were washed with 350 µl RW1. Then an on-column DNase digest using the RNase-Free DNase Set (Qiagen) was performed according to the manufacturer's instructions and incubated at room temperature for 20 min. Subsequently, a second wash step with 350 µl RW1 was performed, followed by two wash steps with 500 µl RPE, a drying step and elution in 30 µl H_2O.

DNA purification was performed according to the manufacturer's instruction with elution in 30 µl H_2O. RNA and DNA concentrations were determined by Nanodrop, and the samples were stored at -80°C.

2.2.1.12 cDNA synthesis

Before cDNA synthesis the RNA was again digested with DNase, to ensure absence of DNA contamination in the cDNA synthesis reaction. For this purpose, up to 20 µl of the purified RNA were mixed with 3 µl of DNaseI Reaction Buffer and 1 µl DNaseI (RNase-free) (both NEB) and the volume was adjusted to 30 µl H_2O. The samples were incubated in a thermocycler for 15 min at 37°C followed by a short heat-inactivation step at 75°C for 10 min.

For cell culture derived RNA the cDNA synthesis was performed with 2000 ng RNA using the Tetro cDNA Synthesis Kit (Meridian Bioscience), used according to the manufacturers instruction with the Oligo$(dT)_{18}$ Primer mix. The sample was then incubated in a thermocycler at 45°C for 30 min and the reaction was terminated via heating to 85°C for 5 min.

For RNA derived from murine samples cDNA synthesis was performed with 800 ng RNA using the SuperScript IV VILO Master Mix with ezDNase Kit (Invitrogen). For the ezDNAse treatment, the RNA was adjusted to a volume of 4 µl and mixed with 0.5 µl 10x ezDNase Buffer and 0.5 µl ezDNase and then incubated for 2 min at 37°C. Then 3 µl H_2O and 2 µl

SuperScript VILO Mastermix or for controls 3 µl H_2O and 2 µl SuperScript IV VILO noRT were added to the samples. The reactions were then incubated for 10 min at 25°C followed by 10 min at 50°C and an inactivation step by heating to 85°C for 5 min. The cDNA was stored at -20°C or directly further processed.

2.2.1.13 Rolling circle amplification

Rolling circle amplification (RCA) was used to enrich circular episomal AAV vector genomes from a DNA input of 60 ng genomic DNA and to amplify circular replication intermediates from 40 ng DpnI pre-digested low molecular weight DNA. For the denaturation mix, 0.5 µl of 10x phi29 DNA Polymerase Reaction Buffer (NEB) were mixed with 1 µl of a 100 µM equimolar mix of suitable phosphothiorated primer (606FB-609FB) and adjusted to 5 µl with H_2O. The mix was then heated to 95°C for 5 min for primer annealing and immediately cooled down on ice. Then 15 µl of the RCA mix, consisting of 1.5 µl 10x phi29 DNA Polymerase Buffer, 1 µl phi29, 2 µl dNTP (all NEB) and 10.5 µl H_2O, were added. The reaction was then incubated for 18 hours at 36°C and inactivated by heating to 65°C for 10 min.

The samples were subsequently adjusted to a volume of 200 µl with H_2O and 10 µl were used for an elongated restriction enzyme digest (12 h) with a single cutter restriction enzyme (EcoRI-HF, BamHI-HF or SpeI-HF) before any further usage, as the restriction digest debranches/dissolves large branched DNA complexes formed during RCA.

2.2.1.14 T5 exonuclease digestion

For digestion of non-circularized vector genomes from genomic DNA (gDNA). 1000 ng gDNA, were mixed with 3 µl NEB4 and 2 µl T5 exonuclease (both NEB) and the volume was adjusted to 30 µl with H_2O. The samples were digested for 23 h at 37°C followed by an inactivation of the T5 exonuclease for 10 min at 95°C.

2.2.2 Specific cloning procedures

2.2.2.1 single ITR constructs

2.2.2.1.1 Dual ITR base construct

To set the basis for generating a single ITR plasmid, a conventional dual-ITR construct had to be modified. To achieve a high modularity, the construct was assembled from fragments from different constructs. Instead of the commonly used pSSV9 construct, which possesses ITR proximal regions from AAV2, a nearly completely synthetic construct was generated. Thus,

instead of pSSV9, the plasmid backbone with ITRs was derived from pX601-AAV-CMV::NLS-SaCas9-NLS-3xHA-bGHpA;U6::BsaI-sgRNA, a gift from Feng Zhang (Addgene ID #61591), and cleaved with XbaI and NotI-HF. The transgene region was assembled from three different fragments: (1) kanamycin (kan) + ori from pEvolvR-enCas9-PolI3M-TBD, a gift from John Dueber & David Schaffer (Addgene ID #113077), (2) CMV promoter from #2887, (3) mCherry + minimal polyA from #2887. Fragment 1 was amplified with primers 38FB and 39FB and cut with XbaI and EcoRI-HF (both NEB). Fragment 2 was amplified with primers 40FB and 41FB and digested with EcoRI-HF and AgeI (both NEB), and fragment 3 was generated by digestion of the original plasmid with AgeI and NotI-HF (both NEB). Upon four fragment assembly ligation, transformation, and co-selection on both kanamycin and ampicillin resistance, the resulting plasmid (FB14) was the basis for the generation of the single ITR plasmids.

2.2.2.1.2 First generation single ITR origin construct

2.2.2.1.2.1 Base construct of the first generation single ITR

For the first generation single ITR construct a single primed PCR was performed with primer 1FB, which results in two PCR products, the plasmid backbone and the transgene region. The PCR product of the transgene region was phosphorylated using T4 PNK (NEB) and ligated. This small plasmid (FB36) contains a single ITR region, with a long hairpin stem and a cloning site at the tip of the hairpin. Sequencing of the hairpin was performed after linearization using the EcoRI-HF restriction cleavage site at the tip of the hairpin.

2.2.2.1.2.2 First generation single ITR barcoded constructs

The single ITR base construct was then equipped with a barcode sequence by digesting the single ITR origin construct with NheI and BsrGI, cleaving out a short (9 bp) fragment. Then a 7 nt barcode sequence was inserted as annealed oligonucleotide (see section 2.2.1.4). The list of the barcodes used in this work are given in Table 14. The oligonucleotides to generate these barcode sequences were designed as follows:

Forward: 5‘ GTAC – AAGTAA[7 nt barcode]

Reverse: 5‘ CTAG – [7 nt barcode (reverse complement)]TTACTT

To ensure hairpin sequence integrity upon cloning, sequencing was performed after plasmid linearization with EcoRI-HF to ensure presence of the cloning site subsequently used for hairpin alteration.

Table 14: Barcode sequences for first generation sITR base constructs.

plasmid	Barcode	plasmid	Barcode
FB64	AACAAGT	FB84	GCTTAAC
FB68	AGAATAC	FB85	GGGAAAC
FB69	AGAGTTG	FB86	GGGATTG
FB72	ATAACGC	FB87	GGTGCTA
FB77	CCTTTAT	FB90	TAAGAGG
FB78	CGTGGAA	FB98	AACTTAC
FB80	GAAGTGC	FB99	ACCGACA
FB82	GATTTAG	FB111	CTAGGTC
FB83	GCATGGT		

2.2.2.1.2.3 First generation barcoded single ITR hairpin variant generation

With the barcode plasmids established, hairpin sequence variants were inserted as annealed oligonucleotides, with the list of primers provided in Table 15. The oligonucleotides to generate the hairpins were designed as follows:

Forward: 5‘ AGGC – [Hairpin sequence]

Reverse: 5‘ AGGC – [Hairpin sequence (reverse complement)]

This cloning step was performed partially as one pot assembly using Golden Gate cloning. The hairpin and barcode sequences were confirmed using Sanger sequencing after digest with suitable restriction enzyme or the T7E1 based ITR sequencing protocol, generating the listed plasmids.

Table 15: First generation sITR plasmids hairpin sequence and barcode combination.

Plasmid	ITR sequence	Barcode
FB51 AptCV	AACGACCACCGGTGCGCCGTACAGGTAACTAGCGTCGTCGTT	AGAATAC
FB52 AptMG	CTCAGATCTAACCTTGTTAAATTGAG	AGAGTTG
FB53 wtAAV2	CGGGCGACCAAAGGTCGCCCGACGCCCGGGCTTTGCCCGGGCG	ATAACGC
FB54 2loops	AAGCCGCCGATAGACGATCAGATGCCGCCGACAATAATCTGATCGTGTATACAATATT	CCTTTAT
FB55 TA Hairpin	AAATATATATTAAATATATTTTAATTAATTTATAAAATTAATT	CGTGGAA
FB56 AAV2-loop-indent	ACGGGCGACCAAAGGTCGCCCGATACGAATACGTCGCCCGGGCTTTGCCCGGGCGCGATC	GAAGTGC
FB57 AAV2-AT-loop	CGGGCGACCAAAGGTCGCCCGATAATAACGCCCGGGCTTTGCCCGGGCG	GATTTAG
FB58 AAV2-loop	CGGGCGACCAAAGGTCGCCCGAGCACAGAGCGCCCGGGCTTTGCCCGGGCG	GCATGGT

Plasmid	ITR sequence	Barcode
FB59 AAV2triple	CGGGCGACCAAAGGTCGCCCGAGCCGATACGGCTCGCCC GGGCTTTGCCCGGGCG	GCTTAAC
FB60 wtAAV1	CGGCAGAGGAGACCTCTGCCGTCTGCGGACCTTTGGTCC GCAG	GGGAAAC
FB61 PorcParvo	TTAAATCAGTACTGCCAATTTTTACTGGAATATACTGTAA CATGGGCTGTTGTGAGGCGAAA	TAAGAGG
FB134 wtAAV3b	CTGGCGACCAAAGGTCGCCAGACGGACGTGCTTTGCACG TCCG	CTAGGTC
FB135 wtAAV4	CTGGAGACCAAAGGTCTCCAGACTGCCGGCCTCTGGCCG GCAG	AACTTAC
FB125 wtAAV5	CGACGGCCAGAGGGCCGTCGTCTGGCAGCTCTTTGAGCT GCCA	GGGATTG
FB136 wtAAV6	CGGCAGAGGAGACCTCTGCCGTCTGCGGACCTTTGGTCC GCAG	ACCGACA
FB137 wtAAV7	CTGCGGACCAAAGGTCCGCAGACGGCAGAGCTCTGCTCT GCCG	CACTGAA
FB126 AeDV	TACAGTTTCTATTAGAAACGATGTATTACATCTGTA	GGTGCTA

2.2.2.1.3 Second Generation single ITR construct

The first generation single ITR constructs had limited applicability due to the short barcode sequence and its proximity to the polyA sequence. Additionally, all first generation single ITR constructs possessed a point mutation at the 5' end of the right hairpin stem. The second generation single ITR construct was designed to circumvent all of these problems.

2.2.2.1.3.1 Second generation base construct

The generation of the second generation single ITR base construct involved multiple cloning steps. First, the first generation single ITR base construct (FB36) was digested with BsrGI and NheI and the annealed oligos 176FB/177FB were ligated into the cloning site, generating FB112. The plasmid was subsequently digested with SalI and SpeI and the annealed oligos 284FB/285FB were inserted, generating FB146. To remove the mutation in the plasmid stem, this plasmid was then used as a template for a PCR with the primer 289FB, followed by a second PCR on the generated amplicon using primer 93FB. The PCR product was then phosphorylated using PNK and ligated, creating plasmid FB185. This construct was the second-generation sITR base construct with two Esp3I sites for hairpin insertion and two BsaI sites for barcode insertion. After each cloning step, the integrity of the hairpin was evaluated by EcoRI-HF digest as well as by Sanger sequencing of the linearized plasmid.

2.2.2.1.3.2 Second generation Barcode insertions

To insert barcodes into the second-generation plasmids, the plasmid was digested using BsaI, the barcode sequences were inserted as annealed oligonucleotides. The oligonucleotides to generate these barcode sequences were designed as follows:

Forward: 5‘ GTAC – [15 nt barcode]

Reverse: 5‘ CTAG – [15 nt barcode (reverse complement)]

A list of generated sITR base constructs with their associated barcode is given in the following Table 16. The hairpin integrity was verified using linearization with EcoRI-HF, the barcode sequence as well as the sequence within the hairpin were verified by Sanger sequencing of the linearized plasmid.

Table 16: Barcode sequences for second generation sITR base constructs.

plasmid	Barcode	plasmid	Barcode
FB270	AGACTCGTTGTATAT	FB302	TTGGACTCACAGATG
FB271	TAGAGATTTAAACCG	FB303	AAGGTGACCTAGTGT
FB272	CGTGACAGCGGATGG	FB304	CCCTCATGAGGTCCG
FB273	TGGGCGGTCAGGGTC	FB305	ATGACAATGTGCAGG
FB274	TTGCCGTCCTTCGAG	FB306	GCGAGGTCGTTAGTT
FB275	TTCAGCGGACGGGCC	FB307	TAAGACTGTTCCGGG
FB276	GTCAGTCCGCTCTTT	FB308	GTTTGTAATCTCTAC
FB277	TTAAGATCCTGGTCG	FB309	GTTAACGCGGCCATT
FB278	TCAACATGGGCAACG	FB310	AGCGGCGTTTATCGT
FB279	CTTGATCGACGCCCA	FB311	TTGGTATGTGTCAAT
FB280	TACGCTATTCAATCT	FB313	GAGCGTAATTGTGAG
FB281	GTGCTTCTGGCGGAT	FB315	GTGACATGCAGGTAG
FB282	CGGCTGTCGGTCGCC	FB316	ACGATCGTACGTCTT
FB283	ATCGTACGTTACTGA	FB317	GTTCAGGTCAGGTCT
FB284	GATTCGAAAGCATAG	FB364	TAAGGAGGGCTGTAG
FB285	CGTATCGGGTCCGGA	FB365	TATCAAGCTAACGTT
FB286	TGGTTGGGTTTGTGG	FB366	GCTCTGGATGTAGTA
FB287	TCGTTGTAACGGTAC	FB367	TAGATGTGGCGGACA
FB289	GACCACTAGAAGGGC	FB368	GTCAACATCGTTACA
FB290	CTGCATGGCGGAGTT	FB369	GGGCCCTAGCGCGTG
FB291	TCAACGATTGTCTGG	FB370	GATAGGCTGGTCCAA
FB292	ACGTCGCACCGTTTG	FB371	TATTTGTGTCGTTCC
FB293	TAACGTTGGGTTGCC	FB372	AGTTAGGGCGCTGCG
FB294	CAGGCTTAACGCGGG	FB373	GCGGAACATAGGCGG
FB295	ACCATAGCGCCACGA	FB374	GCCCTTCAGTCAGCT
FB296	GTCCCGACTAGGACT	FB375	CGGTCGCGTGACGTG
FB297	GTCTTGATTGCTTCG	FB377	CGAGTCGTATGTGGC
FB298	ATTTGGCACAGGATG	FB382	TCCACGGAGGCTGCG
FB299	GGCCACCGTGTGTGA	FB386	TGGTTTACAAATTAT
FB300	ATGAGCAGCGAATGA	FB387	TGTCCGGAAAGGACA
FB301	ATGTTTAACGGCATA		

2.2.2.1.3.3 Second generation Hairpin variations

The hairpin region was inserted as annealed oligonucleotides into the barcoded constructs upon digestion with Esp3I. The primers used for the hairpin containing oligonucleotide formation were generated using the following design:

Forward: 5‘AGGC – [Hairpin sequence]

Reverse: 5‘AGGC – [Hairpin sequence (reverse complement)]

A list of the used hairpins generated, along with the final plasmid ID and the associated barcode is provided in Tables 17-19. The hairpin and barcode sequences were confirmed using Sanger sequencing using the T7E1 based ITR sequencing protocol. ITR integrity of FB555 was examined by Sanger sequencing after digest with EcoRI-HF.

Table 17: Library 1 (V1-V54) - hairpin and barcode sequences. Mismatching nucleotides against the wtITR2 sequence are highlighted in red.

Plasmid	Hairpin	Barcode
FB325 WT	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	AGACTCGTTGTATAT
FB393 V1	CGCCCGGGG AAA CCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	TAGAGATTTAAACCG
FB394 V2	CGCCCGGGA AAA TCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	CGTGACAGCGGATGG
FB395 V3	CGCCCGGGT AAA ACCCGGGCG T CGGGCGACC TTT GGTCGCCCG	TGGGCGGTCAGGGTC
FB396 V4	CGCCCGGCC AAA GGCCGGGCG T CGGGCGACC TTT GGTCGCCCG	TTGCCGTCCTTCGAG
FB397 V5	CGCCCGGAC AAA GTCCGGGCG T CGGGCGACC TTT GGTCGCCCG	TTCAGCGGACGGGCC
FB398 V6	CGCCCGGTC AAA GACCGGGCG T CGGGCGACC TTT GGTCGCCCG	GTCAGTCCGCTCTTT
FB399 V7	CGCCCGCGC AAA GCGCGGGCG T CGGGCGACC TTT GGTCGCCCG	TTAAGATCCTGGTCG
FB400 V8	CGCCCGAGC AAA GCTCGGGCG T CGGGCGACC TTT GGTCGCCCG	TCAACATGGGCAACG
FB401 V9	CGCCCGTGC AAA GCACGGGCG T CGGGCGACC TTT GGTCGCCCG	CTTGATCGACGCCCA
FB402 V10	CGCCCCGGC AAA GCCGGGGCG T CGGGCGACC TTT GGTCGCCCG	TACGCTATTCAATCT
FB403 V11	CGCCCAGGC AAA GCCTGGGCG T CGGGCGACC TTT GGTCGCCCG	GTGCTTCTGGCGGAT
FB404 V12	CGCCCTGGC AAA GCCAGGGCG T CGGGCGACC TTT GGTCGCCCG	CGGCTGTCGGTCGCC
FB405 V13	CGCCGGGGC AAA GCCCCGGCG T CGGGCGACC TTT GGTCGCCCG	ATCGTACGTTACTGA
FB406 V14	CGCCAGGGC AAA GCCCTGGCG T CGGGCGACC TTT GGTCGCCCG	GATTCGAAAGCATAG
FB407 V15	CGCCTGGGC AAA GCCCAGGCG T CGGGCGACC TTT GGTCGCCCG	CGTATCGGGTCCGGA

Plasmid	Hairpin	Barcode
FB408 V16	CGCGCGGGC AAA GCCCGCGCG T CGGGCGACC TTT GGTCGCCCG	TGGTTGGGTTTGTGG
FB409 V17	CGCACGGGC AAA GCCCGTGCG T CGGGCGACC TTT GGTCGCCCG	TCGTTGTAACGGTAC
FB410 V18	CGCTCGGGC AAA GCCCGAGCG T CGGGCGACC TTT GGTCGCCCG	TGGTTTACAAATTAT
FB411 V19	CGGCCGGGC AAA GCCCGGCCG T CGGGCGACC TTT GGTCGCCCG	GACCACTAGAAGGGC
FB412 V20	CGACCGGGC AAA GCCCGGTCG T CGGGCGACC TTT GGTCGCCCG	CTGCATGGCGGAGTT
FB413 V21	CGTCCGGGC AAA GCCCGGACG T CGGGCGACC TTT GGTCGCCCG	TCAACGATTGTCTGG
FB414 V22	CCCCCGGGC AAA GCCCGGGGG T CGGGCGACC TTT GGTCGCCCG	ACGTCGCACCGTTTG
FB415 V23	CACCCGGGC AAA GCCCGGGTG T CGGGCGACC TTT GGTCGCCCG	TAACGTTGGGTTGCC
FB416 V24	CTCCCGGGC AAA GCCCGGGAG T CGGGCGACC TTT GGTCGCCCG	CAGGCTTAACGCGGG
FB417 V25	GGCCCGGGC AAA GCCCGGGCC T CGGGCGACC TTT GGTCGCCCG	ACCATAGCGCCACGA
FB418 V26	AGCCCGGGC AAA GCCCGGGCT T CGGGCGACC TTT GGTCGCCCG	GTCCCGACTAGGACT
FB419 V27	TGCCCGGGC AAA GCCCGGGCA T CGGGCGACC TTT GGTCGCCCG	GTCTTGATTGCTTCG
FB420 V28	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACG TTT CGTCGCCCG	ATTTGGCACAGGATG
FB421 V29	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACA TTT TGTCGCCCG	GGCCACCGTGTGTGA
FB422 V30	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACT TTT AGTCGCCCG	ATGAGCAGCGAATGA
FB423 V31	CGCCCGGGC AAA GCCCGGGCG T CGGGCGAGC TTT GCTCGCCCG	ATGTTTAACGGCATA
FB424 V32	CGCCCGGGC AAA GCCCGGGCG T CGGGCGAAC TTT GTTCGCCCG	TTGGACTCACAGATG
FB425 V33	CGCCCGGGC AAA GCCCGGGCG T CGGGCGATC TTT GATCGCCCG	AAGGTGACCTAGTGT
FB426 V34	CGCCCGGGC AAA GCCCGGGCG T CGGGCGGCC TTT GGCCGCCCG	CCCTCATGAGGTCCG
FB427 V35	CGCCCGGGC AAA GCCCGGGCG T CGGGCGCCC TTT GGGCGCCCG	ATGACAATGTGCAGG
FB428 V36	CGCCCGGGC AAA GCCCGGGCG T CGGGCGTCC TTT GGACGCCCG	GCGAGGTCGTTAGTT
FB429 V37	CGCCCGGGC AAA GCCCGGGCG T CGGGCCACC TTT GGTGGCCCG	TAAGACTGTTCCGGG
FB430 V38	CGCCCGGGC AAA GCCCGGGCG T CGGGCAACC TTT GGTTGCCCG	GTTTGTAATCTCTAC
FB431 V39	CGCCCGGGC AAA GCCCGGGCG T CGGGCTACC TTT GGTAGCCCG	GTTAACGCGGCCATT
FB432 V40	CGCCCGGGC AAA GCCCGGGCG T CGGGGGACC TTT GGTCCCCCG	AGCGGCGTTTATCGT
FB433 V41	CGCCCGGGC AAA GCCCGGGCG T CGGGAGACC TTT GGTCTCCCG	TTGGTATGTGTCAAT
FB434 V42	CGCCCGGGC AAA GCCCGGGCG T CGGGTGACC TTT GGTCACCCG	TGTCCGGAAAGGACA

Plasmid	Hairpin	Barcode
FB435 V43	CGCCCGGGC AAA GCCCGGGCG T CGGCCGACC TTT GGTCGGCCG	GAGCGTAATTGTGAG
FB436 V44	CGCCCGGGC AAA GCCCGGGCG T CGGACGACC TTT GGTCGTCCG	GTGACATGCAGGTAG
FB437 V45	CGCCCGGGC AAA GCCCGGGCG T CGGTCGACC TTT GGTCGACCG	ACGATCGTACGTCTT
FB438 V46	CGCCCGGGC AAA GCCCGGGCG T CGCGCGACC TTT GGTCGCGCG	GTTCAGGTCAGGTCT
FB439 V47	CGCCCGGGC AAA GCCCGGGCG T CGAGCGACC TTT GGTCGCTCG	TAAGGAGGGCTGTAG
FB440 V48	CGCCCGGGC AAA GCCCGGGCG T CGTGCGACC TTT GGTCGCACG	TATCAAGCTAACGTT
FB441 V49	CGCCCGGGC AAA GCCCGGGCG T CCGGCGACC TTT GGTCGCCGG	GCTCTGGATGTAGTA
FB442 V50	CGCCCGGGC AAA GCCCGGGCG T CAGGCGACC TTT GGTCGCCTG	TAGATGTGGCGGACA
FB443 V51	CGCCCGGGC AAA GCCCGGGCG T CTGGCGACC TTT GGTCGCCAG	GTCAACATCGTTACA
FB444 V52	CGCCCGGGC AAA GCCCGGGCG T GGGGCGACC TTT GGTCGCCCC	GGGCCCTAGCGCGTG
FB445 V53	CGCCCGGGC AAA GCCCGGGCG T AGGGCGACC TTT GGTCGCCCT	GATAGGCTGGTCCAA
FB446 V54	CGCCCGGGC AAA GCCCGGGCG T TGGGCGACC TTT GGTCGCCCA	TATTTGTGTCGTTCC

Table 18: Library 2 (V55-V90) - hairpin and barcode sequences. Mismatches with the wtITR2 hairpin are highlighted in red.

Plasmid	Hairpin	Barcode
FB325 WT	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	AGACTCGTTGTATAT
FB454 V55	CGCCCGGGG AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	AGTTAGGGCGCTGCG
FB455 V56	CGCCCGGCC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	GCGGAACATAGGCGG
FB456 V57	CGCCCGCGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	GCCCTTCAGTCAGCT
FB457 V58	CGCCCCGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	CGGTCGCGTGACGTG
FB458 V59	CGCCGGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	CGAGTCGTATGTGGC
FB459 V60	CGCGCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	TAAGACTGTTCCGGG
FB460 V61	CGGCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	TAGAGATTTAAACCG
FB461 V62	CCCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	CGTGACAGCGGATGG
FB462 V63	GGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	TGGGCGGTCAGGGTC
FB463 V64	CGCCCGGGC AAA CCCCGGGCG T CGGGCGACC TTT GGTCGCCCG	TTGCCGTCCTTCGAG
FB464 V65	CGCCCGGGC AAA GGCCGGGCG T CGGGCGACC TTT GGTCGCCCG	TTCAGCGGACGGGCC

Plasmid	Hairpin	Barcode
FB465 V66	CGCCCGGGC AAA GCGCGGGCG T CGGGCGACC TTT GGTCGCCCG	GTCAGTCCGCTCTTT
FB466 V67	CGCCCGGGC AAA GCCGGGGCG T CGGGCGACC TTT GGTCGCCCG	TTAAGATCCTGGTCG
FB467 V68	CGCCCGGGC AAA GCCCCGGCG T CGGGCGACC TTT GGTCGCCCG	TCAACATGGGCAACG
FB468 V69	CGCCCGGGC AAA GCCCGCGCG T CGGGCGACC TTT GGTCGCCCG	CTTGATCGACGCCCA
FB469 V70	CGCCCGGGC AAA GCCCGGCCG T CGGGCGACC TTT GGTCGCCCG	TACGCTATTCAATCT
FB470 V71	CGCCCGGGC AAA GCCCGGGGG T CGGGCGACC TTT GGTCGCCCG	GTGCTTCTGGCGGAT
FB471 V72	CGCCCGGGC AAA GCCCGGGCC T CGGGCGACC TTT GGTCGCCCG	TATTTGTGTCGTTCC
FB472 V73	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT CGTCGCCCG	ATCGTACGTTACTGA
FB473 V74	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GCTCGCCCG	GATTCGAAAGCATAG
FB474 V75	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGACGCCCG	CGTATCGGGTCCGGA
FB475 V76	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTGGCCCG	TGGTTGGGTTTGTGG
FB476 V77	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCCCCCG	TCGTTGTAACGGTAC
FB477 V78	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGGCCG	TAACGTTGGGTTGCC
FB478 V79	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCGCG	GACCACTAGAAGGGC
FB479 V80	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCGG	CTGCATGGCGGAGTT
FB480 V81	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACC TTT GGTCGCCCC	TCAACGATTGTCTGG
FB481 V82	CGCCCGGGC AAA GCCCGGGCG T CGGGCGACG TTT GGTCGCCCG	ACGTCGCACCGTTTG
FB482 V83	CGCCCGGGC AAA GCCCGGGCG T CGGGCGAGC TTT GGTCGCCCG	GTTTGTAATCTCTAC
FB483 V84	CGCCCGGGC AAA GCCCGGGCG T CGGGCGTCC TTT GGTCGCCCG	CAGGCTTAACGCGGG
FB484 V85	CGCCCGGGC AAA GCCCGGGCG T CGGGCCACC TTT GGTCGCCCG	ACCATAGCGCCACGA
FB485 V86	CGCCCGGGC AAA GCCCGGGCG T CGGGGGACC TTT GGTCGCCCG	GTCCCGACTAGGACT
FB486 V87	CGCCCGGGC AAA GCCCGGGCG T CGGCCGACC TTT GGTCGCCCG	GTCTTGATTGCTTCG
FB487 V88	CGCCCGGGC AAA GCCCGGGCG T CGCGCGACC TTT GGTCGCCCG	ATTTGGCACAGGATG
FB488 V89	CGCCCGGGC AAA GCCCGGGCG T CCGGCGACC TTT GGTCGCCCG	GGCCACCGTGTGTGA
FB489 V90	CGCCCGGGC AAA GCCCGGGCG T GGGGCGACC TTT GGTCGCCCG	ATGAGCAGCGAATGA

Table 19: Alternative hairpin and barcodes sequences in second generation sITR plasmids (Y1-Y17). Hairpin sequences used for the generation of the second generation sITR library with divergent

hairpin sequences Y1-Y17. For variant Y17, no hairpin was inserted but the barcoded plasmid FB271 with the Esp3I cleavage sites was used without modification.

Plasmid	Hairpin	Barcode	Type
FB539 Y1	CTCAGATCTAACCTTGTTAAATTGAG	TTCAGCGGACGGGCC	Aptamer (AptMG)
FB540 Y2	ATTTTACTCTTCTCCCCTCG	TCAACATGGGCAACG	Loop
FB541 Y3	GAGAGGAGGAGGAGAG	TACGCTATTCAATCT	Loop
FB542 Y4	AAATATATATTAAATATATTTTAATTAATTTATAAAATTAATT	CGGCTGTCGGTCGCC	TA-rich hairpin
FB543 Y5	ACGGGCGACCAAAGGTCGCCCGATACGAATACGTCGCCCGGGCTTTGCCCGGGCGCGATC	ATCGTACGTTACTGA	AAV2 bent hairpin
FB544 Y6	TAGCGACCAAAGGTCGACTAGACGGGCTTTGCCCGGAT	GATTCGAAAGCATAG	AAV2 short BC arms
FB545 Y7	CGGGCGACCAAAGGTCGCCCGATAATAACGCCCGGGCTTTGCCCGGGCG	CGTATCGGGTCCGGA	AAV2 + loop
FB546 Y8	CGGGCGACCAAAGGTCGCCCGAGCACAGAGCGCCCGGGCTTTGCCCGGGCG	TGGTTGGGTTTGTGG	AAV2 + loop
FB547 Y9	CGGGCGACCAAAGGTCGCCCGAGCCGATACGGCTCGCCCGGGCTTTGCCCGGGCG	GACCACTAGAAGGGC	AAV2 Triple Hairpin
FB548 Y10	TACAGTTTCTATTAGAAACGATGTATTACATCTGTA	ACGTCGCACCGTTTG	Densovirus (*A. aegypti*)
FB549 Y11	GCGACCAAAGGTCGCCCGGGCTTTGCCCGG	CAGGCTTAACGCGGG	AAV2 short BC
FB550 Y12	TTGGGGTATGGGGTATGGGGTATGGGGTT	ATTTGGCACAGGATG	G4
FB551 Y13	TTGGGGTATGGGGAA	GGCCACCGTGTGTGA	Partial G4
FB552 Y14	TCCGCCTACGGCGAGCAGCGGGCCTTCGGCCCCCCTTCGGGGCTGCTG	ATGAGCAGCGAATGA	Hepandensovirus (*P. monodon*)
FB553 Y15	ATGACCACGACGCGCGAAGCGCGTCACTTCGTGTGGTCACTACGTAT	ATGTTTAACGGCATA	Densovirus (*S. fusca*)
FB554 Y16	CTGGCGACCAAAGGTCGCCAGACGGACGTGCTTTGCACGTCCG	TCCACGGAGGCTGCG	wtITR3 hairpin
FB555 Y17	None (= FB271)	TAGAGATTTAAACCG	Esp3I sites

2.2.2.1.3.4 Second generation Stem mutants

Alterations in the stem sequence of the ITR in single ITR plasmids were incorporated using mismatching primers. For mutations in the terminal regions of the ITR sequence, the mutation was incorporated using a single primer PCR on a barcoded base construct, the rest of the ITR as well as Esp3I cleavage sites were then attached in a second single primer PCR. The PCR product was then digested with Esp3I and the wtITR2 hairpin inserted as annealed oligonucleotide with primers 98FB and 99FB. For stem mutants deeper within the ITR, a single PCR with a single primer was sufficient to insert the mutation as well as the attachment of the cloning site. This single PCR followed by the insertion of the wtITR2 hairpin as annealed

oligonucleotide into the Esp3I cleavage site. A list of used primers for the PCRs is provided in Table 20. The barcode and the hairpin sequences were verified for all plasmids using the T7E1 based ITR sequencing protocol. This approach was used for the plasmids FB562 – FB564.

Table 20: Stem mutant primer combinations. Primer used for the generation of the stem variants VS2-VS4 with the barcode in the plasmid and the number and indication of the generated plasmid (n.a. = not applicable).

Plasmid	1st PCR Primer	2nd PCR Primer	Barcode
FB562 VS2	196FB	n.a.	TAGAGATTTAAACCG
FB563 VS3	201FB	n.a.	GTCAGTCCGCTCTTT
FB564 VS4	290FB	93FB	CTTGATCGACGCCCA

2.2.2.2 Plasmids for circular AAV genome generation

Plasmids with two equally oriented A+D sequences to observe potential ligation events mediated by the AAV replication machinery were cloned as described in the subsequent chapters.

2.2.2.2.1 Dual RBE plasmids

The mCherry expression cassette, including CMV promoter and short synthetic polyA signal were amplified from plasmid FB185 using primers 523FB & 524FB. The plasmid backbone, *i.e.* ampicillin resistance and ori were amplified from pBlueScript using primers 477FB & 478FB. The amplicons were digested with Esp3I, the prokaryotic backbone fragment dephosphorylated, and the two fragments were ligated, generating the plasmid FB503 with the backbone and the mCherry expression cassette separated by two equally oriented AAV A+D sequences (RBE*).

2.2.2.2.2 Luciferase activity switching reporter plasmids

To follow AAV-Rep mediated ligation events, dual luciferase assay plasmids were designed that may switch the expressed transgene in presence of the AAV replication machinery. For the first-generation Luciferase switching reporter the ori-amp-RBEW fragment was amplified from FB503 using 653FB and 663FB, the Renilla luciferase (*R*luc) gene with poly-A followed by the TK promoter was amplified using 654FB and 655FB from pSiCheck2 (Promega). The Firefly luciferase (FFluc) gene followed by a polyA site was amplified from pSiCheck2 using the primers 656FB and 657FB. The PCR amplicons were then assembled in a Golden Gate cloning using Esp3I, generating FB535. The data from this plasmid was not included in this thesis work.

For the second-generation luciferase switching reporter, the *R*luc was split, additionally synthetic splice donor and splice acceptor sites were included to aid expression. For the fragment with splice acceptor, RBE and splice donor a short fragment was amplified from FB535 with primers 671FB and 672FB. The fragment with FFLuc, polyA site, ori, ampicillin resistance cassette, RBE and a splice acceptor was generated by amplification from FB535 with primers 673FB and 674FB. The C-terminal part of the *R*luc and the TK-promoter were amplified from FB535 with primers 675FB and 676FB. The fragment encoding the N-terminal part of *R*luc was generated by PCR with primers 677FB and 678FB from FB535. The fragments were then assembled by Golden Gate cloning with Esp3I, generating the reporter plasmid FB538.

2.2.2.3 CMV-Rep and Rep mutant plasmids

The AAV2 Rep gene was amplified from the WHc2 (SpeI) Rep2-Cap2 plasmid using the primers 16FB and 17FB. The PCR product as well as the acceptor plasmid pcDNA3.1 (Invitrogen) were digested with NheI and XbaI. After ligation, the plasmid FB4 was formed. To prevent the expression of the large Rep isoforms, the plasmid was used as template for a PCR with 304FB and 305FB, the PCR product was digested with Esp3I and ligated. This new construct FB18 prevents the expression of the larger Rep isoforms. To additionally incorporate a Kozak sequence and make the expression vector comparable to other plasmids (not presented in this work), the Rep68 gene cassette was amplified from FB18 with 427FB and 428FB, digested with Esp3I and inserted into pcDNA 3.1 linearized with NheI and EcoRI-HF, generating FB319 which was used for experiments.

To generate the AAV2 Rep68 Y156F mutant plasmid, FB319 was amplified using 659FB and 660FB, the fragment was digested with Esp3I and ligated forming FB536. For the AAV2 Rep68 K340H mutant, the primers 661FB and 662FB were used to amplify FB319, followed again by digestion with Esp3I and ligation, forming FB537.

2.2.3 Cell culture

2.2.3.1 Maintenance of eukaryotic cells

Human embryonic kidney cells (Hek239T) and HuH7 cells were kept in 75 cm^2 or 175 cm^2 CELLSTAR® cell culture flasks (Greiner Bio-One) in Dulbecco's Modified Eagle Medium (DMEM) with GlutaMAX+4,5 g/l D-Glucose (Gibco) supplemented with 10% fetal bovine serum (Capricorn Scientific) and 100 Units/ml Penicillin and 100 µg/ml Streptomycin with

Penicillin-Streptomycin (Gibco). Cells were incubated in a humidified incubator at 37°C and 5% CO_2. Depending on confluency, cells were washed with 1x PBS (Gibco) and detached using 0.25% or 2.5% Trypsin-EDTA (Gibco) every 2-4 days and split 1:3 to 1:20 in fresh cell culture medium. For cell quantification, cells were mixed 1:1 with Trypan Blue Solution 0.4% (Thermo Fisher) and the life cell number was determined using the Countess automated cell counter (Invitrogen) or the Neubauer counting chamber (Brand).

2.2.3.2 Small and medium scale PEI transfection

For transfection in 96-well plates 12.500 cells per well were seeded one day prior to transfection in a volume of 100 µl culture medium. Cells were transfected with a total of 100 ng DNA. The DNA was diluted in 1.633 µl H_2O and 1.633 µl 300 mM NaCl per transfected well. This DNA mix was then mixed with a premixed PEI solution consisting of 0.73 µl PEI, 1.633 µl 300 mM NaCl and 0.9 µl H_2O per well. The transfection mixes were vortexed, spun down, incubated at room temperature for 10 min and added to the cells.

For transfection in 6-well plates 300.000 cells were seeded per well in 2 ml of culture medium one day prior to transfection. For each well 2600 ng DNA were transfected, which was diluted in 49 µl H_2O and 49 µl 300 mM NaCl. The DNA mix was then mixed with a PEI solution consisting of 22 µl PEI, 27 µl H_2O and 49 µl 300 mM NaCl. The transfection mix was mixed by vortexing, spun down and incubated at room temperature for 10 min before adding to the cells.

2.2.3.3 Dual Luciferase assay

The dual luciferase assay was used to verify the formation of circular replication intermediates which should increase the expression of the *R*Luc while reducing the expression of Firefly luciferase (FFLuc) compared to the input plasmid control. Upon transduction Hek293T cells with the AAV vectors or transfection with the plasmid. The assay was performed using the Dual-Luciferase® Reporter Assay System (Promega) with the provided solutions for cell lysis and luciferase activity quantification. In brief, cells were lysed 72 hours post-transfection using 50 µl of 1x passive lysis buffer. 10 µl of the lysate were transferred to a white flat-bottom 96-well plate (Greiner). The luciferase signal was measured using the GloMax Navigator Microplate Luminometer with dual injectors (Promega) using the following program: 1) injection of 35 µl LAR II solution, 2) 2 sec incubation time, 3) integration of FFLuc luminescence signal over 10 sec, 4) injection of 35 µl Stop & Glo solution, 5) 2 sec incubation

time, 6) 2 sec delay, 7) 10 sec integration of the *R*Luc luminescence unit. For the analysis, the values of *R*Luc and FFLuc activity, denominated in photon counts, were normalized and the ratio of these values was indicated.

2.3.3.4 Extraction of low molecular weight DNA from mammalian cells

Low molecular weight extraction from mammalian cells was performed via an adapted Hirt extraction protocol. 800 µl of Hirt lysis buffer (10 mM Tris-HCl, 10 mM EDTA, 0.6% SDS) were added to one well of a 6 well plate and swirled gently for 2 min. After 10 min of incubation at room temperature cell lysate was collected and 200 µl of 5.0 M NaCl were added and mixed gently by inverting. The tubes were place on ice for 1-16 hours and then centrifuged in a pre-cooled centrifuge (4°C) for 30 min at 20.000 g. The supernatant was collected and mixed with 666 µl Propan-2-ol to precipitate the DNA and centrifuged for 15 min at 20.000 g. The supernatant was discarded, and the precipitated DNA was washed with 500 µl 70% ethanol and centrifuged for 5 min at 20.000 g. Then the ethanol was discarded and after a total of two washing steps, the DNA pellet was dried at room temperature for at least 10 minutes, resuspended in 100 µl H_2O and the DNA concentration was determined via the Nanodrop 2000 Spectrophotometer (ThermoScientific).

2.2.4 Virological methods

2.2.4.1 Transfections for AAV vector production

rAAV vectors were produced using polyethyleneimine (PEI) as transfection reagent in a triple-transfection setup. Thus, three types of plasmids were used for the production, namely, the adenoviral helper plasmid (AdH) encoding the adenoviral E4, E2a and VA genes, the AAV-helper encoding the AAV Rep and Cap genes with the p5 promoter downstream of the Cap gene. The third plasmid was the vector plasmid, containing the ITR sequence and the sequence that was to be packaged in the vector particles. Alternatively, ITRs were replaced with RBE sequences to observe the potential packaging of circular DNA fragments.

Large scale vector production refers to production in 150 mm tissue culture dishes (Sarstedt). For each dish, $4x10^6$ cells were seeded in a volume of 22 ml. 48 h after seeding, the cells were transfected either at a 1:1:1 mass ratio or a 1:1:1 molar ratio. For the transfection of one plate, 44 µg of plasmid DNA was diluted with H_2O to a total volume of 790 µl and mixed with 790 µl 300 mM NaCl. Then a mix of 790 µl 300 mM NaCl, 438 µl H_2O and 352 µl PEI-Max added

to the diluted DNA, thoroughly vortexed and incubated for 10 min at room temperature. 3.2 ml of the mix were added dropwise to each dish.

For larger ITR variant libraries which were produced separately, transfections were performed in 6 well dishes (Greiner). A total of 2600 ng of plasmid were used for the transfection with a 1:1:1 mass ratio and was performed as described in 2.2.3.2.

Cells were then harvested 72 h post-transfection by scraping them off using a cell lifter (Corning) and resuspended in their culture medium. For the production of ITR variant libraries, in which the vector particles were generated in separate wells, the cell suspensions were pooled at this step. The cell suspension was then centrifuged at 800 g for 15 min and the supernatant was discarded.

2.2.4.2 AAV vector purification

2.2.4.2.1 Cell Lysis and vector solution clearance

Cell pellets from vector productions were resuspended in 5 ml Benzonase buffer. Cells were lysed by subjecting them to five freeze thaw cycles consisting of incubation in liquid N_2 for 5 min followed by thawing in a water bath at 37°C for 10-15 min. Next, samples were subjected to a short sonication for 1 min 20 sec in a Sonorec Digitex DT (Bandelin) to assist later removal of cellular DNA by fragmentation. Non-packaged DNA was then removed by digestion with 250U/ml EMPROVE Benzonase (Merck) for 1 hour at 37°C in a water bath and inverted every 10 min. To minimize DNA contamination from non-packaged DNA, for vector samples that were later analyzed by Nanopore sequencing, the incubation time was elongated to 6 hours at 37°C. In the last 5 hours of the incubation, the tube was only inverted every 45 min.

The vector particle solution was then cleared at least three times by centrifugation 5000 g for 15 min, each time the supernatant was collected, and the debris pellet was discarded. The cleared lysate was then used for the iodixanol gradient preparation.

2.2.4.2.2 Iodixanol gradient ultracentrifugation

The cleared lysates were transferred into 16 x 75 mm Re-seal polyallomer centrifuge tubes (Seton Scientific). A 230 mm glass Pasteur pipette (Brand) was used to sub-layer the solution with 1.5 ml of 15%, 25%, 40% and 60% iodixanol phases. Tubes were heat-sealed using the Tube Sealer (Beckman Coulter) and balanced pairwise. The tubes were centrifuged using the 70.1 Ti rotor in the OptimaTM L90K ultracentrifuge (both Beckman Coulter) at 50.000 rpm (~229,000g) for 2 h 30 min at 4°C. After centrifugation, the 40% iodixanol phase was

collected. For pressure release, the top of the tube was punctured and the 40% phase, containing the assembled full capsids, was collected using a syringe. AAVs were aliquoted and stored at -80°C or -20°C for short-term storage.

2.2.4.2.3 Buffer exchange

For vectors used for *in vivo* experiments the samples were rebuffered. Here, vector samples from separate productions were pooled to increase the total vector yield. Thus, up to 2.4 ml of the vector in iodixanol was mixed with 1x PBS to reach a final volume of 15 ml and transferred into an Amicon Ultra-15 Centrifugal Filter Unit (MWCO 100000; Merck). Following the manufacturer's instruction, the filter units were centrifuged at 4000 g until around 1 ml remained in the filter unit. The flow through was discarded and the filter unit was refilled with 14 ml 1x PBS and centrifuged until reaching a volume of 300-600 µl which was then collected, aliquoted and stored at -80°C or -20°C for short-term storage.

2.2.4.3 AAV titration by qPCR

Purified rAAVs in Iodixanol or PBS lysed by alkaline lysis. For this purpose, 10 µl of the rAAV sample was mixed with 10 µl TE buffer and subsequently lysed with 20 µl 2M NaOH and incubated for 30 min at 56°C. The solution was neutralized by the addition of 38 µl 1M HCl. To obtain a 1:100 dilution 922 µl H_2O were added as well. Before measurement, the samples were further diluted 1:10 in H_2O.

Titration was then performed by qPCR using the SensiMixII Probe kit (Bioline). Each sample was measured in triplicates, containing 5 µl SensiMix, 0.4 µl of 10 µM forward and reverse primer, 0.1 µl of 10 µM probe, 2.671 µl H_2O and 1.428 µl AAV sample or plasmid standard. The plasmid standard was prepared from a plasmid stock with $3.5x10^{11}$ molecules/ml and 1:10 serial dilutions ranging to $3.5x10^{6}$ molecules/µl were measured as well. The qPCR was performed in a Corbett RG6000 instrument and was initiated with a 10 min step at 95°C followed by 40 cycles of 95°C for 10 sec and 60°C for 20 sec. After each elongation step the fluorescent signal was acquired at 510 nm. The Ct values of the plasmid standard and the samples were determined, and a standard curve was generated using the RotorGene 6000 Series Software 1.7 to determine the concentration of the AAV samples. The AAV concentration in the sample was extrapolated from the standard curve and by inclusion of the dilution factors through alkaline lysis and pre-dilution. For ssAAV all samples were additionally multiplied by two, as correction factor.

2.2.4.4 AAV Transduction and microscopic observation

For transductions of cell lines in 6-well plates, Hek293T or HuH7 cells were seeded one day prior to transduction at $1.25x10^5$ / well in 2 ml. The cells were then transduced with the indicated MOI. For volumes lower than 20 µl the vector was adjusted to 20 µl with 1x PBS to ensure a more uniform distribution in the wells. Cell lines were then incubated for 72 h in a humidified incubator at 37°C with 5% CO_2 before further processing.

For transductions in 96 well plates, $1.25x10^4$ cells were seeded per well in a volume of 100 µl. The vector was added in the indicated amount and the cells were incubated for 72 h in a humidified incubator at 37°C with 5% CO_2 before further processing. Microscopy images were taken using an Olympus CKX41 with a 20x magnification objective using the Olympus PEN lite E-PL3 camera and the Olympus U-RFL-T burner setup. Scale bars were added manually using ImageJ V1.53k.

2.2.4.5 ddPCR quantification of vector genome content in gDNA and gene expression

For AAV titration using ddPCR, vector samples were diluted $1:10^4$ – $1:10^7$. For the determination of the vector genome content as ratio to host genome. Differently labeled primers and probes that target the vector genome (CMV, FAM-labeled) and the host genome (ddPCR GEX HEX Assay Rpp30, Mmu (#10031255) mix, Bio-Rad) were used. For measuring transgene expression in comparison to a housekeeping gene, primers and probes were used that target the transgene in the transcribed region (mCherry, FAM-labeled) and a spliced region in the host genome that is in close proximity on cDNA but distant on gDNA (muRPP30_c9, HEX-labeled). For detection of vector genome copies in host gDNA, the gDNA was diluted to 5 ng/µl of which 5.5 µl were used as input for the ddPCR reaction. Per reaction, 11 µl ddPCR Supermix for Probes (No dUTP; Bio-Rad) as well as 1.1 µl of each 20x primer/probe mix (final concentration: 900 nM of both primers and 250 nM of probe) were added together with 1.1 µl of a 1:4 in Diluent B diluted HindIII-HF (both NEB) and adjusted to 22 µl with H_2O. The reaction mix was then incubated for 15 min at room temperature to pre-digest the gDNA, which enhances ddPCR reliability.

For detection of the transcription levels, the cDNA transcribed with the SuperScript IV VILO Master Mix with ezDNase Kit (Invitrogen), as described in 2.2.1.12 was diluted 1:10 and 5.5 µl were used as ddPCR reaction template. The ddPCR reaction mastermix was prepared the same,

without adding HindIII-HF to the reaction mix. Therefore, also no pre-digest incubation was performed.

20 µl of the ddPCR reaction mixes were then used for droplet generation using the QX200 Droplet Generator using the Droplet Generation Oil (both Bio-Rad) Droplets were transferred into 96-well plates and sealed with a pierceable PCR Plate Heat Seal aluminum foil using the PX1 PCR Plate Sealer (both Bio-Rad). The ddPCR reactions were then run in the C1000 Touch Thermal Cycler (Bio-Rad) using the following cycling conditions (Table 21).

Table 21: ddPCR cycling conditions.

Step	Temperature	Time [min:sec]	
Initial Denaturation	94	10:00	
Denaturation	94	00:30	x40
Annealing/Elongation	58	01:00	
Droplet Stabilization	98	10	
Hold	12	hold	

Plates were incubated for 2-16 hours at 4°C before analysis, which increased the number of readable droplets in the analysis. The droplets were then analyzed using the QX200 Droplet Reader, measuring both the FAM-signal and HEX-signal. To determine the transduction efficiency, the measured transgene concentration ($C_{Transgene}$) and host genome concentration (C_{Host}) values were used to calculate the vector copies per host diploid genome (vg/dg) as follows:

$$\text{transduction efficiency}\left[\frac{\text{vg}}{\text{dg}}\right] = \frac{C_{Transgene}}{2 \times C_{Host}}$$

The relative expression values were similarly calculated using the following formula:

$$\text{rel. expression} = \frac{C_{Transgene}}{C_{Host}}$$

2.2.4.6 AAV titration by ddPCR

For AAV titration using ddPCR, vector samples were diluted $1:10^4$ – $1:10^7$. Of these dilutions 5.5 µl were used as template for the ddPCR reaction pre-mix with 11 µl ddPCR Supermix for Probes (no dUTP; BioRad) as well as 1 µl of 20x primer/probe mix (final concentration: 900 nM of both primers and 250 nM of probe) for a region in the packaged vector genome (CMV, FAM-labeled). Droplet generation with 20 µl of the ddPCR reaction pre-mix as well as the ddPCR and the droplet measurement was performed as described in 2.2.4.5.

To calculate the vector concentrations as vector genomes per ml (vg/µl) the measured transgene concentration ($C_{Transgene}$) in vg/µl in the sample and the dilution factor (DF) were used. To take into account the 5 µl input template the value was divided by a correction factor, as described in the following formula:

$$c\left[\frac{vg}{\mu l}\right] = \frac{C_{Transgene} \times DF}{5}$$

2.2.5 Next- and third-Generation sequencing methods

Vector genomes of rAAV were analyzed with a focus on ITR integrity was determined by Nanopore sequencing as well as PacBio Sequencing. An Illumina Next generation sequencing (NGS) based amplicon-sequencing approach was used to determine the distribution of barcoded ITR variants in viral vectors as well as DNA, RNA and enriched episomal DNA upon transduction. Furthermore, an Illumina NGS-based integration assay was performed, to evaluate the integration capacity of different ITR variants.

2.2.5.1 PacBio sequencing of AAV vector genomes

PacBio sample preparation and sequencing was performed by Suk Namkung, Mitchell Yip and Phillip Tai at the University of Massachusetts Medical School (UMMS) according to their previously published workflow [363]. In brief, $1x10^{12}$ vg were treated with DNase I before treating the sample with Pronase (Sigma Aldrich). The vector DNA was isolated by phenol/chloroform extraction and precipitated with ethanol before resuspending in H_2O. 10% linearized lambda phage DNA was spiked to the sample before end-repair and A-tailing followed by ligation of SMRTbell adapters (SMRTbell express template prep Kit, PacBio). Sequencing was performed on a Sequel I (PacBio) at the UMMS Deep Sequencing Core Facility.

2.2.5.2 Nanopore sequencing of plasmids and AAV vector genomes

2.2.5.2.1 Vector genome isolation for Nanopore sequencing

For the isolation of the rAAV vector genomes for Nanopore sequencing the DNA had to be extracted from the vector particles. Up to $1x10^{12}$ vg were diluted with H_2O to a volume of 600 µl or a maximum of 600 µl of AAV vectors in iodixanol solution were treated with 200 µg Proteinase K (Roche). The samples were incubated in a thermoshaker at 55°C for 45 minutes at 400 rpm and then at 95°C for 20 minutes for proteinase K inactivation. For clean-up of the vector genomes, the DNA Clean & Concentrator-5 Kit (Zymo) was employed according to the manufacturer's instruction for ssDNA. Thus, the 610 µl from the digest were mixed with

4270 µl of DNA binding buffer, added to the spin column. After two washes with DNA wash buffer, the DNA was eluted in 10 µl of H_2O, the DNA concentration was measured using the Nanodrop 2000 Spectrophotometer (ThermoScientific) or with the Qubit dsDNA HS Assay Kit (ThermoScientific).

2.2.5.2.2 Nanopore sequencing library preparation

Sequencing library preparation was performed using the Rapid sequencing Kit with barcodes (SQK-RBK004; Oxford Nanopore) according to the manufacturer's instruction. Briefly, up to 400 ng of plasmid DNA/isolated vector DNA or a maximum volume of 7.5 µl were mixed with 2.5 µl of the barcoded Fragmentation Mix on ice, incubated for 1 min at 30°C and then heated to 80°C for 1 min. To the fragmented DNA with inserted transposase adapters 1 µl Rapid Adapter was directly added, mixed and incubated for 5 min at room temperature. The final library was then stored on ice until loading on the flow cell.

2.2.5.2.3 Priming and loading the SpotON Flow cell

For sequencing the SpotON Flow Cell, R9 Version (Oxford Nanopore) was used, as the SQK-RBK004 library preparation kit is only compatible with V9 chemistry. The flow cell was placed in the MinION Mk1B and flow cell priming and sample loading was performed according to the manufacturer's instructions. Data Acquisition, real-time basecalling, and barcode demultiplexing were conducted via the MinKNOW (v22.05.5; Oxford Nanopore) software during sequencing and through post-run analysis.

2.2.5.3 Barcode sequencing by amplicon NGS

The sequencing of the barcode region in the AAV vector genome was performed by NGS. Two different sequencing approaches were used for barcode sequencing (BC-seq). Sequencing using the Illumina MiSeq was initially performed to verify the NGS pipeline using dual dephasing adapters. Subsequent sequencing runs were performed with NextSeq using a single sided dephasing adapter and sequencing approach.

2.2.5.3.1 Amplicon sequencing with dual sided dephasing adapters

For the amplification of the barcode region encoded in the vector genome. The prepared cDNA derived from transduced cells was further diluted 1:2 with H_2O and 2.5 µl were used for the PCR reaction. From the T5 exonuclease digested gDNA, 12.25 µl (~400 ng undigested gDNA)

were used as PCR input. Around 10^9-10^{10} copies of the vector and plasmid mixes were used as input for the PCR. The primers, both forward and reverse primers (268FB-283FB), contained adapters for the secondary PCR (Read 1 and Read 2 adapter) as well as 0-7 random nucleotides (N) which were mixed at a 1:1:1:1:1:1:1:1 molar ratio. The PCR was performed with Q5 High-Fidelity Polymerase (NEB) using 500 nM of the forward and reverse primer pools, as described in 2.2.1.2 Amplicon formation was validated on an agarose gel, the amplicon was purified using ProNex Size Selective Purification System (Promega) with a 2x ratio according to the manufacturer's instructions. The size of the fragments was verified using the DNA 1000 Kit on the Agilent 2100 Bioanalyzer. Illumina indices and adapters were added in a secondary PCR with the Nextera XT Index Kit v2 (Illumina) primers and the KAPA HiFi HotStart ReadyMix PCR Kit (Roche) via a PCR reaction with eight amplification cycles in a volume of 25 µl. Therefore 12.5 µl 2x KAPA HiFi HotStart Ready Mix were mixed with 2.5 µl of the purified first PCR reaction, 2.5 µl of Nextera Index Primer 1, 2.5 µl of Nextera Index Primer 2 and 5 µl H_2O with different Primer 1/Primer 2 combinations for each sample. The PCR reaction in a thermocycler was initiated with a 95°C step for 3 min, followed by eight cycles with 95°C for 30 sec, 55°C for 30 sec and 72°C for 30 sec. After a final elongation step at 72°C for 5 min, the samples were again purified using the ProNex Size Selective Purification System (Promega) with a 2x ratio according to the manufacturer's instructions. The fragment size was again verified using the DNA 1000 Kit on the Agilent 2100 Bioanalyzer, the concentration was determined using the Qubit dsDNA HS Kit and then pooled equimolarly before submission to the EMBL Genomics Core facility. Sequencing was performed on a MiSeq sequencer using the MiSeq 50 kit (Illumina) with the 10-8-8-50 format, with 10% of phiX spike in.

