Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Natural Sciences and Mathematics

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M.Sc. Conradin Baumgartl

born in: Passau, Germany

Oral examination: 22.07.2025

The hidden syntax of AAV: Insights into the epigenetic regulation and discovery of a dinucleotide pattern in Adeno-associated virus (AAV)

Referees: Prof. Dr. Benedikt Brors

Prof. Dr. Dirk Grimm

Abstract

Recombinant Adeno-associated Virus (rAAV) is an auspicious gene-delivery vector for human gene therapy. Even though there are multiple approved rAAV-based drugs available, many underlying molecular mechanisms remain uncertain. This thesis is dedicated to the investigation of two related areas of AAV's biology: (i) investigation of epigenetic regulation of delivered DNA comparing two distinct AAV vector capsids, and (ii) exploration of a sequence pattern conserved for the genus *Dependoparvovirus*, which may contribute to improved transgene cassette design.

AAV-delivered DNA mainly persists as circular episomal DNA within the nuclei of transduced cells. The extent of epigenetic regulation that the transduced cells enact on the delivered vector DNA and the influence of the capsid are severely understudied. In a long-term mouse model experiment, I found that the DNA methylation in the formed episome is negligible using nanopore sequencing. However, histone modifications were differentially deposited on AAV2- and AAV9-delivered episomes. At later timepoints, expression-beneficial histone acetylations on the AAV9-delivered transgene became lost, and its expression was repressed. AAV2 appeared to have an expression-permissive regulation, but was being held back by its unpacking speed.

In the second part of this work, I describe the discovery of a pronounced 15 bp YY/RR dinucleotide pattern unique to the genus *Dependoparvovirus*. It resembles the 10 bp histone-positioning code in eukaryotic genomes, but the longer repeat length likely enables a different binding partner, potentially those involved in viral replication or packaging. The biological importance of this pattern was demonstrated through its enrichment during the selection of shuffled AAV *cap* sequences. Additionally, I could show that rAAV vectors that contained sequences with the wild-type periodicity produced higher titers in a small-scale production system, suggesting a functional role necessary for vector production. Based on the evidence provided, an increased interaction of the DNA with Rep proteins is a likely explanation for the presence of this pattern. The described pattern could be harnessed in the future to improve rAAV transgene cassette designs.

Zusammenfassung

Rekombinante Adeno-assoziierte Viren (rAAV) stellen vielversprechende Vektoren für die gentherapeutische Anwendung beim Menschen dar. Obwohl bereits mehrere rAAV-basierte Arzneimittel zugelassen sind, bleiben viele zugrunde liegende molekulare Mechanismen bislang unzureichend verstanden. Diese Arbeit widmet sich zwei miteinander verbundenen Aspekten der AAV-Biologie: (i) der Untersuchung der epigenetischen Regulation des übertragenen DNA-Materials in Abhängigkeit zweier unterschiedlicher AAV-Kapsidvarianten sowie (ii) der Analyse eines konservierten Sequenzmotivs innerhalb der Gattung *Dependoparvovirus*, welches zur Optimierung von Transgen-Kassetten beitragen könnte.

Nach der Transduktion liegt die AAV-vermittelte DNA überwiegend als zirkuläres episomales DNA-Molekül im Zellkern der Zielzellen vor. Die epigenetischen Veränderungen, welche die