2.5.2.3.2. Amplicon sequencing with single sided dephasing adapters

For the amplification of the barcode region encoded in the second generation sITR plasmids the approach with single sided dephasing adapters was used. For samples derived from cell culture material, the prepared cDNA derived from transduced cells was further diluted 1:2 with H_2O and 2.5 µl were used for the PCR reaction. For cDNA generated from mouse tissues, 5 µl of the cDNA were used input to the PCR reaction. Rolling circle amplified DNA was linearized by a restriction digest for debranching as described before (see section 2.2.1.13) and 1 µl was used as amplicon. For purified AAV vectors, about 10^9 vector genomes were used as PCR template. For gDNA extracted from murine tissue, 90 ng of gDNA were used for the PCR reaction. The forward primers with Read 1 adapter and 0-7 N dephasing (448FB - 455FB)

sequence were pooled at 1:1:1:1:1:1:1:1 molar ratio, and a single reverse primer (456FB) were used for the barcode amplification. PCR, purification, secondary PCR, quality control and sample pooling were performed as described in the previous section 2.2.5.3.1. Sequencing after submission was performed at the EMBL Genomics Core facility using a NextSeq 500/550 High Output Reagent Kit v2 (75 cycles) in the 75-8-8-0 format or the NextSeq 1000/2000 P2 (100 cycles) Kit in the 100-8-8-0 format, both with a spike in of 10% phiX.

2.2.5.4 Adapted ITR-seq for integration region – barcode identification

Two slightly different approaches for library preparation to trace both barcode variant and integration region by NGS were used to determine the integration efficiency. For sequencing using a commercial sequencing the primer used for amplification contained primer binding sites for TruSeq indexing. For sequencing on a single flow cell and thus increased Read recovery primers with dephasing adapters and binding sites for Nextera indices were used.

2.2.5.4.1 Amplicon generation

Isolated gDNA was fragmented using the NEBNext UltraShear (NEB) with 1000-5000ng of gDNA input. The digest time was optimized by determining size distribution of fragmented DNA using the DNA 1000 Kit on the Agilent 2100 Bioanalyzer. When the majority of the fragments were between sizes of 300-1000 bp the samples were purified using the magnetic bead based ProNex Size Selective Purification System (Promega) with a 1.3x ratio to remove fragments smaller than 350 bp. The fragmented DNA was then subjected to end-repair using the NEBNext End Repair Module (NEB) following the manufacturer's recommendation and again purified using magnetic beads. Then, dA-tailing was achieved using the Klenow Fragment (3'-5' exo$^-$) supplied in the NEBNext dA-Tailing module (NEB) according to the manufacturer's instructions, followed by another magnetic bead-based purification step using the ProNex Size Selective Purification System (Promega). Oligonucleotide annealing, as described in section 2.2.1.4, was performed to generate Y-adapter oligonucleotides, with different Y-adapter primer pairs depending on the following processing. For sample generation without dephasing adapters, oligos 667FB and 668FB were annealed, for samples with subsequent dephasing adapter attachment the Y-adapter was generated with 667FB and 679FB.

2.2.5.4.2 Library finalization without dephasing adapters

For samples sequenced using a commercial amplicon sequencing service, ligation of Y-adapters to the purified DNA after dA-tailing was performed using the Quick Ligation

Kit (NEB). Therefore, 1 µl of the annealed Y-adapter (10 µM) was ligated to the end-repaired fragmented DNA according to the manufacturer's recommendation in a volume of 40 µl. The ProNex Size Selective Purification System (Promega) was used with a 1.3x ratio to remove non-ligated adapters.

As commercial amplicon sequencing services permits sequencing amplicons with low sample diversity, as a variety of diverse samples were pooled on one flow cell. Thus, the TruSeq Read 1 adapter was already partially included in the Y-adapter. The TruSeq Read 2 sequencing adapter was incorporated during the PCR on the ligation reaction using the Phusion Flash High-Fidelity PCR Master Mix (Thermo Fisher Scientific) with primers 669FB and 670FB according to the manufacturer's protocol and as described in 2.2.1.2. The PCR reactions were then again purified using ProNex Size Selective Purification System (Promega) at a 1.3x ratio, the concentration was determined using Qubit and the size distribution was analyzed via the DNA 1000 Kit on the Agilent 2100 Bioanalyzer. The samples were then diluted to 20 ng/µl in 25 µl H_2O and sent for sequencing using the Amplicon-EZ NGS service by Azenta/Genewiz.

2.2.5.4.3 Library finalization for NGS with dephasing adapters

When sequencing a single amplicon type on a flow cell, the inclusion of dephasing regions in the primers is essential. Despite variable genomic regions present in the sequencing library in this experiment, the sequence diversity with the Y-adapter sequence was limited and might not be sufficient for successful cluster identification during NGS. To finalize the amplicon generation the annealed Y-adapter was ligated to the purified DNA after dA-tailing using the Quick Ligation Kit (NEB) in a volume of 30 µl. Equimolar forward and reverse primer (680FB-694FB) pools with dephasing adapters that facilitate cluster identification, were used for the amplification of the prepared library and the attachment of Read1 and Read2 adapters. The PCR was performed with Q5 High-Fidelity 2x Master Mix (NEB) as described in 2.2.1.2 using 500 nM of the forward and reverse primer pools (pooled at equimolar ratios) and 3 µl of the ligation reaction as reaction input in a volume of 25 µl. The PCR reactions were purified with the Size Selective Purification System (Promega) at a 1.3x ratio according to the manufacturer's instructions. The amplification was verified using the using the DNA 1000 Kit on the Agilent 2100 Bioanalyzer. Attachment of Illumina indices and adapters was performed in a secondary PCR with the Nextera XT Index Kit v2 (Illumina) primers and the KAPA HiFi HotStart ReadyMix PCR Kit (Roche) via an eight cycle PCR reaction in a volume of 25 µl. 12.5 µl 2x KAPA HiFi HotStart Ready Mix were mixed with 2.5 µl of the purified first PCR

reaction, 2.5 µl of Nextera Index Primer 1, 2.5 µl of Nextera Index Primer 2 and 5 µl H_2O with different Primer 1/Primer 2 combinations for each sample. The PCR reaction in a thermocycler was initiated with a 95°C step for 3 min, followed by eight cycles with 95°C for 30 sec, 55°C for 30 sec and 72°C for 30 sec. After a final elongation step at 72°C for 5 min, the samples were again purified using the ProNex Size Selective Purification System (Promega) with a 1.3x ratio according to the manufacturer's instructions, the fragment length distribution determined using the DNA 1000 Kit on the Agilent 2100 Bioanalyzer. The samples were mixed at equimolar ratios and sequenced at the EMBL Genomics Core facility using the Illumina MiSeq PE250 in the 250-8-8-250 format.

2.2.6 Animal experiments

Experimental procedures in mice (*i.e.* viral injection, euthanasia and organ harvest) were conducted by Guilherme L. Gabriel, Rui Jorge Nobre, Kevin Leandro and Luís Pereira de Almeida at the Center for Neuroscience and Cell Biology, University of Coimbra. These experiments were previously approved by ORBEA (Animal Welfare Body of the Faculty of Medicine and the Center for Neuroscience and Cell Biology, University of Coimbra) and were conducted in accordance with the European Directive 2010/63/EU on the protection of animals used for scientific purposes. All efforts were made to minimize animal suffering. All researchers received appropriate training (FELASA-certified course) and were accredited by the Portuguese authorities (Direção-Geral de Alimentação e Veterinária). Systemic administration of either the ITR variant library 1 (ITR-var 1-54 + WT ITR2) or library 2 (ITR-var 55-90 + WT ITR2) was performed in female, 6-week-old C57BL/6 mice by retroorbital injection. The animals were sacrificed 14 days or 60 days post-injection by cervical dislocation. The animals were immediately perfused with cold 1x PBS before extracting the following tissues: Eye (whole), Brain (Cerebellum and Brainstem; CBS, whole), Brain (Midbrain, Thalamus and Forebrain; MTF, whole), Spinal cord (portion), Heart (portion), Lungs (portion), Diaphragm (portion), Liver (portion), Pancreas (portion), Spleen (portion), Kidney (whole) and Muscle (Quadriceps femoris, portion). After extraction samples were frozen in Liquid N_2 and stored at -80°C before shipping to Heidelberg on dry ice for subsequent RNA and DNA extraction.

2.2.7 Bioinformatics and statistical analyses

2.2.7.1 PacBio sequencing data analysis

PacBio sequencing data analysis was performed by Mitchell Yip, Suk Namkung and Phillip Tai at the University of Massachusetts Medical School (UMMS) according to their previously published workflow [363]. In brief, read files were generated in SMRT Link with the option --minSnr=3.75 --minPasses=2 –minZScore=-10 –bystrand. Then the data was further analyzed using the UMass High Performance Computer Cluster and the Gao Lab Galaxy platform. First samples were demultiplexed by PacBio Barcodes and mapped to the vector reference genome as well as the lambda phage reference using BWA-MEM with the –x pacbio option. The barcode sequence in the vector genome was isolated using cutadapt (80 bp sequence preceding the 7 nt barcode with 15% maximum error rate) and sorted using Barcode Splitter. Mapped and filtered reads were visualized using IGV (v2.18.4).

2.2.7.2 Nanopore sequencing data analysis

Nanopore sequencing data were already demultiplexed automatically by the MinKNOW™ (v22.05.5) (Oxford Nanopore) software during sequencing and post-run analysis. Reads with a Q score below 7 were excluded. Adapter trimming and re-basecalling using guppy (v6.1.5) with the high accuracy model (dna_r9.4.1_450bps_hac) was performed by Conradin Baumgartl (Universität Heidelberg, AG Grimm). The reads were then aligned to a reference vector genome, a partial reference genome or linearized plasmid sequence using Minimap2(v2.24) with the default setting (map-ont) [364]. For analysis of the plasmid integrity and the formation of circular replication intermediates, Minimap2 was used with the option for spliced alignment (*-ax splice*) to allow mapping with elongated gaps. For samples with ITR hairpin regions of the same length with high similarity to the wtITR, a consensus sequence of the wtITR in flip and flop configuration based on IUPAC ambiguity codes was used as reference genome. For the sequencing data from pools of ITRs with different lengths and divergent sequences, for each ITR variant a separate reference sequence with the consensus IUPAC ambiguity codes was generated and used for the alignment.

For the analysis of barcode and ITR integrity, after the alignment, sequence alignment map (sam) files were compressed (bam) and indexed using samtools and further processed using a custom R (v4.3.3) script. In brief, Rsamtools (v2.18.0) was used to import the bam file and their index. The GenomicAlignments (v1.38.2) package [365] was used to generate a GAlignments object. Using the GenomicRanges (v1.43.1) package [365] the region ranging from the barcode

to end of the ITR hairpin as defined as region of interest. Then the aligned sequences were extracted as DNAStringSet. The Barcode sequence in the sequence was identified and matched with the actual barcodes used for vector production, with a tolerance of up to three mismatches within the 15 nt barcode. For plotting of the ITR sequence base composition, sequences lacking the hairpin region of the ITR or sequences with preliminary read termination were excluded. Then the sequence in the hairpin region at each position was tabulated and counted.

For analysis of the integrity of the ITR region, after matching the barcode the whole ITR region was extracted. Reads terminating in the ITR region were excluded. Then the length of deletions and „splice gaps“ as well as the longest consecutive gap consisting of both types of gaps in each read was determined and visualized. The coverage was determined using a custom R (v4.3.3) script using the bamsignals (v1.34.0) package. All data was visualized using ggplot2 (v3.5.0). For overview of the whole alignment, the data was examined via IGV (v2.18.4), for analysis of few selected long reads, the reads were aligned against a concatenated reference plasmid sequence in Geneious (V7.1.7).

2.2.7.3 Barcode sequencing (BC-seq) analysis

Barcode amplicon NGS results from MiSeq and NextSeq sequencing experiments were received from the EMBL Genomic Core facility as demultiplexed fastq files and further analyzed in two steps. First the barcodes were extracted and matched to the reference barcode sequences in a custom R (v4.3.3) script based on previous publications[190,366]. In brief, reads were extracted using the ShortRead (v1.60.0) package [367]. The barcode was recovered by identifying an 8 nt flanking sequence upstream and downstream of the barcode, with one mismatch tolerated. The barcode sequence was then matched with the barcodes used in the experiment with a mismatch tolerance of two nucleotides (15 nt barcodes) or one nucleotide (7 nt barcodes) and the barcode occurrences were counted. For barcode sequences extracted from Nanopore sequencing data that were used as independent verification of the productivity NGS data, the barcode sequences from the nanopore read alignment processed as in section 2.2.7.2 were extracted and the barcodes were identified and counted with a tolerance for three mismatches.

The barcode count (n_{ITR}) in a sample of interest α was then used to calculate an enrichment score for each barcode/ITR variant. The enrichment score represents the factor distinguishing

the relative barcode frequency in sample α: $\frac{n_{ITR\propto}}{\sum_{\propto} n_{ITR\propto}}$ from the relative barcode frequency in a reference sample β: $\frac{n_{ITR\beta}}{\sum_{\beta} n_{ITR\beta}}$. Thus, the enrichment score was defined as:

$$ES_{\propto} = \frac{\frac{n_{ITR\propto}}{\sum_{\propto} n_{ITR\propto}}}{\frac{n_{ITR\beta}}{\sum_{\beta} n_{ITR\beta}}}$$

If the sequencing of the reference samples β, *e.g.*, for the input vector, was performed in duplicates or multiplicates (x). Then the relative barcode frequency in sample α was set in relation to the mean of the relative barcode frequency in the reference samples β: $\frac{1}{x_{ITR}} \sum_{i=1}^{x_{ITR}} \frac{n_{ITR\beta}}{\sum_{\beta} n_{ITR\beta}}$. In these samples, the enrichment score was then defined as:

$$ES_{\propto} = \frac{\frac{n_{ITR\propto}}{\sum_{\propto} n_{ITR\propto}}}{\frac{1}{x}\sum_{i=1}^{x} \frac{n_{ITR\beta}}{\sum_{\beta} n_{ITR\beta}}}$$

For samples in which there was no reference sample, *e.g.*, the vector production from separate productions, no enrichment score could be calculated. In this case the relative barcode distribution given as $\frac{n_{ITR\propto}}{\sum_{\propto} n_{ITR\propto}}$ was used to indicate, *e.g.*, the productivity of different ITR variants directly.

Plots were generated with ggplot2 (v3.5.0) and ggpubr (v0.6.0).

2.2.7.4 Integration analysis

Identification of integration sites was performed by two different sequencing approaches, with and without dephasing adapters. Thus, two slightly different analysis pipelines needed to be employed to identify insertions sites.

2.2.7.4.1 Identification of integrated vector genomes from reads without dephasing adapters

The fastq files of Read 1 from AmpliconEZ sequencing experiments were processed on the Galaxy web platform. Using the Je-clip tool (v1.2.1), the UMI section of Read 1 was clipped off and stored in the read name additionally the rest of the adapter sequence was clipped as well (settings: *Len8, Add1, XT16*). Then, the clipped read was aligned to a suitable reference genome (mm10 for in vivo experiments in C57BL/6 mice) using BBMap (v39.08) with standard settings except: *minid = 0.3* and *local = t*. Generated bam files were then further

processed in a custom R (v4.3.3) script. In brief, using the Rsamtools package (v2.18.0) bam files were imported and matching raw read sequences were loaded using the ShortRead (v1.60.0) package [367] and pre-filtered. Reads were removed where Read 2 does not contain the primer sequences used for amplification or the barcode region, this step was accompanied by the extraction of the barcode sequence for each read. Additionally, reads were removed if Read 1 matches to the Read 2 sequencing adapter, avoiding short amplicons and potential artifact integrations. An additional filter was set to the mapq score of the Read 1 alignment, with a minimum mapq of 10. For mapq < 10, the read was retained when a partial match of the last 16 nt of Read 1 to the vector genome downstream of the barcode sequence was detectable. This was to ensure that reads with short alignment to the host genome with low mapq scores, which contain larger fragments of the vector genome, were still recognized as integrated vector genome, albeit their precise mapping might be ambiguous. The remaining reads were then considered as pre-filtered reads.

2.2.7.4.2 Identification of reads from integrated vector genomes with dephasing adapters

Read 2 of the de-multiplexed fastq files received from the EMBL Genomics Core Facility were processed using the Galaxy web platform (https://usegalaxy.eu/). The first 46 nt of Read 2 were trimmed using Trim (v0.0.2). Due to the variable length of the adapters, this can remove up to three nucleotides aligning to the reference genome or leave up to four nucleotides of the Y-adapter attached to the read. Then Read 2 was aligned to the reference genome (mm10) using BBMap (v39.08) with standard settings, except: *minid = 0.3* and *local = t*. The data was then further processed in a custom R script. After importing and combining bam files and raw read sequences, as described 2.2.7.4.1. First the dephasing adapter was removed from the raw Read 1 and Read 2 after matching the 8 nt sequence downstream of the adapter, with one nucleotide mismatch tolerance. Then, the UMI was extracted from Read 2. Reads were then pre-filtered. Read 1 should contain the PCR primer sequence used for the amplicon generation as well as the region surrounding the barcode sequence in the vector genome, which was subsequently extracted. The raw Read 2 was required to match an 11 nt sequence upstream of the UMI. If Read 2 matched to the Read 1 adapter, the read was discarded due to the possibility of artifacts from very short amplicons. All reads below a mapq of 10 were discarded, except if they were showing a partial match of a 16 nt terminal region to the vector genome. All remaining reads were designated as pre-filtered.

2.2.7.4.3 Integration site analysis

As the read count is not indicative of unique integration sites, further analysis of the pre-filtered reads was required using a custom R script. First, the barcode sequence was matched to a list of the barcodes present in the vector genomes of the given sample. If the barcode region that was extracted during pre-filtering did not match to a barcode sequence in the input library (max. 2 mismatches) the read was removed. Reads mapping to the same region (+/- 10 nt) with the same UMI but contain different barcodes were removed, as these were likely artifacts. Reads with the same Barcode at a given insertion site (+/-100 nt) were counted and associated with unique insertion events. Additionally, clonal expansion was determined by counting the occurrence of unique UMIs for each barcode at the insertion sites. Plots were generated with ggplot2 (v3.5.0). Mappings to Blacklisted Regions were identified using the DAC Blacklisted Regions bed file (ENCFF547MET) generated for the ENCODE project [368]. Annotation of insertion sites was performed using the annotatr (v1.28.0) package [369], using the txDb.Mmusculus.UCSC.mm10.knownGene(v3.20) and org.Mm.eg.db (v3.20) annotation database. For randomization of the integration regions, the regioneR (v1.34.0) package [370] was employed using the BSgenome.Mmusculus.UCSC.mm10 (v3.20) full genome sequences. The data was then plotted using ggplot2 (v3.5.0). For visualization of vector-genome junction, reads were aligned using Geneious (V7.1.7) to the reference sITR genome.

2.2.7.5 Statistical analysis

Plotting of the data was performed in R (v4.3.3) with the ggplot2 package or using Prism5 (v5.03), statistical analysis was performed using R (v4.3.3) using base R and the ggpubr package. Normal distribution was determined using the Shapiro-Wilk test. When the null hypothesis was rejected ($p>0.05$), normal distribution was assumed. Pairwise comparison of normal distributed data was performed using the Student's t-test. Pairwise comparison, when the Shapiro-Wilk test did not detect normality, was done using the Wilcoxon test. Statistical analysis on ordinal scale values (*e.g.*, mapq scores) was performed using the Kruskal-Wallis test.

3. Results

3.1 A novel pipeline to engineer and screen ITR variant vectors

Despite their great potential to enhance rAAV-based gene therapy, the ITRs' potential is still not fully harnessed, limited by the lack of screening pipelines for ITR variants. The most extensive ITR variant screen was performed with D-sequence variants, with a focus on productivity [357]. In contrast, a parallel screen of ITR hairpin variants during transduction has so far never been performed. To facilitate this, alternative approaches for ITR engineering were adopted and developed in this study, harnessing approaches from high-throughput capsid engineering and screening, and translating them to ITR variant screens.

3.1.1 Single-ITR plasmid design for the screening of ITR variants

The first major challenge during ITR engineering was the generation of appropriate vector genome plasmid constructs. Reliance on the classical vector plasmids with two ITRs would require inefficient and expensive cloning strategies that would make large ITR libraries unattainable, thus I decided to harness an alternative vector plasmid design. Xiao and colleagues described the use of a plasmid with a single ITR (sITR) of 165 bp in length [256], which leads to the packaging of the whole plasmid sequence as vector genome (Figure 7A). Despite the elongated ITR, this construct design allowed me to introduce alterations in the ITR sequence using simple restriction enzyme-based cloning procedures. More specifically, the generation of a base construct with a dual type IIS restriction enzyme (TIIS) in the center of the ITR facilitated an efficient insertion of ITR hairpin variants as annealed oligonucleotides. Mutations within the stem of the ITR (D- and A- sequence) could be introduced via PCR using a single primer due to the palindromic nature of the ITRs, followed by the subsequent insertion of the hairpin as annealed oligonucleotide. To be able to measure the functionality of vectors generated with the sITR constructs, I included a CMV promoter-driven *mCherry* transgene with a short synthetic polyA-signal. Conventional vector production protocols could be applied with the sITR plasmid, relying on co-transfection of the standard helper plasmids encoding AAV *rep* and *cap* as well as the Adenoviral helper genes. The *mCherry* expression after transduction could then be used to evaluate the functionality of the generated vectors.

The first tested sITR plasmid construct contained the full-length wtAAV2 sITR sequence described by Xiao et al. (Figure 7B) [256], which should lead to the generation of vector genomes enclosed by the wtITR2 and would therefore be indistinguishable from conventionally produced rAAV. The large secondary structure also highlights the challenge of sequencing the

sITR, which was only possible after linearization of the ITR by restriction enzyme cleavage within the hairpin sequence[357]. After rAAV production using this wtITR2 sITR plasmid, the vector was used to transduce Hek293T cells at different MOIs to confirm its general functionality (Figure 7C). Fluorescence microscopy revealed a clear dose-dependent increase in mCherry signal, confirming that sITR constructs generate infectious rAAV particles.

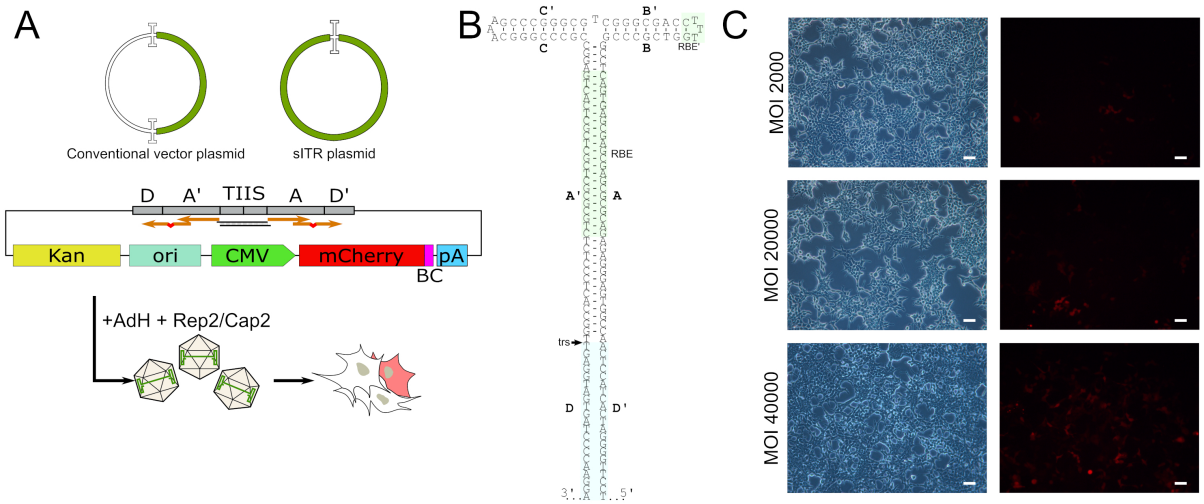


Figure 7: sITR construct design for ITR engineering and proof of function. Top: Comparison of the conventional vector plasmid and the sITR plasmid. The part of the plasmid that is packaged in the AAV particles is indicated in green. Center: The used sITR contains the prokaryotic kanamycin (Kan) resistance cassette, ori and CMV promoter-driven *mCherry* transgene with a short polyA (pA). In the 3' UTR of the *mCherry* gene, a short 7 nt barcode (BC) was incorporated. The base sITR construct contains both D/D'- and A/A'- sequences in the stem, with a dual TIIS cleavage site at the tip of the stem. Using Golden Gate cloning, the hairpin can be inserted as simple annealed oligonucleotide. For modification of the stem sequence, PCRs with single primers (orange arrows) can be used for amplification of the stem. Incorporation of mutations distant from the hairpin tip via mismatches in the primer (red indent in orange arrow) requires two single-primed PCR reactions. Bottom: Vector production can then be performed using the conventional *rep/cap* and Adenoviral genes. Vector function can finally be traced by visualizing mCherry expression in the transduced cells. **(B)** The sequence of the 165 nt plasmid wtITR2 with all elements possessing complementary sequences leading to a D-A'-C-C'-B'-B-A-D' configuration. RBE and RBE' region are indicated in green, the D-sequence in blue. **(C)** Fluorescence microscopy images after transduction of Hek293T cells at the indicated MOI with vectors generated with *rep/cap* from AAV2 with a wtITR2 sITR plasmid. A brightfield image is of the same is shown on the left. Scale bar is 50 µm.

3.1.2 Tracing barcoded ITR hairpin variants during production and transduction

Following the successful production of functional vectors using the sITR plasmid system, the next objective was to establish a method for tracking ITR variants during transduction. ITRs can be sequenced in rAAV vector genomes, but they are frequently lost after transduction or truncated during episome formation [355] or integration [214]. Thus, to trace the ITR variants after transduction, an independent identifier distinguishing the ITR variants was required. For small AAV capsid library screens, this is typically achieved by linking each capsid variant with a

unique barcode in the vector genome, which can then be tracked post-transduction [190,371]. I adapted this strategy for ITR variants by inserting unique, seven nucleotide barcodes into the 3' untranslated region (UTR) of the mCherry transgene, as previously indicated in Figure 7A, each assigned to a different ITR variant. This design enabled the quantification of each variant's abundance in the vector library as well as the subsequent tracking of their presence post-transduction, independent of ITR sequence integrity.

For a proof-of-concept screen, I created a small ITR variant library consisting of the wtITR2 and eleven alternative hairpin sequences with varying degrees of divergence. Since little is known about structural features that may affect transduction, these variants comprised diverse structural motifs, including known DNA aptamers [372,373] and hairpins of different shapes and structures (Figure 8A). For verification of the ITR sequence in the sITR plasmids, I again relied on classic restriction enzymes for plasmid linearization and sequencing via conventional Sanger sequencing. The barcoded sITR variant plasmids were then pooled equimolarly and used for vector production with AAV2 *rep* and *cap*. The purified vector library was subsequently used to transduce Hek293T cells. The barcodes in the produced vector library and in the RNA/cDNA extracted at 24 h and 48 h post-transduction were amplified by PCR and sequenced using the AmpliconEZ NGS platform. Barcode analysis revealed a strong skew toward the wtITR2 in the produced vector (Figure 8B). This suggested that, despite the use of equimolar plasmid ratios during production, the wtITR2 substantially outcompeted all other variants during production. The only ITR variants that still exhibited elevated productivity had a high similarity to the wtITR2. This includes the wtITR1 and the AAV2-TA-loop, an AAV2 ITR with an TA-rich loop region in the hinge between the B- and C-hairpin. The strongly skewed distribution also persisted after transduction in the recovered RNA/cDNA. Only minor changes were detectable between time points and no major changes were found compared to the barcode distribution in the input library. A slight increase in the relative abundance of the wtAAV1 hairpin was observed, which could indicate a modest advantage during transduction. However, the dominant influence of wtITR2 on vector production overshadowed any minor differences in transduction efficiency.

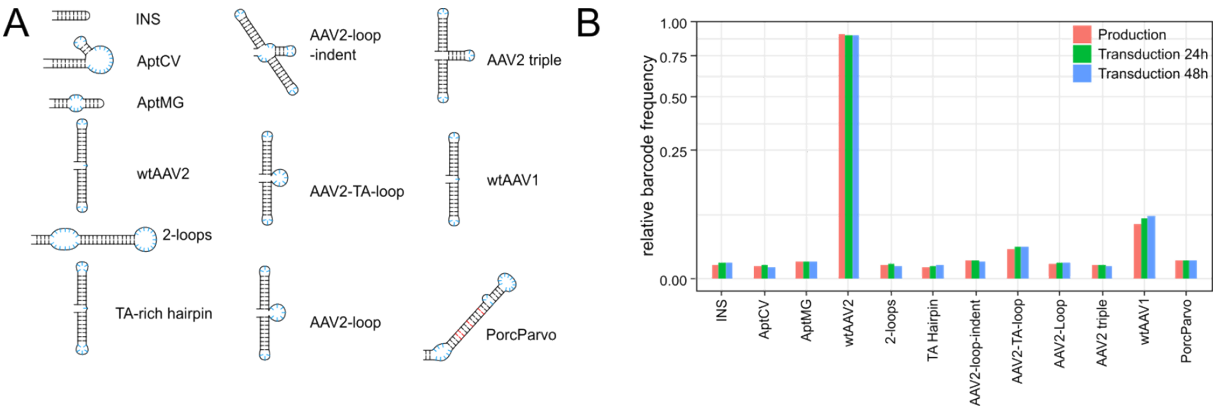


Figure 8: Tracing multiplexed production and transduction of barcoded ITR hairpin variants. **(A)** Simplified depiction of the ITR hairpin structures attached to the AAV2 ITR stem (D+A sequence). Structures were predicted with mfold[236] and manually simplified for visualization. Matching nucleotides are depicted in black, mismatched nucleotides in red and unmatched nucleotides in blue. **(B)** Barcode distribution after sequencing using the AmpliconEZ sequencing service. The barcodes were determined in the produced vector library as well as 24 h and 48 h after transduction of Hek293T cells with an MOI of 40000 using RNA/cDNA as template for barcode amplification, showing an overrepresentation of the wtAAV2 hairpin. INS = insertion sequence, AptCV = crystal violet binding aptamer, AptMG = malachite green binding aptamer, PorcParvo = porcine Parvovirus hairpin.

3.1.3 A novel approach for sequencing ITRs in plasmid DNA

Following the production of barcoded sITR variants, the findings of a high underrepresentation of the alternative hairpin sequences and of the potential of elevated transduction with the wtITR1 hairpin implied to initially focus on wtITR hairpin variants and derivatives. However, some hairpins (*e.g.*, wtAAV5) lack restriction sites that would be compatible with linearization and subsequent two-sided Sanger sequencing and thus could not be included in the initial library. Existing commercial sequencing methods that can resolve ITR sequences, such as Nanopore-based whole-plasmid sequencing and ITR-plasmid sequencing, which relies on adapted Sanger sequencing protocols [240], are cost-intensive and require a high turn-around time in comparison to conventional Sanger sequencing services. Additionally, although the cloning efficiency for the sITR variants was reliably high, the screening of multiple clones was often necessary, especially after plasmid retransformation. Therefore, a fast, low-cost and dependable method for ITR sequencing was crucial for the parallel sequence confirmation of larger ITR variant libraries. To this end, I evaluated different nucleases for their ability to mediate a structure-dependent but sequence-independent resolution of the ITR hairpin, enabling subsequent conventional Sanger sequencing of the ITRs. Among the tested enzymes, T7e endonuclease I (T7eI) emerged as a promising candidate. T7eI is known for its use in the detection of genomic mismatches by resolving heteroduplexes at the mismatched position [374,375], but a structure-sensitive cleavage activity also enables the resolution of

Holliday-junctions [375]. I hypothesized that these mechanisms may enable the resolution of the ITR by cleavage at the three-way junctions in the ITR plasmid as well as at mismatched and potentially at unpaired nucleotides (Figure 9A). This wide array of cleavage sites including junctions, loops and bulges would generate a heterogeneous pool of linearized plasmid fragments, which should then be amenable to conventional Sanger sequencing (Figure 9B). The heterogeneity of fragments was expected to lead to a continuous drop in peak height during Sanger sequencing due to fragment dropout, which was indeed partially visible in the exemplary single-sided sequencing read. I first employed this new method, termed T7eI-ITR sequencing, for the sequencing of the 95 nt ITRs in conventional vector plasmids (Figure 9C). Sequencing from both sides allowed for full ITR coverage, generating an overlap in the central region of the hairpin where the peak height was reduced. This overlapping sequence boosts confidence in the sequencing result despite the reduced peak height. Standard Sanger sequencing input amounts (250-500 ng plasmid DNA) proved sufficient for ITR sequence resolution, although higher quantities of digested plasmid DNA increased the peak height in the central region and thus read quality. The mismatching sequence in the terminal part of the read near the end of the dual hairpin was the result of a too low peak height to properly determine the sequence. Occasionally, template switching also led to apparent read inversions with a similarly mismatching terminal part (Supplementary Figure 1), which was readily detectable and excluded from the read.

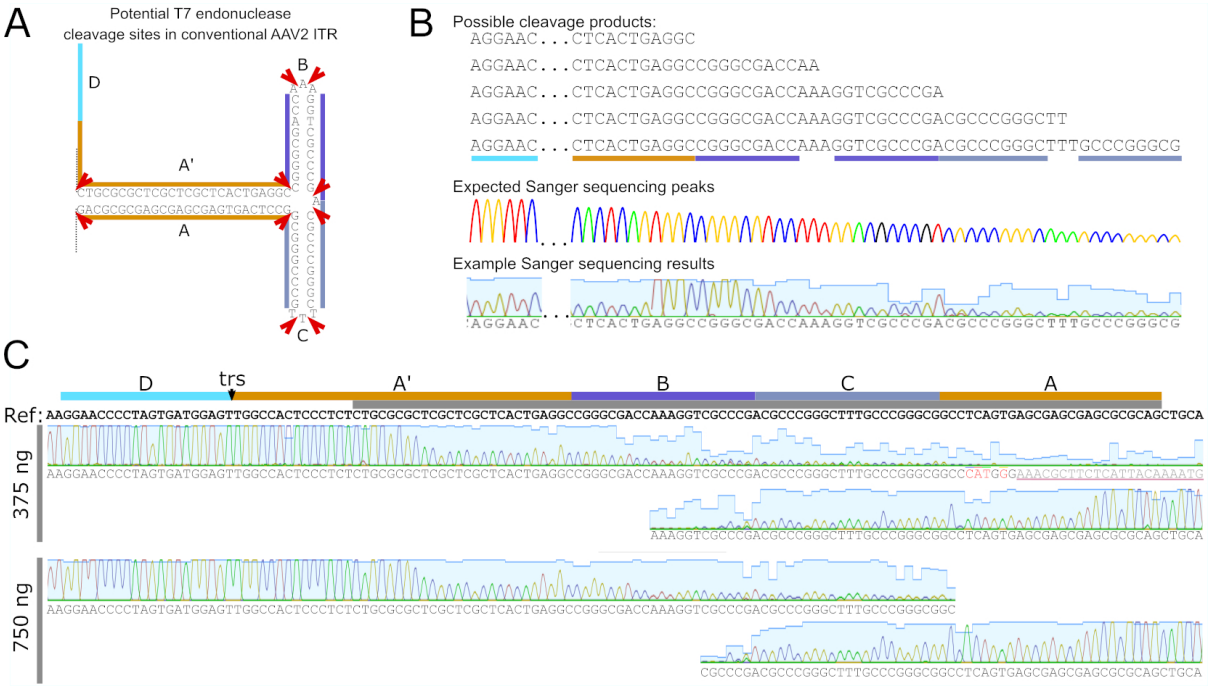


Figure 9: Sequencing of conventional vector plasmid ITRs via T7eI-ITR sequencing. (A) Hairpin resolution mechanism of T7eI with potential cleavage sites in the ITR of a pSSV9/pSub201 plasmid with indicated ITR regions. Red arrows indicate the sites at which T7eI may cleave the DNA. The

different regions of the hairpin are color-coded. **(B)** Top: Potential cleavage products of T7eI-digested ITR sequences, only showing sequences generated by cleavage events in the central region. Hairpin elements are color-coded as in panel A. Center: Expected Sanger sequencing peaks, with a continuous reduction in peak height during progression through the hairpin, due to the distribution of DNA fragments of different length after T7eI cleavage. Bottom: Exemplary Sanger sequencing results of a T7eI-ITR sequencing experiment, showing the peak heights during read progression through the wtITR2 hairpin. The light blue bars in the background of the peaks indicate the read quality score for each nucleotide. **(C)** Two T7eI-ITR sequencing results for one ITR of a pSSV9/pSub201 plasmid, sequenced with different plasmid input quantities. Reads aligned to the ITR reference sequence (bold) using Geneious, with indicated and color-coded ITR-regions. The 95 nt hairpin region is additionally indicated by a gray bar. Mismatching nucleotides are depicted in red, elongated stretches of nucleotides with too low peak height or quality to properly align are depicted in gray and underlined in red. Quality scores for each nucleotide are indicated by light blue bars in the background of the peaks.

The successful validation of the T7eI-ITR sequencing approach for the 95 nt hairpin structure present in conventional rAAV vector plasmids motivated subsequent efforts to apply the approach to the substantially larger hairpins present in sITR plasmids. This could eliminate the reliance on sequence-dependent endonuclease cleavage for hairpin resolution, while also significantly accelerating the throughput of ITR sequencing compared to commercial services. To exemplify the application spectrum of the T7eI-ITR sequencing approach, I not only compared the new method with a commercially available sequencing service, but also assessed the functionality on different hairpin structures, with their sequences and structures indicated in Figure 10. Besides the conventional wtITR2 sITR plasmid (see Figure 7B) a wtITR2 hairpin with a single paired point mutation (Figure 10A) and an AT-rich dual hairpin (Figure 10B) at the tip of the stem were examined. These should be resolved due to their similar secondary structure to the wtITR2, which proved to provide sufficient cleavage sites for the T7eI. To test whether alternative hairpin structures in sITR plasmids also enable T7eI-ITR sequencing, I moreover included a simple hairpin loop (Figure 10C) as well as the terminal hairpin of the *Sibine fusca* (*S. fusca*) densovirus ITR (Figure 10D). In theory, both should enable T7eI-ITR sequencing as they contain unpaired nucleotides as well as junctions that facilitate T7eI cleavage.

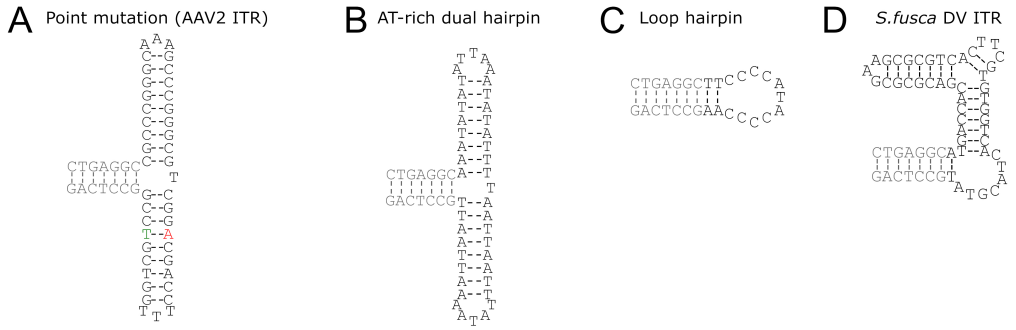


Figure 10: ITR hairpins with diverse sequence and structure to challenge T7eI-ITR sequencing. sITR plasmids with an AAV2 ITR stem sequence (last seven nucleotides of the stem are indicated in gray), but variable hairpin tip secondary structures were employed to exemplify the application spectrum of T7eI-sITR sequencing. The secondary structures of **(A)** the wtAAV2 with a point mutation (colored nucleotides) in the B-hairpin (ITR variant V44), **(B)** an AT-rich dual hairpin, **(C)** a simple loop hairpin, and **(D)** the terminal hairpin of the *S. fusca* densoviral ITR were predicted with mfold [236] and adapted for better visualization.

The commercial ITR-sequencing service reliably covers the whole ITR in conventional ITR plasmids with a 95 nt hairpin with high sequencing quality (data not shown). I also compared the commercial ITR sequencing service with T7eI-ITR sequencing of the wtITR2 in the sITR plasmid, with exemplary results shown in Figure 11A. While the commercial sequencing approach covered the whole 165 bp ITR in the sITR plasmid, the read quality deteriorated during read progression through the hairpin as visible by the emergence of overlapping peaks. The terminal 31 nucleotides of the alignment had such a low read quality that the consensus sequence did not match the reference sequence. The same plasmid could be reliably sequenced using the T7eI-ITR sequencing approach, again generating the read overlap in the central B- and C- region. Despite the drop in read quality near the hinge between the B- and C-hairpin, the complete ITR sequence could be confirmed. Like any sequencing method, T7eI-ITR sequencing also suffers from variations in read quality likely based on plasmid quality or impurities. This can be well exemplified by the sequencing results for the AAV2 ITR with the paired mutation in the B-hairpin (Figure 11B). There, the overlap of forward and reverse sequencing reads was reduced to six nucleotides. Also here, in the hinge region between B- and C-hairpin, a mismatching nucleotide in the forward read could be detected. Despite the reduced overlap and the nucleotide mismatch, the sequence could still be confidently determined to match the reference, aided by the correct background peak and high read clarity in the reverse read. Successful T7eI-ITR sequencing of a TA-rich dual hairpin proved the sequence-independence of the T7eI hairpin resolution approach (Figure 11C). The T7eI-ITR forward and reverse sequencing reads again generated an overlap in the central part of the ITR, thereby yielding a clear consensus sequence matching the reference. In this example, the

hairpin hinge appeared to affect the read quality only in the forward read with a drop in peak height and read quality, after which the read quickly terminated. T7eI-ITR sequencing of a simple loop region as hairpin tip (Figure 11D), studied as an example for an ITR with limited secondary structures, led to reduced peak heights and read quality throughout the ITR. Nevertheless, the reads still overlapped, thus enabling the determination of a complete ITR consensus sequence. This suggested that hairpin tips lacking Holliday-junction-like elements can still be resolved by T7eI-mediated endonucleolytic cleavage, potentially due to the ability to cleave at unpaired nucleotides. However, T7eI sequencing has difficulties resolving perfectly matched linear hairpins, as evidenced by the very few successful sequencing attempts that yielded low peak heights and read quality (Supplementary Figure 2). The T7eI-ITR sequencing results for a sITR plasmid with the hairpin of the *S. fusca* densovirus ITR, containing both loop regions and hairpin structures, also showed a successful sequence recovery (Figure 11E). It is noteworthy that this hairpin exceeds the length of the conventional 165 bp sITR, proving that even larger hairpins can be sequenced using this method. Despite local drops in peak height and even base mismatches near hinge sites, the information of the opposite sequencing read was used to determine a match with the reference sequence. These exemplary datasets demonstrate the ability of the method to resolve long and diverse hairpins, provided the structure elements contained sufficient cleavage sites for the T7eI enzyme.

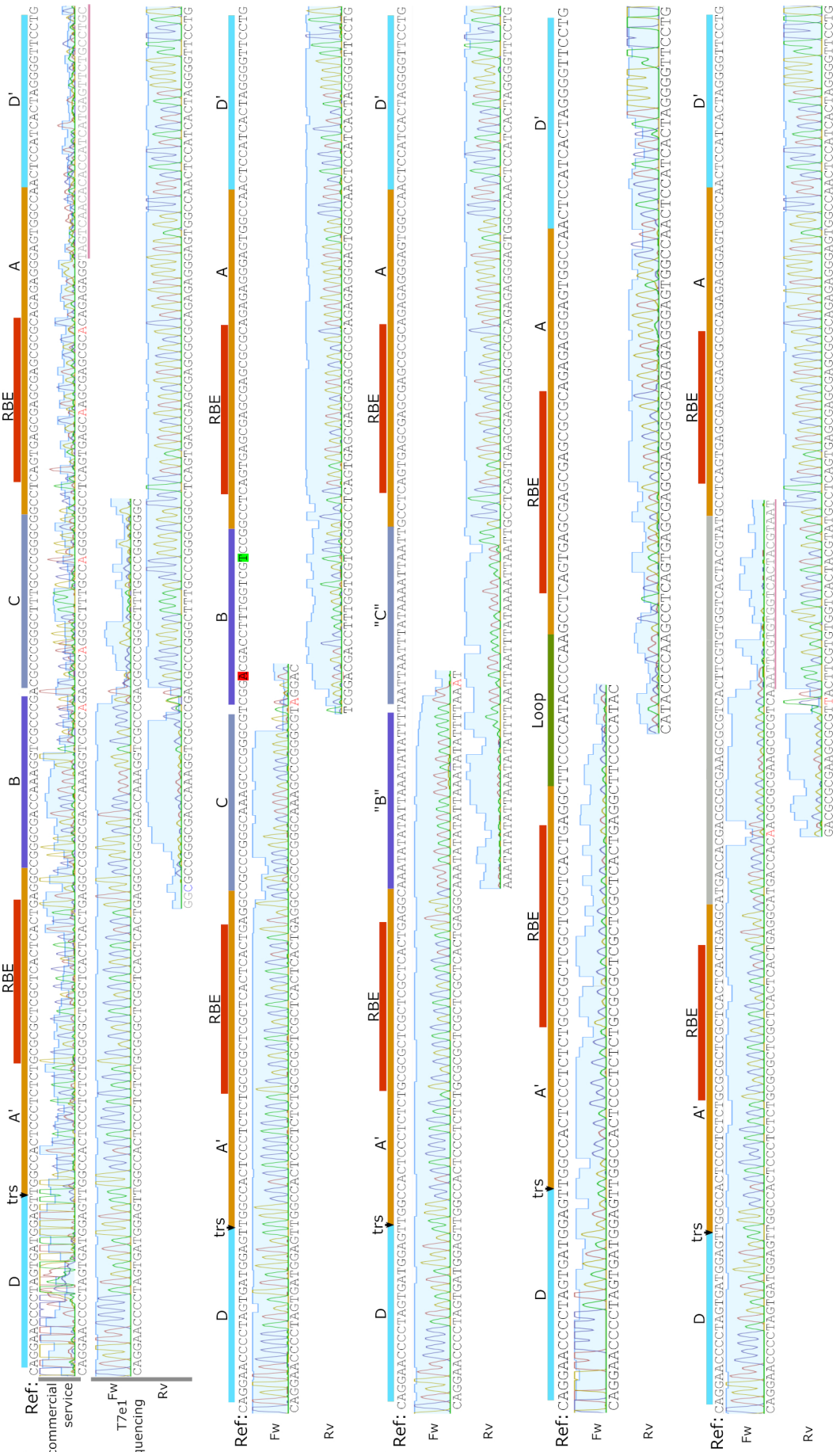

A
Ref:
commercial service
T7e1
sequencing
Fw
Rv
D
trs
A'
RBE
B
C
RBE
A
D'
B
"B"
"C"
Loop

Figure 11: T7eI-ITR sequencing with sITR plasmids – comparison with a commercial service and sequencing of alternative hairpin sequences. (A) Comparison of sequencing results for a 165 bp wtAAV2 ITR in an sITR plasmid using a commercial ITR sequencing service and T7eI-ITR sequencing. Shown are T7eI sequencing results for an sITR plasmid with **(B)** wtAAV2 ITR with paired mutation in the B hairpin, **(C)** AAV2 ITR stem and an AT-rich dual hairpin, **(D)** AAV2 ITR stem with a loop region at the tip of the hairpin, or **(E)** AAV2 ITR stem and the *S. fusca* densovirus ITR hairpin tip. All reads were aligned to the reference plasmid sequence using Geneious. The ITR regions are annotated in the reference sequence and relevant reference sequence nucleotides (*e.g.*, point mutation compared to the wtITR2) are highlighted. Sequencing reads indicate both peak height and read quality for each nucleotide. Mismatched nucleotides in the reads are indicated in red, elongated stretches of mismatched nucleotides due to low read quality and peak height are given in gray and underlined in red.

3.1.4 Barcoded wtITR hairpin variants during production and transduction

Relying on this novel ITR sequencing technique, I could now verify the incorporation of the wtITR1-7 terminal hairpins as well as the unrelated terminal hairpin of the *Aedes aegypti* densovirus (*Ae*DV) within an AAV2 stem in sITR plasmids. These wtITR hairpins may retain a higher replicative ability, compared to alternative hairpin sequences, and thereby reduce the impact of variant underrepresentation which was previously observed in the first vector library pool (see section 3.1.2). The screening strategy to measure performance of the hairpin variants during production and transduction is shown in Figure 12A. Equimolar sITR plasmid pools were used for rAAV production using AAV2 Rep and AAV1-8 VP, in order to determine if the capsid has an impact on the packaging of different ITR-variants. Hek293T cells were then transduced with the ITR variant vector library with a MOI of 25000 and DNA/RNA was extracted 48 h post-transduction. The barcode distribution was determined in the plasmid pool, in the vector libraries after production, in the RNA/cDNA and in T5 exonuclease-digested gDNA that contained enriched episomal DNA [376], using an NGS protocol for amplicon sequencing with dephasing adapters, which was used similarly for all subsequent barcode sequencing (BC-seq) experiments. The barcode distributions in the vector libraries were normalized to the barcode distribution in the plasmid pool to determine an enrichment score as a measure of productivity and are indicated with the vector library titers measured by qPCR (Figure 12B). As expected, vector productivity varied across capsid serotypes, yet the ratio of ITR variants was relatively uniform with the wtAAV2 hairpin strongly overrepresented in all capsids. Again, productivity seemed to correlate well with sequence similarity to the wtITR2 hairpin, as the wtITR3b and wtITR4 hairpin ranked highest, while wtITR7 and the *Ae*DV hairpin showed the lowest productivity. No capsid-dependent effects on packaging and replication of ITR hairpin variants could be detected. Functional transduction of the ITR variants was then examined by assessing the barcode distribution on the RNA level, which was

then normalized to the barcode distribution in the vector libraries to calculate an enrichment score (Figure 12C). There, the wtITR2 hairpin showed the highest transgene expression post-transduction, with the wtITR3b and wtITR4 hairpins following suit. Differences between the capsids were observable, with the major outliers likely due to their lower transduction efficiency in Hek293T cells [377], but the overall trend of ITR variant behavior across the capsids was similar. In episomal DNA, the barcode distribution was likewise dominated by the variant containing the wtITR2 hairpin (Figure 12D), although the influence of outlier variants, particularly those with weaker transduction efficiency, was more pronounced.

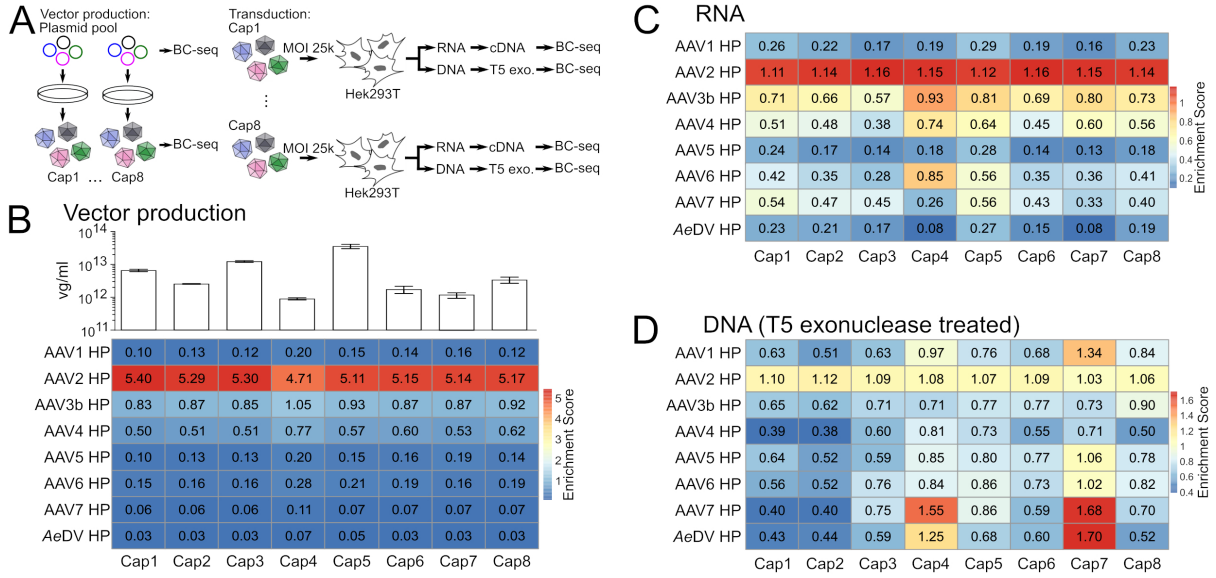


Figure 12: Tracing production and transduction of wtITR hairpin variants. (A) Pipeline for tracing the ITR hairpin variants from the plasmid pool to the vector library after production and subsequent transduction in Hek293T cells. The gDNA and RNA extracted 48 h post-transduction was then processed to determine the barcode distribution. **(B)** Top: Titration of the ITR variant libraries in different capsids via qPCR. Bottom: Barcode distribution enrichment scores for each ITR variant based on the enrichment in the vector compared to the plasmid library. **(C)** Enrichment scores for the ITR variants in the different capsids after transduction on the RNA level, determined by comparing the barcode distribution in RNA and the vector library. **(D)** Enrichment scores for ITR variants in different capsids after transduction in episomal DNA. Enrichment scores were calculated based on the barcode distribution in the T5 exonuclease-digested gDNA compared to the vector library.

3.1.5 Analyzing ITR integrity using third generation sequencing

Having demonstrated the feasibility to track barcoded ITR variants during vector production and transduction, it was imperative to confirm the integrity of the ITR variants within the vector genomes. Third-generation sequencing approaches represent ideal tools to not only sequence vector genomes, including the ITRs, but also for plasmid sequencing to confirm the integrity of the ITRs identified by T7eI sequencing.

3.1.5.1 Verification of the presence of ITR sequences using PacBio sequencing

To address the issue of ITR integrity in vector genomes, I collaborated with the group of Phillip Tai and Guangping Gao at the University of Massachusetts Medical School (UMMS), who kindly applied their PacBio-based AAV genome population sequencing approach [363] to my initial ITR variant library containing the diverse set of ITRs (see section 3.1.2) as well as the wtITR hairpin library (see section 3.1.4) packaged in capsids from AAV1 and AAV2. To this end, the vector genomes were purified and end-repaired followed by attachment of the PacBio sequencing adapters, in accordance with their previous publication [363]. The barcode distribution for the wtITR hairpin library determined by PacBio (Figure 13A) matched the distribution previously determined via NGS, with the wtITR2 hairpin being highly overrepresented. The alignment of the sequences to the reference vector genome did not reveal substantial vector genome aberrations, exemplified with the sequences containing the barcode associated to the wtITR2 hairpin (Figure 13B). The configuration of both ITRs in the plasmid was also analyzed for the ITR variants packaged in capsids from AAV1 (Figure 13C) and AAV2 (Figure 13D). In theory, the four ITR configurations (Flip-Flop, Flip-Flip, Flop-Flop, Flop-Flip) would be expected at equal ratios (1:1:1:1) [34,363]. This expected distribution was indeed observed for vector genomes containing the barcode of the AAV3b hairpin. However, substantial deviations could be noted in vector genomes with barcodes linked to the AAV2 or AAV4 hairpin, which displayed an enrichment of the Flop-Flip and Flop-Flop configuration. In contrast, genomes linked with AAV1 hairpins showed a bias for the Flip-Flip and Flip-Flop configuration, which could suggest an aberrant replication dynamic at the ITRs when using sITR constructs.

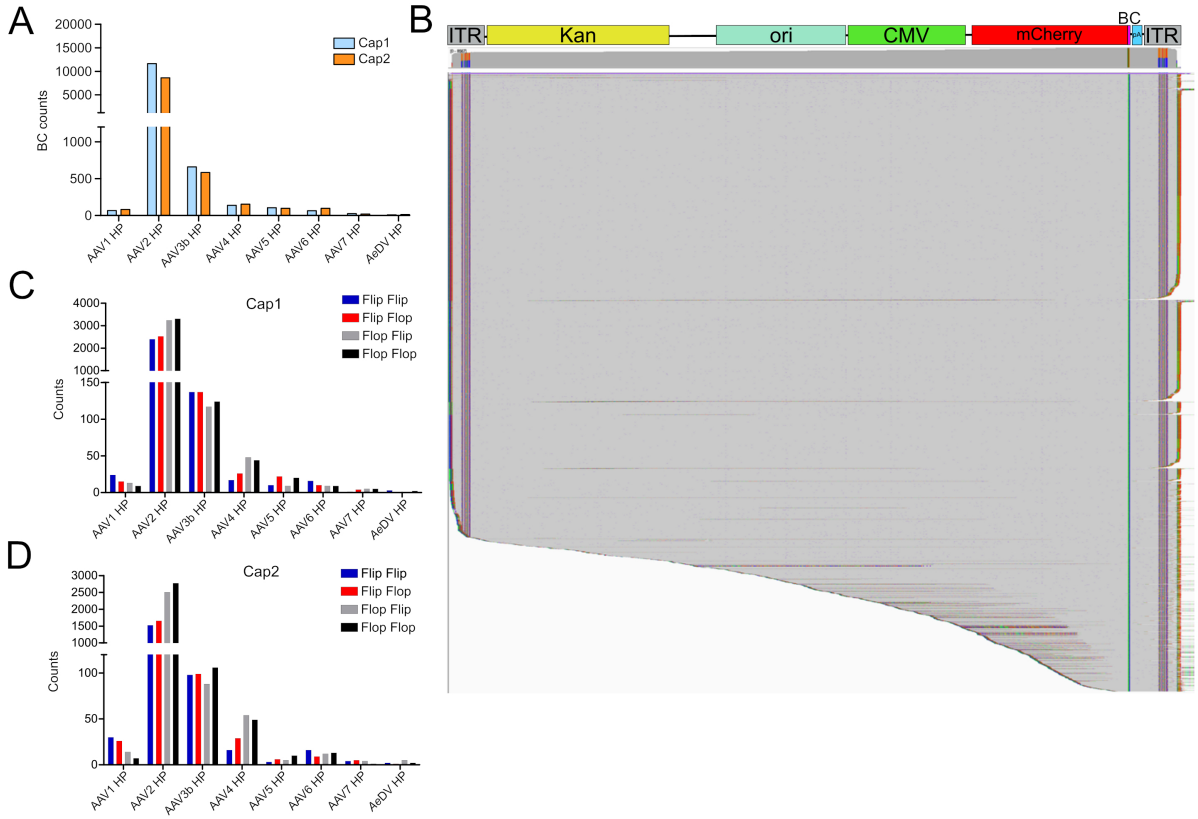


Figure 13: PacBio sequencing of the wtITR variant library to assess genome and ITR integrity. **(A)** Distribution of barcodes encoding the ITR hairpin variants (wtAAV1-7 + *AeDV*) in the vector libraries determined by PacBio sequencing. **(B)** IGV display of the PacBio sequencing reads containing the wtITR2 barcode aligned to the rAAV reference genome, revealing the expected alignment pattern with a majority of intact genomes and no major vector genome aberrations. **(C-D)** Comparison of the ITR orientation of the different ITR hairpin variants in vector genomes packaged in AAV1 **(C)** or AAV2 capsids **(D)**. Data was acquired and analyzed by Phillip Tai, Suk Namkung and Mitchell Yip (UMMS).

3.1.5.2 PacBio sequencing reveals a repair event present in ITR sequences

The overall integrity of the rAAV genomes as well as the aberrant ITR configuration in the vector genomes suggested the necessity to perform a more detailed analysis of the ITR sequences. This revealed that certain ITR sequences had undergone sequence alterations, particularly reversions towards the wtITR2 hairpin.

This was well observable in the alignment of reads linked to the AAV1 hairpin to the 3' ITR AAV1 reference in flop configuration. In genomes packaged in AAV2 (Figure 14A) or AAV1 capsids (Figure 14B), the ITR hairpin structure was preserved overall, except for relatively rare deletions. Despite the slight over-representation of the flip orientation in the 3' ITR, as suggested by the previous analysis, a high percentage of reads was expected to match the reference wtITR1 hairpin in flop configuration. However, many sequences actually diverged from the AAV1 reference in flop configuration and also did not match the AAV1 sequence in

flip configuration. Instead, many reads exhibited a high similarity to the wtITR2 sequence in both flip and flop configuration. Furthermore, a small subset of reads neither matched the hairpin of AAV2 nor that of AAV1 but instead suggested the presence of the AAV4 hairpin. This exchange from the hairpin of one serotype to another could indicate the presence of an inter-ITR repair mechanism that uses ITR sequences from one genome as template to correct another ITR sequence. This would represent a novel, as-of-yet undescribed repair mechanism for ITRs and would complicate the previously described NGS screening of barcodes after pooled production due to a risk of disruption of the barcode-ITR association.

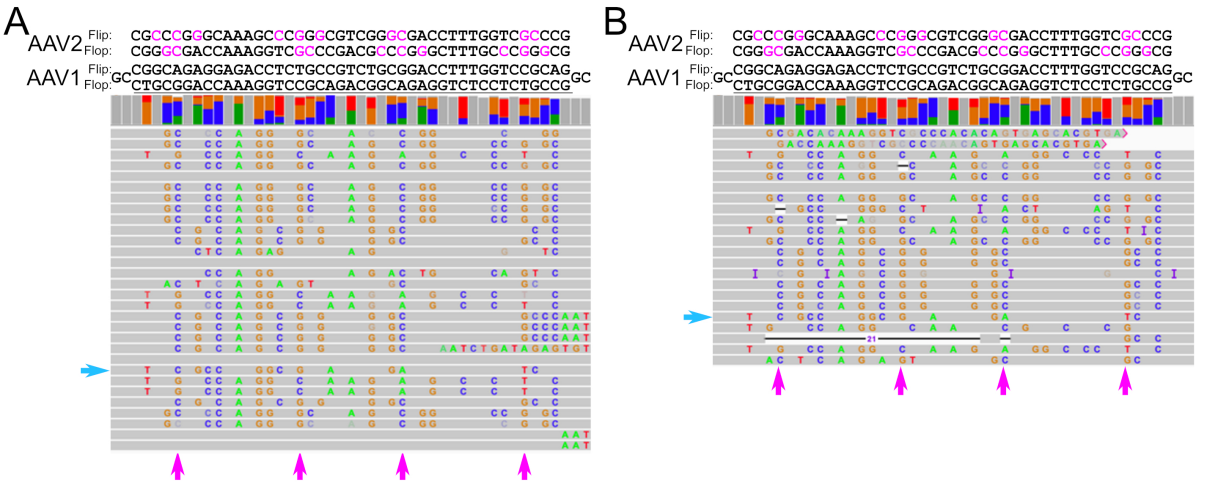


Figure 14: PacBio sequencing of wtITR hairpin variant pools reveals a repair event in the hairpin. **(A-B)** IGV display of the 3' ITR hairpin region of a random selection of PacBio sequencing reads containing the wtITR1 barcode. Vectors genome sequences, packaged in capsids of AAV2 (30/84 aligned reads) **(A)** or AAV1 (22/64 aligned reads) **(B)**, were aligned to the reference genome with wtITR1 in flop configuration. The wtITR1 hairpin sequences in flip and flop (underlined) configuration and two additional nucleotides on each side are indicated above the alignment. The wtITR2 hairpin sequence in flip and flop are given above, with nucleotides differing from the AAV1 hairpin indicated in pink. Reads are colored gray when matching the reference, mismatching nucleotides are colored depending on the base. Pink arrows indicate positions where the nucleotides of the AAV2 and AAV1 hairpin in flip and flop configuration do not match, thus facilitating identification of converted sequences. The blue arrow indicates reads matching the wtITR4 hairpin sequence. Data was acquired and analyzed by Phillip Tai, Suk Namkung and Mitchell Yip (UMMS).

To further dissect this mechanism, the ITRs in the first ITR variant library with more diverse ITR sequences (section 3.1.2) were also examined in more detail. There, the repair effect was even more pronounced than observed with wtITR hairpin variants. For instance, the alignment of genomes with barcodes originally assigned the malachite green binding DNA aptamer (AptMG) hairpin sequence to its reference sequence revealed a near complete absence of the expected sequences matching the AptMG reference in both flip and flop (Figure 15A). Instead, the majority of reads matched the wtITR2 reference in both flip and flop configurations. Other reads did not align with either the AptMG or wtITR2 sequences but instead matched other

hairpin variants present in the screen, such as the TA-rich hairpin variants. In ITR variants with higher similarity to the wtITR2, such as the one containing an additional TA-rich loop sequence between the B- and C-hairpins of the wtITR2, a higher proportion of reads retained the expected sequence (Figure 15B). These reads occasionally also showed the presence of the alternative ITR in both flip and flop configuration, implying the vector DNA had undergone replication rather than merely reflecting residual plasmid contamination. However, the repair rate for other highly divergent ITR hairpins appeared so high that the limited sequencing depth did not allow the detection of a single intact variant hairpin read. One such example was the crystal violet DNA aptamer (AptCV) hairpin sequence (Figure 15C). There, the wtITR2 hairpin, detectable in both flip and flop configuration, dominated the alignment. This emphasized that pooled production of ITR variants is not a feasible strategy due to the extensive sequence corrections. The clear tendency for repair towards the wtITR2 hairpin may indicate a dependency of the conversion mechanism on replication efficiency. Furthermore, the repair towards other alternative hairpins detected for some variants suggested that the effect is likely independent of the AAV Rep protein and only dependent on the presence of a repair template ITR. Nevertheless, I deemed a further characterization of this mechanism necessary to exclude the presence of an artifact.

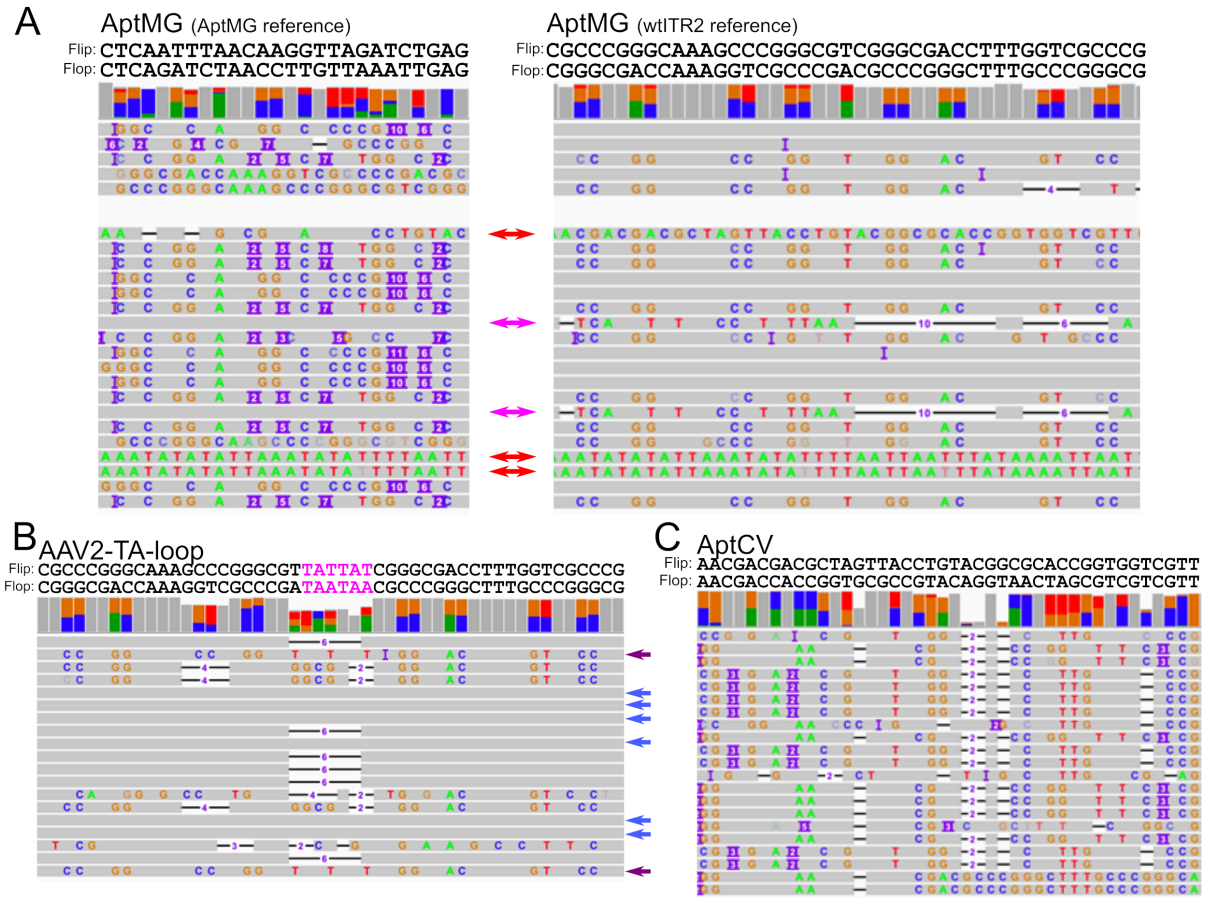


Figure 15: Repair events in pools of artificial ITR hairpins identified by PacBio sequencing. **(A)** IGV display of the 3'ITR hairpin region with all reads containing the AptMG barcode aligned to a reference genome with AptMG (left) and flop wtITR2 (right) hairpin. Red arrows indicate reads matching neither the wtITR2 nor the AptMG hairpin. Pink arrows indicate reads matching the AptMG reference.The gap between the fifth and eight read is derived from reads that did not cover the 3'ITR hairpin. **(B)** IGV screenshot of the 3' ITR hairpin region of 19 randomly selected reads with AAV2-AT-loop barcodes (total reads 81) in an alignment to their reference sequence in flop configuration. In the reference sequence provided above the alignment, the pink nucleotides indicate the nucleotides inserted into the wtITR2 sequence. Blue arrows indicate the reads matching the hairpin reference in flop configuration, and violet arrows those in flip configuration. Reads matching the wtITR2 hairpin contain a deletion of six nucleotides. **(C)** IGV screenshot of the 3' ITR region with all detected reads with AptCV barcode aligned to their reference genome in flop configuration, with none of the reads matching the AptCV hairpin sequence but most indicating the wtITR2 hairpin sequence. Data was acquired and analyzed by Phillip Tai, Suk Namkung and Mitchell Yip (UMMS).

3.1.5.3 Second generation sITR plasmid design facilitates third generation sequencing

The presence of a *cis*-repair mechanism has been previously described for AAV ITRs [270], but a *trans*-repair mechanism, as suggested by the PacBio sequencing data, represents a novel observation. Given the prohibitive cost of PacBio sequencing, I decided to adapt the sITR plasmid to facilitate sequencing using Nanopore technology and generate a large point mutant hairpin library (Figure 16A). This modified sITR plasmid contained a larger 15 bp barcode sequence that can improve barcode recognition in the lower-quality Nanopore sequencing reads, also enabled the screening of larger sets of ITR variants by the NGS-based BC-seq

approach. Additionally, the 3' UTR region was extended to improve primer binding for amplicon generation during BC-seq, avoiding an overlap with the polyA site that could have previously impaired amplicon generation, especially from cDNA. Due to the limited productivity of largely divergent ITR sequences in the previous screens, I decided to perform a comprehensive screen using two libraries of 90 ITR variants (V1-V90) featuring either paired point mutations that preserved the secondary structure (V1-V54), or single mismatches in the hairpin (V55-V90). These variants could not only enable the identification of regions in the ITR that are more tolerant to sequence modification, but they could also be used to gain further insights into the *trans*-repair mechanism, by maintaining a relatively high replication rate even in competition with the wtITR2.

The transition from PacBio sequencing to Nanopore was not only motivated by costs and accessibility, but also by methodological concerns. Specifically, the PacBio sequencing library preparation includes an end- and damage-repair step before the ligation of the sequencing adapters to the vector genomes. While this terminal ligation of sequencing adapters allowed end-to-end sequencing of the vector genome, the end-repair step could theoretically lead to artificial corrections of mismatching regions in the ITRs of annealed viral genomes, albeit this was unlikely. Nevertheless, to avoid an end-repair step during sample preparation, a transposase-based Nanopore sequencing protocol was used (Figure 16B). This approach was previously established for rAAV genome sequencing [378], and relies on transposase-mediated insertion of barcode adapters and subsequent ligation of the Nanopore sequencing adapters. Furthermore, this sequencing protocol can be directly applied to plasmids and thus enables the verification of the ITR sequence and structural integrity in sITR plasmids, in turn allowing confirmation of the novel T7eI-ITR sequencing methodology. However, this sequencing strategy also has disadvantages, as it does not permit the end-to-end sequencing required for the parallel analysis of the ITR configuration previously performed with the PacBio sequencing data (compare Figures 13C and 13D). This was clearly visible in the read length distributions for Nanopore sequencing runs with sITR plasmids and rAAV vectors (Figure 16C). The read lengths recovered from rAAVs confirmed that practically no reads covered the complete vector genome from ITR to ITR. Notably, the read lengths from sITR plasmid samples, which could theoretically be dimeric due to the cloning strategy, confirmed a monomeric state of the plasmids.

The subsequent analysis approach that was regularly employed in this work to determine the sequence within the ITRs is illustrated in Figure 16D. After alignment of the Nanopore reads

to the reference sequence, reads that span both ITR and barcode region were used for the subsequent analysis. For each detected barcode sequence, the ITR sequence composition was determined and visualized as an integrated heatmap/dotplot.

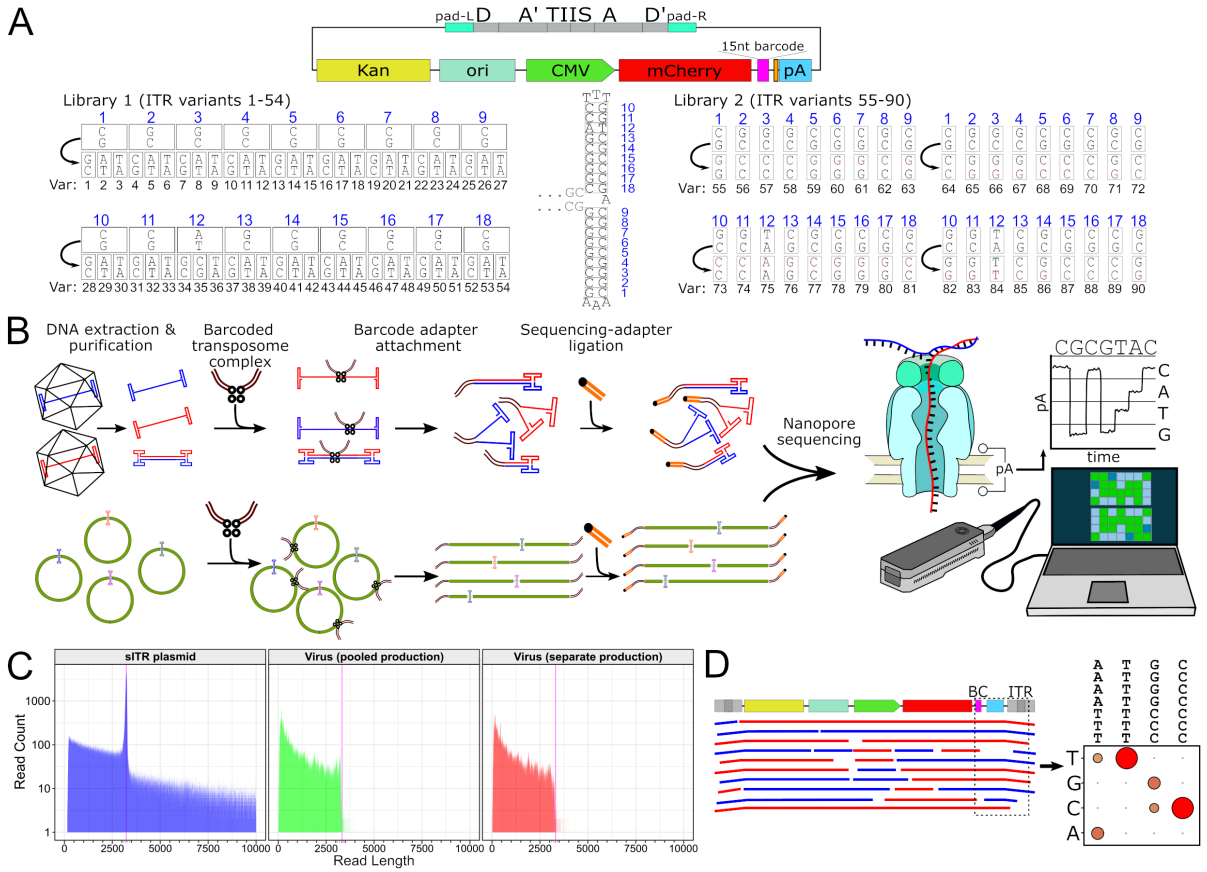


Figure 16: Second generation sITR plasmids for Nanopore sequencing of plasmids and vectors.(Top) Design of the second generation sITR plasmids with increased barcode length, elongated upstream 5'UTR (orange) for improved barcode amplification and two primer-binding pads adjacent to the ITR which facilitates replacement of the complete ITR (turquoise). Again, the hairpin could be inserted via dual TIIS cleavage sites at the tip of the hairpin. (Bottom) Design of two wtITR2 point mutant libraries comprising a total of 90 ITR variants. A reference wtITR2 is shown in the center, with blue numbers indicating the positions in the hairpin at which mutations were introduced. Two large sets of ITR variants were generated with either paired mutation (bottom left), where the hairpin sequence is mutated to all possible nucleotides while maintaining the hairpin structure. Alternatively, the second library (bottom right) contained mismatching point mutations (dark red) at each position in both hairpin arms. For both sets of variants, the original hairpin nucleotide is depicted at the top and the alternative nucleotides are shown below, together with the variant identifier number. **(B)** Sequencing approach for vector genomes from rAAV vectors (top) and sITR plasmids (bottom) using transposase-based adapter insertion for Nanopore sequencing. **(C)** Nanopore read length distributions determined from a 55 variant sITR plasmid library (left) and rAAV vector genomes from pooled (center) and separate (right) production with 55 different ITR variant plasmids. The pink line indicates the unit length of the sITR plasmid or vector genome. **(D)** To analyze the ITR sequence integrity, the reads were aligned to the reference genome. Then reads spanning barcode and ITR were extracted and for each barcode, a nucleotide profile of the ITR region of interest was generated and visualized in an integrated heatmap/dotplot.

3.1.5.4 Nanopore sequencing to examine the ITR integrity in sITR plasmids

Following the cloning of the library consisting of 55 sITR plasmids with 54 ITR point mutation variants and ITR sequence validation by T7eI-sequencing, I used Nanopore sequencing to confirm the ITR sequence integrity before using these plasmids for vector production. Furthermore, as the loss of ITRs represents a significant risk during plasmid propagation, especially since the ITR comprises roughly 5% of the whole sITR plasmid, the Nanopore sequencing data could be used for an unbiased determination of the fraction of plasmids that lost the full or parts of the ITRs.

There, due to the expected large gaps in reads from plasmids that lack the ITRs, a splice-aware aligner was employed to improve the gap recognition in the alignment. The cumulative gap length of deletions and "splice junctions" in the ITR region were then measured for all reads while the barcode sequence identified the ITR variants (Figure 17A). This revealed that the complete loss of 165 nt ITRs during plasmid replication was a rare event; instead, the loss of around 155 nt appeared to be the more common event. For instance, nearly 1% of the reads derived from plasmids with ITR variant V16 exhibited deletions of this size. One sITR plasmid (V31) showed reduced gap formation levels, but the average gap length per read was high with nearly 20 deletions and far exceeded the number of less than five deletions in other non-hairpin plasmid regions of equal size (data not shown). A large number of gaps stemmed from Nanopore sequencing inaccuracies due to the high GC content, secondary structures, and repetitive motifs inherent to ITR sequences as well as from a strand polarity-dependent artifact in the 3'regions of reads in the A-D sequence (Supplementary Figure 3). As these short gaps accumulate throughout the ITR region, it was difficult to estimate the ratio of completely intact plasmids. Therefore, the maximum consecutive gap size per read was determined, as this would reduce the impact of reads with multiple short gaps derived from sequencing inaccuracies. Exemplary pie charts indicating the consecutive gap length distribution in selected plasmids (Figure 17B) support the assumption that most gaps in the alignment were relatively short. Interestingly, substantial variability was observed among ITR variants. For instance, among the reads from sITR variant V31 nearly 60% had less than five consecutive gaps, while this was the case for only 28% of reads from sITR variant V37. The average number of plasmids with less than five consecutive gaps was around 35%. Considering the low quality of Nanopore reads in the ITR region and the common deletions of the hairpin sequences with a length of 21 nt, it was concluded that reads with a maximum consecutive gap size of up to ten nucleotides were likely still derived from intact sITR plasmids. Using this criterion, around 60-

70% of sITR plasmids would be classified as containing intact ITRs. A deletion that would ensure the complete loss of ITR function, *i.e.*, a consecutive gap size >100 nt, was on average detectable in less than 1% of all reads, with outliers nearing 2%. This further highlighted that ITR stability is a non-neglectable issue and varied across different plasmid preparations.

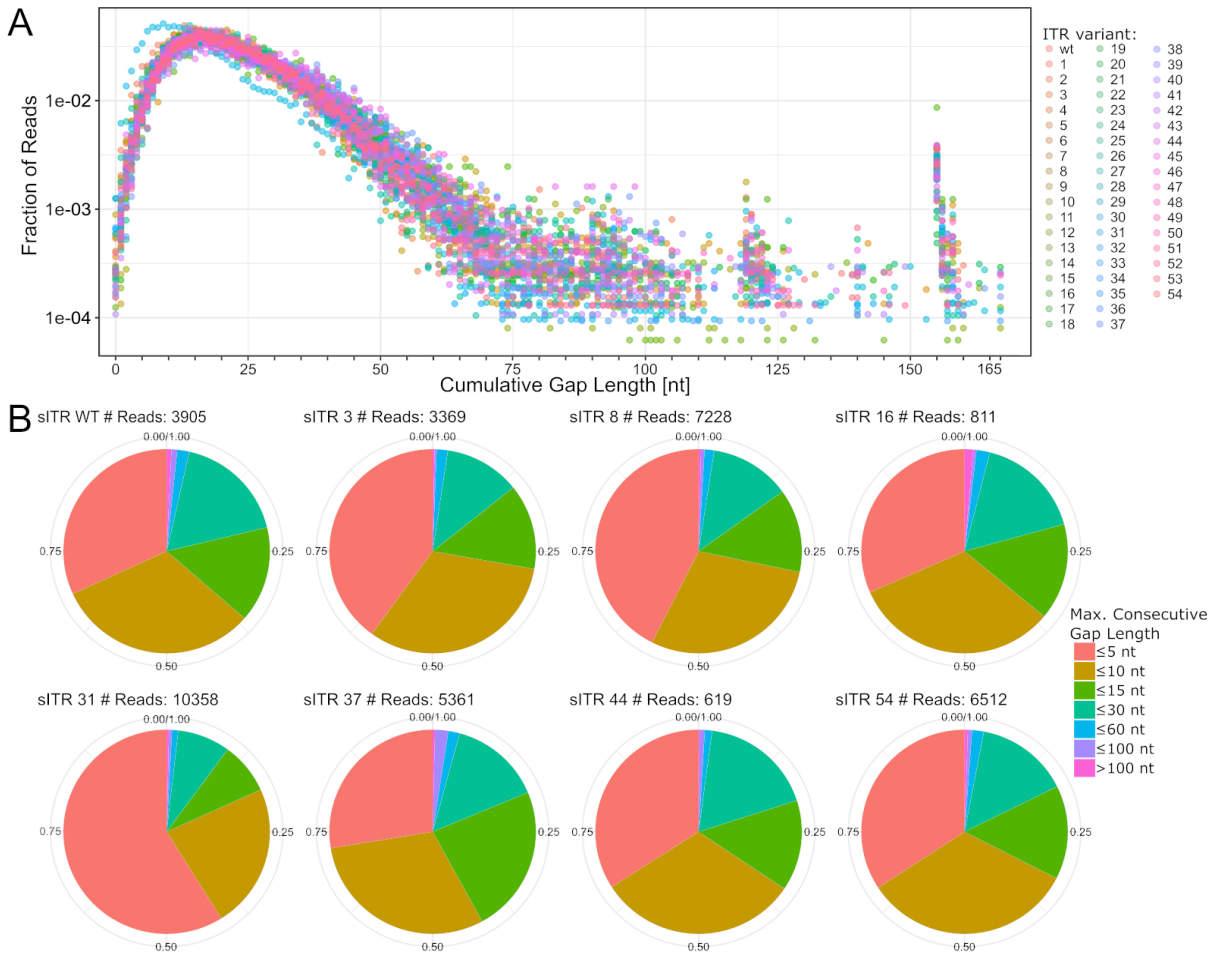


Figure 17: Stability of the 165 bp ITR sequences in sITR plasmids. (A) Nanopore sequencing reads from 55 different sITR plasmids containing 165 bp ITRs were identified by the barcode and aligned to a flip/flop consensus reference plasmid sequence using a splice-aware aligner. The cumulative gap length, both deletions and "splice gaps", was determined for all reads, and the fraction of reads with this gap length is shown. Any deletions larger than 167 bp were counted as 167 bp deletion. **(B)** The maximum length of consecutive gaps in the aligned reads was determined for all reads and its fraction for each ITR variant was determined. Data from eight selected ITR variants (wtITR2, V3, V8, V16, V31, V37, V44, V54) is provided to showcase the varying levels of ITR integrity in the sITR plasmids. For each plasmid, the number of reads used for the analysis is indicated.

Despite the inherent ITR-associated sequencing challenges, Nanopore reads were of sufficient quality to infer a consensus sequence of the 165 nt ITR region in sITR plasmids. Examination of the nucleotide distribution at each position in the alignment, as previously indicated (see Figure 16D), also enabled the identification of regions prone to deletions. The nucleotide distribution in the 165 bp ITR is shown for seven selected ITR variant plasmids in Figure 18A-G. The consensus sequence matched the expected ITR variant sequence in all cases, although

at some positions, alternative nucleotides appeared at substantial levels the reads. In particular ITR variants V3 (Figure 18B) and V37 (Figure 18F) showed alternative nucleotides at the position of the intended mutation in around 20% of the aligned reads. This effect was unlikely due to barcode misassignment, as a more stringent barcode filtering did not eliminate it (data not shown). Thus, this could either indicate a reverse mutation, a contamination with the wtITR or alignment/sequencing errors at these positions. The fact that the occurrence of gaps in the aligned reads at these positions is also elevated may suggest the latter. For most variants, such as ITR variants V16 (Figure 18D) and V31 (Figure 18E), the intended mutations were clearly recognizable with minimal background noise. As previously indicated, the varying quality of different plasmid preparations could explain some of the variations in the number of gaps in the alignment, as exemplified with ITR variant V37 (Figure 18F) that exhibited a low ITR integrity. Other noteworthy accumulations of mismatches can be observed in several regions of all sequenced ITRs, especially the regions with tetra-cytosine or tetra-guanine repeats that showed increased levels of mismatches near the beginning of the hairpin. This type of mismatch was likely derived from the difficulties of Nanopore sequencing to properly distinguish the nucleotides in the repetitive motif. Another region showed even higher levels of mismatches, namely, the region between the trs and the RBE sequence where up to 40% of all reads supposedly contained nucleotide deletions. This was based on the strand-polarity dependent sequencing artifact, accumulating deletions in the 3’ region of the read (Supplementary Figure 3). The bidirectional Nanopore sequencing then led to the accumulation of these deletions on both sides of the ITR in the alignment. Despite the technical challenges associated with the use of Nanopore sequencing for plasmids with 165 bp hairpin sequences, the data indicated that the ITRs remained predominantly intact and matched the ones observed using the T7eI-sequencing approach, therefore supporting their suitability for use in vector productions.

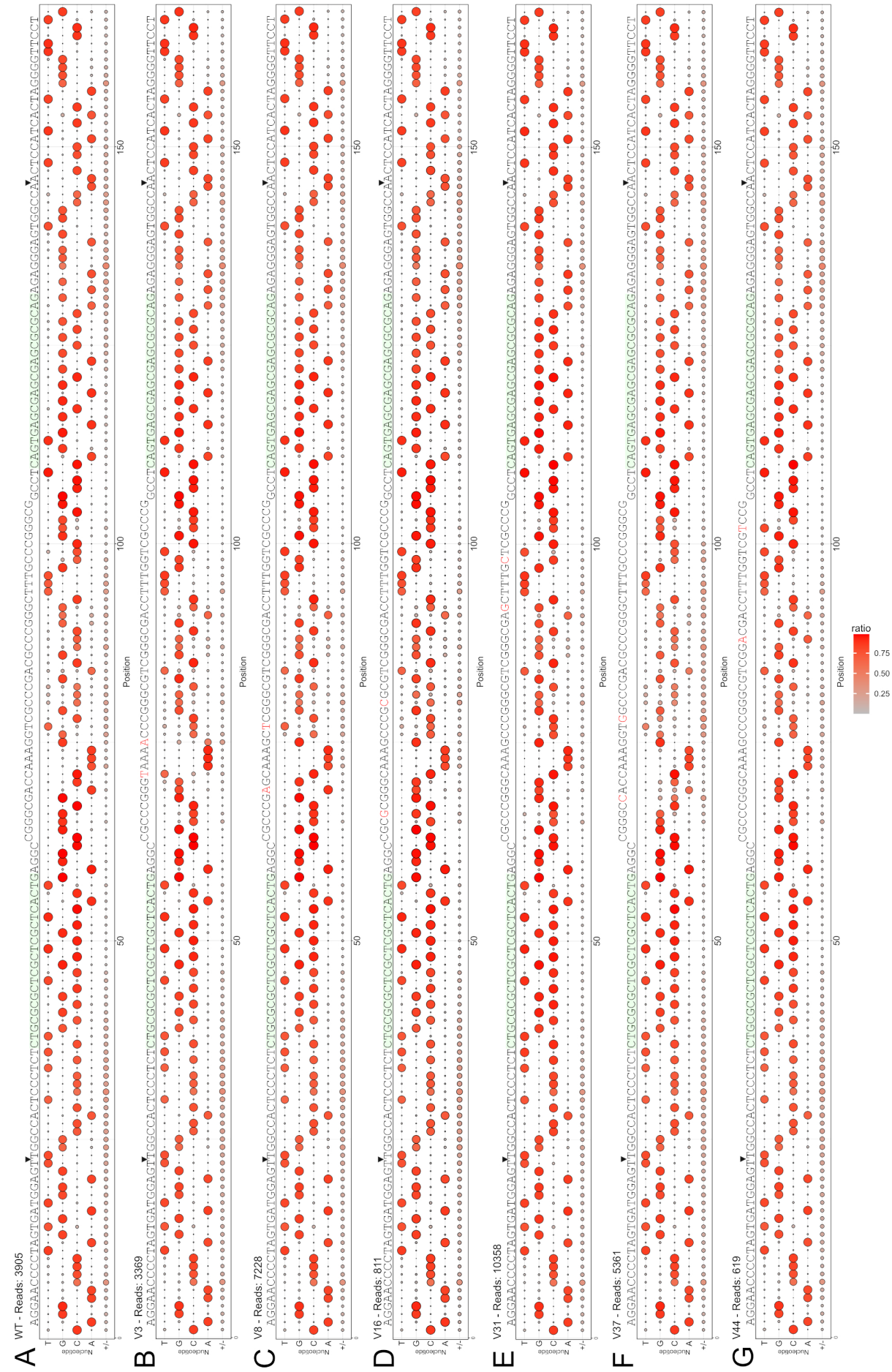

A
WT - Reads: 3905
B
V3 - Reads: 3369
C
V8 - Reads: 7228
D
V16 - Reads: 811
E
V31 - Reads: 10358
F
V37 - Reads: 5361
G
V44 - Reads: 619
Nucleotide
Position
ratio
0.75
0.50
0.25

Figure 18: Determination of sequence integrity in sITR plasmids using Nanopore sequencing. (A-G) Shown are nucleotide distribution plots determined for Nanopore sequencing reads of seven selected sITR variant plasmids aligned to a wtITR flip/flop reference sequence. Reads were aligned using a splice aware aligner. For each nucleotide in the alignment against the reference sequence, the ratio is indicated by both dot size and color. Identified splice gaps (+) are shown together with deletions (-). The sequence above the nucleotide distribution indicates the reference sequence, with the central terminal dual-hairpin delimited from the stem sequence. Red-colored nucleotides indicate the intended ITR mutations. The green highlighted sequence is the RBE, while the black triangle indicates the trs and the trs complementary position. **(A)** wtITR2, **(B)** ITR variant V3, **(C)** V8, **(D)** V16, **(E)** V31, **(F)** V37, and **(G)** V44.

3.1.5.5 Nanopore sequencing of rAAV reveals ITR repair dependency on *trans*-repair templates

Following the confirmation of the sequential and structural integrity of the sITR plasmids using both T7eI-ITR- as well as Nanopore sequencing analysis, I next examined the integrity of the ITRs within the rAAV particles using the same library of 55 different ITRs. Since previous vector productions with sITR plasmid pools resulted in the occurrence of a repair mechanism and the conversion of the intended ITR hairpin variant to alternative hairpins, mostly the wtITR2, a specifically interesting question was whether this would be confirmed using the transposase-based Nanopore sequencing approach. To exclude effects of random mutagenesis or Rep-mediated sequence conversion, and to prove the reliance on the presence of repair templates, ITR variant vectors were also produced separately. In this case, each ITR variant was transfected separately, and the vectors were pooled only before purification. For both production strategies, the vector genomes were isolated after vector purification and subjected to Nanopore sequencing. Sequencing reads were then aligned to a reference vector genome containing a consensus ITR sequence representing both flip and flop ITR configurations. Using the barcode for variant identification, only reads covering the hairpin region of the ITR were used to evaluate the integrity of the ITR sequences. Representative results for selected variants are shown in Figure 19A-F.

Despite the shorter hairpin structure of ITRs in vector genomes compared to plasmid DNA, error rates remained elevated in proximity to repetitive sequences, tri-or poly-nucleotide repeats, the hairpin tips and hinges, as evidenced by the accumulation of deletions in these regions. Additionally, data interpretation is challenging due to the presence of hairpins in both flip and flop configuration. The previously introduced visualization using the heatmap/dotplot helped to distinguish the nucleotides derived from the two configurations, as seen for the wtITR variant (Figure 19A). There, no notable differences were observed between pooled and separate productions, with the expected 50:50 distribution of nucleotides derived from flip or

flop configuration. Crucially, clear evidence of the *trans*-repair mechanism could be observed in the sequencing data from ITR variant vectors from pooled productions. In particular for variants where the hairpin variant contains the same nucleotide in both hairpin configurations, the occurrence of repair events is easily trackable, such as in ITR variants V8, V16 and V31 (Figure 19C-E). In the absence of a repair event, all nucleotides would be the same, while the occurrence of the *trans*-repair mechanism led to a partial reversion to the wtITR2 nucleotide in the pooled production. In the other exemplary ITR variants V3 and V37 (Figure 19B + F), the *trans*-repair led to an uneven distribution of the nucleotides at the mutated site, due to the reversion to the wtITR2 nucleotide. In contrast, for ITR variant V44 (Figure 19G), the *trans*-repair was associated with the emergence of a third nucleotide, again derived from the wtITR2 hairpin. This reversion was visible in all variants from the pooled production, confirming the presence of the *trans*-repair mechanism. Across all pooled vector preparations, approximately 10–15% of reads at the mutated positions reverted to the wtITR2 nucleotide. Accounting for the two possible ITR configurations, this implies total repair rates of around 20-30% of the vector genomes.

In contrast, sequencing data from separately produced vectors showed that ITR integrity was largely preserved. The mutations introduced into the ITR were retained, although the high background error rate at centrally located nucleotides can complicate interpretation in these regions. Some apparent reverse mutations to the wtITR2 sequence, that would also be associated with *trans*-repair, were attributed to alignment errors. This was particularly evident in ITR variants V3 (Figure 19B) and V37 (Figure 19F, where the hairpin mutations remained intact in reads from vector ITRs in flip configuration but appeared partially mutated in reads from flop orientation. Thus, it was concluded that this likely represented an artifact caused by the alignment to the flip/flop consensus reference and that *trans*-repair was absent. For the second ITR variant library (V55-V90), the higher similarity of the variants to the wtAAV2 reference genomes reduced the occurrence of false reverse mutations during alignment and the hairpin sequences were also found to remain intact after separate production (Supplementary Figure 4). Given the confirmed integrity of the barcode-ITR association, these vectors were deemed suitable for use in transduction experiments.

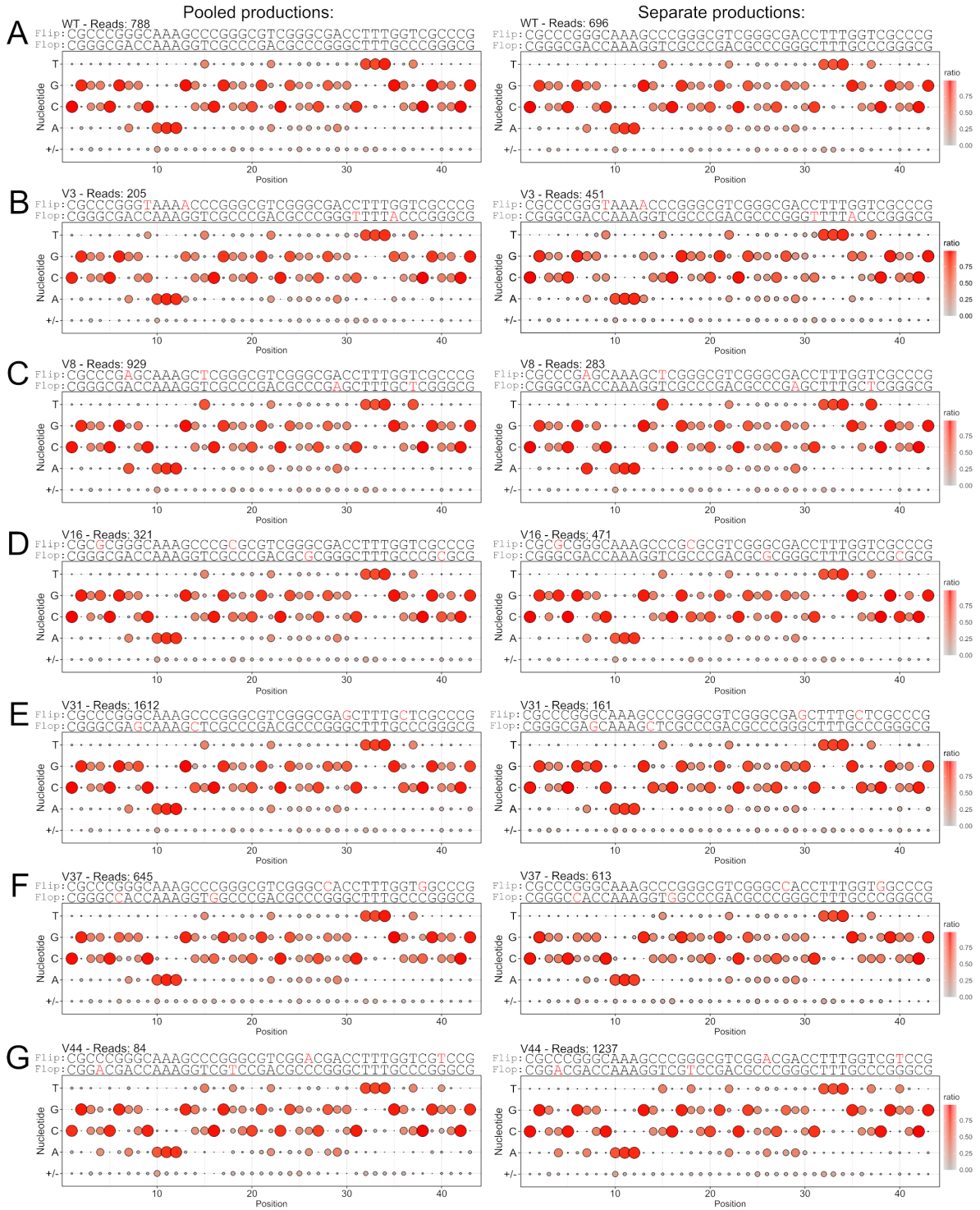


Figure 19: ITR-sequence integrity in vectors from pooled and separate productions. Nanopore sequencing reads were aligned to a consensus flip/flop reference sequence, and the nucleotide distribution in the 3' ITR hairpin region was determined. Shown on the left is the data from pooled productions, on the right the data from separate productions. The ITR variant is indicated together with the number of reads detected for a given variant that were used for the nucleotide distribution. The expected ITR sequence is given above in flip and flop configuration reference with red nucleotides indicating the mutations distinguishing the variants from the wtITR2. Both the color and the size of the circle indicate the ratio of nucleotides at each position in the alignment of the wtITR2 **(A)**, ITR variant V3 **(B)**, V8 **(C)**, V16 **(D)**, V31 **(E)**, V37 **(F)** and V44 **(G)**.

3.1.5.6 Evaluating the fidelity and range of the ITR *trans*-repair mechanism

Initial Nanopore sequencing results seemed to indicate that the *trans*-repair in pooled production resulted primarily in reversions to the wtITR2. However, this apparent preference likely relied on the use of ITR variants whose consensus sequence is the wtITR sequence. Manual inspection of the data suggested that repair events towards other hairpin variants indeed took place as well (data not shown). However, an unbiased sequence-based quantification of each variant ITR read was impossible, due to the intrinsically high error rate of Nanopore sequencing and the very similar sequences. The occurrence of conversion to alternative hairpin sequences was also substantiated by the PacBio sequencing data (see section 3.1.5.2). Thus, Nanopore sequencing was employed to resolve these repair events towards alternative hairpin sequences and to estimate their frequency in smaller ITR variant pools with a maximum of four different ITR sequences. Furthermore, ITR variants that allow for a better distinction from the wtITR sequence as well as mutations in other regions than the terminal hairpin were used to obtain a better insight into the extent of the *trans*-repair mechanism.

In more detail, three barcoded point mutants with nucleotide variation at the same position in the B-hairpin were used for a pooled production in the presence of the wtITR2, facilitating the distinction of the used repair template. After pooled production, the vector was again sequenced, the reads aligned to the flip/flop consensus wtITR2 sequence, and the nucleotide distribution plots for the four ITR variants generated (Figure 20A). In reads derived from both wtITR and ITR variants, alternative nucleotides at the mutated position appeared to be over the background error noise. This effect was best discernable in variant V34 that carries the same nucleotide in both flip and flop configuration at the mutated site. There, the preference of repair towards the wtITR could also be confirmed, with a roughly two-fold higher conversion rate as compared to the two alternative ITR variants. Considering that around 15 % of reads in flip configuration and 15 % in flop configuration contained converted nucleotides, each representing independent repair events, the total repair rate was estimated at 30%. PacBio sequencing has previously revealed even higher repair rates, with some aberrant ITR hairpin sequences not even detectable anymore. Instead, these ITRs seemed to be completely converted to the wtITR2 (see Figure 15C), although admittedly with low read counts. To replicate this using Nanopore sequencing with a higher sequencing depth, I decided to use an equimolar pool of plasmids with the wtITR2 and the highly divergent ITR hairpin variant from *Ae*DV (Figure 20B) for vector production. This sequence divergence enabled a sequence-based differentiation of the two ITRs after alignment to a flip/flop consensus reference sequence of

each of the ITRs. The ITRs were then counted based on sequence matches with a 10% mismatch tolerance (Figure 20C). Only about 10% of the reads matched the reference ITR, despite the relatively high mismatch tolerance, further underlining the low quality of Nanopore reads within the ITRs. Among the reads containing the *Ae*DV barcode, more indicated the presence of the wtITR2 hairpin than the *Ae*DV hairpin. Concurrently, among the reads containing the wtITR2 barcode, the large majority also contained the wtITR2. Nevertheless, a minor fraction of the detectable reads contained the *Ae*DV ITR.

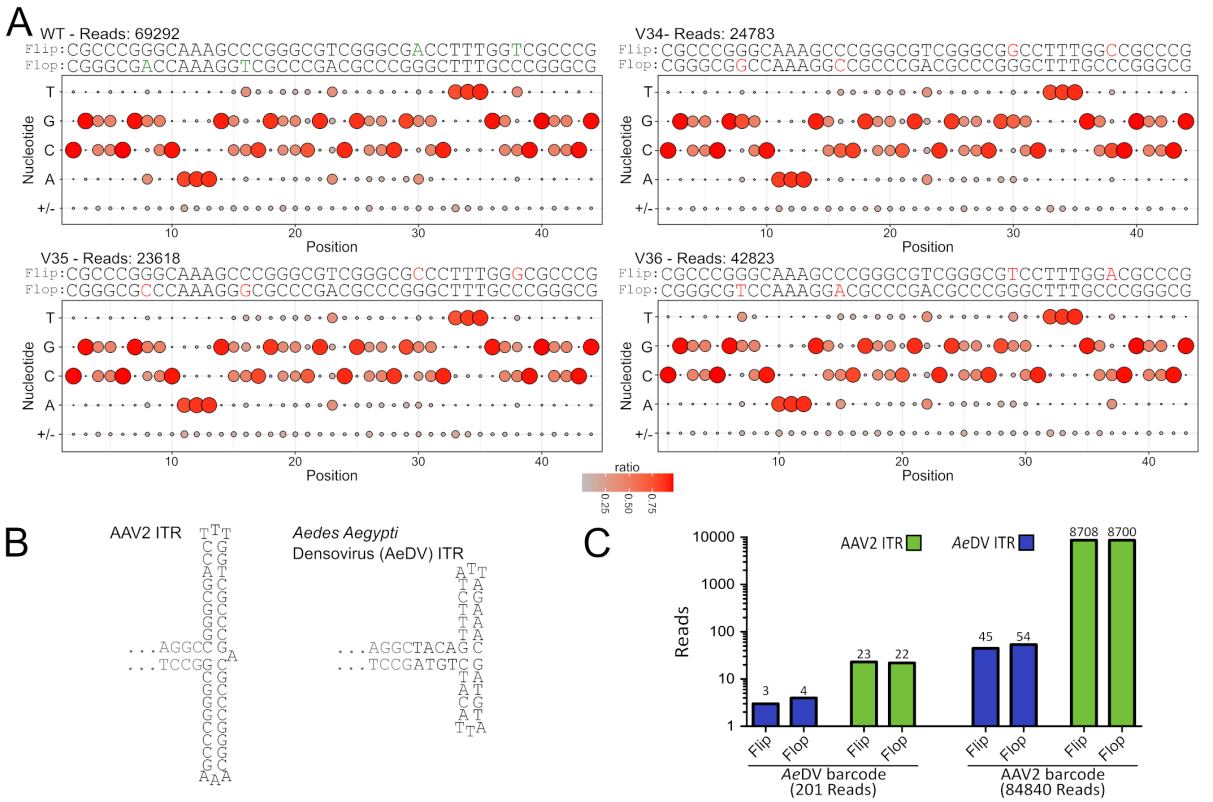


Figure 20: *Trans*-repair in similar and divergent ITR hairpin sequences. (A) Nanopore sequencing read distribution in the 3' ITR region of vector genomes generated from a pool of four different ITR sequences (wtITR2 and ITR variants V34-V36). Reads were aligned to a consensus flip/flop reference sequence, and the nucleotide distribution for each barcode embedded in the vector genome was determined. The ITR variant is indicated together with the number of reads detected for a given variant. The expected ITR sequence is provided in flip and flop configuration as reference, with the position of the mutation indicated in green in the wtITR2 sequence and the alternative nucleotides given in red for the variant hairpin sequences. Both color and size of the circle in the plot indicate the ratio of nucleotides at each position in the alignment. **(B)** Structures of the wtITR hairpin from AAV2 (left) and the *Ae*DV ITR hairpin (right) with the AAV2 stem sequence (gray) used for a pooled production. **(C)** Barplot indicating the detected 3' ITR hairpins in Nanopore sequencing data of vector genomes from a pooled production with *Ae*DV hairpin and wtITR2. Reads were aligned to both flip/flop consensus *Ae*DV and wtITR2 reference genomes, barcodes in the reads were identified and the associated hairpins were identified (10% mismatch tolerance) and counted.

The evidence for *trans*-repair events during pooled production of different ITR hairpin variants raised the question whether the repair mechanism extends across the entire ITR sequence or is

limited to the hairpin termini, since all previously examined sequences were ITR hairpin variants. To this end, I generated three additional wtITR2 stem sequence variants (VS), namely, VS2; deletion of the terminal nucleotide of the RBE; VS3: additional nucleotide inserted at the end of the RBE; and VS4: point mutation in the D-sequence known to interfere with transcription factor binding to the ITR without impacting productivity [343]. Figure 21A compares the sequences of these variants with the wtITR2 in the distal stem region. After production from equimolar plasmid pools, each vector was purified, and the genomes were isolated and sequenced via Nanopore. The reads were then aligned to an adapted reference sequence, containing an additional "N"-nucleotide at the end of the RBE to enable the alignment of all variants to the same reference sequence. The recovered nucleotide distributions for all four ITR variants show that the stem sequence does not change its sequence in flip and flop configuration (Figure 21B), simplifying the recognition of sequence alterations. Interestingly, in reads from both wtITR and the VS4 variant, neither the insertion of the cytosine from VS3 nor the removal of an additional nucleotide as present in VS2 could be detected. The apparent insertion of a guanine nucleotide in about 15% of the reads is likely an artifact from the alignment to the ambiguous reference sequence. The VS2 variant with the nucleotide deletion showed a conversion to the wtITR in nearly 33% of the reads. Similarly, for variant VS3 with the additional nucleotide in VS3, the removal of this nucleotide could be detected in around 28% of the reads. While the conversion of the wtITR to the alternative hairpin could not be detected here, the presence of the *trans*-repair was shown to also persist in the distal part of the ITR stem.

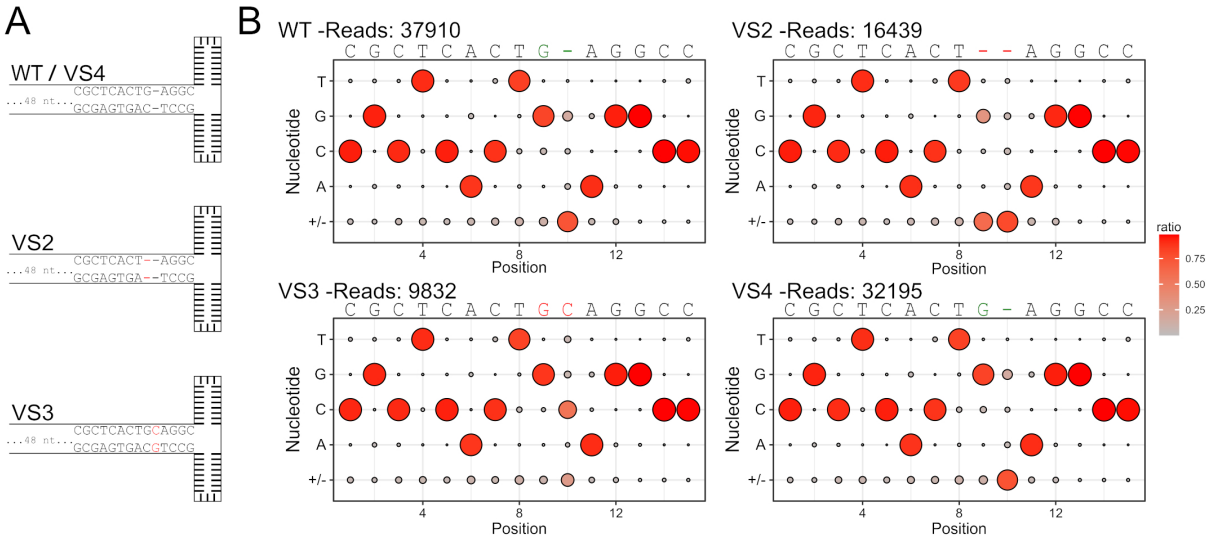


Figure 21: ITR *trans*-repair in the distal stem region. (A) ITR stem sequence variants used for pooled production with two variants containing the wtITR2 sequence (WT + VS4), a variant with a deletion (VS2) and a variant with an insertion (VS3). Both insertion and deletion are highlighted in red. **(B)** Nucleotide distribution in the distal stem region in aligned Nanopore sequencing reads of vector

genomes from a pooled production with the four indicated ITR variants. The number of reads used to generate the plot as well as the expected sequences are indicated above the plot. The wtITR sequence elements are indicated in green, the mutants in red. Both color and size of the dots indicate the nucleotide distribution at each position of the alignment.

The analysis of potential *trans*-repair in the proximal stem sequence was also possible, because the mutation in the D-sequence distinguished VS4 from wtITR, VS2 and VS3 (Figure 22A). The nucleotide distribution in the region of this mutation suggested that the wtITR as well as VS2 and VS3 were kept nearly completely intact (Figure 22B), albeit conversion towards the VS4 variant may persist at levels barely distinguishable from the background error rate. For the VS4 variant, with the mutation in the D-sequence, the *trans*-repair effect was clearly detectable and led to a reversion of the introduced mutation in around 19% of the reads, proving that the repair mechanism affects all parts of the ITR.

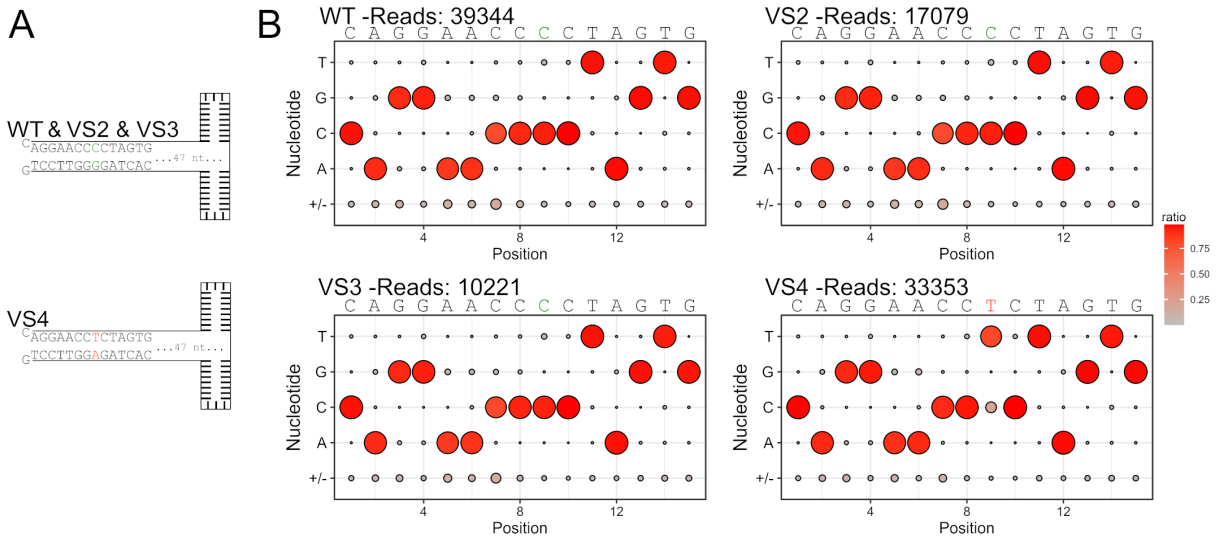


Figure 22: ITR *trans*-repair in the proximal stem region. (A) D-sequence mutation variants used for pooled production with three variants containing the wtITR2 D-sequence (WT/VS2/VS3), and one D-sequence mutant (VS4) with the altered nucleotide indicated in red and the wt nucleotide in green. **(B)** Nucleotide distribution in the region of interest determined in aligned Nanopore sequencing reads of genomic DNA extracted from viral vectors from a pooled production with the four indicated ITR variants. The number of reads that were used for the plot as well as the expected sequence are indicated above the plot. The wtITR nucleotides at the variant position are indicated in green, the mutant in red. Both color and size of the dots indicate the nucleotide distribution at each position of the alignment.

3.1.6 Dissecting the effects of two ITR variant libraries *in vitro*

Based on the verified integrity of the ITR hairpin variants as well as the barcode-to-ITR correlation in vectors after separate productions (see Figure 19), the generated ITR variant libraries, comprising a total of 90 variants (see Figure 16A), could now be used to track the effects of these ITRs during transduction based on the recovery of the barcode sequence.

The workflow of the experiment is depicted in Figure 23, starting with the generation of the ITR variant vectors by separate production. The barcode distribution in the vector library was then determined using NGS and additionally validated by a re-analysis of the Nanopore sequencing data previously used for ITR integrity analysis. I then set out to analyze how the ITRs affect the downstream functionality of the vector upon transduction of two human cell lines. In particular, two key aspects were examined, *i.e.*, the delivery of the transgene and episome formation. The impact of the ITRs on the successful delivery of the transgene was determined by analyzing the barcode distribution in the 3' UTR on the RNA-level. The effect of ITR variants on episome formation has previously been assessed after digestion of non-circular DNA with T5 exonuclease, which led to large variability between samples (see 3.1.4). Thus, I now explored an alternative approach to enrich circular episomal DNA using rolling circle amplification (RCA) with vector genome-specific primers, a method commonly performed to enrich circular viral genomes[379,380]. The RCA product consisting of amplified circular vector genomes was then used for BC-seq, to infer potential effects of the ITR variants on vector genome circularization.

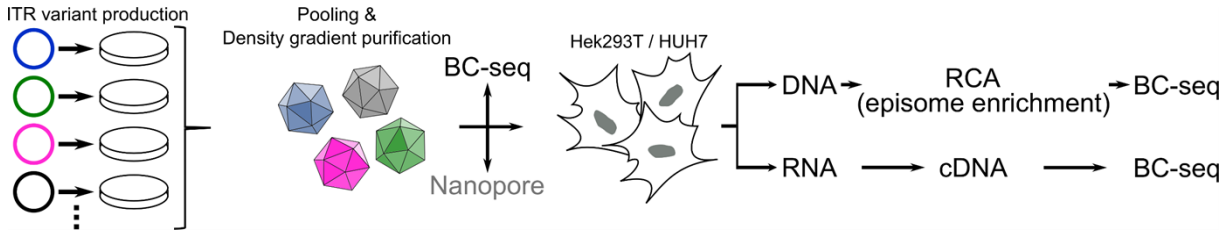


Figure 23: Workflow for *in vitro* ITR variant analysis screen. After separate production and pooling of the vector, the samples were purified before the barcode composition was determined using NGS (BC-seq) and Nanopore sequencing data (see 3.1.5.5). The vector was then used to transduce Hek239T and HuH7 cells, and DNA and RNA were extracted from the cells 72 h later. The DNA was subjected to RCA for episome enrichment with vector genome-specific primers, and the product was used as template for subsequent NGS to determine the barcode composition in episomal DNA. The RNA was used for cDNA generation from which the barcode composition could then be determined using BC-seq.

3.1.6.1 Impact of ITR variants on productivity during separate production

To evaluate the impact of the ITR variants on vector productivity, I re-examined the library composition of both ITR variant libraries after pooled purification of the individually produced variants using BC-seq. Each vector library was sequenced in duplicates to assess the barcode distribution. Additionally, the barcode distribution from the Nanopore sequencing data of the vector genomes was included as an amplification-independent control. For the first set of 54 different ITR variants (Figure 24A), a notable variability between the two productions could be observed. Several variants appeared to increase productivity compared to the wtITR,

predominantly those containing mutations within the B-hairpin (ITR variants V28-V54). In contrast, mutations in the C-hairpin (V1-V27) did not lead to similar increases in the recovered barcode proportions. The barcode distribution observed by Nanopore sequencing mostly matched the NGS data from the analyzed production, increasing confidence in the BC-seq results. For the second ITR variant library (V55-V90; Figure 24B), the barcode composition was determined in the vectors of three separate productions as duplicates and supplemented by the barcode distribution in the Nanopore sequencing data of one of the productions. There, a more consistent pattern between the three productions could be observed, with lower variability between productions. The barcode distribution derived from Nanopore sequencing correlated well with the NGS data. Similar to the first library, ITR mutations in the B-hairpin (V73-V90) indicated higher productivity than mutations in the C-hairpin (V55-V72). This was particularly interesting, considering the ITR B-hairpin is more conserved among wtITR sequences (see Figure 4C).

To further investigate whether the differences in productivity could be directly attributed to the ITR sequence variation, I performed a side-by-side comparison between the wtITR with one of the best performing ITR mutants (V81) via separate production and quantification of each vector (Figure 24C). ITR variant V81 was chosen due to its consistent increase in barcode recovery in three separate productions during the ITR variant library production (see Figure 24B). Despite overall low yields in the productions and considering the outlier during quantification, the titers for both variants were nearly identical (Figure 24D).

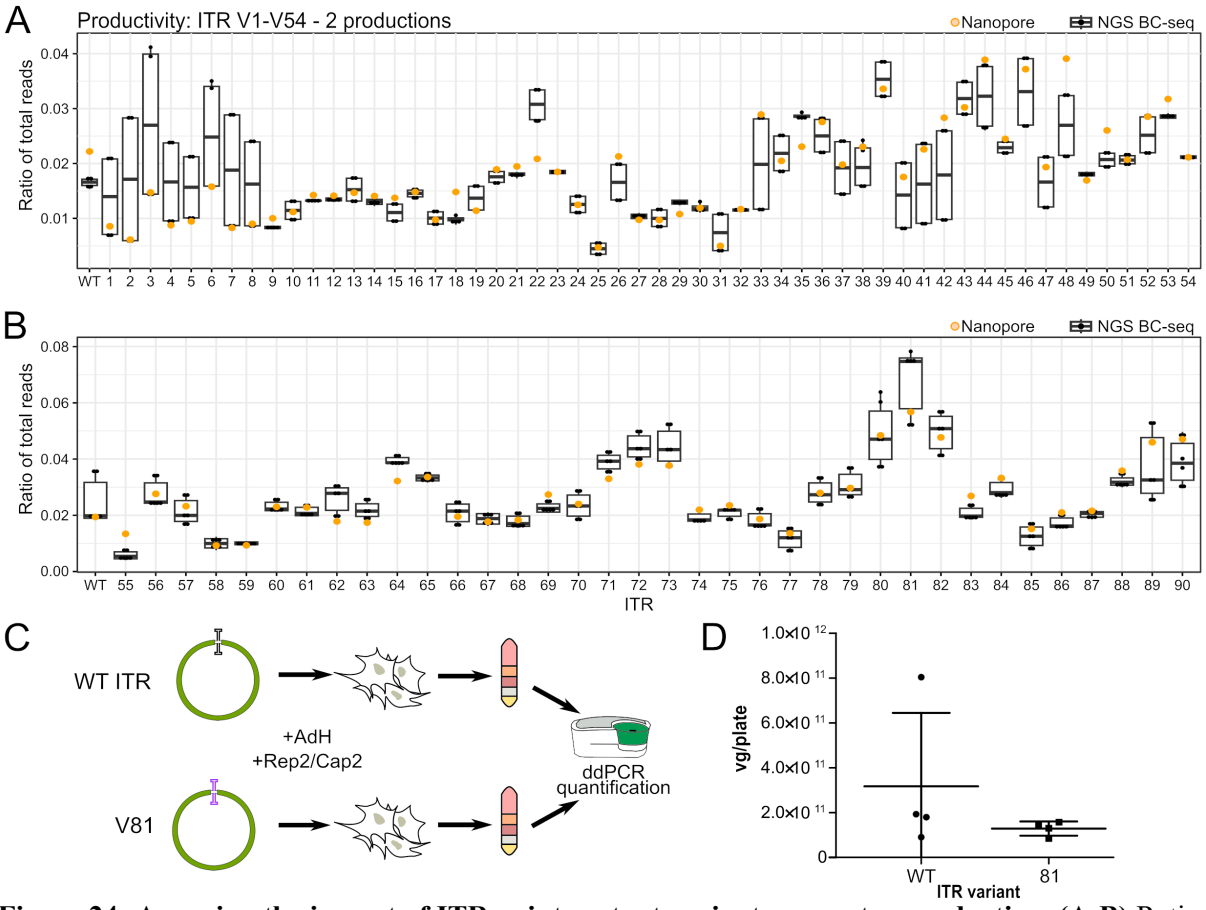


Figure 24: Assessing the impact of ITR point mutant variants on vector production. (A-B) Ratio of reads recovered by BC-seq in the vector libraries after separate production. Distribution from two libraries with the wtITR2 and variants V1-V54 **(A)** and three libraries with wtITR2 and variants V55-V90 **(B)** given as boxplot. Each produced library was sequenced with BC-seq in duplicates, the distribution determined from Nanopore sequencing data from a single library production is supplemented as orange dots. The boxplots only rely on the BC-seq data with lower/upper hinges of the boxes corresponding to first/third quartiles, median is indicated as horizontal line. Whiskers extend to no further than 1.5 inter-quartile ranges (IQR) from the hinges. **(C)** Experimental setup to observe the impact of the best producers from the ITR library by comparing separate production with a wtITR plasmid. After transfection with production plasmid, the vector was purified and then quantified by ddPCR. **(D)** Quantification of the separate production of wtITR in comparison with variant V81 from two different productions analyzed in duplicates by ddPCR and normalized to the productivity per plate. Error bars indicate standard deviation, horizontal lines are means.

3.1.6.2 Impact of ITR variants on functional transduction in cell lines

To assess how ITR variants affect functional transduction, I next transduced Hek293T and HuH7 cells with the ITR libraries and determined enrichment scores based on the recovered barcode distribution in RNA normalized to the mean barcode distribution in the input vector libraries. To increase data reliability, four independent transductions in Hek293T cells and two independent transductions in HuH7 cells were performed with each produced library and analyzed using BC-seq.

For the first ITR variant library (V1-V54) with paired mutations in the hairpin, the determined enrichment scores for both cell lines were remarkably consistent (Figure 25A), despite the highly variable barcode distribution in the input libraries (see Figure 24A). The activity of the ITR variants was very similar in both cell lines with a consistent reduction of gene expression from vectors with the paired ITR point mutation compared to the wtITR2. Differences in transgene expression from different ITR variants were also clearly observable, although a pattern of tolerated mutations compatible with functional transduction could not be identified. The analysis of the impact of the second ITR variant library (V55-V90) with mismatching ITR mutations on functional transduction was performed analogously (Figure 25B). With production as triplicates and subsequent transduction in quadruplicates (Hek293T) and duplicates (HuH7) for each production, variability of the enrichment scores was even lower compared to the first library. Similar to the first library, the wtITR consistently outperformed all variant ITRs, indicating that the introduced mismatch mutations led to a disadvantage during transduction of these two cell lines. Also here, no clear pattern in type and location of mutations emerged that could indicate an increased tolerance for ITR mutagenesis.

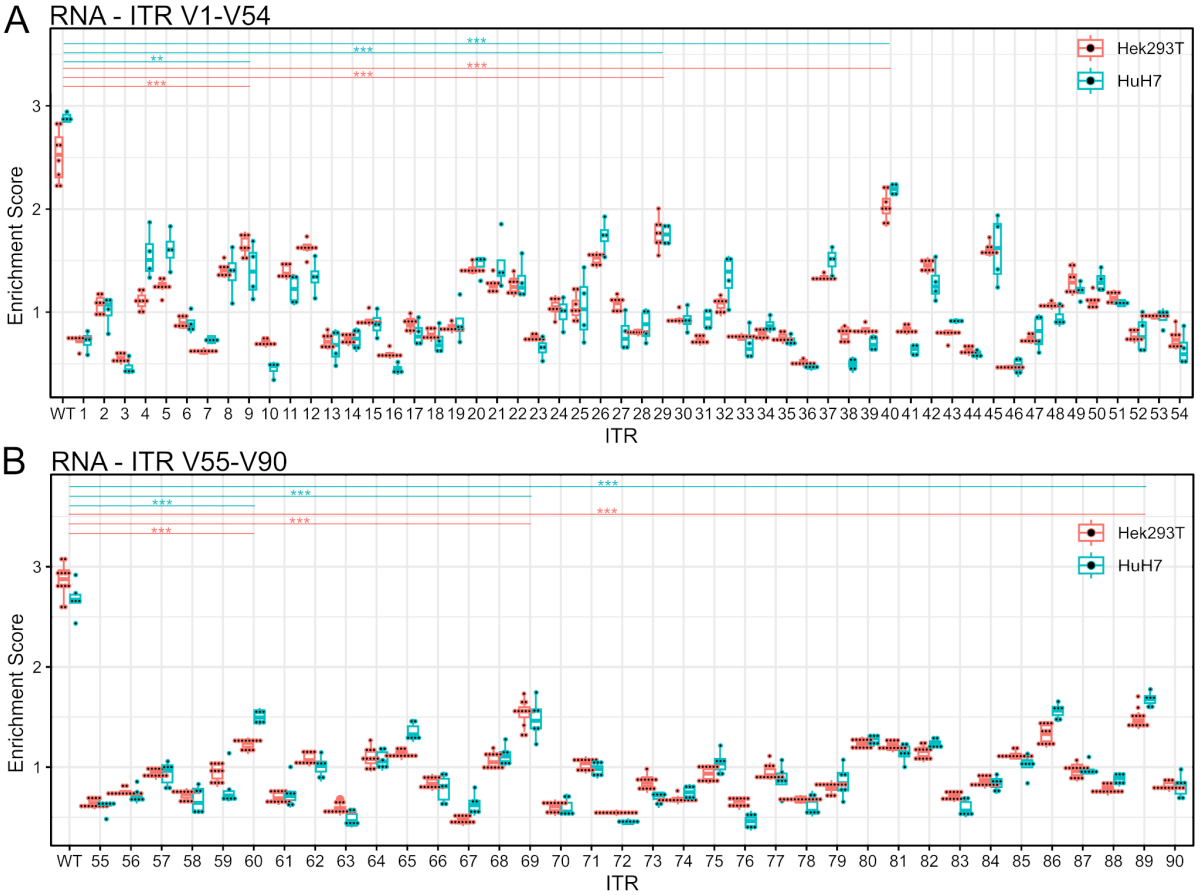


Figure 25: Dissecting functional transduction of ITR variant libraries in cell lines. (A-B) Cell lines were transduced with ITR variant library 1 containing variants V1-V54 **(A)** or library 2 containing variants V55-V90 **(B)** with an MOI of 10^5. Total cellular RNA was extracted 72 h after transduction.

Barcodes were sequenced after cDNA synthesis, and the barcode distribution was normalized to the barcode distribution in the input library, generating an enrichment score shown in the plots for transduction in Hek293T cells (red) and HuH7 cells (turquoise). Lower/upper hinges of the box correspond to first/third quartiles. Median is shown as horizontal line. Whiskers extend no further than 1.5 IQR from the hinges. Pairwise comparison of selected groups was analyzed by Student's t-test (ns: $p>0.05$, *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$).

3.1.6.3 Influence of ITR variants on episome formation in cell lines

To examine the impact of ITR variants on episome formation, I applied RCA to enrich circularized vector genomes in the extracted gDNA (Figure 26A). Selectivity was achieved through the use of vector-specific primers, which mediates the amplification of circular vector genomes but not of the linear, single-stranded and non-transduced vector genomes. Notably, covalently closed rAAV genomes after secondary strand synthesis could theoretically also be amplified by RCA. The RCA amplicons were then debranched, *i.e.*, digested with a single cutter enzyme, and used for BC-seq.

The enrichment scores were determined by normalizing the barcode distribution in the RCA amplified samples to the barcode distribution in the input library. The results for the ITR variant library with variants V1-V54 (Figure 26B) were contrasting the expectation, as RCA-based enrichment did not reduce data variability, signified by the presence of several prominent outliers. Again, the wtITR2 exhibited the highest relative abundance, and the behavior of the ITR variants was consistent between both cell lines. The only notable exception was variant V16, which slightly surpassed the wtITR2 in HuH7 cells but not in Hek293T cells. The enrichment scores determined for the second ITR library V55-V90 (Figure 26C) exhibited slightly less variability but still contained several outliers. The observed enrichment scores between the two cell lines mostly aligned, and the wtITR2 was among the top-performing variants.

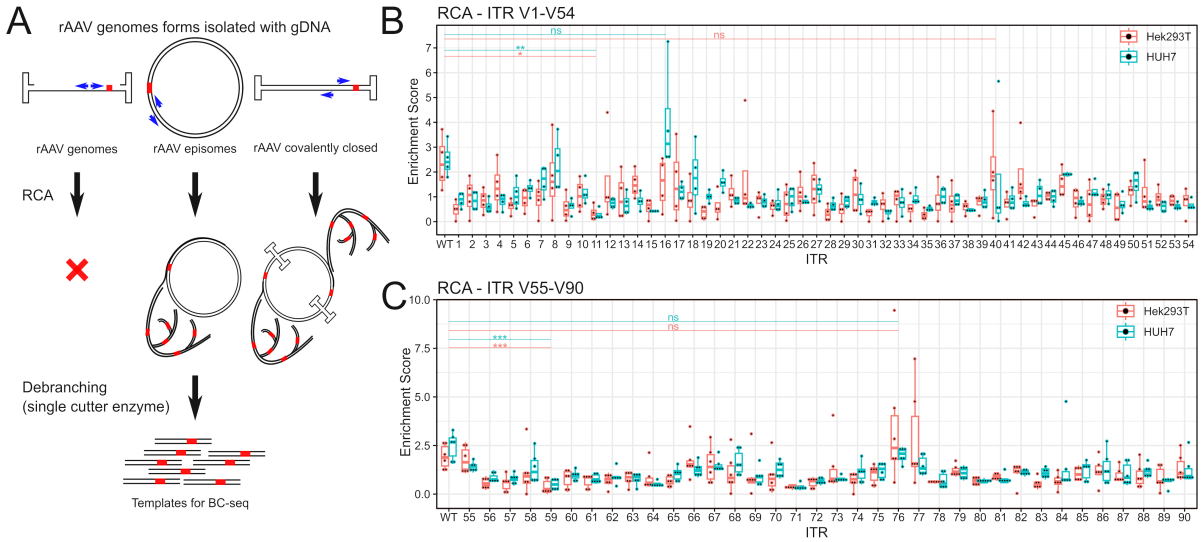


Figure 26: Using RCA to trace episome formation of ITR variants during transduction in two cell lines. (A) Rationale for enriching circular and potentially covalently closed linear vector genomes after second strand synthesis through RCA. The barcode sequence is indicated in red; RCA primers are indicated as blue arrows. The RCA product is then debranched using a restriction enzyme with a single cleavage site in the vector genome, generating vector genome unit-length fragments which can then serve as template for barcode amplification for NGS. **(B-C)** Enrichment scores determined from BC-seq data after RCA-mediated enrichment of episomal DNA in Hek293T (red) and HuH7 cells (turquoise). Data is provided for transduction with the first ITR variant library including variants V1-V54 **(B)** and with the second library with variants V55-V90 **(C)**. Lower/upper hinges of the boxes correspond to first/third quartiles. Median is shown as horizontal line. Whiskers extend no further than 1.5 IQR from the hinges. Pairwise comparison of selected groups was performed using Student's t-test (ns: p>0.05, *: p<0.05, **: p<0.01, ***: p<0.001).

3.1.7 *In vivo* screen of two ITR variant libraries

The *in vitro* screen of the two ITR point mutant variant libraries did not identify any ITR variants which clearly improved functional transduction or episome formation. Instead, wtITR2 was superior to all tested ITR variants. AAV2, originally identified as contamination in human cell culture [16], has likely undergone multiple cycles of replication and transduction, allowing it to adapt for efficient transduction in cell culture. Based on this premise, I speculated that by using the generated ITR variant libraries *in vivo*, an environment to which the serotype has not yet adapted, the performance of the variant ITRs relative to the wtITR2 may change.

To test this hypothesis, I collaborated with the group of Luís Pereira de Almeida at the University of Coimbra, who offered to perform the vector injection, animal handling, euthanasia, and organ extraction as part of our joint ARDAT project. Prior to injection, the production batches of each ITR variant libraries were pooled, re-quantified and the barcodes re-sequenced, before being injected in six female C57BL/6 mice (Figure 27). The intended dose for retro-orbital injection was $5x10^{11}$ vg per mouse, but due to a later identified titration

error the actual dose was substantially lower with only $5x10^9$ vg per mouse. Three mice were sacrificed after 14 days and the remaining three 60 days post-injection. Biodistribution analysis as well as the determination of the barcode composition at both the DNA and RNA level by BC-seq were performed in all tissues. Furthermore, the formation of circular episomes was also analyzed in DNA from liver and spleen, again relying on the approach using RCA amplification. In addition, I developed a novel approach for determining the host genome integration propensity of barcoded vector genomes, which was tested there.

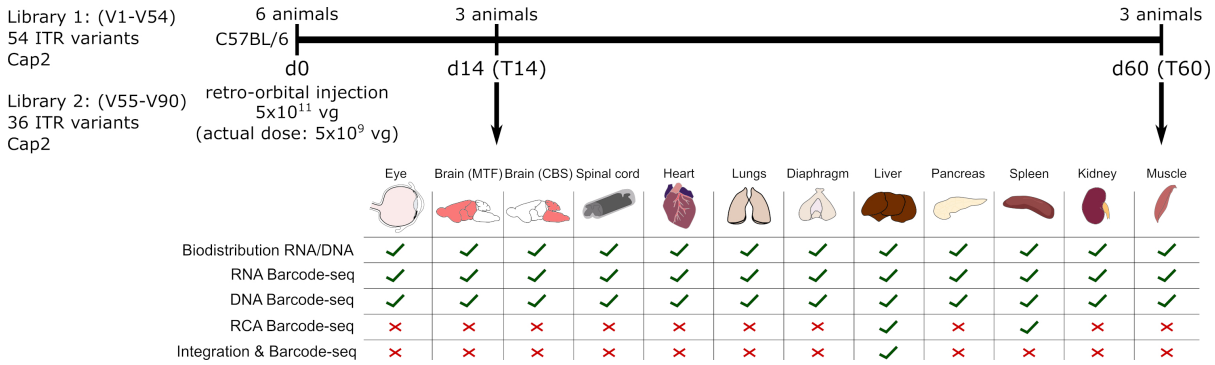


Figure 27: Experimental outline for *in vivo* screening of two ITR variant libraries. Two libraries were administered by retro-orbital injection in C57BL/6 mice with dose and number of animals as well as the time of euthanasia indicated in the timeline. The table below indicates the collected tissue samples and assays for which they were used. Animal injection, sacrifice and organ extraction was performed by the group of Luís Pereira de Almeida at the University of Coimbra. (MTF = Midbrain + thalamus + forebrain, CBS = Cerebellum + Brainstem)

3.1.7.1 Biodistribution of ITR variant vectors *in vivo*

Consistent with the known liver tropism of the AAV2 capsid used for packaging the ITR variant libraries, the biodistribution on the DNA level showed a predominant transduction of liver tissue (Figure 28A). Despite the low vector dose, vector genomes could be detected at both time points, with roughly one vector genome per diploid genome (vg/dg) remaining in liver tissues after 60 days. Interestingly, in thoracic diaphragm and eye tissue, generally not a target tissue of the AAV2 capsid, vector genomes could be detected in some samples. Vector genomes could also be detected in the spleen, yet by comparing the expression of the *mCherry* transgene with a housekeeper gene, this could be identified as non-functional transduction (Figure 28B). This data also revealed that transgene expression was detectable in liver tissue at both time points, with higher relative expression at the later time point. Expression in diaphragm and eye tissue was also observed, albeit not in all examined tissues, suggesting a potential contamination during tissue extraction or injection.

Next, the barcode distribution on the DNA level was determined and the enrichment score calculated. For ITR variant library 1 with variants V1-V54, the median enrichment scores across tissues were displayed as a heatmap (Figure 28C). Already at the 14-day time point, the scores determined for diaphragm, liver and spleen were clearly distinguishable from other tissues due to a much more uniform distribution. This pattern became even clearer on day 60 and was consistent with the observed presence of vector genomes in these tissues. Similarly, the median RNA enrichment scores for library 1 with variants V1-V54 were determined (Figure 28D). There, the enrichment score in most tissues matched the one observed in the liver, although it needs to be noted that most off-target tissues had very low read counts, suggesting a misleading distribution due to a very low expression. The barcode distribution pattern similar to liver may suggest an origin from demultiplexing errors, counting liver-derived reads to the other organs. For the second ITR variant library with variants 55-90, similar issues and patterns could be observed (data not shown). Given these shortcomings of the analysis, especially in tissues with very low transduction, further analysis focused on liver tissue, as it was regarded as the most reliable dataset.

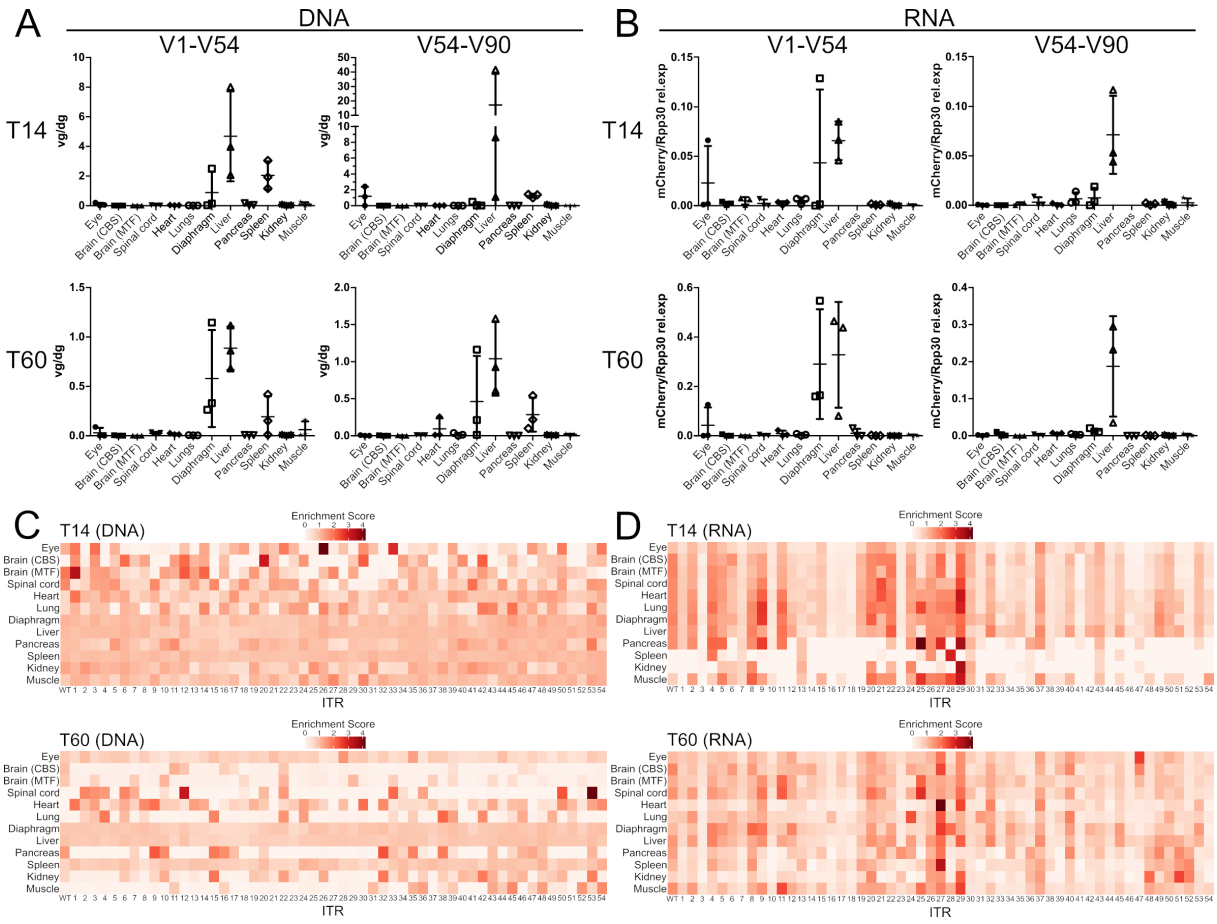


Figure 28: Biodistribution of ITR variant libraries murine tissue. (A-B) Biodistribution of two ITR variant libraries determined by ddPCR. Distribution of vector genome content in the cells on the DNA level measuring vector genome content (CMV) against murine diploid genomes (*Rpp30* gene) **(A)** and

on the RNA level measuring the distribution of transgene expression (*mCherry*) against a host RNA (*Rpp30* cDNA) **(B)**. Data points for all three animals are provided with mean (horizontal line) and standard deviation (whiskers). **(C)** Heatmaps showing the median DNA enrichment scores from all three animals for library 1 (V1-V54) at both time points in all analyzed tissues, calculated by normalizing the barcode recovery in tissue DNA to the barcode distribution in the input library. **(D)** Heatmaps showing the median RNA enrichment scores from all three animals for library 1 (V1-V54) at both time points in all analyzed tissues. Enrichment scores were calculated by normalizing the barcode distribution in tissue RNA to the barcode distribution in the input library. (CBS = cerebellum + brain stem, MTF = midbrain + thalamus + forebrain).

3.1.7.2 Liver transduction and episome formation

Murine liver represents a biologically relevant context where a divergent behavior of the variant ITRs from the one observed in cell culture could be established. To deepen the understanding of the ITRs' actions in the tissue, the barcode distribution on the DNA level was examined in more detail in the liver tissues for the first library (V1-V54; Figure 29A). This indicated that most ITR variants were present at levels similar to those in the input library. Albeit several ITR variants (V11 + V13-V23) with mutations in the C-hairpin displayed a modestly reduced abundance. The distribution appeared largely independent of the time point post-injection, but the variability was increased in the 60-day sample. On the RNA level, the enrichment scores indicated a larger variation from the input library, indicating an impact of the ITRs on functional transduction of the liver tissue (Figure 29B). The wtITR2 did not outperform the ITR variants with paired mutations, contrasting the previous observations *in vitro*. Several ITR mutants even showed slightly elevated transgene expression compared to the wtITR2. Again, no pattern could be identified that connected the variants with higher transgene expression.

To assess episome formation, I also enriched episomal DNA by pre-amplification via RCA. The gel image of the RCA product after debranching showed amplification of unit-length DNA fragments in two of six mice from the 14-day time point (Figure 29C), with one band being rather faint. All six mice at the 60-day time point showed the clear presence of a unit-length DNA fragment, indicating that episome formation after delivery using the AAV2 capsid in murine liver is a slow process requiring more than 14 days. Using the RCA amplicon as template for BC-seq, the enrichment score was determined for the animals injected with library 1 (V1-V54; Figure 29D). Without a consistent formation of an RCA product in the 14-day sample, the data variability was expectedly very high. However, also in the 60-day samples, where RCA generated a clear episome-derived amplicon, substantial data variability was detected between the different animals. A single ITR variant, V14, showed consistently high

enrichment scores, which could indicate an improved episome formation. However, the overall variability, likely driven by the low episome copy numbers, undermined the reliability of the dataset and prevented a clear conclusion.

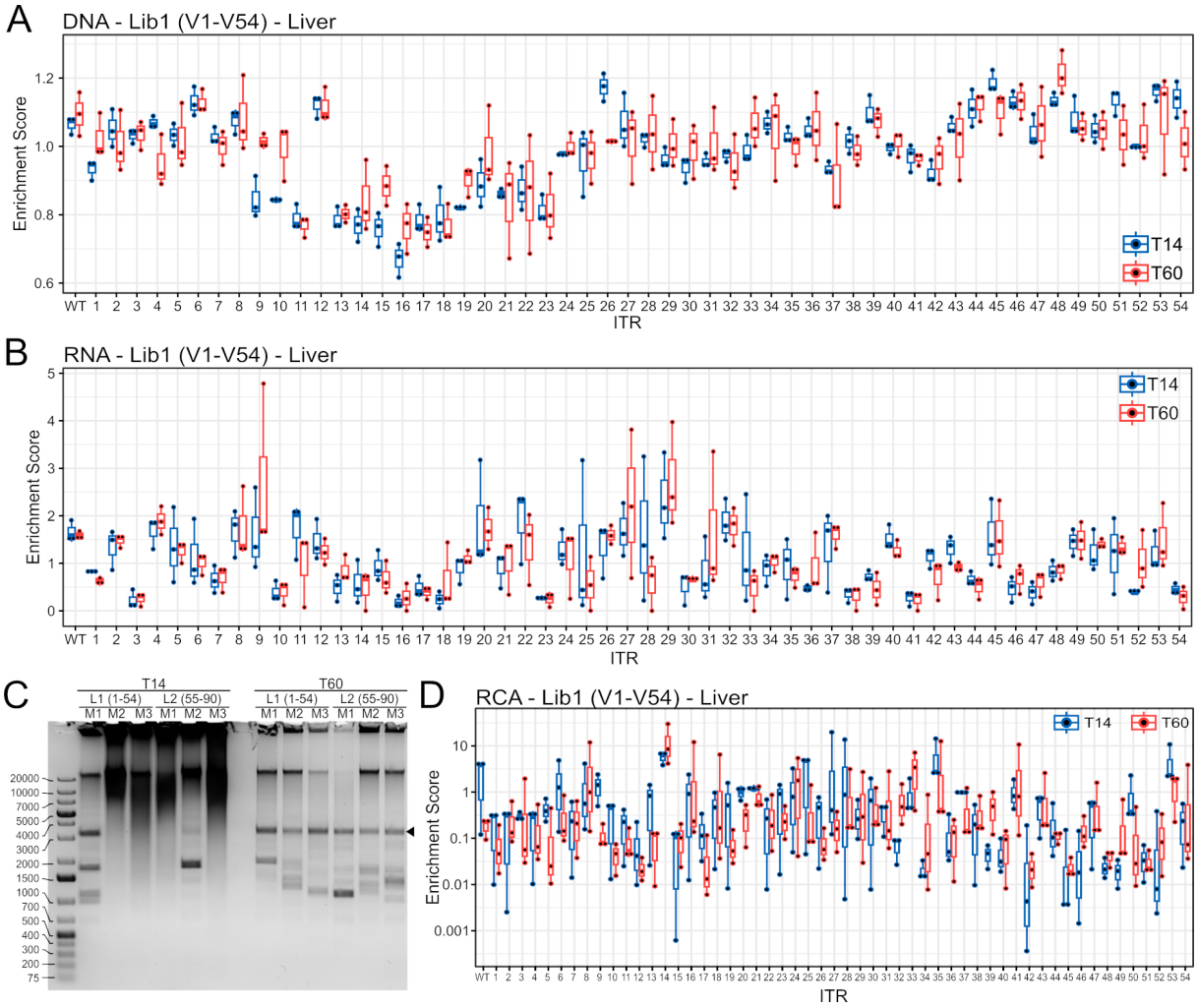


Figure 29: Transduction of ITR variant library 1 (V1-V54) in murine liver. (A) Enrichment scores indicating transduction efficiency of the ITR variants (wtITR2 + V1-V54) on the DNA level in murine liver. The scores were determined by normalizing the barcode distribution in isolated DNA, determined by BC-seq, to the barcode distribution in the input vector library. **(B)** Enrichment scores indicating the ITR variants (wtITR2 + V1-V54) efficiency for functional transduction of murine liver. The scores were determined by normalizing the recovered barcode frequencies determined by BC-seq in RNA/cDNA to the barcode distribution in the input library. **(C)** Gel image of RCA-amplified gDNA from all three mice of both libraries [Library 1 = L1 (V1-V54), Library 2 = L2 (V55-90)] at both time points after debranching with BamHI. The expected unit-length vector genome size of 3.3 kb is indicated by a black triangle. **(D)** Enrichment scores indicating episome formation efficiency of ITR variants (wtITR2 + V1-V54) in murine liver tissue. The scores were determined by normalizing the barcode distribution in RCA-amplified gDNA to the barcode distribution in the input library. **(A, B, D)** Lower/upper hinges of the boxes correspond to first/third quartiles. Median is shown as horizontal line. Whiskers extend no further than 1.5 IQR from the hinges.

For the second ITR variant library (V55-V90), the enrichment scores determined for the transduction of murine liver on the DNA level indicated that most ITR variants maintained

their relative abundance consistent with the input library across both time points (Figure 30A). For C-hairpin mutant variants (ITR variants V55-V72) a slightly reduced abundance compared to variants with mutation in the B-hairpin (ITR variants V73-V90) could be detected, albeit this effect was subtle. Contrasting the previous observations *in vitro*, on the RNA level, some variants exhibited transgene expression on par with or even exceeding that of wtITR2 in murine liver (Figure 30B). A pattern that distinguished well- and low-performing variants could not be determined. The enrichment scores for RCA-enriched episomes (data not shown) exhibited very high variability and were therefore concluded to be unreliable.

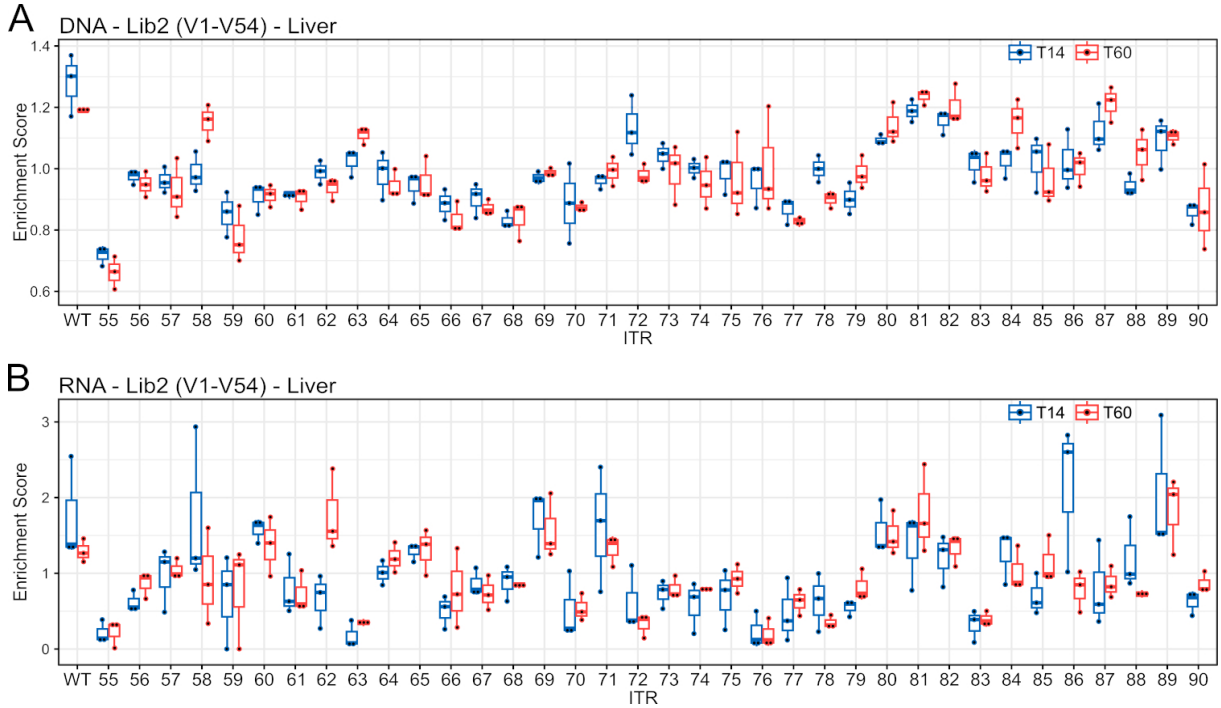


Figure 30: Transduction of murine liver with ITR variant library 2 (V55-V90). (A) Enrichment scores for transduction on DNA level in murine liver with library 2 (ITR variant V55-90). Enrichment scores were determined by normalizing the barcode recovery in isolated DNA to the barcode distribution in the input vector library. **(B)** Enrichment scores for transduction of murine liver with library 2 (ITR variant V55-90) on the RNA level. Enrichment scores were determined by normalizing the barcode recovery in RNA to the barcode distribution in the input library. **(A+B)** Lower/upper hinges of the boxes correspond to first/third quartiles. Median is shown as horizontal line. Whiskers extend no further than 1.5 IQR from the hinges.

3.1.7.3 Tracing the integration of ITR variants in mouse liver

Beyond affecting transgene expression and episome formation, ITRs may also modulate the integration of vector genomes into the host genome. In order to determine the integration propensity of different ITR variants, both the barcode and the integration locus should be determined. For this purpose, I adapted the established ITR-seq method [381] to facilitate the simultaneous detection of the integration site and the barcode sequence. The strategy for this

adapted ITR-seq is depicted in Figure 31A, along with the two sequencing approaches that were used in this work. The difference between the two approaches was the addition of a short, randomized dephasing nucleotide sequence, which facilitates cluster identification when sequencing is performed on a single flow cell. The approach without these dephasing adapters was conducted using a commercial NGS amplicon sequencing service. For both approaches, the gDNA is randomly fragmented followed by end repair and dA-tailing. Then, the Y-adapter was ligated, which, depending on the sequencing approach, already contained part of the Read1 adapter sequence (without dephasing adapters) or a simple primer binding sequence (with dephasing adapters). Both Y-adapters contained a unique molecular identifier (UMI), to determine whether reads mapping to the same region were multiplied due to clonal expansion or PCR amplification. After ligation, primer binding sites in the Y-adapter as well as upstream of the barcode sequence in the vector genome were used to amplify the junction between vector genome and the host genome. Of note, conventional ITR-seq uses a primer binding site located within the ITR sequence [381]. In a subsequent PCR reaction, the sequencing adapters were attached, and the samples could be sequenced using paired-end sequencing. Then, the sequence upstream of the Y-adapter in the NGS reads was mapped to the reference genome. The mapped reads finally underwent a stringent filtering, reducing the number of reads by two orders of magnitude (Figure 31B). To be included for analysis, the forward read had to span the barcode region and contain one of the predefined barcodes. The reverse read must include the Y-adapter sequence to enable UMI identification, but must not map to the forward sequencing adapter, which was observed to lead to false integration events. Additionally, reads that mapped to the genome with a low mapping quality (mapq<10) were excluded, unless the reverse read contained a substantial segment of the vector genome. This retained reads with short genomic fragments and ambiguous integration site, which still represent genuine integration events. Given that the primary objective was to identify barcodes of integrated vector genomes rather than map integration sites with base-pair precision, the liberal mapping quality score and inclusion of reads with ambiguous mapping was considered justified.

After adjusting for PCR amplification and clonal expansion, the total number of unique integration events was obtained (Figure 31C). It should be noted that the DNA input levels using the dephasing adapter strategy for the 14-day samples far exceeded the ones for the 60-day samples, as the fragmentation time was optimized on these samples, and all fragmented DNA was used for subsequent processing. For the strategy without dephasing adapters, the DNA input levels were equal across time points, thereby better representing the distribution of

detected insertions, thus suggesting that integration of vector genomes after transduction was delayed by more than 14 days.

Next, the actual integration sites were further analyzed. Since this is a new sequencing approach and may include a substantial number of ambiguously mapped integration sites, this allows a comparison of the detected integration sites with published literature and can provide validity to the method. To visualize the integration sites, a chromosome map was generated for the sequencing approach with dephasing adapters (Figure 31D). There, the most obvious example of a read with ambiguous mapping position was the single integration event mapped to the Y-chromosome, which is impossible in female animals. Interestingly, the genomic fragment in the read had a length of 83 nt and mapped to a long interspersed nuclear element (LINE) in a total of 42 different genomic regions with 100% identity (data not shown). The presence of a partial vector genome sequence in the read further confirmed that the read stems from an integrated vector genome. Thus, despite the flagrant mapping error based on the multiple exact matches in the genome sequence, the integration event was not discarded. Another conspicuous feature was the accumulation of integrations in specific regions of the genome, among them in the unplaced scaffold chrJH584304 and in specific regions of chromosome 2, 9 and 14. Interestingly, a common element of these regions was the presence of repetitive elements such as LINE, short interspersed nuclear elements (SINE) and rRNA or satellite repeats. This further emphasized the necessity to explore the genomic regions in which the rAAV genomes were integrated, especially the unexpected accumulation in specific genomic regions containing repetitive elements.

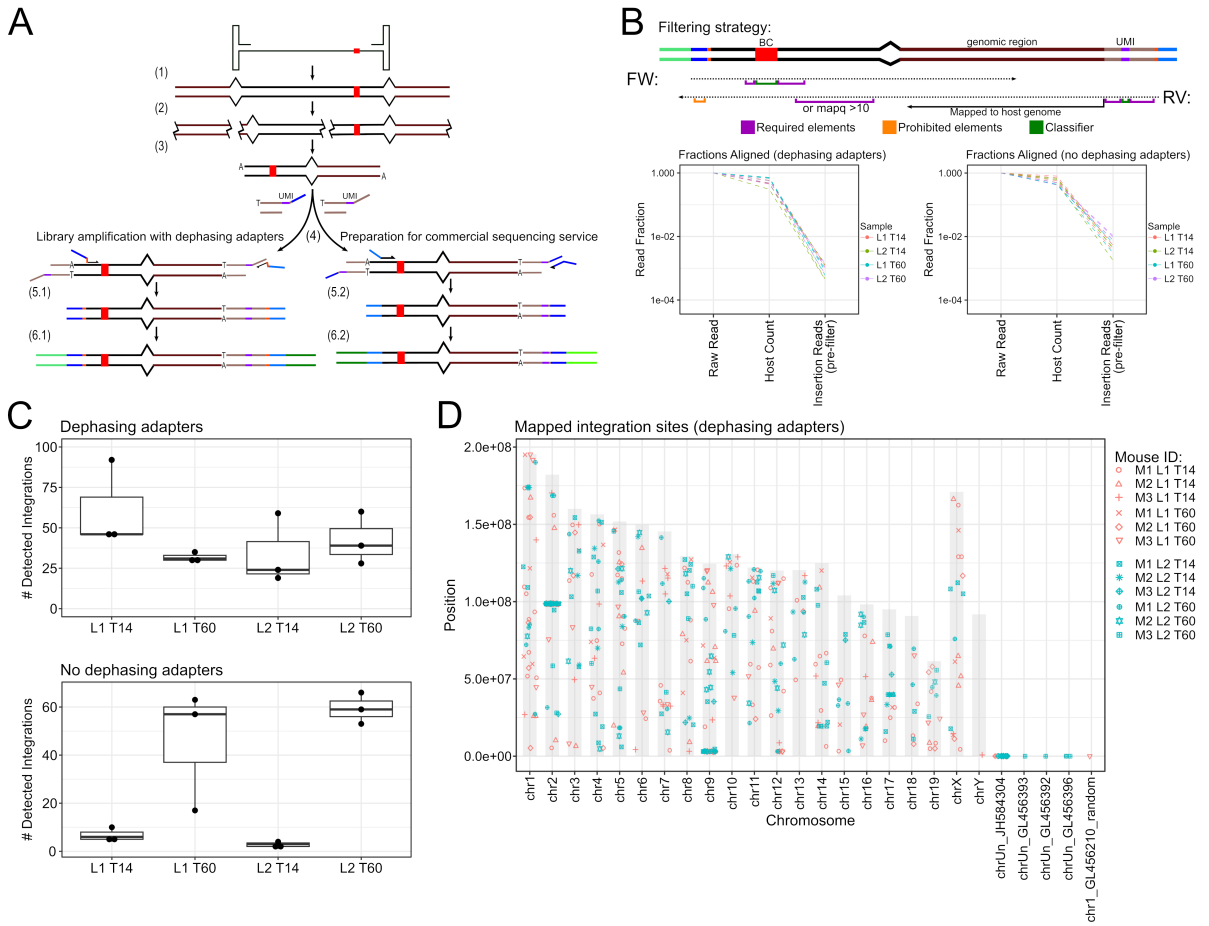


Figure 31: Identification of genomic integration sites of barcoded rAAV *in vivo* using adapted ITR-seq.Strategy to sequence barcodes and integration site via the adapted ITR-seq strategy. (1) The barcoded (red square) vector genome (black lines) integrates into the host genome (brown lines). (2) The genomic DNA is isolated and fragmented followed by (3) end repair, A-tailing and (4) ligation of the Y-adapters equipped with a UMI. Depending on the strategy for sequencing, the subsequent amplification of the insertion with primers binding in the Y-adapters and upstream of the barcodes differs. (5.1) When samples were sequenced on a single flow cell, primers with dephasing adapters were used to increase sequence diversity and then (6.1) the sequencing adapters were attached via a second PCR. (5.2) For commercial sequencing services, the primers contain sequencing adapters that are required for (6.2) the secondary PCR attaching the sequencing adapters. **(B)** Filtering of NGS reads. (Top) After mapping to the host genome, several sequences in the forward and reverse reads are required to be present. Purple bar = reads are required to contain these sequences (or alternatively fulfill the exception indicated below the bar). Red bar = reads were discarded if these sequences were present in the read. Green bar = sequence needs to contain these variable sequences. (Bottom) Plots indicating the loss of sequences by aligning to the host genome (Host count) and by then performing the raw filtering of the reads with the above listed criteria (Insertion reads). **(C)** Total number of identified unique integration events from the two sequencing experiments for all mice at both time points. Lower/upper hinges of the boxes correspond to first/third quartiles. Median is shown as horizontal line. Whiskers extend no further than 1.5 IQR from the hinges. **(D)** Genome ideogram with the identified integrations sites mapped in the mouse genome for both ITR libraries (red and turquoise) from sequencing results with dephasing adapters. Data labels: M1-3 = animal ID, L1 = library 1 (V1-V54), L2 = library 2 (V55-V90).

AAV and rAAV are known to integrate in regions with secondary structures in mice [330], while in humans and primates, AAV integration has also been associated with repetitive regions,

especially rRNA regions, albeit less in satellite repeats [206]. The common denominator of the regions in which I could detect an accumulation of integrations were repetitive elements such as satellite repeats, LINE, SINE and rRNA. Regions enriched with such elements are often excluded from functional genomics data, such as chromatin immunoprecipitation sequencing (ChIP-seq), as they can introduce a bias during normalization. For this purpose, lists of these regions which are problematic for these experiments, so-called blacklisted regions, have been curated, which mainly contain rRNA, alpha satellites, and other repetitive elements [368]. These regions matched well with the regions with accumulations of integration sites, as illustrated in the chromosome map depicting the integration sites determined from the sequencing approach without dephasing adapters (Figure 32A). This further confirmed that a large number of integration events took place within such repetitive regions, as seen for samples generated with dephasing adapters (Figure 32B) and without them (Figure 32C). Except for samples with a very low number of integration events, approximately 20% of all integration events were associated with blacklisted regions at both time points. Due to the inclusion of integrations with low mapq score when the vector genome was detected in the read, this raised the question whether only reads with low mapping quality were aligned to blacklisted regions. Notably, the average mapq score in proximity to blacklisted regions was reduced compared to reads mapped to normal regions, as shown for the sequencing approaches with dephasing adapters (Figure 32D) and without them (Figure 32E). Despite the lower mapq scores near the blacklisted regions, a substantial number of reads still mapped to these repetitive regions with high mapq scores. This suggests that the elevated integration levels in these regions were not merely an artifact due to low mapping quality.

With only 0.003% of the genome annotated as blacklisted but nearly 20% of all integrations mapped into these regions, I next explored how the observed rAAV integration differs from random integration in different genomic regions. For this purpose, the annotation of the detected integration sites was compared with 40 sets of randomized integration sites of equal size, as shown for the data from experiments with dephasing adapters (Figure 32F) and without them (Figure 32G). The observed integration pattern between the two sequencing approaches in comparison to the randomized datasets was strikingly similar. As expected, the strongest overrepresentation of integration events was within the blacklisted regions. Integration in enhancer and promoter regions, or in exons and intron/exon boundaries was also elevated compared to the randomized dataset. Integration in introns was comparable to the randomized dataset, while rAAV integration in intergenic regions was substantially lower than random

integration would suggest. This pattern indicated a bias for rAAV integration in regions associated with transcription and thus potentially open chromatin.

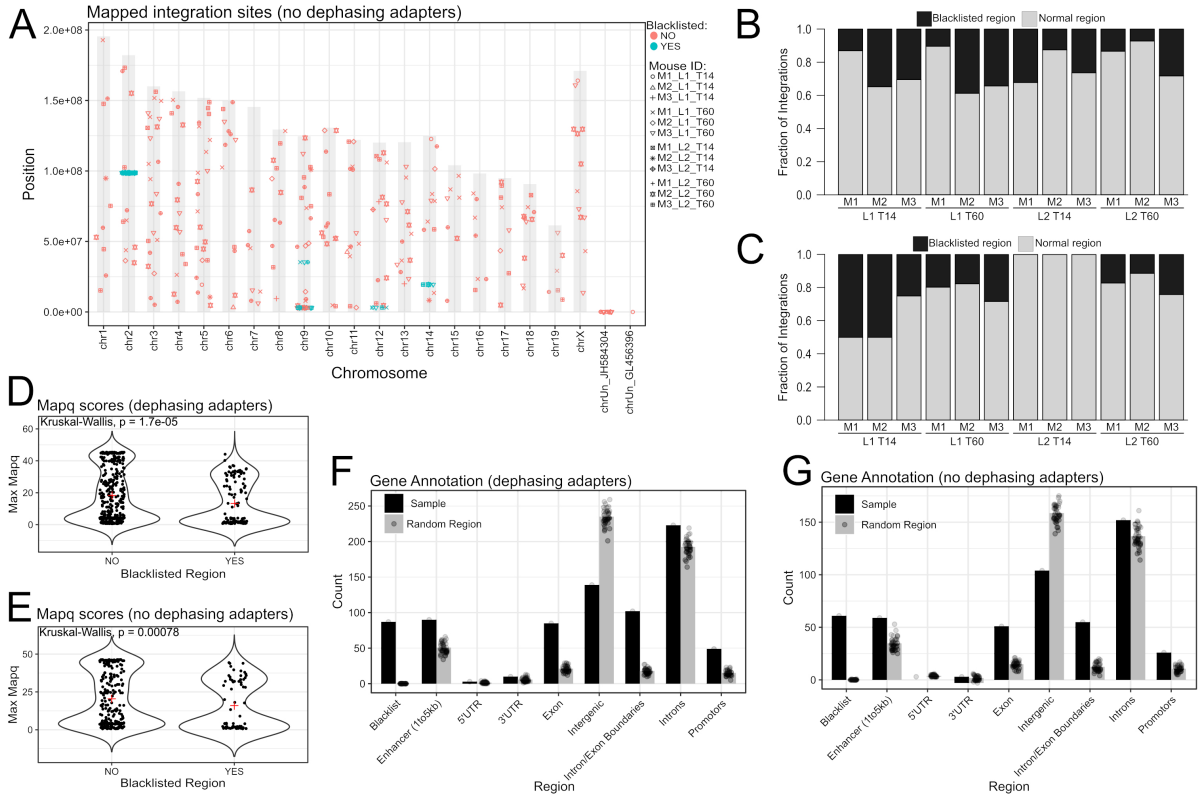


Figure 32: Preferential rAAV integration regions identified by adapted ITR-seq in murine liver. **(A)** Genome ideogram with integration sites from sequencing without dephasing adapters showing the accumulation of integrations in blacklisted regions. **(B-C)** Fraction of identified integrations in blacklisted regions in sequencing results from experiments with dephasing adapters **(B)** and without them **(C)**. **(D-E)** Mapq scores for integrations in normal and blacklisted regions in sequencing results from experiments with dephasing adapters **(D)** and without them **(E)**. The maximum mapq score at an integration site was plotted, and the red cross indicates the mean. Statistical analysis was performed using Kruskal-Wallis test. **(F-G)** Gene annotations of integration sites when compared with 40 equally sized sets of randomly selected genomic regions from experiments with dephasing adapters **(F)** and without them **(G)**.

The primary objective of adapting the ITR-seq approach was to attribute integration events to specific ITR variants based on their associated barcode sequence. Achieving this required the relocation of the primer binding site from the ITRs, as used in conventional ITR-seq, into the transgene region to ensure inclusion of the barcode in the sequencing reads. This, in turn, complicated the clear identification of the vector-genome junction and thus the exact mapping of the read. This led me to investigate whether the alternative primer location might have other advantages, particularly given previous reports of partial vector genome loss during integration [314]. In a random set of 110 reads that passed the initial filtering after alignment to the sITR reference sequence, this loss was clearly recognizable (Figure 33A). While most of

the junctions between vector and host genome were located at the edge of the A-sequence within the ITR, several reads showed extended deletions of vector genome sequences. Integration events with such partial vector genome deletions could not be detected using the conventional ITR-seq approach due to the loss of the primer binding site. The presence of the barcode sequence within these reads could also be clearly observed, and were subsequently used to identify the ITR variant originally present at the termini of the vector genome. Due to the low injection dose and resulting limited integration events, only by pooling the data from all mice from both time points and both sequencing approaches I could identify at least a single integration event for each ITR variant. The barcode distribution in integrated vector genomes could then be normalized to the distribution in the input library, allowing the calculation of enrichment scores that reflect the relative integration propensity of individual ITR variants. For both library 1 (V1-V54; Figure 33B) and library 2 (V55-V90; Figure 33C), the enrichment scores displayed a relatively random pattern, largely due to the limited number of detectable integration events. As a result, definitive conclusions about the integration propensities of specific ITR variants could not be drawn. Nevertheless, these findings validate the feasibility of using the adapted ITR-seq approach to explore whether specific ITR sequences can influence the likelihood of vector genome integration.

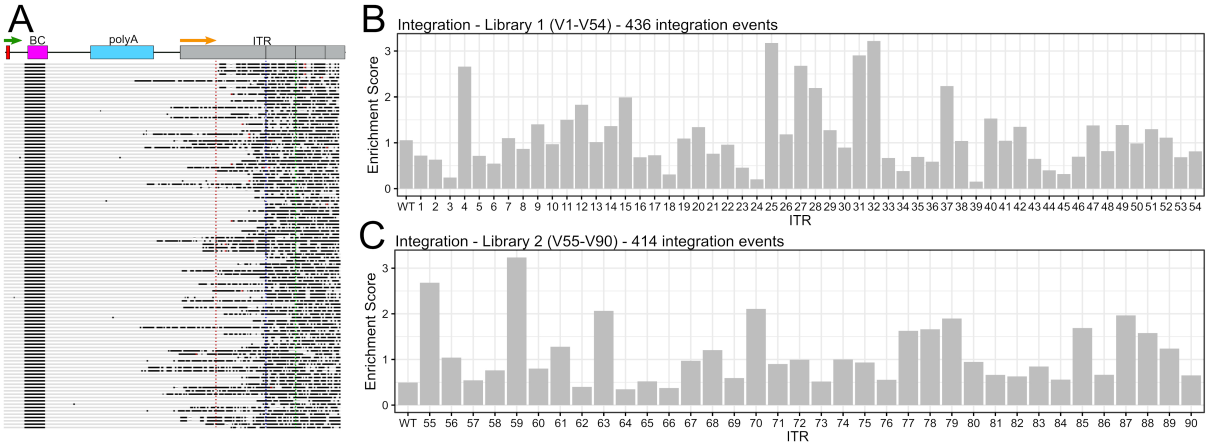


Figure 33: Evaluating the integration propensity of ITR variants via adapted ITR-seq in murine liver. (A) Set of 110 forward reads with identified genomic integration in the reverse read aligned to the vector genome reference, showing the extent of ITR sequence loss and the partial loss of vector genome elements. Nucleotides matching the reference are indicated in gray, mismatches in black. The orange arrow indicates the primer binding site for conventional ITR-seq, the green arrow the primer binding site for the adapted ITR-seq. The red line indicates further indicates this position, the blue line indicates the end of the ITR stem (D-A sequence), and the green line indicates the hinge between B- and C-hairpin. **(B-C)** Barcharts showing the integration enrichment scores, a measure for the integration propensity of the different ITR variants. The scores were calculated by normalizing the barcode distribution in all identified integration sites from both sequencing approaches, time points, and all animals to the barcode distribution in the injected vector input library. Data is given for ITR variant library 1 (V1-V54) **(B)** and ITR variant library 2 (V55-V90) **(C)** and the total number of detected integration events is indicated**.**

3.1.8 Integrity and *in vitro* functionality of aberrant ITR hairpin variants

After observing that point mutations in the ITR hairpins were maintained during separate rAAV production (see Figure 19), I also explored whether this strategy can be transferred to highly divergent ITR variants and maintains their hairpin sequences in the vector genomes. To this end, I generated a new library, based on the second-generation sITR plasmid, comprising structurally diverse ITR hairpin variants but excluding the wtITR2 hairpin (Y1-Y17; see Table 19). These variants ranged from modified wtITR2-derived sequences, including loops or shortened arms, to ITRs derived from distantly related parvoviruses and entirely synthetic sequences.

All sITR plasmid sequence verifications were again performed using the T7eI-ITR sequencing approach before the rAAV were produced via separate transfections, pooled and purified. The barcode distribution in the purified vector pool was determined using both NGS-based BC-seq and Nanopore sequencing (Figure 34A). Several ITR variants dominated the vector pool, all of them having some sequence similarity to the wtITR2 sequence. This included variants with modifications such as shortened hairpin arms (Y6 + Y11), additional loops between the hairpin arms (Y5 + Y7-9) and the wtAAV3b hairpin (Y16). Intriguingly, an alternative hairpin variant with similar dual-hairpin structure as wtITRs, but a TA-rich hairpin sequence (Y4), also retained relatively high productivity.

The Nanopore sequencing data was also used to examine the maintenance of the hairpin sequences, as shown in Figures 34B-F with the nucleotide distribution in the hairpin sequences after alignment to a flip/flop consensus reference sequence of each ITR variant. The variant with the TA-rich dual hairpin (Y4; Figure 34B) showed an intact ITR sequence with both flip and flop nucleotides present at the expected 50:50 ratio. The aptMG sequence (Y1; Figure 34C), which showed a high conversion to the wtITR2 during pooled production (see Figure 15A), could now be completely maintained. However, the flip-flop ratio was skewed with one configuration predominating, as visible by the imbalanced nucleotide distribution. The hairpin of the *S. fusca* densoviral ITR (Y15; Figure 34D) and the *P. monodon* hepandensoviral ITR hairpin (Y14; Figure 34E) were sequenced with only slightly more reads and were not only found to be maintained but also exhibited the ideal 50:50 ratio of flip and flop sequences. Another variant Y13 (Figure 34F), that contained a short loop as hairpin sequence representing a partial G4 sequence [382], also exhibited a skewed flip/flop distribution, despite an overall maintenance of the hairpin sequence.

Interestingly, after transduction in cell culture, in both Hek293T and HuH7 cells (Figure 34G), the variant Y13 with the partial G4 had the second highest enrichment score in the recovered RNA. The wtITR3b variant (Y16) showed the highest enrichment score, however, it should also be noted that the barcode composition of the input library showed a major inconsistency between NGS and Nanopore sequencing data for this variant, potentially inflating the enrichment score. The lowest enrichment score was found for variant Y17, which contained a linear ITR with a TIIS restriction enzyme cleavage site used for cloning of hairpin sequences. Again, RCA amplification of the isolated DNA was performed to enrich episomal DNA followed by BC-seq to determine enrichment scores (Figure 35H). The data variability was more limited, although the number of samples was low. Thus, observations such as the apparent cell line-dependent episome formation efficiency for variant Y10, which contained the *Ae*DV hairpin, could not be definitively clarified.

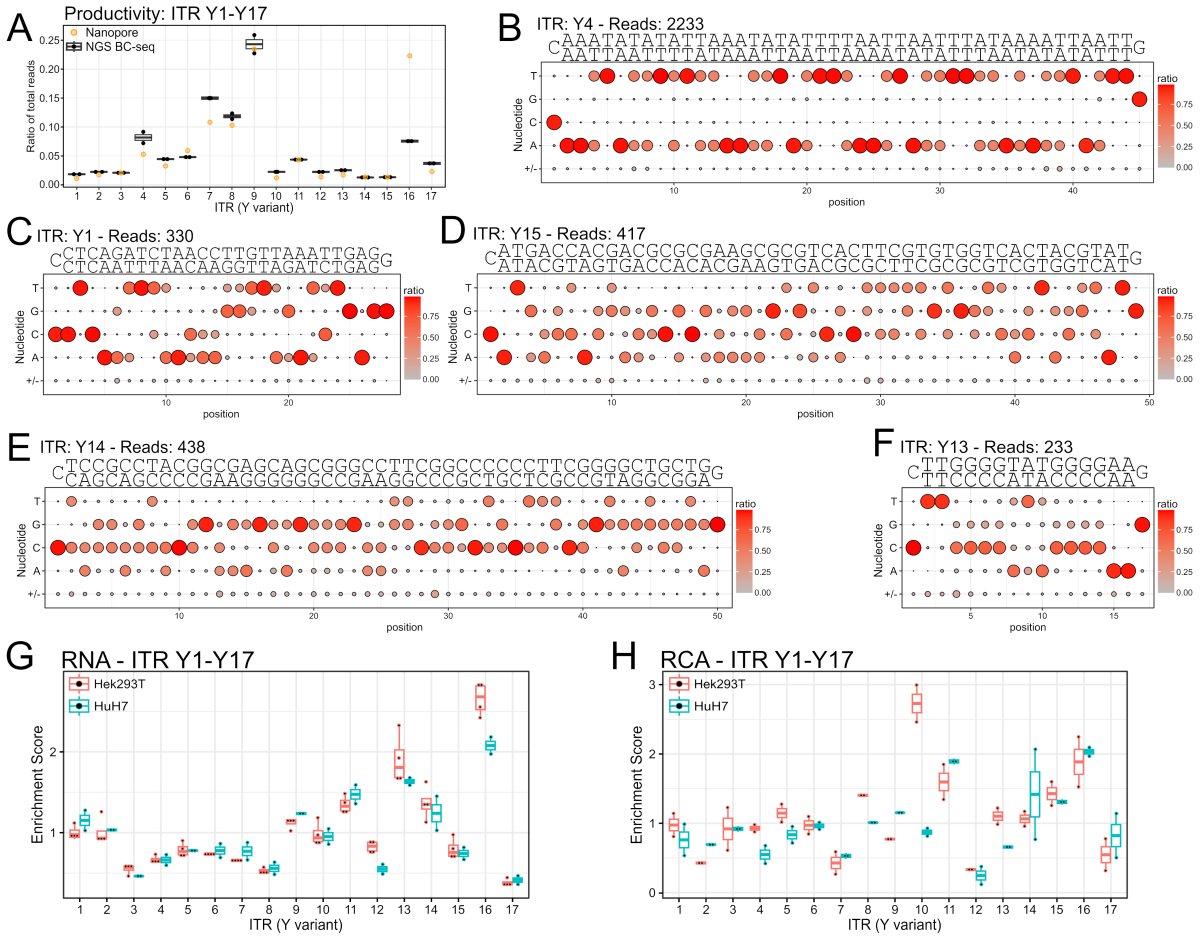


Figure 34: Sequence integrity and functionality of a library with diverse hairpin sequences. **(A)** Barcode distribution after separate production of a diverse set of barcoded ITR variants Y1-Y17. Black dots indicate BC-seq data (sequenced in duplicates) and the orange dots Nanopore sequencing data. **(B-F)** Nucleotide distribution in the hairpin region of aligned Nanopore sequencing reads in vector genomes with **(B)** a TA-rich dual hairpin (Y4), **(C)** a malachite binding aptamer sequence as hairpin (aptMG) (Y1), **(D)** the *S. fusca* densoviral ITR hairpin (Y15), **(E)** the *P. monodon* hepandensoviral ITR hairpin (Y14), or **(F)** a partial G4-sequence forming a loop structure (Y13). **(G-H)** Hek293T cells (red)

and HuH7 cells (turquoise) were transduced with the ITR variant library Y1-Y17 at an MOI of 10^5 and total RNA/DNA was extracted 72 h after transduction. **(G)** Barcodes were sequenced after cDNA synthesis, and the barcode distribution was normalized to the barcode distribution in the input library, generating an enrichment score for transduction in. **(H)** Barcodes were sequenced after pre-amplification of the DNA by RCA to enrich episomal DNA. The enrichment scores were determined by normalization to the barcode distribution in the input library. **(A+G+H)** Lower/upper hinges of the boxes correspond to first/third quartiles Median is shown as horizontal line. Whiskers extend no further than 1.5 IQR from the hinges.

3.2 Harnessing the AAV replication machinery to generate circular genomes

A key limitation of rAAV as vector for gene therapy is their inefficient second-strand synthesis [91] and episome formation, both of which restrict long-term transgene expression [95]. Although scAAV circumvent the second-strand synthesis, they are more immunogenic and possess a lower packaging capacity [257,259]. Furthermore, the linear strands of rAAV may simultaneously be the main origin of random genomic integration events [96]. Concatenated, potentially circular episomal AAVs can integrate [96] but may be less likely to do so, since most integration events were recorded in the early phase during transduction [339]. Therefore, I wanted to explore the possibility to generate circular vector genomes and the ability to package them in AAV capsids, as such vectors may circumvent bottlenecks during transduction and could substantially reduce the risk of random integration.

3.2.1 AAV-mediated formation of circular replication intermediates

Given the evolutionary linkage of AAVs with circular Rep-encoding single-stranded (CRESS) DNA viruses [383], such as circoviruses, and the known DNA ligation ability of AAV2 Rep [384], I hypothesized that AAV Rep could mediate the formation of circular DNA genomes. The AAV Rep-mediated ligation mechanism was originally proposed to be involved in AAV genome integration. The presence of a similar mechanism in circoviral Rep proteins, where it ligates the genome to form circular viral genomes [385], suggested that a redesigned, synthetic AAV ori may enable AAV Rep to catalyze an analogous reaction. For circoviral replication origins, it has been described that two distant RBE sequences can be linked during replication via a RCR copy-release mechanism, generating two small circular genomes from one large plasmid [386]. To test whether AAV2 Rep can facilitate a similar process, I designed a plasmid (2xRBE*) that contains two repeats of the 61 nucleotides of the AAV2 ITR A and D regions (RBE*), separated by the prokaryotic plasmid backbone and a *mCherry* expression cassette. A circovirus-like replication mechanism by AAV Rep2 would result in the generation of two smaller DNA circular genomes, which can then be detected using an RCA-based assay (Figure

35A). These smaller circular genomes contain either the plasmid backbone or the *mCherry* expression cassette and could be isolated from the low-molecular weight DNA fraction of transfected cells. A digest with DpnI could reduce presence of the parental 2xRBE* plasmid, and RCA using primers specific to one of the circular replication intermediates could amplify the smaller circular genomes. After linearization of the RCA product with a single-cutter enzyme, visualization is possible after gel electrophoresis.

An exemplary gel image from an experiment in which Hek293T cells were transfected with 2xRBE* alongside functional and non-functional Rep proteins revealed the formation of circular replication intermediates (Figure 35B). Despite prolonged DpnI digestion, a band corresponding to the original plasmid was still present in most samples. In all samples transfected with a plasmid encoding a functional AAV Rep protein, the presence of an amplified fragment of around 1500 bp could be observed, matching the expected fragment length of the *mCherry* cassette-containing circular replication intermediate. The fragment was absent in the presence of AAV2 Rep mutants with inactivated nuclease function (Y156F) or inactivated helicase domain (K340H), indicating the reliance on Rep-mediated replication. Adenoviral helper genes, which support genome replication, seemed to enhance the formation of the fragment, but were not essential. Usage of alternative restriction enzyme with cleavage site in the replication intermediate yielded the same fragments (Supplementary Figure 5), indicating that the fragment was indeed derived from the circular replication intermediate. As the fragment matching the circular replication intermediate was also generated in the presence of both Rep and VP, potentially even with higher efficiency, I proceeded to use the 2xRBE* as vector plasmid for rAAV production. The presence of the D-sequence within the RBE* sequence was expected to aid the packaging of the vector genome in the capsid particles. Quantification of these vectors by qPCR confirmed that DNA derived from the 2xRBE* plasmids was successfully packaged (Figure 35C).

Attempts to detect circular genomes among the packaged DNA using RCA were unsuccessful (data not shown). To further characterize the nature of the packaged DNA, I performed Nanopore sequencing of the rAAV vector-derived genomes and the original 2xRBE* plasmid. After aligning the Nanopore sequencing reads from the 2xRBE* plasmid as well as from the vector generated with this plasmid to the 2xRBE* plasmid reference sequence using a splice-aware aligner, the presence of large gaps could be observed. The analysis of the gap length in both vector and plasmid sequencing reads (Figure 35D) revealed that most larger gaps matched the size of the prokaryotic plasmid backbone of around 1800 bp. The existence of these gaps

further supported the notion of a re-ligation of the vector genome during replication. Notably, gaps spanning the *mCherry* transgene cassette could also be detected (data not shown). Interestingly, a minor fraction of around 0.03% of plasmid-derived reads also showed a similar 1800 bp gap. While this was substantially lower than the 1% of reads from vector DNA, this warranted a more in-depth analysis of the gaps to exclude potential artifacts. To this end, I generated a Sashimi plot-inspired visualization of both read junction and read coverage for both vector-derived reads (Figure 35E) and plasmid-derived reads (Figure 35F). In the vector samples, a substantial number of reads contained the expected junction across the prokaryotic backbone. The read coverage peaked near the RBE* sequences, indicating that replication is initiated at the RBE*. The subsequent reduction in the 3' direction may also suggest a preliminary abrogation of packaging before reaching the second RBE* and thus an inefficient replication or packaging mechanism. Subsequent filtering for reads that align both upstream of the left and downstream of the right RBE* maintained very long reads as well as short reads that skip the prokaryotic sequences. This indicated that junction formation is more common than initially anticipated. Still, it should be noted that the number of short reads far exceeded the number of long reads (data not shown), which could lead to an overestimation of the gap formation efficiency. Therefore, both coverage and gap junctions further supported the assumption of circular replication intermediates formation and their use as templates for packaging. However, also Nanopore sequencing provided no indication for packaging of circular genomes. Instead, I could observe few reads of potentially linear concatemerized circular replication intermediates (Supplementary Figure 6). Plotting the coverage and gap junctions for plasmid-derived reads (Figure 35F) revealed that the overall number of reads spanning the prokaryotic plasmid backbone sequence is substantially higher. However, a small subset of reads indeed contained junctions across the prokaryotic plasmid backbone. Filtering for reads that span the complete prokaryotic element further emphasized this gap formation in the 2xRBE* plasmid, indicating recombination during plasmid propagation or an alignment artifact. Nevertheless, the combined data of RCA assay and Nanopore sequencing suggested a Rep-mediated circularization process, although no circular genomes appeared to be packaged in the AAV capsids.

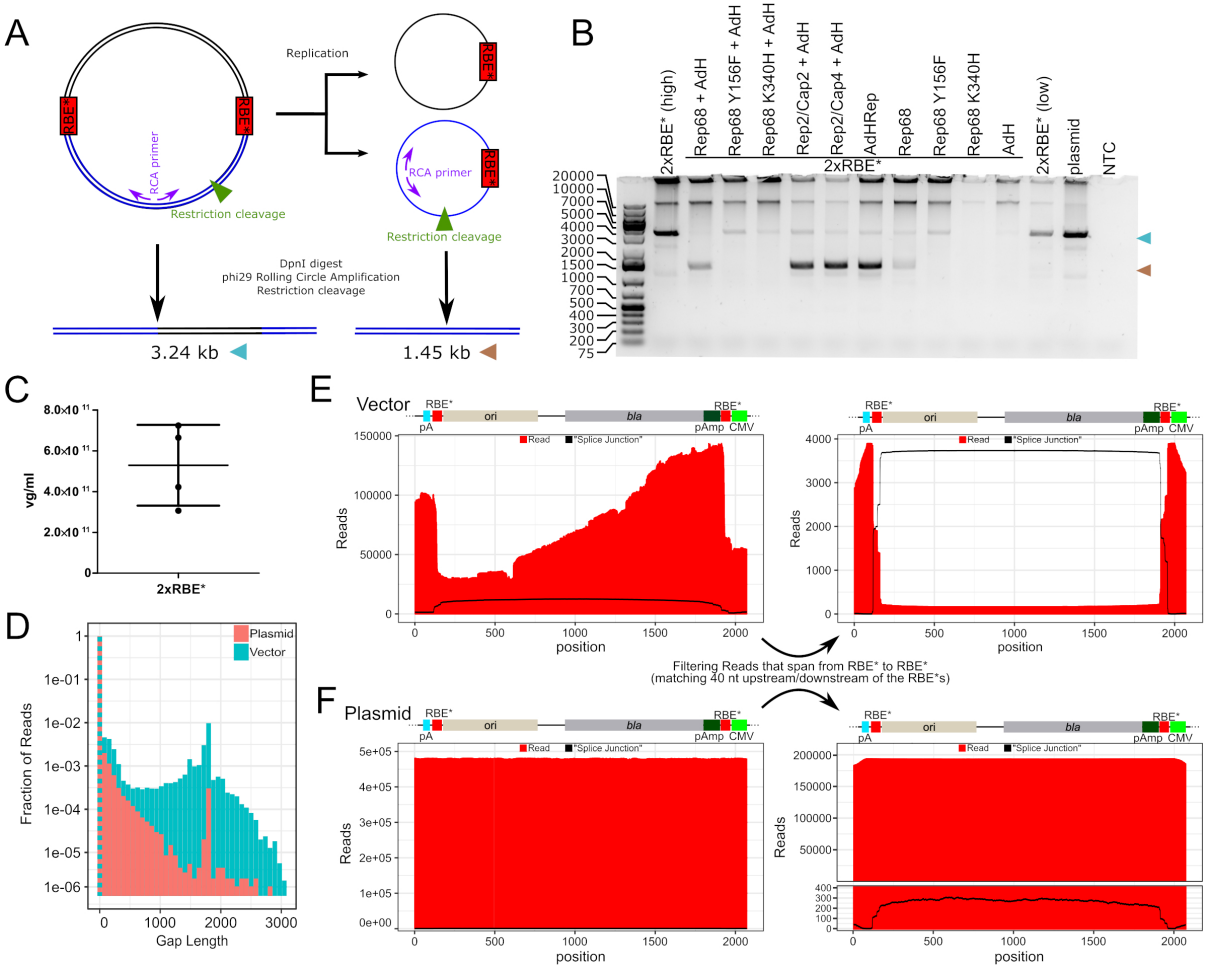


Figure 35: A synthetic AAV replication origin allows circular replication intermediate formation and packaging of genomes in AAV particles. Schematic RCA-assay workflow to verify the AAV Rep-mediated formation of circular replication intermediates. **(B)** Gel image of samples analyzed using the RCA assay. Hek293T cells were transfected with the indicated plasmids, and low-molecular weight DNA was extracted after 72 h and subjected to the RCA assay. For two-plasmid transfections, 2xRBE* was used with the indicated plasmid at a 1:25 molar ratio, whereas in three-plasmid transfections, 1:10:10 molar ratios were used. For samples only transfected with the 2xRBE* plasmid "high" indicates the maximum DNA input, while "low" indicates plasmid input comparable to the triple-transfection samples. During RCA, a plasmid control with 1 ng of the original 2xRBE* plasmid and a no-template control (NTC) were included. Amplicons were debranched by EcoRI-HF digestion and separated on a 1% agarose gel. The triangles indicate the size of the linearized 2xRBE* plasmid (blue) and the size of the linearized replication intermediate (brown) **(C)** qPCR quantification via the CMV promoter sequence in vector produced with 2xRBE* plasmid from four productions with ten plates each. Mean is shown as horizontal line and whiskers indicate the standard deviation. **(D)** Relative distribution of gap sizes in aligned Nanopore sequencing reads from plasmid and vector genomes. The dotted line was used to visualize the overlapping values for the two sample types. **(E-F)** Visualization of Nanopore sequencing read coverage and gaps in vector genomes generated with 2xRBE* plasmid **(E)** and the 2xRBE* plasmid **(F)**. Annotated plasmid region with pA-site (pA), ori, ampicillin resistance gene (bla) and promoter (pAmp). Red peaks indicate the coverage in the prokaryotic plasmid backbone and read junctions are given as black lines. Raw aligned reads (left) and filtered for reads matching upstream and downstream of the RBE* (right).

3.2.2 Circular replication intermediates as templates for rAAV genome packaging

With the previous data showing the presence of potentially recombined 2xRBE* plasmids, I intended to additionally monitor the formation of the circular replication intermediates using a dual-luciferase reporter. This 2xRBE*-based reporter construct was designed to switch the expression from one reporter gene to another upon successful formation of the circular replication intermediates. Similarly, vectors generated from this plasmid were expected to exhibit a reporter gene expression pattern distinct from that of the original, unreplicated plasmid.

To this end, the 2xRBE* plasmid design, containing two segments separated by the RBE* sequence, was modified (Figure 36A). One segment of the plasmid was equipped with a split *Renilla* luciferase (*R*Luc)-encoding sequence with the two fragments separated by the polyA-signal and a promoter. The N-terminal part of the *R*Luc sequence was equipped with a synthetic splice donor site, and a synthetic splice acceptor site was positioned upstream of the C-terminal *R*Luc sequence. In the other half of the plasmid, a sequence encoding Firefly luciferase (FFluc) was incorporated with a splice acceptor site upstream of a Kozak sequence adjacent to the start codon. When transfected, the plasmid was expected to be unable to generate a full-length *R*Luc mRNA due to the interrupting elements, while FFLuc expression remains possible. Upon formation of the circular replication intermediates, the N-terminal *R*Luc fragment with its splice donor would be separated from the splice acceptor for the C-terminal fragment only by the RBE* region, thus facilitating the generation of a functional *R*Luc mRNA. The circular replication intermediate carrying the FFLuc transgene would not contain a promoter, thus no expression should occur. Attempts to harness the reporter for direct observation of Rep-mediated circularization were unsuccessful (data not shown), as no conditions were identified in which the promoter showed a consistent performance. Nonetheless, the reporter plasmid proved useful to verify the packaging of DNA fragments, albeit likely as linear concatemerized rather than circular genomes, derived from circular replication intermediates. Packaged genetic sequences derived from circular replication intermediates should retain similar expression patterns as the intermediates themselves, *i.e.*, elevated *R*Luc expression and diminished FFLuc expression. If vectors predominantly incorporate genetic material derived from unprocessed plasmid DNA, the FFluc signal would be expected to dominate. Consistent with the packaging of circular replication intermediate-derived DNA fragments, the rAAV vector produced with AAV2 Rep and VP showed elevated levels of *R*Luc expression after transduction in Hek293T cells compared to cells directly transfected with the reporter plasmid (Figure 36B).

Additionally, this data proved that vectors generated using 2xRBE* plasmid can deliver functional transgenes.

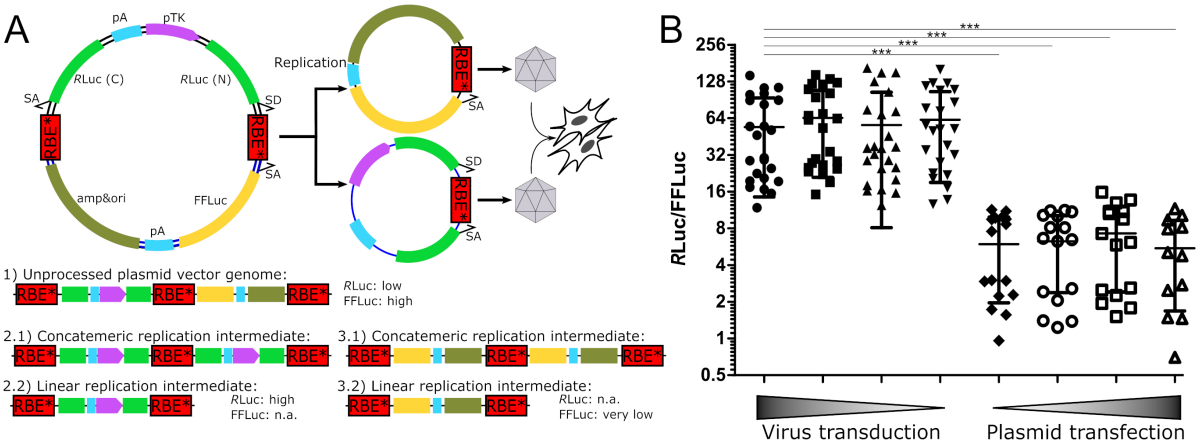


Figure 36: Tracking of circular replication intermediate amplicon packaging using a luciferase reporter construct.(A) (Top) Schematic of the dual luciferase reporter plasmid with split-*R*luc, which is assembled into a functional unit after formation of the circular replication intermediate with the help of splice sites. The intermediates may then be used as templates for packaging of rAAV vector, and the luciferase expression levels can be determined by subsequent transduction. (Bottom) Possible linear vector genomes generated from the 2xRBE* reporter plasmid and its circular replication intermediates may consist of three different types, depending on the template DNA used for replication. The associated changes in luciferase expression are indicated. SA = splice acceptor, pA = polyA signal, pTK = thymidine kinase promoter, SD = splice donor, amp&ori = ampicillin resistance cassette and ori. **(B)** Hek293T cells were transduced with different quantities (20 µl, 15 µl, 10 µl, 5 µl) of the vector generated from the 2xRBE* dual luciferase after iodixanol gradient purification or transfected with the 2xRBE* dual luciferase plasmid (25 ng, 50 ng, 75 ng, 100 ng). Increasing concentrations are indicated by the triangle shapes on the x-axis. Luciferase counts were analyzed 72 h after transduction/transfection and normalized. Data showing the luciferase activity switching is from three different transduction/transfection experiments with vectors from three different productions. Mean is indicated as horizontal line and whiskers represent the standard deviation indicated. Pairwise comparison of selected groups was performed using the Wilcoxon test (ns: $p>0.05$, *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$).

4. Discussion

The development of novel or optimized rAAV vectors with improved characteristics could significantly improve the safety and efficacy of gene therapies, while reducing costs, thereby increasing therapeutic options and accessibility for patients. In this work, I examined two strategies for engineering, the AAV replication origin to generate such a next-generation vector. First, I developed a novel screening pipeline for ITR variants, ranging from plasmid generation for ITR modification and sequence validation of the plasmid ITR, to the analysis of the effects of these ITR variants on transgene expression, episome formation and integration propensity. The second strategy was to directly package circular genomes into AAV particles, harnessing an alternative mode of replication by AAV Rep. Both approaches provide an excellent and versatile basis for future engineering efforts to achieve this seminal goal.

4.1 Engineering and screening AAV vectors with altered ITR sequences

Engineered ITRs have the potential to be a central part of a next-generation rAAV vector platform. To date, AAV ITR engineering has usually been accomplished by generating a specific variant followed by a pairwise comparison with a wtITR vector *in vitro* or *in vivo* [241,292,306]. By leveraging the sITR vector plasmid design in this work and by using cost-effective and simple cloning procedures, it was feasible to generate large libraries of ITR variants, exemplified by the generation of two ITR point mutant libraries comprising a total of 90 ITR hairpin variants. Moreover, a second major former hurdle, *i.e.*, the difficulty of sequencing the extended ITRs within the sITR plasmids, was overcome by the development of a new Sanger sequencing approach for ITRs. Following ITR sequence verification, the sITR plasmids could be used for rAAV production. Due to the newly discovered ITR *trans*-repair mechanism, all productions had to be performed separately to ensure the integrity of the ITRs in the generated rAAV vector genomes. The novel screening pipeline developed in this study included the tracking of ITR variants *in vitro* and *in vivo* post-transduction based on barcodes identifying each ITR variant. This enabled the determination of the ITR's effect on transgene expression, episome formation, and host genome integration. In the following sections, the components, steps and mechanisms as well as the identified pitfalls and challenges involved in this comprehensive engineering and screening strategy will be discussed.

4.1.1 Advantages and challenges of the sITR plasmid design

The sITR plasmid design [256]was pivotal for the generation of vector genome plasmids with alternative ITR sequences. While rarely used, it has already been employed for the generation of single-strand polarity AAV, after deletion of one D-sequence [93,94], and for the creation of ITR variants [306,357]. There, ITR variant generation relied on the assembly of two elongated DNA fragments with the vector backbone in a three-fragment assembly [306,357]. To increase throughput and efficiency while reducing costs, I adapted the cloning strategy. It now encompasses a two-fragment assembly, relying on PCRs with a single primer for stem sequence mutations, harnessing the inverted repeat structure of the ITR, and annealed oligonucleotides to alter the hairpin sequence.

As demonstrated, this facilitated the generation of a variety of sITR plasmids with diverse hairpin structures as well as mutations in the stem sequence. The remaining challenge of ITR engineering is not the cloning of the variant, but the design of the hairpin. Previous works altered specific features in the wtITR, such as binding sites [306] or sequence motifs [241] in the existing wtITR or introduced naturally occurring ITR mutants [292]. Similarly, in this work, every nucleotide of the terminal dual hairpin was mutated, either with a pairing mutation (V1-V54) or a mismatching mutation (V55-V90). However, alternative ITR structures and sequences could have a more pronounced impact on transduction, integration or episome formation compared to structures that maintain a high similarity to wtITRs. Given the limited understanding of structural motifs in ITRs on these processes, my designs for the proof-of-concept ITR library (see Figure 8) and the alternative ITR sequence library with variants Y1-Y17 (see Figure 34) were not guided by a specific design rationale. Instead, variants with increasing difference from the wtITR2 structure, such as inserted loops or shortened hairpin arms, were included along with hairpin termini of other parvoviruses as well as DNA structures that may incorporate functions, such as DNA aptamers or G4 sequences.

Although it permits a relatively easy exchange of the hairpin sequence, especially in comparison to conventional vector plasmids with two ITRs, the sITR plasmid can still be substantially improved. For instance, there is currently no strategy to control the hairpin insertion direction during cloning. Controlling the hairpin orientation during cloning would simplify the identification of hairpin variants, especially in high-throughput sequencing data. With genome replication from sITR plasmids assumed to be bidirectional[256], vector genomes and titers would be indistinguishable regardless of the hairpin orientation.

The low plasmid yield is another major challenge with sITR plasmids that requires improvement, as it limits the scale of vector production and thus the ability to screen ITR variants in larger animals. In line with the literature [387], switching from LB to TB medium and increasing the culture volumes enhanced the yield. Even this optimization did not suffice to achieve the plasmid input concentrations required for commercial ITR-sequencing (200-300 ng/µl) when using small-scale plasmid preparation protocols. Optimized ori sequences [388] and plasmid size could further improve sITR plasmid yield. The here employed 3.3 kb sITR plasmid is sufficiently big to allow for the packaging of intact vector genomes, as verified by both PacBio (see Figure 13) and Nanopore sequencing (data not shown), which is in line with the literature [389]. Still, an increase in plasmid size by 1.5 kb would generate vector genomes matching the wtAAV genome size while simultaneously improving plasmid yields and potentially even ITR stability. Size increases above this limit would not be compatible with rAAV production[123], exceeding the packaging capacity of AAV particles. An associated issue of the sITR plasmid design is the presence of the prokaryotic plasmid sequence in the vector genome which also limits the size available for the transgene. Employing minicircle technology (see section 1.3.3; Figure 5D) it would be possible to remove the prokaryotic sequences from the sITR plasmid product after purification by recombination [262]. This would allow for the use of sITR plasmids with sizes above the packaging limit of rAAV by subsequent removal of the prokaryotic sequences. The recombined minicircle sITR plasmid, which lacks prokaryotic sequences and does not exceed the rAAV packaging limit [123], could then be used to produce rAAV vectors free of prokaryotic sequences. With this approach, once improved ITR variants are identified, it will not be necessary to transfer the alternative ITR sequence to a conventional vector plasmid for therapeutic rAAV production.

The other major limitation of the sITR plasmid is the stability of the elongated ITR hairpin sequence, as was also visible by the number of gaps in the Nanopore sequencing data (Figure 17). The common occurrence of ITR-loss was even more noticeable during cloning, with 10-20% of clones not containing the ITR, and higher loss-rates after plasmid retransformation (data not shown). To foster ITR integrity, I cultured bacteria carrying sITR plasmids at reduced temperatures between 33°C to 35°C, which is a common strategy in the field. In contrast, though, a recent publication suggested that ITR stability increases at higher temperatures for specific *E. coli* strains and declines at lower temperatures [254]. Although this effect may be strain-dependent, any strategy or bacterial strain that aids the maintenance of instable plasmids [390] could further improve the integrity of sITR plasmids and thus facilitate

larger ITR screens. An alternative strategy to improve the stability of the ITRs may be to shorten the sITR hairpin, which could be achieved by using a sITR plasmid without a paired D-sequence leading to a 125 bp hairpin sequence. Such adapted sITR plasmids were previously shown to generate rAAV vectors with single-strand polarity genomes [93,94] and would still enable the use of the cloning strategy established in this work to efficiently generate hairpin variants. Importantly, since single-strand polarity rAAV do not negatively affect transduction efficiency [93,94], these vectors may be similarly suitable to examine the effects of ITR variants as conventional dual-polarity AAV genomes. While the handling of sITR plasmids has challenging aspects, mainly plasmid yield and ITR stability, the fast, simple, and efficient cloning procedures make it the ideal plasmid for ITR engineering. With the here discussed optimization strategies the two main bottlenecks, plasmid yield and ITR integrity, could be overcome, further boosting the applicability of ITR variant screenings.

4.1.2 T7eI-sequencing – a new method for plasmid ITR integrity verification

For conventional vector plasmids, the plasmid ITR quality control for the production of rAAV vectors for research purposes is usually limited to a restriction digest with XmaI and the analysis of the cleavage fragments [391]. However, when examining ITR sequence modifications, sequencing of the ITR is unavoidable. The previously established approach of Sanger sequencing after a hairpin linearization with restriction enzymes [357] cannot be used with synthetic hairpins that may lack such restriction sites. This necessitates the use of alternative approaches or specialized commercial services.

The commercial ITR-sequencing service requires high input concentrations and large quantities of plasmid material, both of which were difficult to achieve given the low yields of sITR plasmids. Additionally, the service is expensive and has a long turnaround time. Therefore, I needed an alternative method to confirm the ITR sequences of the ITR variant libraries. Initially, I employed the previously described approach to sequence ITRs after linearization with restriction enzymes [357]. Interestingly, an incomplete digestion of the wtITR2 with its two XmaI cleavage sites, led to an overlap of the two opposite Sanger sequencing reads. Inspired by this, I examined the possibility of using enzymes with structure-specific cleavage activity that could similarly generate multiple fragments without relying on the hairpin sequence for the hairpin cleavage. This led to the identification of T7eI, which enabled the resolution of the ITR structure and subsequent Sanger sequencing. This proved pivotal for this work, enabling the sequence confirmation of plasmids with diverse sets of hairpin variants.

The ITR resolution mechanism for ITR resolution may be based on the mechanism enabling T7eI to resolve Holliday-junctions [392], as hypothesized in a very recent study describing T7eI cleavage of ITRs [393]. The combination of this hairpin resolution approach in combination with Sanger sequencing has not yet been described.

The conventional ITRs in pSSV9/pSub201 vector plasmids were readily sequenced using the T7eI-ITR sequencing approach (see Figure 9C), but some older vector plasmids still contain guanine-polynucleotide repeats adjacent to the ITR [393]. In these plasmids, the T7eI-ITR sequencing approach was unable to resolve the complete ITR due to dephasing in the repeat sequence, leading to overlapping peaks and premature read termination (data not shown). For such rare cases, the commercial ITR sequencing service still remains useful. Additionally, for some ITR variants, especially those with very little secondary structures at the hairpin tip, sequencing with T7eI-ITR sequencing could also not be reliably achieved. Such structures are relevant, as a similar ITR structure can be found in the published B-B' C-C' deletion ITR, for which a higher transgene expression upon transduction was described [292]. A few low-quality sequencing reads that span into the ITR of such plasmids were indeed observed (Supplementary Figure 2), but in most cases the reads terminated prematurely. Thus, for ITR structures where cleavage by T7eI is very inefficient due to the lack of structural elements at the hairpin tips, alternative sequencing approaches remain important. The minimal secondary structure that allowed reliable T7eI-ITR sequencing appeared to be a loop at the tip of the hairpin (compare Figures 10C and 11D), although even there, the peak heights were lower than for other ITR structures. To improve sequencing results with such ITR structures, it may be possible to improve hairpin cleavage efficiency by engineering the T7eI enzyme, although the commercially available T7eI enzyme was already sufficient for most tested hairpin structures.

Especially long sITR structures with a reasonably large secondary structure element at the hairpin tip, *i.e.*, the AAV T-shaped structure, could be sequenced with better read quality than what was achieved using a commercial ITR-sequencing service (Figure 11A). However, the occurrence of artifact mutations in the reads was also detected regularly, especially near the hinge region between the arms of the T-shaped hairpins (Figure 11B and D). The origin of this artifact that coincided with a drop in read quality and the occurrence of alternative nucleotides could not be determined. In most cases, the opposite read contained the correct nucleotide with higher read quality thus still enabling the determination of a consensus ITR sequence. Due to the regular occurrence of this artifact, a strategy was developed that did not exclude plasmids with such apparent mutations (Figure 37) but included the read quality at the mutation site as

well as the sequence and read quality of the opposing read. In some cases, such mutations could be detected in both reads at different positions, well exemplified with the *S. fusca* hairpin variant (Figure 11E). Based on the decision tree, it was assumed that the sequence was correct. This assumption was later proven to be correct when the vector generated with the sITR plasmid was found to contain an intact hairpin sequence (Figure 34D).

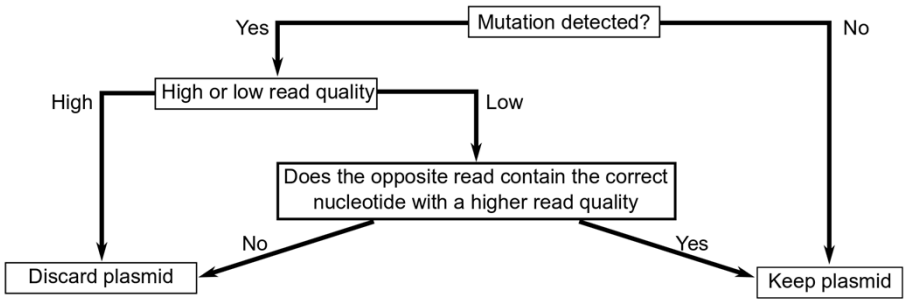


Figure 37: Decision tree for analyzing T7eI-ITR sequencing results. This strategy was used to analyze T7eI-ITR sequencing reads. The presence of mutations in a read did not directly result in discarding the plasmid. Instead, read quality and opposing reads were used to detect potential artifact mutations.

In general, the hinge regions within the hairpin were most susceptible to artifacts usually presenting as apparent point mutations. The drop in peak height and read quality was usually followed by a return to normal peak height and read quality, further complicating any conclusion about the origin of this artifact. Another, less common artifact was the read inversion, which likely resulted from template switching during the chain-termination PCR (Supplementary Figure 1). Since the artifact sequence was easily identifiable due to a short gap in the alignment and low peak heights, it was excluded from the sequencing read as it does not represent a feature of the plasmid DNA.

Despite these minor issues, T7eI-ITR sequencing remained a straightforward and reliable approach for nearly all tested hairpin variants and was foundational for the screening of the large ITR variant libraries in this work. The confirmation of ITR sequence integrity using Nanopore sequencing in 55 sITR plasmids that were previously examined with T7eI-ITR sequencing (see section 3.1.5.4) further boosted the confidence in this approach. While the read depth of the different sITR plasmids varied widely, the sITR were well resolved even at low-read coverage. This suggests that a limited number of Nanopore sequencing reads might be sufficient to confirm sITR plasmid sequences. Thus, commercial Nanopore-based whole-plasmid sequencing represents a reliable competitor to T7eI-ITR sequencing. Nanopore-based plasmid sequencing has the advantage of confirming the sequence integrity of the complete plasmid sequence, while T7eI-ITR sequencing is limited to the ITR sequence and the

surrounding ~200 nucleotides. Still, the cost efficiency and rapid processing of T7eI surpass the Nanopore sequencing approach. This is especially relevant since about 10-20% of the plasmids lose the ITRs during cloning, often requiring the screening of multiple plasmids.

In the future, a combination of T7eI-ITR sequencing and whole-plasmid sequencing may therefore be the ideal strategy for sequence confirmation of sITR plasmids. The presence of the ITR and the barcode sequence can be confirmed with a single T7e-ITR sequencing read. In a subsequent whole-plasmid sequencing reaction, it can be ensured that the complete sITR plasmid sequence is intact. Employing T7eI-ITR sequencing for pre-screening could significantly reduce the number of more costly whole-plasmid sequencing reactions. Considering the higher turn-around time for whole-plasmid sequencing, such a pre-selection would also speed up the ITR verification process compared to iterative screening relying solely on whole-plasmid sequencing.

4.1.3 Third-generation sequencing of ITRs in plasmids and vector genomes

Both PacBio and Nanopore sequencing have been established for rAAV genome or plasmid sequence confirmation [255,363,378,394]. While the analysis of ITR integrity in vector plasmids was limited to conventional wtITR2 [254], an analysis of ITR variant vectors has already been performed using Nanopore sequencing, albeit it was mostly focused on genome integrity and contaminating sequences.

Therefore, this work represents the first instance in which PacBio and Nanopore sequencing were used to specifically confirm the hairpin sequence of variant ITRs in both plasmid and rAAV vectors. Despite its successful use for ITR sequence confirmation, Nanopore sequencing has a distinct disadvantage in resolving the ITR region compared to PacBio sequencing. The ITR region is characterized by a high GC content of nearly 70%, which poses a significant challenge for Nanopore sequencing [395]. This complicated the Nanopore sequencing analysis of the ITRs, especially in plasmid DNA, as will be discussed in the following section. Despite the challenging sequencing template, a novel ITR repair mechanism could be identified along with a strategy to circumvent it, which will be discussed in more detail in section 4.1.3.2. This will be complemented by a comment on the potential impact of sITR plasmids on the hairpin configuration in vector genomes in section 4.1.3.3.

4.1.3.1 Nanopore sequencing for ITR integrity analysis in plasmids

Nanopore sequencing of sITR plasmids not only served as a complementary method to T7eI-ITR sequencing but also enabled the determination of the fraction of plasmids in which the ITR is lost during propagation. First, this chapter details the sequence confirmation of sITR plasmids hairpins, addressing the sequencing artifacts that complicated sequence validation. This is followed by an analysis of the ITR loss frequencies in sITR plasmids and a comparison with the known rates in conventional vector plasmids.

The consensus sequences derived from Nanopore reads matched the intended ITR sequences in the sITR plasmids that were previously confirmed by T7eI-ITR sequencing (Figure 18). The primary challenge in sequence confirmation arose from high error rates in the sequencing reads of the ITR region, attributed to Nanopore-specific artifacts and alignment ambiguities. The B-B' and C-C' hairpin region of the ITR was the main focus of this work but its sequence, containing a total of eight trinucleotide repeats and a GC content of 80% in the wtITR2, makes it particularly prone to sequencing errors [395]. Nevertheless, the alignment of most plasmid ITR variants was surprisingly clear. For others, such as the hairpin variants V3 (Figure 18B) and V37 (Figure 18F), the number of alternative nucleotides was more substantial, especially at the mutated position. These alternative nucleotides could be interpreted as contamination with defective or alternative ITR plasmid sequences. While such contaminations cannot be finally ruled out, subsequent separate vector productions indicated a proper ITR integrity (Figure 19B and 19F) and thus no major contamination problem. Instead, the issue may lie in the alignment to the flip/flop-consensus reference sequence with several ambiguous nucleotides. This effect was highlighted during the alignment of sequencing reads from a wtITR2 rAAV genome to a reference sequence with a single ambiguous nucleotide insertion (Figure 21B). There, the alignment yielded spurious nucleotide insertions at the ambiguous position, with a substantial number of aligned reads containing a nucleotide absent from the ITR variant pool. The additional nucleotide, likely derived from a shift in the alignment, suggests the future use of exactly matching plasmid reference sequences for a more accurate alignment result. However, such an approach was complicated in my work by the current bidirectional cloning strategy used for hairpin insertion into the sITR plasmids, leading to two different possible ITR orientations in the plasmid. A redesigned unidirectional cloning method could eliminate the need for an ambiguous plasmid reference, thereby improving the alignment results for future ITR screens.

An unexpected sITR plasmid-specific artifact was observed in the sequencing reads covering the region between the trs and RBE, leading to a large number of deletions. These deletions were consistently found in the 3' region of the sequencing read mapping to the ITR (Supplementary Figure 3), suggesting strand specificity and thus implying an artifact. For genomic inverted repeats, an artifact has been described that leads to similar errors [396]. There, once the first half of a palindromic region passes through the Nanopore sequencing pore, hairpin formation is initiated. This, in turn, generates sufficient tension on the DNA to actively pull it through the sequencing pore, resulting in erroneous basecalling and false deletions [396]. Although previously only observed in larger genomic inverted repeat sequences, the same mechanism could explain the strand-specific accumulation of deletions in the reads covering the 165 nt sITR plasmid. Notably, such an artifact was not specifically described for the 95 nt conventional plasmid ITR during Nanopore sequencing [254].

To quantify the fraction of plasmids with partial or complete ITR sequence loss despite ITR confirmation using T7eI-ITR sequencing, the reads were aligned to the plasmid reference using a splice-aware aligner to improve gap recognition, even in presence of short overhang sequences adjacent to the ITRs. In a recent study analyzing the ITR loss in conventional ITR plasmids, the same aligner was used without using the splice-sensitive alignment option, suggesting this was not essential [254]. When plotting the cumulative read gap lengths, *i.e.*, the missing regions of the reads spanning the ITR, most reads contained between 15-20 deletions (Figure 17A), whereas plasmid regions of similar size without hairpin sequences contained only around five gaps (data not shown). For conventional 95 nt hairpin ITRs in vector plasmids, hairpin lengths in the reads typically varied by about ten nucleotides [254]. Considering the use of a 1.7-fold longer ITR hairpin in my data, this matches surprisingly well [396]. Subsequently, I used the maximum consecutive gap length in the alignment to classify the integrity of the ITRs. Using this measure should minimize misclassification due to short random gaps in the reads. I then suggested that all plasmids with less than ten consecutive deletions in the reads may represent an intact ITR, as this cut-off would be sufficient to exclude common ITR mutations, such as the deletion of the 21 nt B- or C-hairpin [255]. Considering the template strand-dependent artifact between RBE and trs, which commonly led to deletions between 5-20 nt (Supplementary Figure 3), this stringent cut-off likely overestimated ITR damage. In turn, the estimate of an average sITR integrity of 60-70% (Figure 17B) likely underestimates the actual value. Notably, literature describes an ITR integrity of 60-70% for conventional 95 nt ITRs in the commonly used cloning strain *E. coli* DH5α, albeit after prolonged incubation times [254].

Analysis of defective ITRs, which in light of the high background error rate may be more informative, indicated that a complete loss of the 165 nt ITR was rare, while a partial ITR deletion comprising 155 nt was far more common (Figure 17A). This resembled the observation in conventional ITR plasmids, where the loss of parts of the 95 nt hairpin was more frequently observed than a loss of the complete hairpin [254]. Interestingly, a complete deletion of the 130 bp ITR sequence, far exceeding the hairpin size, was also detected [254]. In my data, such extended deletions would be detected as a 167 bp deletion, which only made up a negligible fraction of the reads.

To simplify the analysis, a complete loss of ITR function was defined as a plasmid with a loss of more than 100 consecutive nucleotides, as the presence of one ITR stem with RBE and trs (61 nt) would still enable replication and packaging of genomes (see section 3.2). Such large deletions were rather infrequent and only detected in 1-2% of the plasmids, representing a less than ten-fold increase compared to the ITR loss in conventional vector plasmids under optimized culture conditions [254]. This would suggest that the sITR may be more stable than initially anticipated, based on the frequent complete loss of ITRs during sITR plasmid re-transformation. Concurrently, this also hints at a major growth advantage of plasmids upon the loss of the ITR. For future, more precise determination of the ITR integrity in sITR plasmids, the use of more accurate sequencing approaches may be indicated, such as PacBio sequencing.

4.1.3.2 ITR sequence integrity in rAAV vectors and the novel ITR *trans*-repair mechanism

ITR sequence integrity in rAAV vectors was first analyzed through a collaboration with the group of Phillip Tai and Guangping Gao (UMMS), using PacBio sequencing of vectors produced from plasmid pools containing diverse ITR hairpins and wtITR hairpins from different serotypes (Figures 13-15). Unexpectedly, the alternative ITR sequences were not maintained in the vector genomes. Instead, hairpin sequences were introduced that were not initially associated with a specific barcode sequence. These apparent hairpin substitutions mostly converted the original hairpin to the wtITR2 hairpin, albeit some alternative hairpins that were also part of the plasmid pool could be detected as well (Figures 14 and 15). This hinted at a novel, previously undescribed ITR repair mechanism during vector production. Notably, for PacBio adapter ligation, an end-repair step is employed on the extracted vector genomes that likely anneal to form duplex vector genomes. After annealing of a vector genome

that has partially lost its ITR with an intact vector genome, such an end-repair step could theoretically fill the missing nucleotides using the intact ITR from the complementary vector.

To exclude that the observed repair was based on such a potential artifact, a Nanopore sequencing protocol was used that does not rely on end-repair but instead randomly inserts the Nanopore sequencing adapters via a transposase. Using this technique made it possible to independently confirm the presence of a repair event during vector productions using a plasmid pool (see Figures 19-22). Importantly, separate production prevented the ITR repair in point mutant variants (Figure 19; Supplementary Figure 4). The analysis of ITR sequence integrity in rAAV genomes was also complicated by the previously discussed low read quality due to the high GC content in the ITR and the alignment to a flip/flop consensus reference sequence (see section 4.1.3.1). Despite these challenges, it was possible to verify that the separate production strategy maintained highly divergent ITR hairpin sequences in the vector genomes (Figure 34B-F).

The repair during pooled production was not limited to the hairpin but could also be observed throughout the whole ITR including the hairpin stem (D- and A–sequence; Figures 21-22). ITR variants with high similarity to the wtITR2 showed lower repair rates compared to non-conventional hairpin sequences, which were additionally outcompeted during replication by the wtITR-like sequences (Figures 14, 15 and 20). Independent of their location in the ITR, the conversion rate of point mutations back to the wtITR2 sequence was relatively uniform, leading to conversions in 20-30% of the genomes. Considering that most of these point mutant hairpins only have a minor impact on the replicative ability of the ITR (Figure 24A-B), this value might represent a background repair rate. Interestingly, this background repair rate was quite consistent throughout the experiments despite the different ratios of the variants in the plasmid pool, such as with the hairpin point mutants at a 1:55 ratio (Figure 19) or a hairpin and stem mutants at a 1:4 ratio (Figures 20A, 21B and 22B).

The link between repair rate and replication efficiency was especially visible during the pooled production of the wtITR2 with three different point mutant ITR variants (Figure 20A). There, repair towards the wtITR hairpin nucleotides was consistently increased compared to the conversion to alternative hairpin sequences, albeit the effect was small. Similarly, the replication efficiency likely also affected the pooled production of a wtITR2 in presence of the *Ae*DV hairpin (Figure 20C), where the conversion of the wtITR2 towards the densoviral hairpin was only detectable due to the substantial sequencing depth (Figure 20C). This likely indicates that the initial repair direction was random but the subsequent increased replication

competence of the "repaired" ITR then led to the overrepresentation of the wtITR2 hairpin. All these observations strongly support the hypothesis that the hairpin conversion was not the result of random modification or a Rep-mediated directed mutagenesis but was mediated by the interaction of genomes with different hairpin sequences.

Therefore, I suggest that this phenomenon represents a novel type of ITR repair mechanism, complementing the previously described *cis*-repair [270]. The assumed mechanism for *cis*-repair, shown in Figure 38A, would be unable to convert the hairpin of two different vector genomes with different barcodes. The exact ITR *trans*-repair mechanism remains unclear, but two possible scenarios are illustrated in Figure 38B. A recombination-based mechanism, in which the two vector genomes anneal and subsequently recombine, would be supported by previous descriptions of severe recombination events during rAAV production, even resulting in the generation of wt-like AAV genomes [397,398]. However, a whole rAAV genome-spanning repair mechanism would also likely lead to the conversion of barcode sequences. Considering the 20-30% conversion rate in ITR point mutant variants with negligible effect on vector replication, a barcode swap at similar rates would have been clearly detectable during vector production with wtITR2 and densoviral hairpin (Figure 21C). Thus, the alternative proposed mechanism relies on the annealing of two vector genomes with the interaction restricted to the ITRs, resembling the mechanism assumed for *cis*-repair. Future research will need to further dissect this *trans*-repair mechanism to examine whether the repair is restricted to the ITRs or whether it is based on the annealing of the complete vector genome.

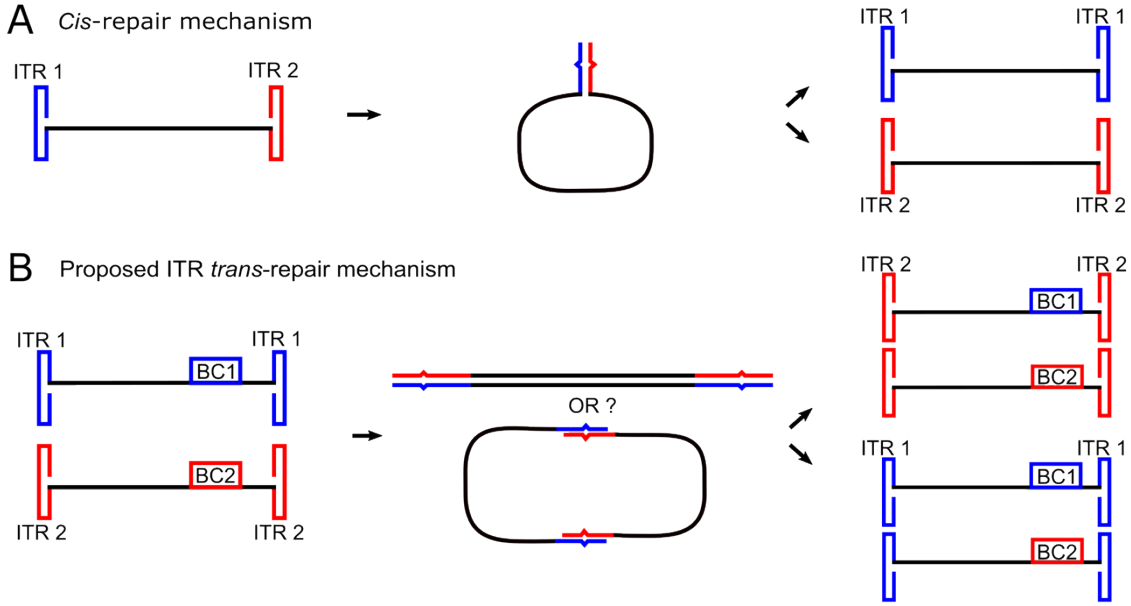


Figure 38: Comparison of the proposed mechanisms for ITR repair during vector prodcution. **(A)** The exact mechanism for ITR *cis*-repair has not been elucidated. The proposed model suggests that in the presence of one alternative ITR, the repair of the ITR takes place by forming a panhandle shape

with subsequent mismatch repair. While never proven, it may be reasonable to speculate that *cis*-repair could lead to the generation of vector genomes with either of the ITRs. **(B)** The novel *trans*-repair mechanism relies on the presence of two different ITRs in two distinct vector genomes, here indicated by two barcodes (BC1 and BC2). The exact mode in which the ITR conversion takes place remains unresolved. The genomes may either anneal completely and a recombination-like mechanism can mediate the exchange of the ITR sequences. Alternatively, the two distinct genomes interact solely via their ITR sequences and a repair mechanism, similar to the one hypothesized for *cis*-repair, can convert the ITR sequences. This repair mechanism can finally lead to the presence of vector genomes with both ITRs being associated with both barcodes.

To determine whether the *trans*-repair phenomenon is based on whole-genome recombination or is restricted to the ITR sequences, I propose using a vector genome containing multiple distant barcode sequences. This could lead to barcode pattern shuffling during vector production if the whole genome is affected by this mechanism. Additionally, completely different vector genome sequences could be used to investigate whether the repair mechanism persists despite non-complementary vector genomes. Furthermore, I have only examined the *trans*-repair mechanism in vector productions using sITR plasmids. Future work should also assess this mechanism during rAAV production with conventional ITR vector plasmids.

Assuming a recombination-based mechanism relies on the presence of vector genomes of both polarities, the *trans*-repair mechanism should be absent in single-polarity parvoviruses. Interestingly, for MVM, the wtITR was already compared with a less replicative ITR variant in a pooled production [109]. There, the less replicative ITR variant was still well detectable even during competition with the wtITR variant, albeit at lower levels, which may hint at the absence of *trans*-repair [109]. Thus, the use of single-polarity rAAV genomes [93] may also be able to reduce *trans*-repair during rAAV production.

The *trans*-repair mechanism could further be harnessed to study unknown processes during AAV production, such as the presence of a possible retro-transduction of rAAV during vector production, a well-described phenomenon for other viral vectors [399]. When rAAVs are released from cells during production, they could transduce other producer cells which, in the presence of helper genes, could further amplify the vector genome and increase productivity. Although such a retro-transduction mechanism has not been explicitly described for rAAV production, it has been reported that genes associated with cell entry improve vector productivity [400]. With regards to ITR variants, a retro-transduction mechanism could provide templates for the ITR *trans*-repair, and thus indirectly confirm the presence of this phenomenon. It may even be the main driver of the ITR *trans*-repair by delivering repair templates to the cells, explaining the constant repair rates, independent of the plasmid pool ratios, during vector production.

The *trans*-repair mechanism severely complicates ITR engineering and the generation of large ITR variant libraries, but the mechanism itself may be fundamental for ensuring ITR integrity during production of therapeutic rAAV. Recently, ITR integrity in commonly used conventional vector plasmid preparations was estimated to be around 90% [254], which in absence of *cis*- and *trans*-repair would likely lead to substantial ITR homogeneity in the rAAV. However, in presence of these repair mechanisms, the damaged ITRs are not only outcompeted but also converted by *trans*-repair. The *Ae*DV hairpin may represent a suitable substitute for a damaged ITR hairpin, as it possesses a similar T-shaped hairpin as AAV. When mixed during production with the wtITR2 plasmid at a 1:1 ratio, around ten-fold fewer reads with the *Ae*DV hairpin would be expected in absence of the *trans*-repair mechanism and competition, which was indicated in a separate production (Y10, Figure 34A) and in line with previous descriptions of the replication competence of damaged ITRs [292]. Due to the competition and *trans*-repair, the previously highly abundant alternative plasmid hairpin could only be detected in 0.1% of rAAV genomes (Figure 20C). This explains why whole-vector genome sequencing approaches never identified ITR integrity as a major problem during rAAV production [363,389,394,398,401], despite unavoidable random plasmid ITR truncations [254,255]. However, if less replicative ITR variants were used for vector production, this issue could arise again and could either require engineered Rep proteins [402], optimized for efficient replication of the used ITR variant, or additional quality control steps ensuring ITR integrity during manufacturing. With both Nanopore and PacBio already being implemented as tools for rAAV vector quality control for DNA contaminants and vector genome integrity [363,389], this would only require an extension of the current analysis pipelines to cover the ITR sequences. Thus, in case alternative ITR hairpins were used for therapeutic vector production, a revised quality control procedure could be quickly implemented.

4.1.3.3 ITR configuration in rAAVs generated with sITR plasmids

The ligation-based protocol used for PacBio sequencing enabled the simultaneous analysis of the configuration of both ITR sequences in the vector genome. Rather than the expected 1:1:1:1 ratio of flip/flip, flip/flop, flop/flip and flop/flop configuration [363], a skewed distribution was detected (Figure 13C-D). While it cannot be excluded that ITR modifications affect the replication of ITRs in a specific configuration, the hairpin transfer model for replication should limit such effects to a minimum. Due to the use of the transposase-based adapter insertion for Nanopore-sequencing in separate productions, I could not simultaneously analyze the

configuration of both ITRs in later experiments as the reads did not cover both ITRs (Figure 16C). In the separately analyzed ITRs, the expected 1:1 ratio of flip and flop hairpins was maintained in all hairpins, except for those with very low sequencing depth and alternative hairpin structures with weak secondary structure (Figure 34C and F). Thus, I can only speculate about the origin of the skewed distribution of ITR configurations in the PacBio data. One probable explanation are artifacts based on *trans*-repair events, as the sequencing has been performed with vectors generated by pooled ITR variant productions, which would lead to the incorporation of alternative ITR sequences. Considering that the identification of the hairpin configuration is based on the number of sequence matches to the expected flip and flop sequence, the hairpin conversion to alternative ITR sequences could have led to misclassification of flip and flop configuration due to sequence divergence. Plasmid contaminations, while often problematic during rAAV genome sequencing, can be excluded as origin of the skewed distribution, since sITR plasmids cannot generate reads that simultaneously map to both ITRs. To ensure a normal ITR configuration distribution after separate production of ITR variants using sITR plasmids, ligation-based sequencing strategies should be again employed in the future. This would provide clarification regarding the presence of an altered hairpin configuration in rAAV generated using sITR plasmids. Therefore, it could guarantee that the effects of rAAV generated with sITR plasmids on transduction, integration, and episome formation are not based on aberrant ITR configurations.

4.1.4. Screening of engineered ITRs from vector production to transduction

The incorporation of DNA barcodes in the vector genome is a long-established approach for the examination of capsid transduction efficiency in pooled capsid screens [190,371], and was thus assumed to be well suited for variant ITR screens. Although pooled transduction of ITR variants could introduce competition for the host's processing machinery, potentially influencing the outcomes, this trade-off was accepted to enable higher throughput and potentially even enhance functional differences. Similarly, during parallel capsid screening competition and synergistic effects may arise, yet this approach has yielded many potent and highly relevant capsid candidates for gene therapy [190,403-405].

In this work, barcoding of the vector genome was essential, providing critical evidence for the *trans*-repair mechanism during pooled productions. Furthermore, barcoding of ITR sequences was especially critical to trace the ITR variants post-transduction, as portions of the ITR sequence are lost during vector genome integration [327,328] and episome formation [406] rendering

the identification of the original ITR sequence impossible. Once it was confirmed that sequence modifications in the ITRs were preserved in rAAV genomes from separate productions (Figure 19 and Supplementary Figure 4), the barcode sequence enabled the assessment of the impact of the ITR variants during production, transduction, episome formation and integration, which will be discussed in the following sections.

4.1.4.1 Engineered ITRs and their impact on vector production

The barcode distribution after production was regarded as a measure for vector productivity of the associated ITR variant. During the initial pooled vector productions, a dominance of wtITR2 was observed, even in the presence of closely related ITR variants that all showed a substantially lower representation (Figures 8B and 12B). This was likely the result of a competitive disadvantage of the variant ITRs compared to wtITR2. Although the *trans*-repair mechanism would convert a fraction of the variants towards wtITR2, this appeared to be insufficient to rescue the reduced productivity of the barcoded genomes with alternative hairpins.

Competition and *trans*-repair could only be circumvented by separately producing ITR variants and pooling them prior to purification. Under these circumstances, several ITR variants with high similarity to the wtITR2 even appeared to increase productivity (Figure 24A-B), particularly those with mutations in the B-sequence. This initially suggested a mechanistic advantage; however, a direct comparison of wtITR2 and the B-sequence mutant variant V81 indicated equal titers (Figure 24D). While V81 was only one of the multiple variants that showed an elevated productivity and may not be representative for all, this points to additional factors, such as transfection efficiency and plasmid quality as being responsible for the observed productivity differences during the screen. In particular, plasmid quality and ITR integrity could be a major determinant, as all ITR variant libraries were generated from a single plasmid DNA preparation. Triplicate plasmid preparations for all 90 ITR variants could have minimized effects from varying plasmid quality but were deemed unfeasible due to the extensive screening efforts required for ITR sequence verification and the limited impact post-transduction. Already the direct comparison of wtITR2 and variant V81 presented a major challenge. A third replicate had originally been planned but was ultimately abandoned after multiple mid-scale plasmid preparations failed to recover intact ITR sequences from re-transformed clones and since the first two productions showed no indication of a different productivity. The loss of ITRs in the vector plasmids could also have led to the effects on

productivity (Figure 24A), but a comparison with the Nanopore sequencing data on sITR plasmid integrity (Figure 17B) revealed no correlation between ITR integrity and vector yield within the ITR variant library screen. This was best exemplified by variant V37 which had the lowest ITR integrity among the sITR plasmids in the library but produced relatively well. Conversely, variant V31 was one of the least productive variants, despite the highest observed plasmid ITR integrity. Thus, a plethora of other factors may have also contributed to the observed productivity ratios in the library screens, including salt or residual organic solvents contaminations from plasmid preparation, both of which can affect transfection efficiency. Moreover, the apparent increase in productivity for ITRs with B-sequence mutation could have also been influenced by procedural variables. Transfection mixes were prepared sequentially by variant number, thus, longer incubation times for later variants could have inadvertently boosted transfection efficiency [407] and productivity.

The separate production of rAAV with highly divergent ITR sequences (Figure 34A) further highlighted the importance of the hairpin secondary structure for vector productivity. Prior studies have suggested that any deviation from the wtITR structure is accompanied by a strong reduction in productivity [239,292]. A similar trend was observed during production of the library (Figure 34A), where sequence and structural similarity to wtAAV2 correlated with higher productivity. Notably, the TA-rich hairpin variant Y4, despite its highly divergent sequence, maintained relatively high productivity likely due to a secondary structure resembling the wtITRs. This supports the previous concept of a hairpin structure-dependent vector productivity [239], albeit the ITR sequence is also of relevance [241]. Nonetheless, vector productivity of ITR variants may ultimately be less critical than transduction efficiency, as Rep proteins could be evolved [402] to enhance the replication of an ITR variant with improved transduction characteristics.

To date, only few published ITR variants have been examined in detail regarding their rAAV productivity. Most chimeric ITR variants led to a reduced productivity compared to the wtITR2, especially when the Rep protein was not modified accordingly [357]. For a CpG-depleted ITR variant, a three-fold reduction in yield was determined [241], while the B-B' C-C' hairpin deletion variant exhibited a ten-fold reduction [292]. The notable exception is the SynITR variant, containing a mutated p53 recognition motif, for which an almost four-fold increase in productivity was reported [306]. This result is particularly surprising, as I included the SynITR variant in a pooled production assay where it was strongly outcompeted by the wtITR2 (data not shown). This discrepancy raises the question of whether ITR variants that are outcompeted

in pooled productions, indicating a lower replicative ability, could ultimately demonstrate a higher productivity when assessed independently.

In conclusion, the BC-seq approach to identify variants with increased replication and packaging efficiency is likely functional. However, the observed effects with engineered ITR sequences were likely obfuscated by extrinsic factors including transfection efficiency and plasmid quality. Further work that minimizes the impact of extrinsic factors such as through replicate plasmid preparations, likely requiring improvements to ITR stability in sITR plasmids (see section 4.1.1), could better support the discovery of variants which genuinely improve productivity.

4.1.4.2 Transgene expression from rAAV vectors with engineered ITRs *in vitro* and *in vivo*

The most direct measure of functional transduction is the expression of the transgene in host cells. In capsid engineering, the transgene expression from different variants can be evaluated using barcode tracking [190,371]. In a more recent development, *i.e.*, RNA-driven biopanning, the whole capsid gene can be expressed in the cell and used to recover the original capsid sequence [408]. This allows the identification of the most efficient functionally transducing capsid variants. Analogously, assessing transgene expression from vector genomes harboring different ITR variants by using barcodes located in the 3' UTR should represent a similarly robust strategy for evaluating ITR functionality.

The major problem with the initial two vector productions (Figures 8 and 12) was the lack of ITR integrity (Figures 13 and 14), later identified during discovery of the *trans*-repair mechanism. Thus, while these libraries were extremely useful for the discovery of this previously undescribed ITR repair mechanism as well as for the development and refinement of the ITR barcoding approach, the effects observed post-transduction cannot be linked to any ITR-related process.

After establishment of independent vector productions as prerequisite for ITR sequence maintenance (Figure 19), *in vitro* transduction of cell cultures using the 90 ITR variant libraries revealed a consistently reduced transgene expression compared to the wtITR (Figure 25). The consistent data, especially for the first libraries with variants V1-V54, was rather surprising, due to the high variability of the barcode distribution in the two input libraries (Figure 24). This indicated that the ITRs behave similarly independent of their proportion in the vector library, further increasing confidence in the data and the normalization to the input vector

library via the enrichment score. This independence of the ITR distribution should even enable the screening of low-producing variants concurrent with high-producing variants, provided the sequencing depth is sufficient to reliably cover the barcode distribution. The superior performance of wtITR2 in the cell culture screen may reflect an adaptation of the ITR to enhance transduction and infectivity *in vitro*. Given that AAV2 was originally isolated from a contaminated simian adenovirus preparation [16], it is plausible that repeated passaging in cell culture unintentionally selected for an ITR with enhanced cell culture performance. This adaptation to cell culture propagation has previously been hypothesized to be the origin of the high affinity of the AAV2 capsid to heparan sulfate proteoglycans, despite being detrimental to transduction of human hepatocytes [409].

The assumption of an adaptation to *in vitro* conditions motivated the analysis of the behavior of the ITR variants *in vivo*, where the effects may differ substantially. To explore this, the ITR variant libraries packaged in the AAV2 capsid were administered by retro-orbital injection in mice. The use of alternative capsids may have been favorable, as other variants possess a broader tropism as well as a higher efficiency depending on the target cells. However, the ITR variant libraries were initially produced using the most researched combination of AAV2 Rep and VP and were intended to be used for cell culture screens, as the AAV2 capsid is efficient in many cell lines [377]. *In vivo*, the AAV2 capsid is known to be mainly limited to the murine liver with relatively slow uncoating kinetics [410], which explains the late onset of transgene expression (Figure 28B). The detected *mCherry* transgene expression in the thoracic diaphragm therefore likely represents contamination with liver tissue during organ extraction. The transgene expression in the eye may be explained through contamination during retro-orbital injection route. With Luxturna, an AAV2 capsid based drug is even used for ocular gene therapy after subretinal injection (Table 1), albeit a systemic delivery would not be expected to mediate transgene expression. This also explains the selective expression in the selected eye samples of mice, likely the eyes near the injection site. The elevated levels of vector genome content in the spleen combined with low expression (Figure 28A-B) suggest an incomplete transduction by AAV2 but is likely not linked to the slow uncoating kinetics[410]. Not only the capsid was a limitation of the experiment but also the use of female mice, which are known for reduced rAAV-mediated liver expression [411]. Finally, a mistakenly low vector dose due to a titration error additionally affected the efficiency of transgene delivery. Nevertheless, *mCherry* expression could be detected in liver tissue at all time points, and the resulting enrichment scores confirmed the hypothesis that ITRs behave differently *in vivo*. Notably, the wtITR did

not outclass all other hairpin variants, since several ITR variants performed comparably or potentially slightly better. This may indicate, that further changes to the ITR may be able to improve transgene expression in the future. However, due to considerable data variability in the liver tissue in this screen, likely a result of the low dose and the resulting low expression levels, a definitive ranking of ITR performance could not be established. Interestingly, the overall distribution of the variant ITRs was rather similar to the one observed *in vitro*. Variants that performed well *in vitro* also showed higher transgene expression *in vitro*, suggesting that cell culture screens retain a predictive value and may be useful for an initial ITR variant pre-screening.

Time-dependent effects in the animals were minimal, apart from variant V61 which may represent an outlier, suggesting that most ITR variants have little influence on the kinetics of transgene expression. It may also be conceivable that the low dose has reduced competition between the variants, considering transduction levels of ~1 vg/dg after 60 days (Figure 28A). Thus, in the future it will be necessary to determine whether ITR competition during transduction could be responsible for the difference between the *in vivo* and *in vitro* data. For this purpose, a second *in vivo* experiment has already been initiated, in which the animals were injected with a high dose of a barcoded 25 variant ITR library packaged in capsid AAV9-P1 for muscle-specific delivery with associated liver-detargeting [190] as well as in AAV8 owing to its broad tropism and high liver transduction [412].

The detection of transgene expression from an ITR variant library both *in vitro* and *in vivo* is a novel approach to test ITR functionality, as the examination of ITR variants has so far only been conducted in separate assays. The recently published SynITR variant demonstrated similar functionality to wtITR2 in Hek293T and HuH7 cells [306], the same cells where the ITR variants in this work showed weaker performance. This either suggests that the SynITR is an exceptional ITR variant that is not weaker during transduction in cell culture than the wtITR2, or that a separate examination of ITRs differently affects transduction, potentially due to competition with other ITR variants in the library. Therefore, a separate transduction of selected ITR variants should also be evaluated, to confirm the effects observed in the library screen. Interestingly, in hESCs, the SynITR did not induce transgene expression, despite avoiding the ITR-mediated toxicity, while wtITR2 was able to induce some expression but also had toxic effects [306]. This further supports the idea that ITR variants might indeed have cell-type-specific effects, which was not possible to study in this work due to the limited tropism of the AAV2 capsid, but will be investigated in the upcoming *in vivo* screen. The study

analyzing the function of the CpG-depleted ITR did not indicate a major effect on transgene expression *in vivo*, but this variant was only examined in muscle tissue and not assessed *in vitro* [241]. The ITR variant lacking the B-B′ and C-C′ hairpin structures of the ITRs was reported to increase transgene expression both *in vitro* and *in vivo*, but organ-specific effects were also not assessed [292]. Thus, so far, ITR variants have never been simultaneously examined *in vivo* nor *in vitro* nor have the transduction efficiencies been studied in different tissues. Additionally, except for the B-B' and C-C' hairpin deletion variant that is forming a linear hairpin [292], no other ITR variant has so far demonstrated increased transduction compared to the wtITR2. Notably, a recent study contradicts this by suggesting this ITR variant is a weak transducer in cell culture assays [255]. Interestingly, during the screen of the divergent hairpin sequence library (Y1-Y17), variant Y17, which possesses a similar hairpin structure as the published B-B' and C-C' deletion variant but has a lower GC content, exhibited very low enrichment scores after transduction (Figure 34G). This could suggest that not the structure but rather the nucleotide sequence of the B-B' and C-C' hairpin deletion variant is responsible for the observed improved transduction initially reported [292]. At the same time, variant Y13, a hairpin forming a G- or C-rich loop region (Figure 34F) depending on the ITR configuration, showed relatively high transgene expression in the cell culture screen. Another indication that high GC content in the terminal hairpin region may be favorable for functional transduction, while the secondary structure may be less relevant. In line with this assumption, the TA-rich hairpin Y4, with its high structural similarity to the wtITR2, performed rather poorly during transduction. Considering the high GC content in AAV ITRs, exceeding that observed in other parvoviral ITRs [226], there might indeed be a correlation between GC content and transduction efficiency, which should be explored further in the future. However, among the point mutant libraries (V1-V90), mutations that maintained or increased the GC content could not be identified as better performing variants, thus additional unknown factors may also play a role.

This new ITR screening approach enabling the parallel assessment of ITR variant functionality not only facilitates the discovery of novel ITR variants with improved transduction, but should also be able to reveal potential cell-type-specific effects that have long been suggested for ITR processing [298]. Additionally, by parallel screening of ITR variants, it reduces workload and use of experimental animals compared to the previously used separate screening and may therefore be the method of choice in the future.

4.1.4.3 Screening engineered ITR variants for episome formation *in vitro* and *in vivo*

Episome formation is commonly equated with successful delivery of the vector genome, but the capsid variant screens in the episomal rAAV fraction have not been widely adopted. Thus, with the goal to examine the impact of ITR variants on episome formation, appropriate assays needed to be established. In this work, two complementary methods were employed: Depletion of linear vector genomes using T5 exonuclease and enrichment of circular episomes via RCA. The use of T5 exonuclease to distinguish episomal DNA from linear vector genomes has been described before [376] and has subsequently also been applied in a study involving ITR variant function [306]. In preliminary tests of this method involving a spike-in of vector genomes into gDNA as well as the use of capsid variants with different transduction efficiencies, the depletion of spiked vector genomes and genomes from non-transducing capsids was clearly observable (data not shown). However, in subsequent NGS-based barcode analyses, the method proved suboptimal due to substantial variability in barcode recovery (Figure 12D). Thus, as an alternative, an RCA-based approach was adopted to enrich episomal rAAV DNA, which is commonly used for amplifying circular viral DNA [379,380]. Still, this also did not reduce the data variability during barcode recovery.

One major problem of analyzing the barcode composition to measure the impact of the ITR variants on episome formation may lie in the inherently slow and inefficient process to generate these circular vector genomes. In cell culture experiments, episomes often make up only a minor fraction of the cells' total vector genome content [413], which in turn increases variation during sampling. In the murine liver gDNA samples, a clear amplification of vector genomes by RCA was observed at the 60-day post-injection time point, but not at 14 days (Figure 29C), suggesting low levels of episome formation early after administration. This aligns with previous reports on the slow uncoating kinetics of AAV2 in murine liver, which can extend up to six weeks [410]. Even for capsids with faster uncoating kinetics, such as AAV5, the process of episome formation was described as slow, with episomes accumulating gradually over several months [413]. The cells' vector genome content in the liver at the 60-day time point was approximately 1 vg/dg (Figure 28A), and the RCA input was equivalent to the gDNA content of roughly 10.000 diploid cells. Despite this, the variability in the subsequent BC-seq data was high, suggesting that only a fraction of vector genomes was converted to episomes. Nevertheless, the high sensitivity of RCA, with its ability to amplify sub-picogram levels of circular input DNA [414], still enabled the generation of RCA amplicons from these episomes (Figure 29C).

The variability observed BC-seq after episome enrichment appeared to correlate with transduction efficiency, as low-efficiency capsids yielded even more divergent results (Figure 12C). Similar observations were made in experiments with lower MOIs (data not shown). Consequently, for later cell culture screens (Figure 26B-C) involving the two large ITR variant libraries, the MOI was increased to mitigate this variability, and the screen was limited to two cell lines (HuH7 and Hek293T) in which the AAV2 capsid is known to be efficient [377], with limited impact. For cell culture experiments, the data variability might be reduced by elongated incubation times far exceeding 72 h, which is likely only achievable using cells that divide very slowly, considering that rAAV transduction is more efficient during S phase [415]. Additionally, extended cultivation of common cell lines would dilute the episomes, further complicating their detection. *In vivo*, the low dose as well as the inefficient AAV2 capsid are likely responsible for the high data variability. Thus, future ITR library screens with higher doses and more efficient capsids may be able to improve the recovery of episomal DNA and a subsequent analysis of the impact of ITRs on episome formation. The data for episome formation *in vitro* with the rather small library with divergent ITRs (variants Y1-Y17; Figure 34H) was the least variable, although one outlier value was also detected and the number of replicates was low. Interestingly, there was also an indication of a different episome formation rate between the cell lines for one variant, but the small sample size complicated interpretation. Another potentially relevant observation during the *in vitro* screens, using both T5 exonuclease (Figure 12D) and RCA (Figure 26B-C), was the consistent ranking of the wtITR2 hairpin among the most enriched variants. This could be attributed to a superior integrity of wtITR2-containing genomes, although both PacBio and Nanopore sequencing data did not clearly indicate major differences in genome integrity (data not shown). Therefore, it is possible that the observed enrichment was based on a more efficient episome formation mediated by the wtITR2 hairpin, at least *in vitro*, potentially reflecting the evolutionary adaptation of AAV2 during passaging in cell culture [409], as previously discussed (see section 4.1.4.2).

Each method for episome enrichment, T5 exonuclease depletion of non-circular DNA and RCA enrichment of circular DNA, offers distinct advantages and limitations. RCA has been reported to exhibit a template-length bias, though primarily for very short (<100 nt) circular fragments [416]. Overall, RCA remains a highly sensitive [414], efficient and relatively unbiased method and is thus also used for whole-genome amplification [417]. T5 exonuclease necessitated extended digestion [376], as both unprocessed vector genomes as well as the linear host gDNA need to be degraded by the enzyme. Furthermore, nicks in the episomes that can occur during

purification, more specifically, by reducing agents used during extraction [418], would lead to degradation of the episome. This makes the T5 exonuclease more stringent than other exonucleases that do not affect nicked circular DNA. Notably, the alternative Exonuclease V did not yield satisfactory rAAV genome removal in the initial spike-in tests (data not shown) and was thus not further used, although this enzyme has been used for viral episome enrichment before [413,419]. In the future, the combined application of T5 exonuclease and RCA, as suggested for biosensor platforms [420], could offer an optimal balance of specificity and sensitivity and may therefore be applicable for ITR variant screens in which barcodes in rAAV episomes are examined.

4.1.4.4 Tracing the integration of engineered ITR variants into the host genome

The random integration of rAAV genomes has recently gained increasing attention as it has been described to participate in long-term transgene expression, but also due to its potential risks of genotoxicity [213,216,217]. In humans, there is no report to date associating rAAV gene therapy with tumorigenesis [220], but the risk remains a concern. Engineered ITR sequences may offer a potential strategy to reduce integration events while simultaneously promoting episome formation via vector genome processing pathways. This could yield a next-generation rAAV vector with sustained transgene expression but reduced potential for genotoxic side effects. Alternatively, ITRs that facilitate a more efficient or even targeted integration into specific genomic loci could also be envisioned, ensuring safe long-term expression. However, realizing this goal requires a screening strategy capable of simultaneously determining the integration profile as well as the ITR variant.

While integration profiling of AAVs and rAAVs has previously been achieved using both third-generation and NGS methods[96,326,339,381], so far, screening of ITR or capsid libraries for differences in integration has not been reported. Also, the examination of the integration propensity of ITR variants has so far not been addressed. Therefore, I adapted a published NGS-based integration profiling approach known as ITR-seq [381]. This allowed me to identify not only the barcode, but also the approximate integration site in the host genome. Due to the positioning of the barcode around 250 bp upstream of the ITR terminus and a maximum read length of 250 nt, the exact vector-genome junction was often impossible to determine. Nonetheless, an approximate integration localization is sufficient to assess relative integration frequencies across different ITR variants. Accordingly, reads with ambiguous mapping were

also not removed if they contained a substantial fraction of the vector genome, as they still represent genuine genomic integration events.

To test the approach, two sequencing strategies were employed: one incorporating dephasing adapters and another without them. Integration events were detected in both workflows, and the minimal integration from the 14-day time point from sequencing without dephasing adapters aligned with the slow uncoating kinetics of the AAV2 capsid in murine liver [410]. Although rAAV vector integration has been described as a rapid post-transduction event [339], the elevated number of integration events at the 14-day time point in samples with dephasing adapter is likely an artifact, given the use of a higher initial DNA input during optimization of the DNA fragmentation which was then included during further processing. However, a sample mix-up between samples from different time points during multiplexing for NGS cannot be fully excluded. Additionally, the DNA was not quantified after adapter ligation and purification. This step would have allowed normalization to the DNA input quantity in the first PCR and also enabled an approximate quantification of integration events per host genome, which should be included in future applications of this method [96].

The detected integration sites were broadly distributed across the genome. Intriguingly, several integration hotspots could be observed, including the *albumin* locus in which integrations were found in several animals (data not shown) and which was previously identified as a hotspot for rAAV integration in murine liver [217]. Integration in other previously defined hotspots, such as the *Rian* locus[217], could not be confirmed. Notably, the high transcriptional activity of the *albumin* gene in murine liver [421] and thus its association with open chromatin was likely responsible for the high integration levels. The open chromatin state of the *albumin* locus is consistent with the enrichment of integration in genomic regions associated with open chromatin (enhancer, promoter, exons, and exon-intron junctions) suggested in my data (Figure 32F-G). At the same time, integration in regions associated with closed chromatin (intergenic regions) was reduced compared to a randomized integration dataset of equal size. A similar pattern of rAAV integration has been observed before[206], supporting the general functionality of the adapted ITR-seq. This analysis could be further complemented by comparing the integration sites with assay for transposase-accessible chromatin sequencing (ATAC-seq) data, as has been done before in a publication in which the preference for open chromatin has been described as well [206,339].

The observation of integration into blacklisted regions (Figure 32A-C) is also not a novel phenomenon, as it has been previously described that rAAV integration in the vicinity of

palindromic regions [330], rRNA repeats, LINEs, low-complexity regions, and simple repeats is increased [206]. Especially highly repetitive regions are included in the blacklisted region [368]. Interestingly, a detargeting of rAAV integration from satellite repeats has been described as well [206], which is notable given that most other repetitive sequence elements tend to be favored integration sites. Upon manual inspection of my data, I could not confirm the previously described detargeting of integrations from satellite repeats in murine liver in my dataset. This could rely on species differences, since the published satellite repeat detargeting was described for rAAV transduction and AAV infection in human and primate tissues [206]. However, satellite repeats represent some of the longest and most repetitive sequences in eukaryotic genomes [422], thus, reads will rarely align with high mapq scores and would be excluded by a high alignment stringency. Therefore, it is plausible that the higher alignment stringency (mapq > 30) in the previous study excluded reads that mapped to satellite repeats [206,381]. In contrast, my approach, employing low alignment stringency (mapq > 10) and retaining ambiguously mapped reads if vector genome sequences were detectable, retained such reads. However, this also inevitably led to the inclusion of clearly misaligned integration sites, such as the one mapped to the Y-chromosome (Figure 31D), which is clearly implausible in female mice. Still, since the focus of this work was not set on the precise mapping of the integrated vector genomes, but rather on including all integration sites, especially given the low transduction levels, this was deemed acceptable.

The main aim of the integration analysis was to detect both barcode and integration region, which was achieved successfully. The low vector dose and integration frequency necessitated the pooling of the data from all animals and both time points in order to determine the enrichment score, representing a measure for the integration propensity. This score was based on the barcode distribution in the input library and the integrated genomes, although a normalization to the ITR distribution within the gDNA could also be employed in the future. Otherwise, ITR variants that have a high integration propensity but may be restricted during transduction could be falsely identified as weakly integrating variants. While the limited number of integration events found in this work did not enable definitive conclusions about the effect of different ITR variants (Figure 33B-C), the consistency of the identified integration regions (Figure 32F-G) with published data suggests that the approach is functional and useful for high-throughput ITR variant integration propensity screening. Attempts to perform similar analyses *in vitro* were less successful due to an even lower number of recoverable integration

events (data not shown). Nonetheless, *in vitro* approaches remain attractive for future development given their accessibility, scalability, and the ability to study a human genome.

Despite its potential to determine the integration propensity of ITR variants in a pool, there are still some caveats for both the conventional as well as the adapted ITR-seq. A substantial proportion of reads only mapped to the host genome without containing vector-derived sequences (Figure 31B) and was thus filtered out. If not stringently filtered, these reads could have resulted in false-positive integration calls. The analysis pipeline of the adapted ITR-seq requires the presence of vector-derived sequences, reducing the likelihood of false integration calls. This issue may not be limited to the adapted ITR-seq protocol but could be even more problematic for the original ITR-seq method. There, the primer binding site is located within the ITR, thus the reads do not contain larger vector genome elements that could be used for filtering (Figure 33A) [381]. Whether this leads to an overestimation of integration events remains to be elucidated. The ability of the adapted ITR-seq protocol to detect vector-to-host genome junctions located deeper within the vector genome (Figure 33A) should, in theory, increase the number of detectable integration events compared to the conventional protocol. These junctions were common, although the preferential vector-to-host genome junction is located at the tip of the ITR stem sequence [327,328]. Despite the potential for broader detection, the conventional ITR-seq approach may ultimately remain more effective due to its bidirectional interrogation of integration sites enabled by the presence of two terminal ITRs. To enable a similar bidirectional approach to assess ITR variant integration, the rAAV would require barcodes close to both ITRs. Subsequently, two primers adjacent to the barcodes would have to be used for the PCR step after Y-adapter ligation, which would also enable the amplification of episome junctions and therefore require substantial optimization efforts.

Validation of the identified integration sites using an orthogonal method, such as long-read sequencing, would be a valuable addition to this work as it would allow the confirmation of the detected integration pattern. Long-read sequencing approaches for the detection of AAV integration using PacBio or Nanopore sequencing following vector genome enrichment, which was recently used to complement conventional ITR-seq data [96], could also be employed here. This would be useful to examine the concatemerization state of integrated vector genomes with alternative ITR sequences. This is especially relevant since concatemerized integrated vector genomes could not be distinguished from episomes by the adapted ITR-seq approach, which is another limitation of this method. In the worst case, ITR variants that increase vector

concatemerization could potentially be falsely identified as variants with low integration propensity.

Nevertheless, the establishment of this protocol represents a major leap towards the screening of ITR variants with altered integration propensity, although several issues remain to be addressed. Additionally, this sequencing strategy could also be used in the future for other gene therapy vectors or genome editing tools in which the integration propensity of different barcode-equipped DNA fragments needs to be assessed.

4.1.5 Strategies to facilitate the screening of synthetic randomized ITR variants

Despite the advancement in ITR engineering achieved in this work, the throughput still lags far behind in comparison to capsid engineering approaches. The ultimate approach for ITR engineering would be a randomized synthetic ITR variant library. However, the creation of fully synthetic ITRs will be hindered by the interdependence of replication proteins and the replication origin within the ITRs. The generation of functional vectors using a synthetic Bocavirus-inspired ITR was possible in the presence of Bocavirus NS1 protein, albeit with low productivity (data not shown). This suggests that a randomized ITR variant screen should mostly maintain the replication origin within the ITRs, otherwise, a co-evolution or the engineering of Rep proteins would likely be necessary, further complicating the process.

While this project was initially aimed at laying the foundation for or even attempting such a large screen ITR screen with partially randomized sequences, in hindsight, the implementation of such a screen will prove extremely difficult. The major obstacle will likely be the newly discovered ITR *trans*-repair mechanism, which nullifies any barcode-ITR association from a large pooled production. Therefore, the generation of a randomized ITR library was not attempted in this work. Nonetheless, I hypothesize that it is possible, especially knowing potential pitfalls and challenges identified in this work, by using a sophisticated barcoding and production strategy. First, a plasmid library with a large set of randomized barcodes with sufficient diversity [423] would be used for the cloning of a much more limited number of ITR variants. The number of ITR variants could be restricted by fixing the sequence in some regions of the ITR, which has the additional advantage of guiding the formation of specific secondary structures and maintaining the replication origin. To reduce the loss and mutagenesis of the ITRs during plasmid propagation, which could compromise the barcode-ITR association, the

aforementioned strategies to ensure ITR stability in vector plasmids might need to be employed, including the shortening of the sITR hairpin (see section 4.1.1).

The most critical step is vector production, where the *trans*-repair pathway must be circumvented at all costs. First, it needs to be ensured that each cell contains at maximum of a single vector plasmid after transfection. Additionally, it needs to be guaranteed that during vector production, the secreted vector is prevented from retro-transduction/re-entry into the producer cells. Otherwise, this could re-introduce the ITR *trans*-repair mechanism by delivering an ITR repair template, thereby preventing a stringent barcode-ITR association. Although vector retro-transduction has never been directly implicated in rAAV production, a recent study has found AAV entry-associated genes to be important for high rAAV productivity [400]. Thus, knocking out essential receptors for vector uptake [74] in the producer cell line could reduce *trans*-repair, but might also negatively impact productivity. After production, the generated vectors need to be sequenced to associate each barcode with an ITR. To maximize sequencing depth, NGS-based sequencing may be favorable over third-generation sequencing. Notably, during integration analysis of my ITR libraries *in vitro*, I observed a large number of reads that contained the viral genome elements but no host genomic sequences (data not shown). These reads were likely derived from non-transduced vector genomes to which the Y-adapter was directly ligated. Using a similar adapter ligation-based NGS approach to determine barcode and ITR sequence in the ITR library could facilitate the required sequencing depth for successful identification of all barcode-ITR combinations. This would allow for the generation of a barcode-ITR reference table, which can then be used to trace ITR functionality after transduction using similar approaches as presented in this work.

While capsid libraries can usually be subcloned after genome extraction from transduced cells, this is not possible with ITRs. Thus, a second screening round would require the generation of a completely new ITR variant library, further increasing the complexity of such a screening procedure and highlighting the specific challenges of ITR engineering in comparison to capsid engineering. Nearly all of the approaches described here would likely require extensive optimization efforts to ensure ITR integrity and proper barcode-ITR association. Nonetheless, the technical foundation for such an attempt has been laid in this doctoral work.

4.2 Circular AAV replication intermediates as a basis for novel rAAV vectors

The hypothesis that AAV Rep proteins can mediate the formation of circular vector genomes was based on the description of ligation events mediated by AAV Rep upon cleavage of ssDNA

by Smith and Kotin [384] and on an equivalent ligation mechanism by Rep proteins of viruses with circular ssDNA [385]. Additionally, the packaging of genome fragments originating at a cryptic RBE in the p5 promoter region suggested that the RBE sequence is sufficient to induce replication and packaging [424]. Consistent with these descriptions, I could observe a similar contamination in the Nanopore sequencing data of my sITR vectors that matched these regions upstream of the p5 promoter region in the *rep*/*cap*-encoding plasmid (data not shown). For CRESS-DNA viruses, the placement of the tandem RBEs separated by spacer sequences has been demonstrated to generate two circular genomes containing the elements between the RBEs within the two circular DNA strands [386]. Thus, I speculated that if transferable to AAV, a similar mechanism could set the basis for the packaging of circular genomes in AAVs, which could reduce host genome integration and reliance on the cellular machinery for circularization and which could concomitantly improve transduction.

The formation of circular genomes from plasmids with tandem RBE* via the RCA-based assay clearly validated the assumed formation of the circular replication intermediates (Figure 35B and Supplementary Figure 5). This also confirmed the initial hypothesis of a possible Rep-mediated ring closure reaction during DNA replication. In subsequent experiments, I could successfully use the 2xRBE* plasmid for production of functional AAV particles. Albeit plasmid-derived genetic material was packaged in the particles (Figure 35C), similar to the packaging of DNA sequences from cryptic RBEs [424], I could not detect any circular DNA using RCA-based assays (data not shown) as well as Nanopore sequencing (Figure 35E). Despite the absence of indications for circular genomes, Nanopore sequencing revealed that the circular replication intermediates have served as replication templates for the packaged DNA. The data also suggested that linear replication products resembling concatenated circular replication intermediate strands were packaged (Supplementary Figure 6). The presence of these linear concatenated sequences should be confirmed in the future using Southern blotting. Unexpectedly, Nanopore sequencing of the plasmid DNA also revealed the presence of very low levels of gaps (Figure 35D and F). I could not exclude those reads to be artifacts, as the generation of chimeric reads is a consistent problem of Nanopore sequencing [425,426]. While this would explain the formation of gaps in the alignment, the consistent length and positioning, matching with the gap expected for circular replication intermediates, rather indicated low levels of plasmid recombination during plasmid propagation. The subsequent dual-luciferase reporter assay (Figure 36) also emphasized that vectors generated using the 2xRBE* plasmids are functional and suggested the packaging of sequences derived from replication intermediate

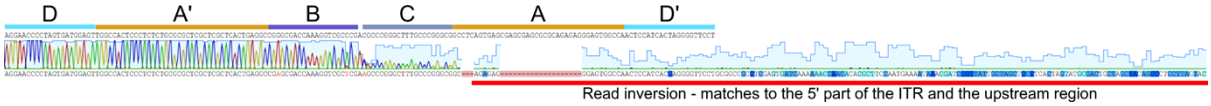
fragments. Of note, the assay could not exclude the involvement of previously recombined plasmid DNA during replication and packaging.

A thorough literature analysis has revealed that the ring closure by AAV Rep has been previously observed with a similar A- and D-sequence element in circular DNA fragments [427]. These circular fragments were shown to serve as templates for genome packaging, and the authors also did not detect packaging of circular but rather linear DNA [428]. This verifies my observation of circular replication intermediate formation. However, in the present work, two distal tandem A- and D-sequence elements in the plasmids were shown to be able to form these circular replication intermediates, distinguishing this work from the previous publications [427,428]. During the work, I also unsuccessfully attempted to use this RBE* replication system to generate vectors carrying circular genomes harnessing not only AAV but also alternative capsids from circo- and polyomavirus, previously described to indiscriminately package DNA [429,430] (data not shown). Rather surprisingly, the circovirus Rep and Rep' were unable to mediate the formation of circular replication intermediates in Hek293T cells with circoviral dual RBE constructs and also did not mediate packaging of DNA in AAV particles (data not shown). Interestingly, though, a chimeric Rep protein containing the DBD of circovirus Rep and the oligomerization and helicase domain of AAV was able to do so, albeit with low efficiency and without detectable packaging of circular DNA (data not shown).

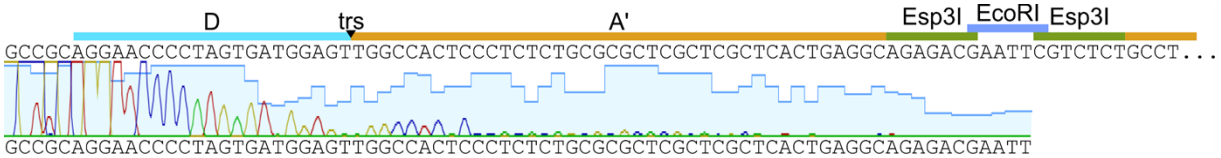
While all approaches to harness these circular replication intermediates to package circular genomes in viral particles failed, the observations during the exploration of this intriguing phenomenon, which could well be a relic from a common ancestor shared by AAV and circoviruses [383], may suggest that the model for AAV packaging needs to be revised. It is currently assumed that the 3' ITR is pumped first into the vector capsid, due to the absence of 5' ITR sequences in capsids with incomplete genomes [111,401]. However, in the case of packaging mediated by cryptic RBEs, the region near the 5' end appears to be packaged preferentially in the capsid [424]. This was consistent with the Nanopore sequencing data in this work, with the highest coverage near but not at the trs (Figure 35E). This could suggest that packaging is initiated after covalent attachment of Rep to the 5' end/ITR before the freshly synthesized strand is pumped through the pore during replication, while the 5' end/ITR remains attached to Rep on the capsid surface. Only after release of the covalent bond between the 5' end of the genome and Rep, the 5' end can be pumped into the capsid as well, which would lead to the observation of initial 3' ITR packaging [111]. Notably, the same mechanism of replication initiation and packaging has been hypothesized for microvirus, which also involves

a DNA ligation step during packaging, generating the circular viral genome [431]. Thus, transferring this system to rAAV, it may be possible to engineer a similar system to facilitate the packaging of circular DNA in rAAV. With the essential ligation step already performed during replication, the only missing step is the same ring closure reaction during DNA packaging. This would likely necessitate extensive engineering of AAV replication origin, Rep protein, and possibly even the capsid protein, but could ultimately generate rAAV vectors that not only transduce more efficiently, circumventing the slow and inefficient episome formation, but are also less prone to induce genotoxic effects through random genomic integration.

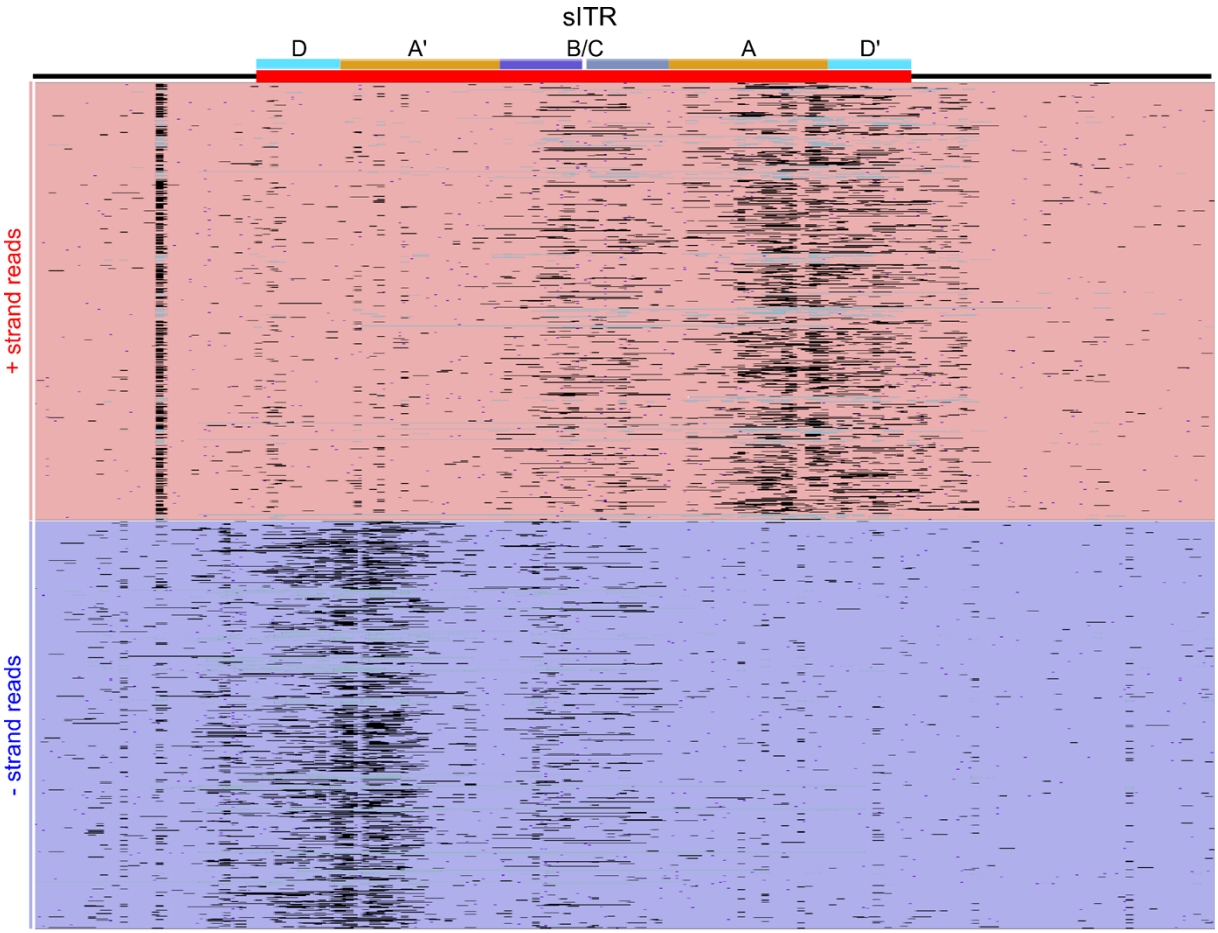
5. Supplementary figures



Supplementary Figure 1: T7eI-ITR sequencing read inversion artifact. Example of a sequencing displaying a read inversion artifact aligned to it a wtITR2 sITR plasmid reference genome in Geneious. Following an sequence gap, the read aligns to the 3‘ portion of the ITR and continues with a segment mismatching the reference. The region underlined in red matches nearly perfect with the region upstream of the ITR, consistent with a read inversion caused by template switching.

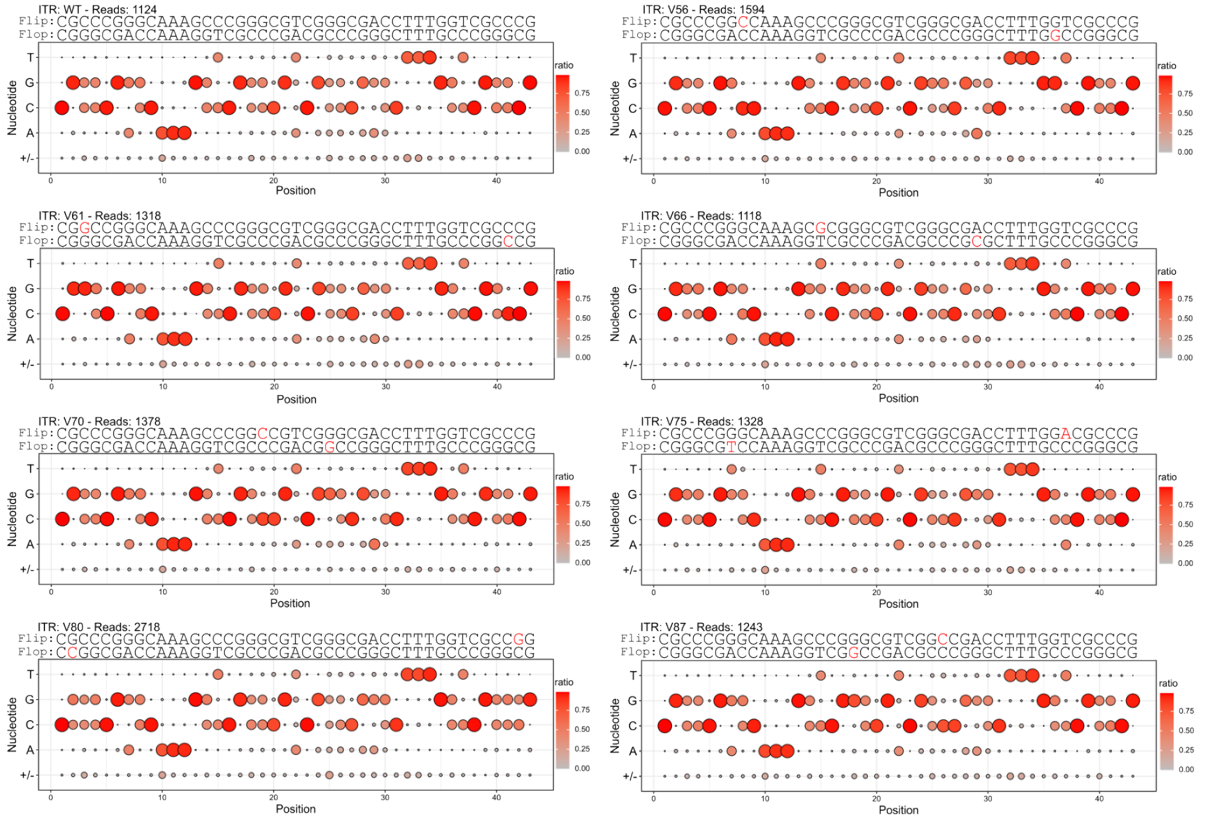


Supplementary Figure 2: T7eI-ITR sequencing read of a perfect linear ITR hairpin. The sequenced sITR plasmid was one of the base constructs with TIIS cleavage sites at the tip of the hairpin. Only a part of the ITR sequence is depicted and the hairpin elements are annotated. Nearly all sequencing results of this plasmid terminate before reaching the tip of the hairpin, thus this read represents an exception.

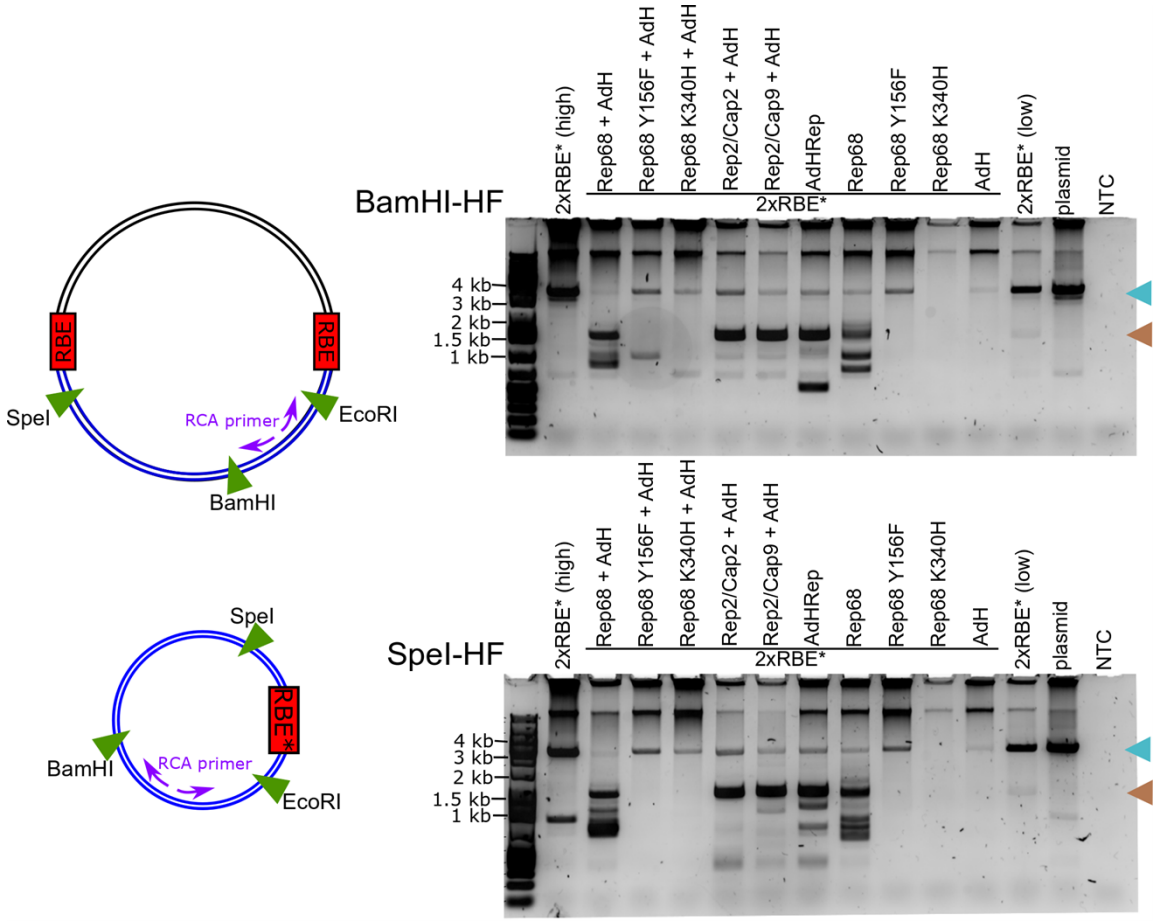


Supplementary Figure 3: Read strand-specific artifact in the A- and D- region in sITR plasmid Nanopore sequencing reads. Random selection of sITR plasmid sequencing reads after splice-aware alignment to the plasmid reference sequence visualized using the IGV and sorted depending on the strand. The ITR region is annotated and color coded, the colors in the read indicate the alignment: red/purple colors are matching nucleotides, black color indicates deletions while “splice” gaps are light

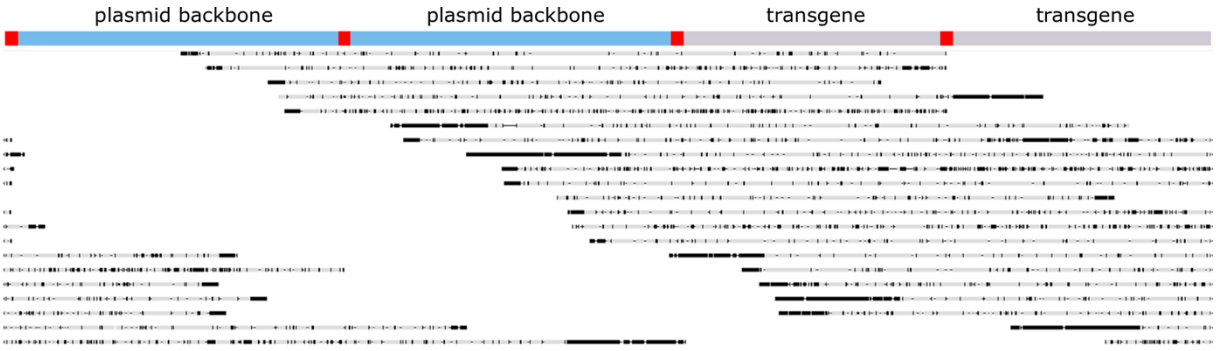
turquoise. The accumulation of mismatches and gaps in the 3‘region of the reads within the A-D junction region is clearly visible.



Supplementary Figure 4: ITR sequence integrity in selected variants of a separately produced mismatch mutant library. Nanopore sequencing reads of eight hairpin variants of the mismatch mutant ITR library (V55-V90) after separate production were aligned to a consensus flip/flop reference sequence and the nucleotide distribution was determined and depicted as integrated heatmap/dotplot. The expected ITR sequence is given both in flip and flop configuration as reference above together with the variant number and the number of reads present in the dataset. Red nucleotides indicate mutant bases compared to the wtITR2 sequence. Both the color and the size of the circle indicate the ratio of nucleotides at each position in the alignment.



Supplementary Figure 5: RCA assay to assess circular replication intermediate formation with alternative enzymes for linearization. (Left) Schematic of 2xRBE* plasmid and the generated circular replication intermediate with primer binding sites (purple) and restriction enzyme cleavage sites (green). (Right) Gel images of samples analyzed using the RCA assay. Hek293T cells were transfected with the indicated plasmids, and the low-molecular weight DNA was extracted after 72 h and subjected to the RCA assay. The "high" indicates the maximum input amount of the 2xRBE* plasmid during transfection, while the "low" indicates an amount comparable to the input in the other samples. For two plasmid transfections 2xRBE* was used with the indicated plasmid at a 1:25 molar ratio, in three plasmid transfections at 1:10:10 molar ratio. During RCA, a plasmid control with 1 ng of the original 2xRBE* plasmid and a no template control (NTC) were included, and amplicons were debranched with the indicated enzyme.



Supplementary Figure 6: Concatemeric reads detected in Nanopore sequencing data from 2xRBE* derived vectors. Alignment of the long sequencing reads to a concatemeric reference genome using Geneious revealed concatenation of sequences, even including aberrant concatemerization. There, reads can also cover the same plasmid element twice before switching to the other element. The

reference sequence is annotated above, indicating the region of the 2xRBE* plasmid in red. Reference sequence matches in the reads are indicated in grey, mismatches in black.

6. References

1 Atchison, R. W., Casto, B. C. & Hammon, W. M. Adenovirus-Associated Defective Virus Particles. *Science* **149**, 754-756 (1965). https://doi.org/10.1126/science.149.3685.754

2 Hoggan, M. D., Blacklow, N. R. & Rowe, W. P. Studies of small DNA viruses found in various adenovirus preparations: physical, biological, and immunological characteristics. *Proc Natl Acad Sci U S A* **55**, 1467-1474 (1966). https://doi.org/10.1073/pnas.55.6.1467

3 Rose, J. A., Berns, K. I., Hoggan, M. D. & Koczot, F. J. Evidence for a single-stranded adenovirus-associated virus genome: formation of a DNA density hybrid on release of viral DNA. *Proc Natl Acad Sci U S A* **64**, 863-869 (1969). https://doi.org/10.1073/pnas.64.3.863

4 Mayor, H. D., Torikai, K., Melnick, J. L. & Mandel, M. Plus and minus single-stranded DNA separately encapsidated in adeno-associated satellite virions. *Science* **166**, 1280-1282 (1969). https://doi.org/10.1126/science.166.3910.1280

5 Berns, K. I. & Rose, J. A. Evidence for a single-stranded adenovirus-associated virus genome: isolation and separation of complementary single strands. *J Virol* **5**, 693-699 (1970). https://doi.org/10.1128/JVI.5.6.693-699.1970

6 Berns, K. I. & Adler, S. Separation of two types of adeno-associated virus particles containing complementary polynucleotide chains. *J Virol* **9**, 394-396 (1972). https://doi.org/10.1128/JVI.9.2.394-396.1972

7 Schlehofer, J. R., Ehrbar, M. & zur Hausen, H. Vaccinia virus, herpes simplex virus, and carcinogens induce DNA amplification in a human cell line and support replication of a helpervirus dependent parvovirus. *Virology* **152**, 110-117 (1986). https://doi.org/10.1016/0042-6822(86)90376-4

8 Ogston, P., Raj, K. & Beard, P. Productive replication of adeno-associated virus can occur in human papillomavirus type 16 (HPV-16) episome-containing keratinocytes and is augmented by the HPV-16 E2 protein. *J Virol* **74**, 3494-3504 (2000). https://doi.org/10.1128/jvi.74.8.3494-3504.2000

9 Yalkinoglu, A. O., Heilbronn, R., Burkle, A., Schlehofer, J. R. & zur Hausen, H. DNA amplification of adeno-associated virus as a response to cellular genotoxic stress. *Cancer Res* **48**, 3123-3129 (1988).

10 Bossis, I. & Chiorini, J. A. Cloning of an avian adeno-associated virus (AAAV) and generation of recombinant AAAV particles. *J Virol* **77**, 6799-6810 (2003). https://doi.org/10.1128/jvi.77.12.6799-6810.2003

11 Farkas, S. L. *et al.* A parvovirus isolated from royal python (Python regius) is a member of the genus Dependovirus. *J Gen Virol* **85**, 555-561 (2004). https://doi.org/10.1099/vir.0.19616-0

12 Penzes, J. J., Pham, H. T., Benko, M. & Tijssen, P. Novel parvoviruses in reptiles and genome sequence of a lizard parvovirus shed light on Dependoparvovirus genus evolution. *J Gen Virol* **96**, 2769-2779 (2015). https://doi.org/10.1099/vir.0.000215

13 Belyi, V. A., Levine, A. J. & Skalka, A. M. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. *J Virol* **84**, 12458-12462 (2010). https://doi.org/10.1128/JVI.01789-10

14 Bohmer, A. *et al.* Novel application for isothermal nucleic acid sequence-based amplification (NASBA). *J Virol Methods* **158**, 199-201 (2009). https://doi.org/10.1016/j.jviromet.2009.02.010

15 Cotmore, S. F. & Tattersall, P. Encapsidation of minute virus of mice DNA: aspects of the translocation mechanism revealed by the structure of partially packaged genomes. *Virology* **336**, 100-112 (2005). https://doi.org/10.1016/j.virol.2005.03.007

16 Issa, S. S., Shaimardanova, A. A., Solovyeva, V. V. & Rizvanov, A. A. Various AAV Serotypes and Their Applications in Gene Therapy: An Overview. *Cells* **12** (2023). https://doi.org/10.3390/cells12050785

17 Zinn, E. *et al.* In Silico Reconstruction of the Viral Evolutionary Lineage Yields a Potent Gene Therapy Vector. *Cell Rep* **12**, 1056-1068 (2015). https://doi.org/10.1016/j.celrep.2015.07.019

18 Navarro-Oliveros, M. *et al.* Seroprevalence of adeno-associated virus types 1, 2, 3, 4, 5, 6, 8, and 9 in a Basque cohort of healthy donors. *Sci Rep* **14**, 15941 (2024). https://doi.org/10.1038/s41598-024-66546-4

19 Chhabra, A. *et al.* Global seroprevalence of neutralizing antibodies against adeno-associated virus serotypes used for human gene therapies. *Mol Ther Methods Clin Dev* **32**, 101273 (2024). https://doi.org/10.1016/j.omtm.2024.101273

20 Pabinger, I. *et al.* Multicenter assessment and longitudinal study of the prevalence of antibodies and related adaptive immune responses to AAV in adult males with hemophilia. *Gene Ther* **31**, 273-284 (2024). https://doi.org/10.1038/s41434-024-00441-5

21 Kashiwakura, Y. *et al.* The seroprevalence of neutralizing antibodies against the adeno-associated virus capsids in Japanese hemophiliacs. *Mol Ther Methods Clin Dev* **27**, 404-414 (2022). https://doi.org/10.1016/j.omtm.2022.10.014

22 van Olden, R. W. *et al.* Adeno-associated virus serotype 9 antibodies in neonates and young children: Seroprevalence and kinetics. *Mol Ther Methods Clin Dev* **32**, 101344 (2024). https://doi.org/10.1016/j.omtm.2024.101344

23 Klamroth, R. *et al.* Global Seroprevalence of Pre-existing Immunity Against AAV5 and Other AAV Serotypes in People with Hemophilia A. *Hum Gene Ther* **33**, 432-441 (2022). https://doi.org/10.1089/hum.2021.287

24 de la Maza, L. M. & Carter, B. J. Inhibition of adenovirus oncogenicity in hamsters by adeno-associated virus DNA. *J Natl Cancer Inst* **67**, 1323-1326 (1981).

25 Coker, A. L. *et al.* Adeno-associated virus is associated with a lower risk of high-grade cervical neoplasia. *Exp Mol Pathol* **70**, 83-89 (2001). https://doi.org/10.1006/exmp.2000.2347

26 Hermonat, P. L. Adeno-associated virus inhibits human papillomavirus type 16: a viral interaction implicated in cervical cancer. *Cancer Res* **54**, 2278-2281 (1994).

27 Hermonat, P. L., Plott, R. T., Santin, A. D., Parham, G. P. & Flick, J. T. Adeno-associated virus Rep78 inhibits oncogenic transformation of primary human keratinocytes by a human papillomavirus type 16-ras chimeric. *Gynecol Oncol* **66**, 487-494 (1997). https://doi.org/10.1006/gyno.1997.4789

28 Ho, A. *et al.* Adeno-associated virus 2 infection in children with non-A-E hepatitis. *Nature* **617**, 555-563 (2023). https://doi.org/10.1038/s41586-023-05948-2

29 Servellita, V. *et al.* Adeno-associated virus type 2 in US children with acute severe hepatitis. *Nature* **617**, 574-580 (2023). https://doi.org/10.1038/s41586-023-05949-1

30 Mitchell, M. M. *et al.* Signatures of AAV-2 immunity are enriched in children with severe acute hepatitis of unknown etiology. *Sci Transl Med* **15**, eadh9917 (2023). https://doi.org/10.1126/scitranslmed.adh9917

31 Pereira, D. J., McCarty, D. M. & Muzyczka, N. The adeno-associated virus (AAV) Rep protein acts as both a repressor and an activator to regulate AAV transcription during a productive infection. *J Virol* **71**, 1079-1088 (1997). https://doi.org/10.1128/JVI.71.2.1079-1088.1997

32 Chiorini, J. A. *et al.* The roles of AAV Rep proteins in gene expression and targeted integration. *Curr Top Microbiol Immunol* **218**, 25-33 (1996). https://doi.org/10.1007/978-3-642-80207-2_2

33 Srivastava, A., Lusby, E. W. & Berns, K. I. Nucleotide sequence and organization of the adeno-associated virus 2 genome. *J Virol* **45**, 555-564 (1983). https://doi.org/10.1128/JVI.45.2.555-564.1983

34 Lusby, E., Fife, K. H. & Berns, K. I. Nucleotide sequence of the inverted terminal repetition in adeno-associated virus DNA. *J Virol* **34**, 402-409 (1980). https://doi.org/10.1128/JVI.34.2.402-409.1980

35 Samulski, R. J., Chang, L. S. & Shenk, T. A recombinant plasmid from which an infectious adeno-associated virus genome can be excised in vitro and its use to study viral replication. *J Virol* **61**, 3096-3101 (1987). https://doi.org/10.1128/JVI.61.10.3096-3101.1987

36 McLaughlin, S. K., Collis, P., Hermonat, P. L. & Muzyczka, N. Adeno-associated virus general transduction vectors: analysis of proviral structures. *J Virol* **62**, 1963-1973 (1988). https://doi.org/10.1128/JVI.62.6.1963-1973.1988

37 Yan, Z., Zak, R., Zhang, Y. & Engelhardt, J. F. Inverted terminal repeat sequences are important for intermolecular recombination and circularization of adeno-associated virus genomes. *J Virol* **79**, 364-379 (2005). https://doi.org/10.1128/JVI.79.1.364-379.2005

38 Hickman, A. B., Ronning, D. R., Perez, Z. N., Kotin, R. M. & Dyda, F. The nuclease domain of adeno-associated virus rep coordinates replication initiation using two distinct DNA recognition interfaces. *Mol Cell* **13**, 403-414 (2004). https://doi.org/10.1016/s1097-2765(04)00023-1

39 Carter, B. J. & Rose, J. A. Transcription in vivo of a defective parvovirus: sedimentation and electrophoretic analysis of RNA synthesized by adenovirus-associated virus and its helper adenovirus. *Virology* **61**, 182-199 (1974). https://doi.org/10.1016/0042-6822(74)90253-0

40 Laughlin, C. A., Westphal, H. & Carter, B. J. Spliced adenovirus-associated virus RNA. *Proc Natl Acad Sci U S A* **76**, 5567-5571 (1979). https://doi.org/10.1073/pnas.76.11.5567

41 Wang, L. *et al.* Productive life cycle of adeno-associated virus serotype 2 in the complete absence of a conventional polyadenylation signal. *J Gen Virol* **96**, 2780-2787 (2015). https://doi.org/10.1099/jgv.0.000229

42 Im, D. S. & Muzyczka, N. Partial purification of adeno-associated virus Rep78, Rep52, and Rep40 and their biochemical characterization. *J Virol* **66**, 1119-1128 (1992). https://doi.org/10.1128/JVI.66.2.1119-1128.1992

43 Holscher, C., Kleinschmidt, J. A. & Burkle, A. High-level expression of adeno-associated virus (AAV) Rep78 or Rep68 protein is sufficient for infectious-particle formation by a rep-negative AAV mutant. *J Virol* **69**, 6880-6885 (1995). https://doi.org/10.1128/JVI.69.11.6880-6885.1995

44 Jalsic, L. *et al.* Inducible HEK293 AAV packaging cell lines expressing Rep proteins. *Mol Ther Methods Clin Dev* **30**, 259-275 (2023). https://doi.org/10.1016/j.omtm.2023.07.002

45 Hedrick, J. L. & Smith, A. J. Size and charge isomer separation and estimation of molecular weights of proteins by disc gel electrophoresis. *Arch Biochem Biophys* **126**, 155-164 (1968). https://doi.org/10.1016/0003-9861(68)90569-9

46 Smith, R. H., Spano, A. J. & Kotin, R. M. The Rep78 gene product of adeno-associated virus (AAV) self-associates to form a hexameric complex in the presence of AAV ori sequences. *J Virol* **71**, 4461-4471 (1997). https://doi.org/10.1128/JVI.71.6.4461-4471.1997

47 Zarate-Perez, F. *et al.* The interdomain linker of AAV-2 Rep68 is an integral part of its oligomerization domain: role of a conserved SF3 helicase residue in oligomerization. *PLoS Pathog* **8**, e1002764 (2012). https://doi.org/10.1371/journal.ppat.1002764

48 Smith, R. H. & Kotin, R. M. The Rep52 gene product of adeno-associated virus is a DNA helicase with 3'-to-5' polarity. *J Virol* **72**, 4874-4881 (1998). https://doi.org/10.1128/JVI.72.6.4874-4881.1998

49 Becerra, S. P., Koczot, F., Fabisch, P. & Rose, J. A. Synthesis of adeno-associated virus structural proteins requires both alternative mRNA splicing and alternative initiations from a single transcript. *J Virol* **62**, 2745-2754 (1988). https://doi.org/10.1128/JVI.62.8.2745-2754.1988

50 Cassinotti, P., Weitz, M. & Tratschin, J. D. Organization of the adeno-associated virus (AAV) capsid gene: mapping of a minor spliced mRNA coding for virus capsid protein 1. *Virology* **167**, 176-184 (1988).

51 Zadori, Z. *et al.* A viral phospholipase A2 is required for parvovirus infectivity. *Dev Cell* **1**, 291-302 (2001). https://doi.org/10.1016/s1534-5807(01)00031-4

52 Popa-Wagner, R., Sonntag, F., Schmidt, K., King, J. & Kleinschmidt, J. A. Nuclear translocation of adeno-associated virus type 2 capsid proteins for virion assembly. *J Gen Virol* **93**, 1887-1898 (2012). https://doi.org/10.1099/vir.0.043232-0

53 Buller, R. M. & Rose, J. A. Characterization of adenovirus-associated virus-induced polypeptides in KB cells. *J Virol* **25**, 331-338 (1978). https://doi.org/10.1128/JVI.25.1.331-338.1978

54 Johnson, F. B., Ozer, H. L. & Hoggan, M. D. Structural proteins of adenovirus-associated virus type 3. *J Virol* **8**, 860-863 (1971). https://doi.org/10.1128/JVI.8.6.860-863.1971

55 Worner, T. P. *et al.* Adeno-associated virus capsid assembly is divergent and stochastic. *Nat Commun* **12**, 1642 (2021). https://doi.org/10.1038/s41467-021-21935-5

56 Xie, Q. *et al.* The atomic structure of adeno-associated virus (AAV-2), a vector for human gene therapy. *Proc Natl Acad Sci U S A* **99**, 10405-10410 (2002). https://doi.org/10.1073/pnas.162250899

57 Mietzsch, M., Penzes, J. J. & Agbandje-McKenna, M. Twenty-Five Years of Structural Parvovirology. *Viruses* **11** (2019). https://doi.org/10.3390/v11040362

58 Sonntag, F., Schmidt, K. & Kleinschmidt, J. A. A viral assembly factor promotes AAV2 capsid formation in the nucleolus. *Proc Natl Acad Sci U S A* **107**, 10220-10225 (2010). https://doi.org/10.1073/pnas.1001673107

59 Elmore, Z. C. *et al.* The membrane associated accessory protein is an adeno-associated viral egress factor. *Nat Commun* **12**, 6239 (2021). https://doi.org/10.1038/s41467-021-26485-4

60 Cao, M., You, H. & Hermonat, P. L. The X gene of adeno-associated virus 2 (AAV2) is involved in viral DNA replication. *PLoS One* **9**, e104596 (2014). https://doi.org/10.1371/journal.pone.0104596

61 Earley, L. F. *et al.* Adeno-associated Virus (AAV) Assembly-Activating Protein Is Not an Essential Requirement for Capsid Assembly of AAV Serotypes 4, 5, and 11. *J Virol* **91** (2017). https://doi.org/10.1128/JVI.01980-16

62 Aksu Kuz, C., Ning, K., Hao, S., Cheng, F. & Qiu, J. Role of the membrane-associated accessory protein (MAAP) in adeno-associated virus (AAV) infection. *J Virol* **98**, e0063324 (2024). https://doi.org/10.1128/jvi.00633-24

63 Huang, L. Y., Halder, S. & Agbandje-McKenna, M. Parvovirus glycan interactions. *Curr Opin Virol* **7**, 108-118 (2014). https://doi.org/10.1016/j.coviro.2014.05.007

64 Summerford, C. & Samulski, R. J. Membrane-associated heparan sulfate proteoglycan is a receptor for adeno-associated virus type 2 virions. *J Virol* **72**, 1438-1445 (1998). https://doi.org/10.1128/JVI.72.2.1438-1445.1998

65 Halbert, C. L., Allen, J. M. & Miller, A. D. Adeno-associated virus type 6 (AAV6) vectors mediate efficient transduction of airway epithelial cells in mouse lungs compared to that of AAV2 vectors. *J Virol* **75**, 6615-6624 (2001). https://doi.org/10.1128/JVI.75.14.6615-6624.2001

66 Schmidt, M. *et al.* Adeno-associated virus type 12 (AAV12): a novel AAV serotype with sialic acid- and heparan sulfate proteoglycan-independent transduction activity. *J Virol* **82**, 1399-1406 (2008). https://doi.org/10.1128/JVI.02012-07

67 Kaludov, N. *et al.* Production, purification and preliminary X-ray crystallographic studies of adeno-associated virus serotype 4. *Virology* **306**, 1-6 (2003). https://doi.org/10.1016/s0042-6822(02)00037-5

68 Wu, Z. *et al.* Single amino acid changes can influence titer, heparin binding, and tissue tropism in different adeno-associated virus serotypes. *J Virol* **80**, 11393-11397 (2006). https://doi.org/10.1128/JVI.01288-06

69 Huang, L. Y. *et al.* Characterization of the Adeno-Associated Virus 1 and 6 Sialic Acid Binding Site. *J Virol* **90**, 5219-5230 (2016). https://doi.org/10.1128/JVI.00161-16

70 Shen, S., Bryant, K. D., Brown, S. M., Randell, S. H. & Asokan, A. Terminal N-linked galactose is the primary receptor for adeno-associated virus 9. *J Biol Chem* **286**, 13532-13540 (2011). https://doi.org/10.1074/jbc.M110.210922

71 Akache, B. *et al.* The 37/67-kilodalton laminin receptor is a receptor for adeno-associated virus serotypes 8, 2, 3, and 9. *J Virol* **80**, 9831-9836 (2006). https://doi.org/10.1128/JVI.00878-06

72 Qing, K. *et al.* Human fibroblast growth factor receptor 1 is a co-receptor for infection by adeno-associated virus 2. *Nat Med* **5**, 71-77 (1999). https://doi.org/10.1038/4758

73 Summerford, C., Bartlett, J. S. & Samulski, R. J. AlphaVbeta5 integrin: a co-receptor for adeno-associated virus type 2 infection. *Nat Med* **5**, 78-82 (1999). https://doi.org/10.1038/4768

74 Pillay, S. *et al.* An essential receptor for adeno-associated virus infection. *Nature* **530**, 108-112 (2016). https://doi.org/10.1038/nature16465

75 Zane, G. *et al.* Cryo-EM structure of adeno-associated virus 4 at 2.2 A resolution. *Acta Crystallogr D Struct Biol* **79**, 140-153 (2023). https://doi.org/10.1107/S2059798322012190

76 Havlik, L. P. *et al.* Receptor Switching in Newly Evolved Adeno-associated Viruses. *J Virol* **95**, e0058721 (2021). https://doi.org/10.1128/JVI.00587-21

77 Duan, D. *et al.* Dynamin is required for recombinant adeno-associated virus type 2 infection. *J Virol* **73**, 10371-10376 (1999). https://doi.org/10.1128/JVI.73.12.10371-10376.1999

78 Bartlett, J. S., Wilcher, R. & Samulski, R. J. Infectious entry pathway of adeno-associated virus and adeno-associated virus vectors. *J Virol* **74**, 2777-2785 (2000). https://doi.org/10.1128/jvi.74.6.2777-2785.2000

79 Nonnenmacher, M. & Weber, T. Adeno-associated virus 2 infection requires endocytosis through the CLIC/GEEC pathway. *Cell Host Microbe* **10**, 563-576 (2011). https://doi.org/10.1016/j.chom.2011.10.014

80 Weinberg, M. S. *et al.* Recombinant adeno-associated virus utilizes cell-specific infectious entry mechanisms. *J Virol* **88**, 12472-12484 (2014). https://doi.org/10.1128/JVI.01971-14

81 Dhungel, B. P., Bailey, C. G. & Rasko, J. E. J. Journey to the Center of the Cell: Tracing the Path of AAV Transduction. *Trends Mol Med* **27**, 172-184 (2021). https://doi.org/10.1016/j.molmed.2020.09.010

82 Kronenberg, S., Bottcher, B., von der Lieth, C. W., Bleker, S. & Kleinschmidt, J. A. A conformational change in the adeno-associated virus type 2 capsid leads to the exposure of hidden VP1 N termini. *J Virol* **79**, 5296-5303 (2005). https://doi.org/10.1128/JVI.79.9.5296-5303.2005

83 Venkatakrishnan, B. *et al.* Structure and dynamics of adeno-associated virus serotype 1 VP1-unique N-terminal domain and its role in capsid trafficking. *J Virol* **87**, 4974-4984 (2013). https://doi.org/10.1128/JVI.02524-12

84 Stahnke, S. *et al.* Intrinsic phospholipase A2 activity of adeno-associated virus is involved in endosomal escape of incoming particles. *Virology* **409**, 77-83 (2011). https://doi.org/10.1016/j.virol.2010.09.025

85 Nonnenmacher, M. E., Cintrat, J. C., Gillet, D. & Weber, T. Syntaxin 5-dependent retrograde transport to the trans-Golgi network is required for adeno-associated virus transduction. *J Virol* **89**, 1673-1687 (2015). https://doi.org/10.1128/JVI.02520-14

86 Kelkar, S. *et al.* A common mechanism for cytoplasmic dynein-dependent microtubule binding shared among adeno-associated virus and adenovirus serotypes. *J Virol* **80**, 7781-7785 (2006). https://doi.org/10.1128/JVI.00481-06

87 Monahan, P. E. *et al.* Proteasome inhibitors enhance gene delivery by AAV virus vectors expressing large genomes in hemophilia mouse and dog models: a strategy for broad clinical application. *Mol Ther* **18**, 1907-1916 (2010). https://doi.org/10.1038/mt.2010.170

88 Kelich, J. M. *et al.* Super-resolution imaging of nuclear import of adeno-associated virus in live cells. *Mol Ther Methods Clin Dev* **2**, 15047 (2015). https://doi.org/10.1038/mtm.2015.47

89 Cervelli, T. *et al.* Processing of recombinant AAV genomes occurs in specific nuclear structures that overlap with foci of DNA-damage-response proteins. *J Cell Sci* **121**, 349-357 (2008). https://doi.org/10.1242/jcs.003632

90 Gliwa, K. *et al.* Biophysical and structural insights into AAV genome ejection. *J Virol*, e0089924 (2025). https://doi.org/10.1128/jvi.00899-24

91 Ferrari, F. K., Samulski, T., Shenk, T. & Samulski, R. J. Second-strand synthesis is a rate-limiting step for efficient transduction by recombinant adeno-associated virus vectors. *J Virol* **70**, 3227-3234 (1996). https://doi.org/10.1128/JVI.70.5.3227-3234.1996

92 Fisher, K. J. *et al.* Transduction with recombinant adeno-associated virus for gene therapy is limited by leading-strand synthesis. *J Virol* **70**, 520-532 (1996). https://doi.org/10.1128/JVI.70.1.520-532.1996

93 Zhou, X. *et al.* Adeno-associated virus of a single-polarity DNA genome is capable of transduction in vivo. *Mol Ther* **16**, 494-499 (2008). https://doi.org/10.1038/sj.mt.6300397

94 Zhong, L. *et al.* Single-polarity recombinant adeno-associated virus 2 vector-mediated transgene expression in vitro and in vivo: mechanism of transduction. *Mol Ther* **16**, 290-295 (2008). https://doi.org/10.1038/sj.mt.6300376

95 Penaud-Budloo, M. *et al.* Adeno-associated virus vector genomes persist as episomal chromatin in primate muscle. *J Virol* **82**, 7875-7885 (2008). https://doi.org/10.1128/JVI.00649-08

96 Greig, J. A. *et al.* Integrated vector genomes may contribute to long-term expression in primate liver after AAV administration. *Nat Biotechnol* **42**, 1232-1242 (2024). https://doi.org/10.1038/s41587-023-01974-7

97 Kotin, R. M. & Berns, K. I. Organization of adeno-associated virus DNA in latently infected Detroit 6 cells. *Virology* **170**, 460-467 (1989). https://doi.org/10.1016/0042-6822(89)90437-6

98 Kotin, R. M. *et al.* Site-specific integration by adeno-associated virus. *Proc Natl Acad Sci U S A* **87**, 2211-2215 (1990). https://doi.org/10.1073/pnas.87.6.2211

99 Kotin, R. M., Menninger, J. C., Ward, D. C. & Berns, K. I. Mapping and direct visualization of a region-specific viral DNA integration site on chromosome 19q13-qter. *Genomics* **10**, 831-834 (1991). https://doi.org/10.1016/0888-7543(91)90470-y

100 Samulski, R. J. *et al.* Targeted integration of adeno-associated virus (AAV) into human chromosome 19. *EMBO J* **10**, 3941-3950 (1991). https://doi.org/10.1002/j.1460-2075.1991.tb04964.x

101 Kotin, R. M., Linden, R. M. & Berns, K. I. Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. *EMBO J* **11**, 5071-5078 (1992). https://doi.org/10.1002/j.1460-2075.1992.tb05614.x

102 Wistuba, A., Weger, S., Kern, A. & Kleinschmidt, J. A. Intermediates of adeno-associated virus type 2 assembly: identification of soluble complexes containing Rep and Cap proteins. *J Virol* **69**, 5311-5319 (1995). https://doi.org/10.1128/JVI.69.9.5311-5319.1995

103 Wistuba, A., Kern, A., Weger, S., Grimm, D. & Kleinschmidt, J. A. Subcellular compartmentalization of adeno-associated virus type 2 assembly. *J Virol* **71**, 1341-1352 (1997). https://doi.org/10.1128/JVI.71.2.1341-1352.1997

104 Maurer, A. C. *et al.* The Assembly-Activating Protein Promotes Stability and Interactions between AAV's Viral Proteins to Nucleate Capsid Assembly. *Cell Rep* **23**, 1817-1830 (2018). https://doi.org/10.1016/j.celrep.2018.04.026

105 Duan, D. *et al.* Formation of adeno-associated virus circular genomes is differentially regulated by adenovirus E4 ORF6 and E2a gene expression. *J Virol* **73**, 161-169 (1999). https://doi.org/10.1128/JVI.73.1.161-169.1999

106 Lkharrazi, A. *et al.* AAV2 can replicate its DNA by a rolling hairpin or rolling circle mechanism, depending on the helper virus. *J Virol* **98**, e0128224 (2024). https://doi.org/10.1128/jvi.01282-24

107 Ni, T. H., Zhou, X., McCarty, D. M., Zolotukhin, I. & Muzyczka, N. In vitro replication of adeno-associated virus DNA. *J Virol* **68**, 1128-1138 (1994). https://doi.org/10.1128/JVI.68.2.1128-1138.1994

108 Ward, P., Urcelay, E., Kotin, R., Safer, B. & Berns, K. I. Adeno-associated virus DNA replication in vitro: activation by a maltose binding protein/Rep 68 fusion protein. *J Virol* **68**, 6029-6037 (1994). https://doi.org/10.1128/JVI.68.9.6029-6037.1994

109 Costello, E., Sahli, R., Hirt, B. & Beard, P. The mismatched nucleotides in the 5'-terminal hairpin of minute virus of mice are required for efficient viral DNA replication. *J Virol* **69**, 7489-7496 (1995). https://doi.org/10.1128/JVI.69.12.7489-7496.1995

110 Wang, X. S., Ponnazhagan, S. & Srivastava, A. Rescue and replication signals of the adeno-associated virus 2 genome. *J Mol Biol* **250**, 573-580 (1995). https://doi.org/10.1006/jmbi.1995.0398

111 King, J. A., Dubielzig, R., Grimm, D. & Kleinschmidt, J. A. DNA helicase-mediated packaging of adeno-associated virus type 2 genomes into preformed capsids. *EMBO J* **20**, 3282-3291 (2001). https://doi.org/10.1093/emboj/20.12.3282

112 Vandenberghe, L. H. *et al.* Efficient serotype-dependent release of functional vector into the culture medium during adeno-associated virus manufacturing. *Hum Gene Ther* **21**, 1251-1257 (2010). https://doi.org/10.1089/hum.2010.107

113 Samulski, R. J., Berns, K. I., 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 U S A* **79**, 2077-2081 (1982). https://doi.org/10.1073/pnas.79.6.2077

114 Laughlin, C. A., Tratschin, J. D., Coon, H. & Carter, B. J. Cloning of infectious adeno-associated virus genomes in bacterial plasmids. *Gene* **23**, 65-73 (1983). https://doi.org/10.1016/0378-1119(83)90217-2

115 Senapathy, P., Tratschin, J. D. & Carter, B. J. Replication of adeno-associated virus DNA. Complementation of naturally occurring rep- mutants by a wild-type genome or an ori- mutant and correction of terminal palindrome deletions. *J Mol Biol* **179**, 1-20 (1984). https://doi.org/10.1016/0022-2836(84)90303-6

116 Hermonat, P. L. & Muzyczka, N. Use of adeno-associated virus as a mammalian DNA cloning vector: transduction of neomycin resistance into mammalian tissue culture cells. *Proc Natl Acad Sci U S A* **81**, 6466-6470 (1984). https://doi.org/10.1073/pnas.81.20.6466

117 Samulski, R. J., Chang, L. S. & Shenk, T. Helper-free stocks of recombinant adeno-associated viruses: normal integration does not require viral gene expression. *J Virol* **63**, 3822-3828 (1989). https://doi.org/10.1128/JVI.63.9.3822-3828.1989

118 Chao, H. *et al.* Several log increase in therapeutic transgene delivery by distinct adeno-associated viral serotype vectors. *Mol Ther* **2**, 619-623 (2000). https://doi.org/10.1006/mthe.2000.0219

119 Burger, C. *et al.* Recombinant AAV viral vectors pseudotyped with viral capsids from serotypes 1, 2, and 5 display differential efficiency and cell tropism after delivery to different regions of the central nervous system. *Mol Ther* **10**, 302-317 (2004). https://doi.org/10.1016/j.ymthe.2004.05.024

120 Grimm, D., Pandey, K., Nakai, H., Storm, T. A. & Kay, M. A. Liver transduction with recombinant adeno-associated virus is primarily restricted by capsid serotype not vector genotype. *J Virol* **80**, 426-439 (2006). https://doi.org/10.1128/JVI.80.1.426-439.2006

121 Grimm, D., Kay, M. A. & Kleinschmidt, J. A. Helper virus-free, optically controllable, and two-plasmid-based production of adeno-associated virus vectors of serotypes 1 to 6. *Mol Ther* **7**, 839-850 (2003). https://doi.org/10.1016/s1525-0016(03)00095-9

122 Wu, Z., Yang, H. & Colosi, P. Effect of genome size on AAV vector packaging. *Mol Ther* **18**, 80-86 (2010). https://doi.org/10.1038/mt.2009.255

123 Grieger, J. C. & Samulski, R. J. Packaging capacity of adeno-associated virus serotypes: impact of larger genomes on infectivity and postentry steps. *J Virol* **79**, 9933-9944 (2005). https://doi.org/10.1128/JVI.79.15.9933-9944.2005

124 Kruger-Haag, A. *et al.* Evaluation of life cycle defective adenovirus mutants for production of adeno-associated virus vectors. *J Gene Med* **21**, e3094 (2019). https://doi.org/10.1002/jgm.3094

125 Myers, M. W., Laughlin, C. A., Jay, F. T. & Carter, B. J. Adenovirus helper function for growth of adeno-associated virus: effect of temperature-sensitive mutations in adenovirus early gene region 2. *J Virol* **35**, 65-75 (1980). https://doi.org/10.1128/JVI.35.1.65-75.1980

126 Chang, L. S., Shi, Y. & Shenk, T. Adeno-associated virus P5 promoter contains an adenovirus E1A-inducible element and a binding site for the major late transcription factor. *J Virol* **63**, 3479-3488 (1989). https://doi.org/10.1128/JVI.63.8.3479-3488.1989

127 Richardson, W. D. & Westphal, H. Requirement for either early region 1a or early region 1b adenovirus gene products in the helper effect for adeno-associated virus. *J Virol* **51**, 404-410 (1984). https://doi.org/10.1128/JVI.51.2.404-410.1984

128 Ostrove, J. M. & Berns, K. I. Adenovirus early region 1b gene function required for rescue of latent adeno-associated virus. *Virology* **104**, 502-505 (1980). https://doi.org/10.1016/0042-6822(80)90354-2

129 Matsushita, T. *et al.* The adenovirus E1A and E1B19K genes provide a helper function for transfection-based adeno-associated virus vector production. *J Gen Virol* **85**, 2209-2214 (2004). https://doi.org/10.1099/vir.0.79940-0

130 Ward, P., Dean, F. B., O'Donnell, M. E. & Berns, K. I. Role of the adenovirus DNA-binding protein in in vitro adeno-associated virus DNA replication. *J Virol* **72**, 420-427 (1998). https://doi.org/10.1128/JVI.72.1.420-427.1998

131 Carter, B. J., Antoni, B. A. & Klessig, D. F. Adenovirus containing a deletion of the early region 2A gene allows growth of adeno-associated virus with decreased efficiency. *Virology* **191**, 473-476 (1992). https://doi.org/10.1016/0042-6822(92)90213-9

132 Chang, L. S. & Shenk, T. The adenovirus DNA-binding protein stimulates the rate of transcription directed by adenovirus and adeno-associated virus promoters. *J Virol* **64**, 2103-2109 (1990). https://doi.org/10.1128/JVI.64.5.2103-2109.1990

133 Allen, J. M., Halbert, C. L. & Miller, A. D. Improved adeno-associated virus vector production with transfection of a single helper adenovirus gene, E4orf6. *Mol Ther* **1**, 88-95 (2000). https://doi.org/10.1006/mthe.1999.0010

134 Janik, J. E., Huston, M. M., Cho, K. & Rose, J. A. Efficient synthesis of adeno-associated virus structural proteins requires both adenovirus DNA binding protein and VA I RNA. *Virology* **168**, 320-329 (1989). https://doi.org/10.1016/0042-6822(89)90272-9

135 Xiao, X., Li, J. & Samulski, R. J. Production of high-titer recombinant adeno-associated virus vectors in the absence of helper adenovirus. *J Virol* **72**, 2224-2232 (1998). https://doi.org/10.1128/JVI.72.3.2224-2232.1998

136 Grimm, D., Kern, A., Rittner, K. & Kleinschmidt, J. A. Novel tools for production and purification of recombinant adenoassociated virus vectors. *Hum Gene Ther* **9**, 2745-2760 (1998). https://doi.org/10.1089/hum.1998.9.18-2745

137 Matsushita, T. *et al.* Adeno-associated virus vectors can be efficiently produced without helper virus. *Gene Ther* **5**, 938-945 (1998). https://doi.org/10.1038/sj.gt.3300680

138 Louis, N., Evelegh, C. & Graham, F. L. Cloning and sequencing of the cellular-viral junctions from the human adenovirus type 5 transformed 293 cell line. *Virology* **233**, 423-429 (1997). https://doi.org/10.1006/viro.1997.8597

139 Doshi, J., Couto, E., Staiti, J., Vandenberghe, L. H. & Zabaleta, N. E2A, VA RNA I, and L4-22k adenoviral helper genes are sufficient for AAV production in HEK293 cells. *Mol Ther Methods Clin Dev* **32**, 101376 (2024). https://doi.org/10.1016/j.omtm.2024.101376

140 van Lieshout, L. *et al.* An Improved Helper Plasmid Containing Deletions Within the E4 and E2a Genes Results in Increased Adeno-Associated Virus Productivity. *Hum Gene Ther* **35**, 767-776 (2024). https://doi.org/10.1089/hum.2024.059

141 Clark, K. R., Voulgaropoulou, F. & Johnson, P. R. A stable cell line carrying adenovirus-inducible rep and cap genes allows for infectivity titration of adeno-associated virus vectors. *Gene Ther* **3**, 1124-1132 (1996).

142 Gao, G. P. *et al.* High-titer adeno-associated viral vectors from a Rep/Cap cell line and hybrid shuttle virus. *Hum Gene Ther* **9**, 2353-2362 (1998). https://doi.org/10.1089/hum.1998.9.16-2353

143 Qiao, C., Wang, B., Zhu, X., Li, J. & Xiao, X. A novel gene expression control system and its use in stable, high-titer 293 cell-based adeno-associated virus packaging cell lines. *J Virol* **76**, 13015-13027 (2002). https://doi.org/10.1128/jvi.76.24.13015-13027.2002

144 Barajas, D. *et al.* Generation of infectious recombinant Adeno-associated virus in Saccharomyces cerevisiae. *PLoS One* **12**, e0173010 (2017). https://doi.org/10.1371/journal.pone.0173010

145 Cervelli, T., Backovic, A. & Galli, A. Formation of AAV single stranded DNA genome from a circular plasmid in Saccharomyces cerevisiae. *PLoS One* **6**, e23474 (2011). https://doi.org/10.1371/journal.pone.0023474

146 Backovic, A. *et al.* Capsid protein expression and adeno-associated virus like particles assembly in Saccharomyces cerevisiae. *Microb Cell Fact* **11**, 124 (2012). https://doi.org/10.1186/1475-2859-11-124

147 Urabe, M., Ding, C. & Kotin, R. M. Insect cells as a factory to produce adeno-associated virus type 2 vectors. *Hum Gene Ther* **13**, 1935-1943 (2002). https://doi.org/10.1089/10430340260355347

148 Galli, A. *et al.* Characterization of Viral Genome Encapsidated in Adeno-associated Recombinant Vectors Produced in Yeast Saccharomyces cerevisiae. *Mol Biotechnol* **63**, 156-165 (2021). https://doi.org/10.1007/s12033-020-00294-4

149 Wu, J., Liu, X., Yang, H., He, Y. & Yu, D. Advances in biopharmaceutical products for hemophilia. *iScience* **27**, 111436 (2024). https://doi.org/10.1016/j.isci.2024.111436

150 Flotte, T. R. *et al.* Stable in vivo expression of the cystic fibrosis transmembrane conductance regulator with an adeno-associated virus vector. *Proc Natl Acad Sci U S A* **90**, 10613-10617 (1993). https://doi.org/10.1073/pnas.90.22.10613

151 Flotte, T. *et al.* A phase I study of an adeno-associated virus-CFTR gene vector in adult CF patients with mild lung disease. *Hum Gene Ther* **7**, 1145-1159 (1996). https://doi.org/10.1089/hum.1996.7.9-1145

152 Flotte, T. R. *et al.* Phase I trial of intranasal and endobronchial administration of a recombinant adeno-associated virus serotype 2 (rAAV2)-CFTR vector in adult cystic fibrosis patients: a two-part clinical study. *Hum Gene Ther* **14**, 1079-1088 (2003). https://doi.org/10.1089/104303403322124792

153 Wagner, J. A. *et al.* Safety and biological efficacy of an adeno-associated virus vector-cystic fibrosis transmembrane regulator (AAV-CFTR) in the cystic fibrosis maxillary sinus. *Laryngoscope* **109**, 266-274 (1999). https://doi.org/10.1097/00005537-199902000-00017

154 Gaudet, D. *et al.* Efficacy and long-term safety of alipogene tiparvovec (AAV1-LPLS447X) gene therapy for lipoprotein lipase deficiency: an open-label trial. *Gene Ther* **20**, 361-369 (2013). https://doi.org/10.1038/gt.2012.43

155 Morrison, C. $1-million price tag set for Glybera gene therapy. *Nat Biotechnol* **33**, 217-218 (2015). https://doi.org/10.1038/nbt0315-217

156 Senior, M. After Glybera's withdrawal, what's next for gene therapy? *Nat Biotechnol* **35**, 491-492 (2017). https://doi.org/10.1038/nbt0617-491

157 *FDA Approved Cellular and Gene Therapy Products (Access Date: 18.04.2025)*, <https://www.fda.gov/vaccines-blood-biologics/cellular-gene-therapy-products/approved-cellular-and-gene-therapy-products> (

158 *Paul Ehrlich Insititut: EMA approved Gene Therapy Medicinal Products (Access Date: 18.04.2025)*, <https://www.pei.de/EN/medicinal-products/atmp/gene-therapy-medicinal-products/gene-therapy-node.html> (

159 Prasad, S. *et al.* Immune Responses and Immunosuppressive Strategies for Adeno-Associated Virus-Based Gene Therapy for Treatment of Central Nervous System Disorders: Current Knowledge and Approaches. *Hum Gene Ther* **33**, 1228-1245 (2022). https://doi.org/10.1089/hum.2022.138

160 Braun, M. *et al.* Preexisting antibody assays for gene therapy: Considerations on patient selection cutoffs and companion diagnostic requirements. *Mol Ther Methods Clin Dev* **32**, 101217 (2024). https://doi.org/10.1016/j.omtm.2024.101217

161 Chan, Y. K. *et al.* Inflammation in Viral Vector-Mediated Ocular Gene Therapy: A Review and Report From a Workshop Hosted by the Foundation Fighting Blindness, 9/2020. *Transl Vis Sci Technol* **10**, 3 (2021). https://doi.org/10.1167/tvst.10.4.3

162 Shieh, P. B. *et al.* Re: "Moving Forward After Two Deaths in a Gene Therapy Trial of Myotubular Myopathy" by Wilson and Flotte. *Hum Gene Ther* **31**, 787 (2020). https://doi.org/10.1089/hum.2020.217

163 Lek, A. *et al.* Death after High-Dose rAAV9 Gene Therapy in a Patient with Duchenne's Muscular Dystrophy. *N Engl J Med* **389**, 1203-1210 (2023). https://doi.org/10.1056/NEJMoa2307798

164 Philippidis, A. Patient Dies in Beam Trial of Sickle Cell Disease Candidate; Company Cites Conditioning. *Hum Gene Ther* **35**, 951-954 (2024). https://doi.org/10.1089/hum.2024.111924

165 Wilson, J. M. & Flotte, T. R. Moving Forward After Two Deaths in a Gene Therapy Trial of Myotubular Myopathy. *Hum Gene Ther* **31**, 695-696 (2020). https://doi.org/10.1089/hum.2020.182

166 Zhong, L. *et al.* Tyrosine-phosphorylation of AAV2 vectors and its consequences on viral intracellular trafficking and transgene expression. *Virology* **381**, 194-202 (2008). https://doi.org/10.1016/j.virol.2008.08.027

167 Kanaan, N. M. *et al.* Rationally Engineered AAV Capsids Improve Transduction and Volumetric Spread in the CNS. *Mol Ther Nucleic Acids* **8**, 184-197 (2017). https://doi.org/10.1016/j.omtn.2017.06.011

168 Tabebordbar, M. *et al.* Directed evolution of a family of AAV capsid variants enabling potent muscle-directed gene delivery across species. *Cell* **184**, 4919-4938 e4922 (2021). https://doi.org/10.1016/j.cell.2021.08.028

169 Bryant, D. H. *et al.* Deep diversification of an AAV capsid protein by machine learning. *Nat Biotechnol* **39**, 691-696 (2021). https://doi.org/10.1038/s41587-020-00793-4

170 Chang, H. *et al.* Non-canonical amino acid incorporation into AAV5 capsid enhances lung transduction in mice. *Mol Ther Methods Clin Dev* **31**, 101129 (2023). https://doi.org/10.1016/j.omtm.2023.101129

171 Kelemen, R. E. *et al.* A Precise Chemical Strategy To Alter the Receptor Specificity of the Adeno-Associated Virus. *Angew Chem Int Ed Engl* **55**, 10645-10649 (2016). https://doi.org/10.1002/anie.201604067

172 Yao, T. *et al.* Site-Specific PEGylated Adeno-Associated Viruses with Increased Serum Stability and Reduced Immunogenicity. *Molecules* **22** (2017). https://doi.org/10.3390/molecules22071155

173 Ponnazhagan, S., Mahendra, G., Kumar, S., Thompson, J. A. & Castillas, M., Jr. Conjugate-based targeting of recombinant adeno-associated virus type 2 vectors by using avidin-linked ligands. *J Virol* **76**, 12900-12907 (2002). https://doi.org/10.1128/jvi.76.24.12900-12907.2002

174 Puzzo, F. *et al.* Aptamer-programmable adeno-associated viral vectors as a novel platform for cell-specific gene transfer. *Mol Ther Nucleic Acids* **31**, 383-397 (2023). https://doi.org/10.1016/j.omtn.2023.01.007

175 Judd, J. *et al.* Random Insertion of mCherry Into VP3 Domain of Adeno-associated Virus Yields Fluorescent Capsids With no Loss of Infectivity. *Mol Ther Nucleic Acids* **1**, e54 (2012). https://doi.org/10.1038/mtna.2012.46

176 Koerber, J. T., Jang, J. H., Yu, J. H., Kane, R. S. & Schaffer, D. V. Engineering adeno-associated virus for one-step purification via immobilized metal affinity chromatography. *Hum Gene Ther* **18**, 367-378 (2007). https://doi.org/10.1089/hum.2006.139

177 Ried, M. U., Girod, A., Leike, K., Buning, H. & Hallek, M. Adeno-associated virus capsids displaying immunoglobulin-binding domains permit antibody-mediated vector retargeting to specific cell surface receptors. *J Virol* **76**, 4559-4566 (2002). https://doi.org/10.1128/jvi.76.9.4559-4566.2002

178 Zdechlik, A. C., He, Y., Aird, E. J., Gordon, W. R. & Schmidt, D. Programmable Assembly of Adeno-Associated Virus-Antibody Composites for Receptor-Mediated Gene Delivery. *Bioconjug Chem* **31**, 1093-1106 (2020). https://doi.org/10.1021/acs.bioconjchem.9b00790

179 Hoffmann, M. D. *et al.* Multiparametric domain insertional profiling of adeno-associated virus VP1. *Mol Ther Methods Clin Dev* **31**, 101143 (2023). https://doi.org/10.1016/j.omtm.2023.101143

180 Mevel, M. *et al.* Mannose-coupled AAV2: A second-generation AAV vector for increased retinal gene therapy efficiency. *Mol Ther Methods Clin Dev* **32**, 101187 (2024). https://doi.org/10.1016/j.omtm.2024.101187

181 Mevel, M. *et al.* Chemical modification of the adeno-associated virus capsid to improve gene delivery. *Chem Sci* **11**, 1122-1131 (2019). https://doi.org/10.1039/c9sc04189c

182 Leray, A. *et al.* Novel chemical tyrosine functionalization of adeno-associated virus improves gene transfer efficiency in liver and retina. *Biomed Pharmacother* **171**, 116148 (2024). https://doi.org/10.1016/j.biopha.2024.116148

183 Wu, Y. *et al.* Enhanced Targeted Gene Transduction: AAV2 Vectors Conjugated to Multiple Aptamers via Reducible Disulfide Linkages. *J Am Chem Soc* **140**, 2-5 (2018). https://doi.org/10.1021/jacs.7b08518

184 Hoffmann, M. D., Gallant, J. P., LeBeau, A. M. & Schmidt, D. Unlocking precision gene therapy: harnessing AAV tropism with nanobody swapping at capsid hotspots. *NAR Mol Med* **1**, ugae008 (2024). https://doi.org/10.1093/narmme/ugae008

185 Eichhoff, A. M. *et al.* Nanobody-Enhanced Targeting of AAV Gene Therapy Vectors. *Mol Ther Methods Clin Dev* **15**, 211-220 (2019). https://doi.org/10.1016/j.omtm.2019.09.003

186 Zhang, Y. *et al.* Site-specific tethering nanobodies on recombinant adeno-associated virus vectors for retargeted gene therapy. *Acta Biomater* **187**, 304-315 (2024). https://doi.org/10.1016/j.actbio.2024.07.023

187 Olarewaju, O. *et al.* alphaFAP-specific nanobodies mediate a highly precise retargeting of modified AAV2 capsids thereby enabling specific transduction of tumor tissues. *Mol Ther Methods Clin Dev* **32**, 101378 (2024). https://doi.org/10.1016/j.omtm.2024.101378

188 Li, W. *et al.* Engineering and Selection of Shuffled AAV Genomes: A New Strategy for Producing Targeted Biological Nanoparticles. *Mol Ther* **16**, 1252-1260 (2008). https://doi.org/10.1038/mt.2008.100

189 El Andari, J. *et al.* Semirational bioengineering of AAV vectors with increased potency and specificity for systemic gene therapy of muscle disorders. *Sci Adv* **8**, eabn4704 (2022). https://doi.org/10.1126/sciadv.abn4704

190 Weinmann, J. *et al.* Identification of a myotropic AAV by massively parallel in vivo evaluation of barcoded capsid variants. *Nat Commun* **11**, 5432 (2020). https://doi.org/10.1038/s41467-020-19230-w

191 Kim, J. H. *et al.* Bap1/SMN axis in Dpp4+ skeletal muscle mesenchymal cells regulates the neuromuscular system. *JCI Insight* **7** (2022). https://doi.org/10.1172/jci.insight.158380

192 Rhee, J. *et al.* Improved therapeutic approach for spinal muscular atrophy via ubiquitination-resistant survival motor neuron variant. *J Cachexia Sarcopenia Muscle* **15**, 1404-1417 (2024). https://doi.org/10.1002/jcsm.13486

193 Seita, Y. *et al.* Comprehensive evaluation of ubiquitous promoters suitable for the generation of transgenic cynomolgus monkeysdagger. *Biol Reprod* **100**, 1440-1452 (2019). https://doi.org/10.1093/biolre/ioz040

194 Brooks, A. R. *et al.* Transcriptional silencing is associated with extensive methylation of the CMV promoter following adenoviral gene delivery to muscle. *J Gene Med* **6**, 395-404 (2004). https://doi.org/10.1002/jgm.516

195 Cordier, L. *et al.* Muscle-specific promoters may be necessary for adeno-associated virus-mediated gene transfer in the treatment of muscular dystrophies. *Hum Gene Ther* **12**, 205-215 (2001). https://doi.org/10.1089/104303401750061267

196 Chan, Y. C. *et al.* An unbiased AAV-STARR-seq screen revealing the enhancer activity map of genomic regions in the mouse brain in vivo. *Sci Rep* **13**, 6745 (2023). https://doi.org/10.1038/s41598-023-33448-w

197 Chan, Y. K. *et al.* Engineering adeno-associated viral vectors to evade innate immune and inflammatory responses. *Sci Transl Med* **13** (2021). https://doi.org/10.1126/scitranslmed.abd3438

198 Hagedorn, C. *et al.* S/MAR Element Facilitates Episomal Long-Term Persistence of Adeno-Associated Virus Vector Genomes in Proliferating Cells. *Hum Gene Ther* **28**, 1169-1179 (2017). https://doi.org/10.1089/hum.2017.025

199 Gao, Q. Q. & McNally, E. M. The Dystrophin Complex: Structure, Function, and Implications for Therapy. *Compr Physiol* **5**, 1223-1239 (2015). https://doi.org/10.1002/cphy.c140048

200 Zaidman, C. M. *et al.* Delandistrogene Moxeparvovec Gene Therapy in Ambulatory Patients (Aged >/=4 to <8 Years) with Duchenne Muscular Dystrophy: 1-Year Interim Results from Study SRP-9001-103 (ENDEAVOR). *Ann Neurol* **94**, 955-968 (2023). https://doi.org/10.1002/ana.26755

201 Hart, C. C. *et al.* Potential limitations of microdystrophin gene therapy for Duchenne muscular dystrophy. *JCI Insight* **9** (2024). https://doi.org/10.1172/jci.insight.165869

202 Yan, Z., Zhang, Y., Duan, D. & Engelhardt, J. F. Trans-splicing vectors expand the utility of adeno-associated virus for gene therapy. *Proc Natl Acad Sci U S A* **97**, 6716-6721 (2000). https://doi.org/10.1073/pnas.97.12.6716

203 Riedmayr, L. M. *et al.* mRNA trans-splicing dual AAV vectors for (epi)genome editing and gene therapy. *Nat Commun* **14**, 6578 (2023). https://doi.org/10.1038/s41467-023-42386-0

204 Xu, Z. *et al.* Trans-splicing adeno-associated viral vector-mediated gene therapy is limited by the accumulation of spliced mRNA but not by dual vector coinfection efficiency. *Hum Gene Ther* **15**, 896-905 (2004). https://doi.org/10.1089/hum.2004.15.896

205 Tasfaout, H. *et al.* Split intein-mediated protein trans-splicing to express large dystrophins. *Nature* **632**, 192-200 (2024). https://doi.org/10.1038/s41586-024-07710-8

206 Martins, K. M. *et al.* Prevalent and Disseminated Recombinant and Wild-Type Adeno-Associated Virus Integration in Macaques and Humans. *Hum Gene Ther* **34**, 1081-1094 (2023). https://doi.org/10.1089/hum.2023.134

207 Cukor, G., Blacklow, N. R., Kibrick, S. & Swan, I. C. Effect of adeno-associated virus on cancer expression by herpesvirus-transformed hamster cells. *J Natl Cancer Inst* **55**, 957-959 (1975). https://doi.org/10.1093/jnci/55.4.957

208 Ostrove, J. M., Duckworth, D. H. & Berns, K. I. Inhibition of adenovirus-transformed cell oncogenicity by adeno-associated virus. *Virology* **113**, 521-533 (1981). https://doi.org/10.1016/0042-6822(81)90180-x

209 Mayor, H. D., Houlditch, G. S. & Mumford, D. M. Influence of adeno-associated satellite virus on adenovirus-induced tumours in hamsters. *Nat New Biol* **241**, 44-46 (1973). https://doi.org/10.1038/newbio241044b0

210 Nault, J. C. *et al.* Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat Genet* **47**, 1187-1193 (2015). https://doi.org/10.1038/ng.3389

211 Berns, K. I. *et al.* Adeno-Associated Virus Type 2 and Hepatocellular Carcinoma? *Hum Gene Ther* **26**, 779-781 (2015). https://doi.org/10.1089/hum.2015.29014.kib

212 Buning, H. & Schmidt, M. Adeno-associated Vector Toxicity-To Be or Not to Be? *Mol Ther* **23**, 1673-1675 (2015). https://doi.org/10.1038/mt.2015.182

213 Donsante, A. *et al.* Observed incidence of tumorigenesis in long-term rodent studies of rAAV vectors. *Gene Ther* **8**, 1343-1346 (2001). https://doi.org/10.1038/sj.gt.3301541

214 Miller, D. G. *et al.* Large-scale analysis of adeno-associated virus vector integration sites in normal human cells. *J Virol* **79**, 11434-11442 (2005). https://doi.org/10.1128/JVI.79.17.11434-11442.2005

215 Nakai, H. *et al.* Large-scale molecular characterization of adeno-associated virus vector integration in mouse liver. *J Virol* **79**, 3606-3614 (2005). https://doi.org/10.1128/JVI.79.6.3606-3614.2005

216 Donsante, A. *et al.* AAV vector integration sites in mouse hepatocellular carcinoma. *Science* **317**, 477 (2007). https://doi.org/10.1126/science.1142658

217 Chandler, R. J. *et al.* Vector design influences hepatic genotoxicity after adeno-associated virus gene therapy. *J Clin Invest* **125**, 870-880 (2015). https://doi.org/10.1172/JCI79213

218 Ferla, R. *et al.* Low incidence of hepatocellular carcinoma in mice and cats treated with systemic adeno-associated viral vectors. *Mol Ther Methods Clin Dev* **20**, 247-257 (2021). https://doi.org/10.1016/j.omtm.2020.11.015

219 Nguyen, G. N. *et al.* Long-Term AAV-Mediated Factor VIII Expression in Nine Hemophilia A Dogs: A 10 Year Follow-up Analysis on Durability, Safety and Vector Integration. *Blood* **134**, 611 (2019,). https://doi.org/https://doi.org/10.1182/blood-2019-126007.

220 Schmidt, M. *et al.* Molecular evaluation and vector integration analysis of HCC complicating AAV gene therapy for hemophilia B. *Blood Adv* **7**, 4966-4969 (2023). https://doi.org/10.1182/bloodadvances.2023009876

221 Nguyen, G. N. *et al.* A long-term study of AAV gene therapy in dogs with hemophilia A identifies clonal expansions of transduced liver cells. *Nat Biotechnol* **39**, 47-55 (2021). https://doi.org/10.1038/s41587-020-0741-7

222 Kasimsetty, A. *et al.* Modulation of AAV transduction and integration targeting by topoisomerase poisons. *Mol Ther Methods Clin Dev* **32**, 101364 (2024). https://doi.org/10.1016/j.omtm.2024.101364

223 Cataldi, M. P. & McCarty, D. M. Differential effects of DNA double-strand break repair pathways on single-strand and self-complementary adeno-associated virus vector genomes. *J Virol* **84**, 8673-8682 (2010). https://doi.org/10.1128/JVI.00641-10

224 Gomez-Moreno, A. *et al.* Topoisomerase Inhibitors Increase Episomal DNA Expression by Inducing the Integration of Episomal DNA in Hepatic Cells. *Pharmaceutics* **15** (2023). https://doi.org/10.3390/pharmaceutics15102459

225 Song, S. *et al.* DNA-dependent PK inhibits adeno-associated virus DNA integration. *Proc Natl Acad Sci U S A* **101**, 2112-2116 (2004). https://doi.org/10.1073/pnas.0307833100

226 Laugel, M. *et al.* The Diversity of Parvovirus Telomeres. *Recent Advances in Canine Medicine* (2022). https://doi.org/DOI: 10.5772/intechopen.102684

227 Tattersall, P. & Cotmore, S. F. Parvoviruses. *In eLS, (Ed.)* (2003). https://doi.org/https://doi.org/10.1038/npg.els.0000423

228 Penzes, J. J. *et al.* Reorganizing the family Parvoviridae: a revised taxonomy independent of the canonical approach based on host association. *Arch Virol* **165**, 2133-2146 (2020). https://doi.org/10.1007/s00705-020-04632-4

229 Satkunanathan, S., Thorpe, R. & Zhao, Y. The function of DNA binding protein nucleophosmin in AAV replication. *Virology* **510**, 46-54 (2017). https://doi.org/10.1016/j.virol.2017.07.007

230 Shu, H., Zhang, R., Xiao, K., Yang, J. & Sun, X. G-Quadruplex-Binding Proteins: Promising Targets for Drug Design. *Biomolecules* **12** (2022). https://doi.org/10.3390/biom12050648

231 Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**, 8627-8637 (2015). https://doi.org/10.1093/nar/gkv862

232 Guan, W., Wong, S., Zhi, N. & Qiu, J. The genome of human parvovirus b19 can replicate in nonpermissive cells with the help of adenovirus genes and produces infectious virus. *J Virol* **83**, 9541-9553 (2009). https://doi.org/10.1128/JVI.00702-09

233 Baquerizo-Audiot, E. *et al.* Structure and expression strategy of the genome of Culex pipiens densovirus, a mosquito densovirus with an ambisense organization. *J Virol* **83**, 6863-6873 (2009). https://doi.org/10.1128/JVI.00524-09

234 Jackson, E. W., Pepe-Ranney, C., Johnson, M. R., Distel, D. L. & Hewson, I. A Highly Prevalent and Pervasive Densovirus Discovered among Sea Stars from the North American Atlantic Coast. *Appl Environ Microbiol* **86** (2020). https://doi.org/10.1128/AEM.02723-19

235 Day, J. M. & Zsak, L. Determination and analysis of the full-length chicken parvovirus genome. *Virology* **399**, 59-64 (2010). https://doi.org/10.1016/j.virol.2009.12.027

236 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003). https://doi.org/10.1093/nar/gkg595

237 Chiorini, J. A., Afione, S. & Kotin, R. M. Adeno-associated virus (AAV) type 5 Rep protein cleaves a unique terminal resolution site compared with other AAV serotypes. *J Virol* **73**, 4293-4298 (1999). https://doi.org/10.1128/JVI.73.5.4293-4298.1999

238 Earley, L. F. *et al.* Adeno-Associated Virus Serotype-Specific Inverted Terminal Repeat Sequence Role in Vector Transgene Expression. *Hum Gene Ther* **31**, 151-162 (2020). https://doi.org/10.1089/hum.2019.274

239 Wilmott, P., Lisowski, L., Alexander, I. E. & Logan, G. J. A User's Guide to the Inverted Terminal Repeats of Adeno-Associated Virus. *Hum Gene Ther Methods* **30**, 206-213 (2019). https://doi.org/10.1089/hgtb.2019.276

240 Mroske, C., Rivera, H., Ul-Hasan, T., Chatterjee, S. & Wong, K. K. A capillary electrophoresis sequencing method for the identification of mutations in the inverted terminal repeats of adeno-associated virus. *Hum Gene Ther Methods* **23**, 128-136 (2012). https://doi.org/10.1089/hgtb.2011.231

241 Pan, X. *et al.* Rational engineering of a functional CpG-free ITR for AAV gene therapy. *Gene Ther* **29**, 333-345 (2022). https://doi.org/10.1038/s41434-021-00296-0

242 Shitik, E. M., Shalik, I. K. & Yudkin, D. V. AAV- based vector improvements unrelated to capsid protein modification. *Front Med (Lausanne)* **10**, 1106085 (2023). https://doi.org/10.3389/fmed.2023.1106085

243 Cavalier-Smith, T. Palindromic base sequences and replication of eukaryote chromosome ends. *Nature* **250**, 467-470 (1974). https://doi.org/10.1038/250467a0

244 Spear, I. S., Fife, K. H., Hauswirth, W. W., Jones, C. J. & Berns, K. I. Evidence for two nucleotide sequence orientations within the terminal repetition of adeno-associated virus DNA. *J Virol* **24**, 627-634 (1977). https://doi.org/10.1128/JVI.24.2.627-634.1977

245 Snyder, R. O., Im, D. S. & Muzyczka, N. Evidence for covalent attachment of the adeno-associated virus (AAV) rep protein to the ends of the AAV genome. *J Virol* **64**, 6204-6213 (1990). https://doi.org/10.1128/JVI.64.12.6204-6213.1990

246 Brister, J. R. & Muzyczka, N. Rep-mediated nicking of the adeno-associated virus origin requires two biochemical activities, DNA helicase activity and transesterification. *J Virol* **73**, 9325-9336 (1999). https://doi.org/10.1128/JVI.73.11.9325-9336.1999

247 Brister, J. R. & Muzyczka, N. Mechanism of Rep-mediated adeno-associated virus origin nicking. *J Virol* **74**, 7762-7771 (2000). https://doi.org/10.1128/jvi.74.17.7762-7771.2000

248 Ling, C. *et al.* Enhanced transgene expression from recombinant single-stranded D-sequence-substituted adeno-associated virus vectors in human cell lines in vitro and in murine hepatocytes in vivo. *J Virol* **89**, 952-961 (2015). https://doi.org/10.1128/JVI.02581-14

249 Im, D. S. & Muzyczka, N. The AAV origin binding protein Rep68 is an ATP-dependent site-specific endonuclease with DNA helicase activity. *Cell* **61**, 447-457 (1990). https://doi.org/10.1016/0092-8674(90)90526-k

250 Snyder, R. O., Samulski, R. J. & Muzyczka, N. In vitro resolution of covalently joined AAV chromosome ends. *Cell* **60**, 105-113 (1990). https://doi.org/10.1016/0092-8674(90)90720-y

251 Grimm, D. *et al.* Preclinical in vivo evaluation of pseudotyped adeno-associated virus vectors for liver gene therapy. *Blood* **102**, 2412-2419 (2003). https://doi.org/10.1182/blood-2003-02-0495

252 Chiorini, J. A., Kim, F., Yang, L. & Kotin, R. M. Cloning and characterization of adeno-associated virus type 5. *J Virol* **73**, 1309-1319 (1999). https://doi.org/10.1128/JVI.73.2.1309-1319.1999

253 Logan, G. J. *et al.* Identification of liver-specific enhancer-promoter activity in the 3' untranslated region of the wild-type AAV2 genome. *Nat Genet* **49**, 1267-1273 (2017). https://doi.org/10.1038/ng.3893

254 Radukic, M. T. *et al.* Degradation and stable maintenance of adeno-associated virus inverted terminal repeats in E. coli. *Nucleic Acids Res* **53** (2025). https://doi.org/10.1093/nar/gkae1170

255 Chen, Y. *et al.* A Comprehensive Study of the Effects by Sequence Truncation within Inverted Terminal Repeats (ITRs) on the Productivity, Genome Packaging, and Potency of AAV Vectors. *Microorganisms* **12** (2024). https://doi.org/10.3390/microorganisms12020310

256 Xiao, X., Xiao, W., Li, J. & Samulski, R. J. A novel 165-base-pair terminal repeat sequence is the sole cis requirement for the adeno-associated virus life cycle. *J Virol* **71**, 941-948 (1997). https://doi.org/10.1128/JVI.71.2.941-948.1997

257 McCarty, D. M. *et al.* Adeno-associated virus terminal repeat (TR) mutant generates self-complementary vectors to overcome the rate-limiting step to transduction in vivo. *Gene Ther* **10**, 2112-2118 (2003). https://doi.org/10.1038/sj.gt.3302134

258 McCarty, D. M., Monahan, P. E. & Samulski, R. J. Self-complementary recombinant adeno-associated virus (scAAV) vectors promote efficient transduction independently of DNA synthesis. *Gene Ther* **8**, 1248-1254 (2001). https://doi.org/10.1038/sj.gt.3301514

259 Martino, A. T. *et al.* The genome of self-complementary adeno-associated viral vectors increases Toll-like receptor 9-dependent innate immune responses in the liver. *Blood* **117**, 6459-6468 (2011). https://doi.org/10.1182/blood-2010-10-314518

260 Ogbonmide, T. *et al.* Gene Therapy for Spinal Muscular Atrophy (SMA): A Review of Current Challenges and Safety Considerations for Onasemnogene Abeparvovec (Zolgensma). *Cureus* **15**, e36197 (2023). https://doi.org/10.7759/cureus.36197

261 Chadeuf, G., Ciron, C., Moullier, P. & Salvetti, A. Evidence for encapsidation of prokaryotic sequences during recombinant adeno-associated virus production and their in vivo persistence after vector delivery. *Mol Ther* **12**, 744-753 (2005). https://doi.org/10.1016/j.ymthe.2005.06.003

262 Schnodt, M. *et al.* DNA Minicircle Technology Improves Purity of Adeno-associated Viral Vector Preparations. *Mol Ther Nucleic Acids* **5**, e355 (2016). https://doi.org/10.1038/mtna.2016.60

263 Karbowniczek, K. *et al.* Doggybone™ DNA: an advanced platform for AAV production *Cell Gene Therapy Insights* **3**, 731-738 (2017). https://doi.org/10.18609/cgti.2017.074

264 Lusebrink, J. *et al.* Detection of head-to-tail DNA sequences of human bocavirus in clinical samples. *PLoS One* **6**, e19457 (2011). https://doi.org/10.1371/journal.pone.0019457

265 Zhao, L. *et al.* Keeping all secondary structures of the non-coding region in the circular genome of human bocavirus 2 is important for DNA replication and virus assembly, as revealed by three hetero-recombinant genomic clones. *Emerg Microbes Infect* **8**, 1563-1573 (2019). https://doi.org/10.1080/22221751.2019.1682949

266 Zhang, K. *et al.* Diverse Head-to-Tail Sequences in the Circular Genome of Human Bocavirus Genotype 1 among Children with Acute Respiratory Infections Implied the Switch of Template Chain in the Rolling-Circle Replication Model. *Pathogens* **13** (2024). https://doi.org/10.3390/pathogens13090757

267 Ward, P., Elias, P. & Linden, R. M. Rescue of the adeno-associated virus genome from a plasmid vector: evidence for rescue by replication. *J Virol* **77**, 11480-11490 (2003). https://doi.org/10.1128/jvi.77.21.11480-11490.2003

268 Bennett, A., Mietzsch, M. & Agbandje-McKenna, M. Understanding capsid assembly and genome packaging for adeno-associated viruses. *Future Virol* **12**, 283-297 (2017). https://doi.org/10.2217/fvl-2017-0011

269 Wang, X. S., Ponnazhagan, S. & Srivastava, A. Rescue and replication of adeno-associated virus type 2 as well as vector DNA sequences from recombinant plasmids containing deletions in the viral inverted terminal repeats: selective encapsidation of viral genomes in progeny virions. *J Virol* **70**, 1668-1677 (1996). https://doi.org/10.1128/JVI.70.3.1668-1677.1996

270 Samulski, R. J., Srivastava, A., Berns, K. I. & Muzyczka, N. Rescue of adeno-associated virus from recombinant plasmids: gene correction within the terminal repeats of AAV. *Cell* **33**, 135-143 (1983). https://doi.org/10.1016/0092-8674(83)90342-2

271 Luo, Y. *et al.* Parvovirus B19 infection of human primary erythroid progenitor cells triggers ATR-Chk1 signaling, which promotes B19 virus replication. *J Virol* **85**, 8046-8055 (2011). https://doi.org/10.1128/JVI.00831-11

272 Adeyemi, R. O., Landry, S., Davis, M. E., Weitzman, M. D. & Pintel, D. J. Parvovirus minute virus of mice induces a DNA damage response that facilitates viral replication. *PLoS Pathog* **6**, e1001141 (2010). https://doi.org/10.1371/journal.ppat.1001141

273 Luo, Y., Chen, A. Y. & Qiu, J. Bocavirus infection induces a DNA damage response that facilitates viral DNA replication and mediates cell death. *J Virol* **85**, 133-145 (2011). https://doi.org/10.1128/JVI.01534-10

274 Saudan, P., Vlach, J. & Beard, P. Inhibition of S-phase progression by adeno-associated virus Rep78 protein is mediated by hypophosphorylated pRb. *EMBO J* **19**, 4351-4361 (2000). https://doi.org/10.1093/emboj/19.16.4351

275 Berthet, C., Raj, K., Saudan, P. & Beard, P. How adeno-associated virus Rep78 protein arrests cells completely in S phase. *Proc Natl Acad Sci U S A* **102**, 13634-13639 (2005). https://doi.org/10.1073/pnas.0504583102

276 Alexander, I. E., Russell, D. W. & Miller, A. D. DNA-damaging agents greatly increase the transduction of nondividing cells by adeno-associated virus vectors. *J Virol* **68**, 8282-8287 (1994). https://doi.org/10.1128/JVI.68.12.8282-8287.1994

277 Zentilin, L., Marcello, A. & Giacca, M. Involvement of cellular double-stranded DNA break binding proteins in processing of the recombinant adeno-associated virus genome. *J Virol* **75**, 12279-12287 (2001). https://doi.org/10.1128/JVI.75.24.12279-12287.2001

278 Maurer, A. C. *et al.* Double-strand break repair pathways differentially affect processing and transduction by dual AAV vectors. *Nat Commun* **16**, 1532 (2025). https://doi.org/10.1038/s41467-025-56738-5

279 Jurvansuu, J., Raj, K., Stasiak, A. & Beard, P. Viral transport of DNA damage that mimics a stalled replication fork. *J Virol* **79**, 569-580 (2005). https://doi.org/10.1128/JVI.79.1.569-580.2005

280 Raj, K., Ogston, P. & Beard, P. Virus-mediated killing of cells that lack p53 activity. *Nature* **412**, 914-917 (2001). https://doi.org/10.1038/35091082

281 Jurvansuu, J., Fragkos, M., Ingemarsdotter, C. & Beard, P. Chk1 instability is coupled to mitotic cell death of p53-deficient cells in response to virus-induced DNA damage signaling. *J Mol Biol* **372**, 397-406 (2007). https://doi.org/10.1016/j.jmb.2007.06.077

282 Fragkos, M., Breuleux, M., Clement, N. & Beard, P. Recombinant adeno-associated viral vectors are deficient in provoking a DNA damage response. *J Virol* **82**, 7379-7387 (2008). https://doi.org/10.1128/JVI.00358-08

283 Yan, Z. *et al.* Rad52 Restrains Resection at DNA Double-Strand Break Ends in Yeast. *Mol Cell* **76**, 699-711 e696 (2019). https://doi.org/10.1016/j.molcel.2019.08.017

284 Griffith, A. J., Blier, P. R., Mimori, T. & Hardin, J. A. Ku polypeptides synthesized in vitro assemble into complexes which recognize ends of double-stranded DNA. *J Biol Chem* **267**, 331-338 (1992).

285 Van Dyck, E., Stasiak, A. Z., Stasiak, A. & West, S. C. Binding of double-strand breaks in DNA by human Rad52 protein. *Nature* **398**, 728-731 (1999). https://doi.org/10.1038/19560

286 Yong, W. *et al.* Essential role for Co-chaperone Fkbp52 but not Fkbp51 in androgen receptor-mediated signaling and physiology. *J Biol Chem* **282**, 5026-5036 (2007). https://doi.org/10.1074/jbc.M609360200

287 Wallace, N. C. FKBP52 and its role in DNA damage repair; PhD Thesis. *School of Biomedical Sciences, Queensland University of Technology* (2017). https://doi.org/10.5204/thesis.eprints.112769

288 Zhao, W. *et al.* Role of cellular FKBP52 protein in intracellular trafficking of recombinant adeno-associated virus 2 vectors. *Virology* **353**, 283-293 (2006). https://doi.org/10.1016/j.virol.2006.04.042

289 Qing, K. *et al.* Adeno-associated virus type 2-mediated gene transfer: role of cellular FKBP52 protein in transgene expression. *J Virol* **75**, 8968-8976 (2001). https://doi.org/10.1128/JVI.75.19.8968-8976.2001

290 Schwartz, R. A. *et al.* The Mre11/Rad50/Nbs1 complex limits adeno-associated virus transduction and replication. *J Virol* **81**, 12936-12945 (2007). https://doi.org/10.1128/JVI.01523-07

291 Lentz, T. B. & Samulski, R. J. Insight into the mechanism of inhibition of adeno-associated virus by the Mre11/Rad50/Nbs1 complex. *J Virol* **89**, 181-194 (2015). https://doi.org/10.1128/JVI.01990-14

292 Zhou, Q. *et al.* Deletion of the B-B' and C-C' regions of inverted terminal repeats reduces rAAV productivity but increases transgene expression. *Sci Rep* **7**, 5432 (2017). https://doi.org/10.1038/s41598-017-04054-4

293 Shibata, A. & Jeggo, P. A. ATM's Role in the Repair of DNA Double-Strand Breaks. *Genes (Basel)* **12** (2021). https://doi.org/10.3390/genes12091370

294 Sanlioglu, S., Benson, P. & Engelhardt, J. F. Loss of ATM function enhances recombinant adeno-associated virus transduction and integration through pathways similar to UV irradiation. *Virology* **268**, 68-78 (2000). https://doi.org/10.1006/viro.1999.0137

295 Cataldi, M. P. & McCarty, D. M. Hairpin-end conformation of adeno-associated virus genome determines interactions with DNA-repair pathways. *Gene Ther* **20**, 686-693 (2013). https://doi.org/10.1038/gt.2012.86

296 Song, S., Laipis, P. J., Berns, K. I. & Flotte, T. R. Effect of DNA-dependent protein kinase on the molecular fate of the rAAV2 genome in skeletal muscle. *Proc Natl Acad Sci U S A* **98**, 4084-4088 (2001). https://doi.org/10.1073/pnas.061014598

297 Duan, D., Yue, Y. & Engelhardt, J. F. Consequences of DNA-dependent protein kinase catalytic subunit deficiency on recombinant adeno-associated virus genome circularization and heterodimerization in muscle tissue. *J Virol* **77**, 4751-4759 (2003). https://doi.org/10.1128/jvi.77.8.4751-4759.2003

298 Inagaki, K., Ma, C., Storm, T. A., Kay, M. A. & Nakai, H. The role of DNA-PKcs and artemis in opening viral DNA hairpin termini in various tissues in mice. *J Virol* **81**, 11304-11321 (2007). https://doi.org/10.1128/JVI.01225-07

299 Nakai, H., Storm, T. A., Fuess, S. & Kay, M. A. Pathways of removal of free DNA vector ends in normal and DNA-PKcs-deficient SCID mouse hepatocytes transduced with rAAV vectors. *Hum Gene Ther* **14**, 871-881 (2003). https://doi.org/10.1089/104303403765701169

300 Ning, K. *et al.* Inhibition of DNA-dependent protein kinase catalytic subunit boosts rAAV transduction of polarized human airway epithelium. *Mol Ther Methods Clin Dev* **31**, 101115 (2023). https://doi.org/10.1016/j.omtm.2023.101115

301 Hirsch, M. L. *et al.* Viral single-strand DNA induces p53-dependent apoptosis in human embryonic stem cells. *PLoS One* **6**, e27520 (2011). https://doi.org/10.1371/journal.pone.0027520

302 Fragkos, M. & Beard, P. Mitotic catastrophe occurs in the absence of apoptosis in p53-null cells with a defective G1 checkpoint. *PLoS One* **6**, e22946 (2011). https://doi.org/10.1371/journal.pone.0022946

303 Ingemarsdotter, C., Keller, D. & Beard, P. The DNA damage response to non-replicating adeno-associated virus: Centriole overduplication and mitotic catastrophe independent of the spindle checkpoint. *Virology* **400**, 271-286 (2010). https://doi.org/10.1016/j.virol.2010.02.003

304 Qin, H. *et al.* Regulation of apoptosis and differentiation by p53 in human embryonic stem cells. *J Biol Chem* **282**, 5842-5852 (2007). https://doi.org/10.1074/jbc.M610464200

305 Rapti, K. *et al.* Effectiveness of gene delivery systems for pluripotent and differentiated cells. *Mol Ther Methods Clin Dev* **2**, 14067 (2015). https://doi.org/10.1038/mtm.2014.67

306 Song, L. *et al.* AAV vector transduction restriction and attenuated toxicity in hESCs via a rationally designed inverted terminal repeat. *Nucleic Acids Res* **53** (2025). https://doi.org/10.1093/nar/gkaf013

307 Johnston, S. *et al.* AAV ablates neurogenesis in the adult murine hippocampus. *Elife* **10** (2021). https://doi.org/10.7554/eLife.59291

308 Handa, H., Shiroki, K. & Shimojo, H. Establishment and characterization of KB cell lines latently infected with adeno-associated virus type 1. *Virology* **82**, 84-92 (1977). https://doi.org/10.1016/0042-6822(77)90034-4

309 Berns, K. I., Pinkerton, T. C., Thomas, G. F. & Hoggan, M. D. Detection of adeno-associated virus (AAV)-specific nucleotide sequences in DNA isolated from latently infected Detroit 6 cells. *Virology* **68**, 556-560 (1975). https://doi.org/10.1016/0042-6822(75)90298-6

310 Young, S. M., Jr. & Samulski, R. J. Adeno-associated virus (AAV) site-specific recombination does not require a Rep-dependent origin of replication within the AAV terminal repeat. *Proc Natl Acad Sci U S A* **98**, 13525-13530 (2001). https://doi.org/10.1073/pnas.241508998

311 Meneses, P., Berns, K. I. & Winocour, E. DNA sequence motifs which direct adeno-associated virus site-specific integration in a model system. *J Virol* **74**, 6213-6216 (2000). https://doi.org/10.1128/jvi.74.13.6213-6216.2000

312 Huser, D., Weger, S. & Heilbronn, R. Kinetics and frequency of adeno-associated virus site-specific integration into human chromosome 19 monitored by quantitative real-time PCR. *J Virol* **76**, 7554-7559 (2002). https://doi.org/10.1128/jvi.76.15.7554-7559.2002

313 Huser, D. & Heilbronn, R. Adeno-associated virus integrates site-specifically into human chromosome 19 in either orientation and with equal kinetics and frequency. *J Gen Virol* **84**, 133-137 (2003). https://doi.org/10.1099/vir.0.18726-0

314 Yang, C. C. *et al.* Cellular recombination pathways and viral terminal repeat hairpin structures are sufficient for adeno-associated virus integration in vivo and in vitro. *J Virol* **71**, 9231-9247 (1997). https://doi.org/10.1128/JVI.71.12.9231-9247.1997

315 Palombo, F. *et al.* Site-specific integration in mammalian cells mediated by a new hybrid baculovirus-adeno-associated virus vector. *J Virol* **72**, 5025-5034 (1998). https://doi.org/10.1128/JVI.72.6.5025-5034.1998

316 Pieroni, L. *et al.* Targeted integration of adeno-associated virus-derived plasmids in transfected human cells. *Virology* **249**, 249-259 (1998). https://doi.org/10.1006/viro.1998.9332

317 Nony, P. *et al.* Novel cis-acting replication element in the adeno-associated virus type 2 genome is involved in amplification of integrated rep-cap sequences. *J Virol* **75**, 9991-9994 (2001). https://doi.org/10.1128/JVI.75.20.9991-9994.2001

318 Philpott, N. J., Gomos, J., Berns, K. I. & Falck-Pedersen, E. A p5 integration efficiency element mediates Rep-dependent integration into AAVS1 at chromosome 19. *Proc Natl Acad Sci U S A* **99**, 12381-12385 (2002). https://doi.org/10.1073/pnas.182430299

319 Guilbaud, M. *et al.* Relative influence of the adeno-associated virus (AAV) type 2 p5 element for recombinant AAV vector site-specific integration. *J Virol* **82**, 2590-2593 (2008). https://doi.org/10.1128/JVI.01956-07

320 Huser, D. *et al.* Integration preferences of wildtype AAV-2 for consensus rep-binding sites at numerous loci in the human genome. *PLoS Pathog* **6**, e1000985 (2010). https://doi.org/10.1371/journal.ppat.1000985

321 Drew, H. R., Lockett, L. J. & Both, G. W. Increased complexity of wild-type adeno-associated virus-chromosomal junctions as determined by analysis of unselected cellular genomes. *J Gen Virol* **88**, 1722-1732 (2007). https://doi.org/10.1099/vir.0.82880-0

322 Tenenbaum, L., Lehtonen, E. & Monahan, P. E. Evaluation of risks related to the use of adeno-associated virus-based vectors. *Curr Gene Ther* **3**, 545-565 (2003). https://doi.org/10.2174/1566523034578131

323 Schultz, B. R. & Chamberlain, J. S. Recombinant adeno-associated virus transduction and integration. *Mol Ther* **16**, 1189-1199 (2008). https://doi.org/10.1038/mt.2008.103

324 Russell, D. W. & Hirata, R. K. Human gene targeting by viral vectors. *Nat Genet* **18**, 325-330 (1998). https://doi.org/10.1038/ng0498-325

325 Galli, A. & Cervelli, T. Inverted terminal repeats of adeno-associated virus decrease random integration of a gene targeting fragment in Saccharomyces cerevisiae. *BMC Mol Biol* **15**, 5 (2014). https://doi.org/10.1186/1471-2199-15-5

326 Batty, P. *et al.* Vector integration and fate in the hemophilia dog liver multiple years after AAV-FVIII gene transfer. *Blood* **143**, 2373-2385 (2024). https://doi.org/10.1182/blood.2023022589

327 Rutledge, E. A. & Russell, D. W. Adeno-associated virus vector integration junctions. *J Virol* **71**, 8429-8436 (1997). https://doi.org/10.1128/JVI.71.11.8429-8436.1997

328 Nakai, H., Iwaki, Y., Kay, M. A. & Couto, L. B. Isolation of recombinant adeno-associated virus vector-cellular DNA junctions from mouse liver. *J Virol* **73**, 5438-5447 (1999). https://doi.org/10.1128/JVI.73.7.5438-5447.1999

329 Nakai, H. *et al.* AAV serotype 2 vectors preferentially integrate into active genes in mice. *Nat Genet* **34**, 297-302 (2003). https://doi.org/10.1038/ng1179

330 Inagaki, K. *et al.* DNA palindromes with a modest arm length of greater, similar 20 base pairs are a significant target for recombinant adeno-associated virus vector integration in the liver, muscles, and heart in mice. *J Virol* **81**, 11290-11303 (2007). https://doi.org/10.1128/JVI.00963-07

331 Brown, R. E. & Freudenreich, C. H. Structure-forming repeats and their impact on genome stability. *Curr Opin Genet Dev* **67**, 41-51 (2021). https://doi.org/10.1016/j.gde.2020.10.006

332 Calabria, A. *et al.* Intrathymic AAV delivery results in therapeutic site-specific integration at TCR loci in mice. *Blood* **141**, 2316-2329 (2023). https://doi.org/10.1182/blood.2022017378

333 Miller, D. G., Petek, L. M. & Russell, D. W. Adeno-associated virus vectors integrate at chromosome breakage sites. *Nat Genet* **36**, 767-773 (2004). https://doi.org/10.1038/ng1380

334 Landau, D. J. *et al.* In Vivo Zinc Finger Nuclease-mediated Targeted Integration of a Glucose-6-phosphatase Transgene Promotes Survival in Mice With Glycogen Storage Disease Type IA. *Mol Ther* **24**, 697-706 (2016). https://doi.org/10.1038/mt.2016.35

335 Nelson, C. E. *et al.* Long-term evaluation of AAV-CRISPR genome editing for Duchenne muscular dystrophy. *Nat Med* **25**, 427-432 (2019). https://doi.org/10.1038/s41591-019-0344-3

336 Maeder, M. L. *et al.* Development of a gene-editing approach to restore vision loss in Leber congenital amaurosis type 10. *Nat Med* **25**, 229-233 (2019). https://doi.org/10.1038/s41591-018-0327-9

337 Hanlon, K. S. *et al.* High levels of AAV vector integration into CRISPR-induced DNA breaks. *Nat Commun* **10**, 4439 (2019). https://doi.org/10.1038/s41467-019-12449-2

338 Anguela, X. M. *et al.* Robust ZFN-mediated genome editing in adult hemophilic mice. *Blood* **122**, 3283-3287 (2013). https://doi.org/10.1182/blood-2013-04-497354

339 Ismail, A. M. *et al.* The longitudinal kinetics of AAV5 vector integration profiles and evaluation of clonal expansion in mice. *Mol Ther Methods Clin Dev* **32**, 101294 (2024). https://doi.org/10.1016/j.omtm.2024.101294

340 Flotte, T. R. *et al.* Gene expression from adeno-associated virus vectors in airway epithelial cells. *Am J Respir Cell Mol Biol* **7**, 349-356 (1992). https://doi.org/10.1165/ajrcmb/7.3.349

341 Haberman, R. P., McCown, T. J. & Samulski, R. J. Novel transcriptional regulatory signals in the adeno-associated virus terminal repeat A/D junction element. *J Virol* **74**, 8732-8739 (2000). https://doi.org/10.1128/jvi.74.18.8732-8739.2000

342 Qiu, J., Nayak, R., Tullis, G. E. & Pintel, D. J. Characterization of the transcription profile of adeno-associated virus type 5 reveals a number of unique features compared to previously characterized adeno-associated viruses. *J Virol* **76**, 12435-12447 (2002). https://doi.org/10.1128/jvi.76.24.12435-12447.2002

343 Julien, L. *et al.* RFX1 and RFX3 Transcription Factors Interact with the D Sequence of Adeno-Associated Virus Inverted Terminal Repeat and Regulate AAV Transduction. *Sci Rep* **8**, 210 (2018). https://doi.org/10.1038/s41598-017-18604-3

344 Li, C. *et al.* Adeno-associated virus type 2 (AAV2) capsid-specific cytotoxic T lymphocytes eliminate only vector-transduced cells coexpressing the AAV2 capsid in vivo. *J Virol* **81**, 7540-7547 (2007). https://doi.org/10.1128/JVI.00529-07

345 Siders, W. M. *et al.* Cytotoxic T lymphocyte responses to transgene product, not adeno-associated viral capsid protein, limit transgene expression in mice. *Hum Gene Ther* **20**, 11-20 (2009). https://doi.org/10.1089/hum.2008.055

346 Ashley, S. N., Somanathan, S., Giles, A. R. & Wilson, J. M. TLR9 signaling mediates adaptive immunity following systemic AAV gene therapy. *Cell Immunol* **346**, 103997 (2019). https://doi.org/10.1016/j.cellimm.2019.103997

347 Toth, R. *et al.* Methylation Status of the Adeno-Associated Virus Type 2 (AAV2). *Viruses* **11** (2019). https://doi.org/10.3390/v11010038

348 Brown, N., Song, L., Kollu, N. R. & Hirsch, M. L. Adeno-Associated Virus Vectors and Stem Cells: Friends or Foes? *Hum Gene Ther* **28**, 450-463 (2017). https://doi.org/10.1089/hum.2017.038

349 Shao, W. *et al.* Double-stranded RNA innate immune response activation from long-term adeno-associated virus vector transduction. *JCI Insight* **3** (2018). https://doi.org/10.1172/jci.insight.120474

350 Kwon, H. J. *et al.* Adeno-Associated Virus D-Sequence-Mediated Suppression of Expression of a Human Major Histocompatibility Class II Gene: Implications in the Development of Adeno-Associated Virus Vectors for Modulating Humoral Immune Response. *Hum Gene Ther* **31**, 565-574 (2020). https://doi.org/10.1089/hum.2020.018

351 Pegtel, D. M. *et al.* Functional delivery of viral miRNAs via exosomes. *Proc Natl Acad Sci U S A* **107**, 6328-6333 (2010). https://doi.org/10.1073/pnas.0914843107

352 Diggins, N. L. & Hancock, M. H. Viral miRNA regulation of host gene expression. *Semin Cell Dev Biol* **146**, 2-19 (2023). https://doi.org/10.1016/j.semcdb.2022.11.007

353 Stutika, C. *et al.* Comprehensive Small RNA-Seq of Adeno-Associated Virus (AAV)-Infected Human Cells Detects Patterns of Novel, Non-Coding AAV RNAs in the Absence of Cellular miRNA Regulation. *PLoS One* **11**, e0161454 (2016). https://doi.org/10.1371/journal.pone.0161454

354 Stutika, C. *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). https://doi.org/10.1128/JVI.02750-15

355 Schnepp, B. C., Jensen, R. L., Chen, C. L., Johnson, P. R. & Clark, K. R. Characterization of adeno-associated virus genomes isolated from human tissues. *J Virol* **79**, 14793-14803 (2005). https://doi.org/10.1128/JVI.79.23.14793-14803.2005

356 Jiang, Z. & Dalby, P. A. Challenges in scaling up AAV-based gene therapy manufacturing. *Trends Biotechnol* **41**, 1268-1281 (2023). https://doi.org/10.1016/j.tibtech.2023.04.002

357 Hewitt, F. C. & Samulski, R. J. Creating a novel origin of replication through modulating DNA-protein interfaces. *PLoS One* **5**, e8850 (2010). https://doi.org/10.1371/journal.pone.0008850

358 Ryan, J. H., Zolotukhin, S. & Muzyczka, N. Sequence requirements for binding of Rep68 to the adeno-associated virus terminal repeats. *J Virol* **70**, 1542-1553 (1996). https://doi.org/10.1128/JVI.70.3.1542-1553.1996

359 Wang, X. S., Qing, K., Ponnazhagan, S. & Srivastava, A. Adeno-associated virus type 2 DNA replication in vivo: mutation analyses of the D sequence in viral inverted terminal repeats. *J Virol* **71**, 3077-3082 (1997). https://doi.org/10.1128/JVI.71.4.3077-3082.1997

360 Yan, Z., Lei-Butters, D. C., Zhang, Y., Zak, R. & Engelhardt, J. F. Hybrid adeno-associated virus bearing nonhomologous inverted terminal repeats enhances dual-vector reconstruction of minigenes in vivo. *Hum Gene Ther* **18**, 81-87 (2007). https://doi.org/10.1089/hum.2006.128

361 Zhang, J. *et al.* "D" matters in recombinant AAV DNA packaging. *Mol Ther* **29**, 1937-1939 (2021). https://doi.org/10.1016/j.ymthe.2021.05.002

362 Shoti, J., Qing, K. & Srivastava, A. Reply to "D" matters in recombinant AAV packaging. *Mol Ther* **29**, 2628-2630 (2021). https://doi.org/10.1016/j.ymthe.2021.08.012

363 Tran, N. T. *et al.* AAV-Genome Population Sequencing of Vectors Packaging CRISPR Components Reveals Design-Influenced Heterogeneity. *Mol Ther Methods Clin Dev* **18**, 639-651 (2020). https://doi.org/10.1016/j.omtm.2020.07.007

364 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018). https://doi.org/10.1093/bioinformatics/bty191

365 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118 (2013). https://doi.org/10.1371/journal.pcbi.1003118

366 Rapti, K. *et al.* Isolation of Next-Generation Gene Therapy Vectors through Engineering, Barcoding, and Screening of Adeno-Associated Virus (AAV) Capsid Variants. *J Vis Exp* (2022). https://doi.org/10.3791/64389

367 Morgan, M. *et al.* ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607-2608 (2009). https://doi.org/10.1093/bioinformatics/btp450

368 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019). https://doi.org/10.1038/s41598-019-45839-z

369 Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381-2383 (2017). https://doi.org/10.1093/bioinformatics/btx183

370 Gel, B. *et al.* regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**, 289-291 (2016). https://doi.org/10.1093/bioinformatics/btv562

371 Marsic, D., Mendez-Gomez, H. R. & Zolotukhin, S. High-accuracy biodistribution analysis of adeno-associated virus variants by double barcode sequencing. *Mol Ther Methods Clin Dev* **2**, 15041 (2015). https://doi.org/10.1038/mtm.2015.41

372 Wang, H. Y. *et al.* Selection and Characterization of Malachite Green Aptamers for the Development of Light-up Probes. *Chemistryselect* **1**, 1571-1574 (2016). https://doi.org/10.1002/slct.201600154

373 Chen, Y. *et al.* Selection and characterization of a DNA aptamer to crystal violet. *Photochem Photobiol Sci* **17**, 800-806 (2018). https://doi.org/10.1039/c7pp00457e

374 Mashal, R. D., Koontz, J. & Sklar, J. Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases. *Nat Genet* **9**, 177-183 (1995). https://doi.org/10.1038/ng0295-177

375 Guschin, D. Y. *et al.* A rapid and general assay for monitoring endogenous gene modification. *Methods Mol Biol* **649**, 247-256 (2010). https://doi.org/10.1007/978-1-60761-753-2_15

376 Rossi, A. *et al.* Vector uncoating limits adeno-associated viral vector-mediated transduction of human dendritic cells and vector immunogenicity. *Sci Rep* **9**, 3631 (2019). https://doi.org/10.1038/s41598-019-40071-1

377 Borner, K. *et al.* Pre-arrayed Pan-AAV Peptide Display Libraries for Rapid Single-Round Screening. *Mol Ther* **28**, 1016-1032 (2020). https://doi.org/10.1016/j.ymthe.2020.02.009

378 Radukic, M. T., Brandt, D., Haak, M., Muller, K. M. & Kalinowski, J. Nanopore sequencing of native adeno-associated virus (AAV) single-stranded DNA using a transposase-based rapid protocol. *NAR Genom Bioinform* **2**, lqaa074 (2020). https://doi.org/10.1093/nargab/lqaa074

379 Marincevic-Zuniga, Y., Gustavsson, I. & Gyllensten, U. Multiply-primed rolling circle amplification of human papillomavirus using sequence-specific primers. *Virology* **432**, 57-62 (2012). https://doi.org/10.1016/j.virol.2012.05.030

380 Johne, R., Muller, H., Rector, A., van Ranst, M. & Stevens, H. Rolling-circle amplification of viral DNA genomes using phi29 polymerase. *Trends Microbiol* **17**, 205-211 (2009). https://doi.org/10.1016/j.tim.2009.02.004

381 Breton, C., Clark, P. M., Wang, L., Greig, J. A. & Wilson, J. M. ITR-Seq, a next-generation sequencing assay, identifies genome-wide DNA editing sites in vivo following adeno-associated viral vector-mediated genome editing. *BMC Genomics* **21**, 239 (2020). https://doi.org/10.1186/s12864-020-6655-4

382 Edwards, D. N., Machwe, A., Wang, Z. & Orren, D. K. Intramolecular telomeric G-quadruplexes dramatically inhibit DNA synthesis by replicative and translesion polymerases, revealing their potential to lead to genetic change. *PLoS One* **9**, e80664 (2014). https://doi.org/10.1371/journal.pone.0080664

383 Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat Commun* **10**, 3425 (2019). https://doi.org/10.1038/s41467-019-11433-0

384 Smith, R. H. & Kotin, R. M. An adeno-associated virus (AAV) initiator protein, Rep78, catalyzes the cleavage and ligation of single-stranded AAV ori DNA. *J Virol* **74**, 3122-3129 (2000). https://doi.org/10.1128/jvi.74.7.3122-3129.2000

385 Steinfeldt, T., Finsterbusch, T. & Mankertz, A. Demonstration of nicking/joining activity at the origin of DNA replication associated with the rep and rep' proteins of porcine circovirus type 1. *J Virol* **80**, 6225-6234 (2006). https://doi.org/10.1128/JVI.02506-05

6. References

386 Cheung, A. K. Rolling-circle replication of an animal circovirus genome in a theta-replicating bacterial plasmid in Escherichia coli. *J Virol* **80**, 8686-8694 (2006). https://doi.org/10.1128/JVI.00655-06

387 Wood, W. N., Smith, K. D., Ream, J. A. & Lewis, L. K. Enhancing yields of low and single copy number plasmid DNAs from Escherichia coli cells. *J Microbiol Methods* **133**, 46-51 (2017). https://doi.org/10.1016/j.mimet.2016.12.016

388 Lahijani, R., Hulley, G., Soriano, G., Horn, N. A. & Marquet, M. High-yield production of pBR322-derived plasmids intended for human gene therapy by employing a temperature-controllable point mutation. *Hum Gene Ther* **7**, 1971-1980 (1996). https://doi.org/10.1089/hum.1996.7.16-1971

389 Dunker-Seidler, F. *et al.* Recombinant AAV batch profiling by nanopore sequencing elucidates product-related DNA impurities and vector genome length distribution. *Mol Ther Methods Clin Dev* **33**, 101417 (2025). https://doi.org/10.1016/j.omtm.2025.101417

390 Al-Allaf, F. A., Tolmachov, O. E., Zambetti, L. P., Tchetchelnitski, V. & Mehmet, H. Remarkable stability of an instability-prone lentiviral vector plasmid in Escherichia coli Stbl3. *3 Biotech* **3**, 61-70 (2013). https://doi.org/10.1007/s13205-012-0070-8

391 Gray, S. J. *et al.* Production of recombinant adeno-associated viral vectors and use in in vitro and in vivo administration. *Curr Protoc Neurosci* **Chapter 4**, Unit 4 17 (2011). https://doi.org/10.1002/0471142301.ns0417s57

392 Hadden, J. M., Declais, A. C., Carr, S. B., Lilley, D. M. & Phillips, S. E. The structural basis of Holliday junction resolution by T7 endonuclease I. *Nature* **449**, 621-624 (2007). https://doi.org/10.1038/nature06158

393 Wilmott, P. & Lisowski, L. AAV Genome Topology Decides ITR Secondary Structure. *Bioessays* **47**, e202400266 (2025). https://doi.org/10.1002/bies.202400266

394 Namkung, S. *et al.* Direct ITR-to-ITR Nanopore Sequencing of AAV Vector Genomes. *Hum Gene Ther* **33**, 1187-1196 (2022). https://doi.org/10.1089/hum.2022.143

395 Delahaye, C. & Nicolas, J. Sequencing DNA with nanopores: Troubles and biases. *PLoS One* **16**, e0257521 (2021). https://doi.org/10.1371/journal.pone.0257521

396 Spealman, P., Burrell, J. & Gresham, D. Inverted duplicate DNA sequences increase translocation rates through sequencing nanopores resulting in reduced base calling accuracy. *Nucleic Acids Res* **48**, 4940-4945 (2020). https://doi.org/10.1093/nar/gkaa206

397 Wang, X. S. *et al.* Characterization of wild-type adeno-associated virus type 2-like particles generated during recombinant viral vector production and strategies for their elimination. *J Virol* **72**, 5472-5480 (1998). https://doi.org/10.1128/JVI.72.7.5472-5480.1998

398 Yip, M. *et al.* Querying Recombination Junctions of Replication-Competent Adeno-Associated Viruses in Gene Therapy Vector Preparations with Single Molecule, Real-Time Sequencing. *Viruses* **15** (2023). https://doi.org/10.3390/v15061228

399 Banos-Mateos, S. *et al.* Abolishing Retro-Transduction of Producer Cells in Lentiviral Vector Manufacturing. *Viruses* **16** (2024). https://doi.org/10.3390/v16081216

400 O'Driscoll, E. E., Arora, S., Lang, J. F., Davidson, B. L. & Shalem, O. CRISPR screen reveals modifiers of rAAV production including known rAAV infection genes playing an unexpected role in vector production. *Mol Ther Methods Clin Dev* **33**, 101408 (2025). https://doi.org/10.1016/j.omtm.2025.101408

401 Zhang, J. *et al.* Thorough molecular configuration analysis of noncanonical AAV genomes in AAV vector preparations. *Mol Ther Methods Clin Dev* **32**, 101215 (2024). https://doi.org/10.1016/j.omtm.2024.101215

402 Steininger, T. *et al.* Improved Recombinant Adeno-Associated Viral Vector Production via Molecular Evolution of the Viral Rep Protein. *Int J Mol Sci* **26** (2025). https://doi.org/10.3390/ijms26031319

403 Westhaus, A. *et al.* High-Throughput In Vitro, Ex Vivo, and In Vivo Screen of Adeno-Associated Virus Vectors Based on Physical and Functional Transduction. *Hum Gene Ther* **31**, 575-589 (2020). https://doi.org/10.1089/hum.2019.264

404 Kondratov, O. *et al.* A comprehensive study of a 29-capsid AAV library in a non-human primate central nervous system. *Mol Ther* **29**, 2806-2820 (2021). https://doi.org/10.1016/j.ymthe.2021.07.010

405 Stone, D. *et al.* A multiplexed barcode approach to simultaneously evaluate gene delivery by adeno-associated virus capsid variants in nonhuman primates. *Hepatol Commun* **7**, e0009 (2023). https://doi.org/10.1097/HC9.0000000000000009

406 Schnepp, B. C. *et al.* Recombinant Adeno-Associated Virus Vector Genomes Take the Form of Long-Lived, Transcriptionally Competent Episomes in Human Muscle. *Hum Gene Ther* **27**, 32-42 (2016). https://doi.org/10.1089/hum.2015.136

407 Sang, Y. *et al.* Salt ions and related parameters affect PEI-DNA particle size and transfection efficiency in Chinese hamster ovary cells. *Cytotechnology* **67**, 67-74 (2015). https://doi.org/10.1007/s10616-013-9658-z

408 Nonnenmacher, M. *et al.* Rapid evolution of blood-brain-barrier-penetrating AAV capsids by RNA-driven biopanning. *Mol Ther Methods Clin Dev* **20**, 366-378 (2021). https://doi.org/10.1016/j.omtm.2020.12.006

409 Cabanes-Creus, M. *et al.* Restoring the natural tropism of AAV2 vectors for human liver. *Sci Transl Med* **12** (2020). https://doi.org/10.1126/scitranslmed.aba3312

410 Thomas, C. E., Storm, T. A., Huang, Z. & Kay, M. A. Rapid uncoating of vector genomes is the key to efficient liver transduction with pseudotyped adeno-associated virus vectors. *J Virol* **78**, 3110-3122 (2004). https://doi.org/10.1128/jvi.78.6.3110-3122.2004

411 Berraondo, P. *et al.* Intrahepatic injection of recombinant adeno-associated virus serotype 2 overcomes gender-related differences in liver transduction. *Hum Gene Ther* **17**, 601-610 (2006). https://doi.org/10.1089/hum.2006.17.601

412 Walkey, C. J. *et al.* A comprehensive atlas of AAV tropism in the mouse. *Mol Ther* **33**, 1282-1299 (2025). https://doi.org/10.1016/j.ymthe.2025.01.041

413 Sihn, C. R. *et al.* Molecular analysis of AAV5-hFVIII-SQ vector-genome-processing kinetics in transduced mouse and nonhuman primate livers. *Mol Ther Methods Clin Dev* **24**, 142-153 (2022). https://doi.org/10.1016/j.omtm.2021.12.004

414 Garafutdinov, R. R., Sakhabutdinova, A. R., Gilvanov, A. R. & Chemeris, A. V. Rolling Circle Amplification as a Universal Method for the Analysis of a Wide Range of Biological Targets. *Russ J Bioorg Chem* **47**, 1172-1189 (2021). https://doi.org/10.1134/S1068162021060078

415 Russell, D. W., Miller, A. D. & Alexander, I. E. Adeno-associated virus vectors preferentially transduce cells in S phase. *Proc Natl Acad Sci U S A* **91**, 8915-8919 (1994). https://doi.org/10.1073/pnas.91.19.8915

416 Joffroy, B., Uca, Y. O., Presern, D., Doye, J. P. K. & Schmidt, T. L. Rolling circle amplification shows a sinusoidal template length-dependent amplification bias. *Nucleic Acids Res* **46**, 538-545 (2018). https://doi.org/10.1093/nar/gkx1238

417 Lasken, R. S. & Egholm, M. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends Biotechnol* **21**, 531-535 (2003). https://doi.org/10.1016/j.tibtech.2003.09.010

418 Fjelstrup, S. *et al.* The Effects of Dithiothreitol on DNA. *Sensors (Basel)* **17** (2017). https://doi.org/10.3390/s17061201

419 Myers, J. E. *et al.* Detecting episomal or integrated human papillomavirus 16 DNA using an exonuclease V-qPCR-based assay. *Virology* **537**, 149-156 (2019). https://doi.org/10.1016/j.virol.2019.08.021

420 Wang, Y., Wang, X., Gailing, O. & Xi, D. Visual detection of Fusarium proliferatum based on dual-cycle signal amplification and T5 exonuclease. *RSC Adv* **10**, 35131-35135 (2020). https://doi.org/10.1039/d0ra06559e

421 Sharma, R. *et al.* In vivo genome editing of the albumin locus as a platform for protein replacement therapy. *Blood* **126**, 1777-1784 (2015). https://doi.org/10.1182/blood-2014-12-615492

422 Thakur, J., Packiaraj, J. & Henikoff, S. Sequence, Chromatin and Evolution of Satellite DNA. *Int J Mol Sci* **22** (2021). https://doi.org/10.3390/ijms22094309

423 Johnson, M. S., Venkataram, S. & Kryazhimskiy, S. Best Practices in Designing, Sequencing, and Identifying Random DNA Barcodes. *J Mol Evol* **91**, 263-280 (2023). https://doi.org/10.1007/s00239-022-10083-z

424 Brimble, M. A. *et al.* Preventing packaging of translatable P5-associated DNA contaminants in recombinant AAV vector preps. *Mol Ther Methods Clin Dev* **24**, 280-291 (2022). https://doi.org/10.1016/j.omtm.2022.01.008

425 White, R., Pellefigues, C., Ronchese, F., Lamiable, O. & Eccles, D. Investigation of chimeric reads using the MinION. *F1000Res* **6**, 631 (2017). https://doi.org/10.12688/f1000research.11547.2

426 Li, Y. *et al.* A Genomic Language Model for Chimera Artifact Detection in Nanopore Direct RNA Sequencing. *bioRxiv* (2024). https://doi.org/10.1101/2024.10.23.619929

427 Musatov, S. A., Scully, T. A., Dudus, L. & Fisher, K. J. Induction of circular episomes during rescue and replication of adeno-associated virus in experimental models of virus latency. *Virology* **275**, 411-432 (2000). https://doi.org/10.1006/viro.2000.0504

428 Musatov, S., Roberts, J., Pfaff, D. & Kaplitt, M. A cis-acting element that directs circular adeno-associated virus replication and packaging. *J Virol* **76**, 12792-12802 (2002). https://doi.org/10.1128/jvi.76.24.12792-12802.2002

429 Spanielova, H., Fraiberk, M., Suchanova, J., Soukup, J. & Forstova, J. The encapsidation of polyomavirus is not defined by a sequence-specific encapsidation signal. *Virology* **450-451**, 122-131 (2014). https://doi.org/10.1016/j.virol.2013.12.010

430 Sarker, S. *et al.* Structural insights into the assembly and regulation of distinct viral capsid complexes. *Nat Commun* **7**, 13014 (2016). https://doi.org/10.1038/ncomms13014

431 Love, S. D., Posey, S., Burch, A. D. & Fane, B. A. Disenfranchised DNA: biochemical analysis of mutant øX174 DNA-binding proteins may further elucidate the evolutionary significance of the unessential packaging protein A. *J Virol* **98**, e0182723 (2024). https://doi.org/10.1128/jvi.01827-23

7. Acknowledgements

Completing this PhD has been one of the most challenging and rewarding experiences of my life, and it would not have been possible without the support, guidance, and encouragement of many individuals.

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Dr. Dirk Grimm, for his invaluable guidance, continuous support, and patience throughout my research journey. You allowed me the freedom to explore my ideas while providing the structure, critical feedback and encouragement which helped shape both the direction and quality of this work. For your patience, your thoughtful critiques, and your unwavering belief in my abilities, I am sincerely thankful. I am also grateful for the help on several manuscripts, a research grant proposal and the opportunity to present my work at several scientific conferences and meetings.

Furthermore, I want to thank Prof. Dr. Robert Russell for taking over the role of first advisor for my thesis, despite the change of focus away from protein engineering to ITR engineering in the early stages of my thesis, and the examination committee members Prof. Dr. Ulrike Müller and Prof. Dr. Kerstin Göpfrich.

I also want to thank the members of my TAC committee, Prof. Dr. Dirk Grimm, Prof. Dr. Robert Russell, and Prof. Dr. Dominik Niopek for their time, insightful comments, and thought-provoking questions. Their expertise added significant value to this work.

I gratefully acknowledge the financial support provided by the ARDAT consortium for the funding and the (not only) scientific exchange among the consortium members.

This research would not have been possible without the many collaborators and partners who contributed their expertise, time, and energy. I would like to thank Prof. Dr. Luís Pereira de Almeida, Dr. Rui Nobre and Prof. Dr. Phillip Tai and their teams for their vital contributions by performing the animal handling and performing PacBio sequencing. Our collaborations not only enriched the quality of this dissertation but also made the process more dynamic and rewarding. It has been a privilege to work alongside such dedicated and talented individuals.

A heartfelt thanks goes to my colleagues and friends in the AG Grimm. Our discussions, collaborations, and shared struggles created a truly stimulating and supportive research environment. Special thanks to Conradin, for his help with setting up the Nanopore sequencing of my vector genomes. Additionally, Conradin, Jixin, Kristian and Andrej, for the lunch sessions, advice and discussions and camaraderie which meant a great deal to me and kept up

the motivation. I want to thank the technicians Ellen, Emma, Chiara and Lena for keeping the lab running and their help when needed.

To my family — your unwavering love and support have been my foundation. To my parents, Ruth and Jürgen and my brother Johannes, thank you for encouraging my curiosity and teaching me resilience and humility. You have been my role models in more ways than you know.

To my wife, Katya, thank you for walking this path with me. Your belief in me, even when I doubted myself, your quiet sacrifices, your understanding through the long nights and weekends, and your enduring patience have been my anchor. This accomplishment is as much yours as it is mine. And to my child, Alexis, thank you for being my light. You were born during this journey, and your presence brought me a new sense of purpose. Your laughter reminded me of what truly matters, and your very existence taught me more about patience, resilience, and love than any textbook ever could. Writing this thesis often meant long hours away from you, and yet you filled every day with joy and perspective.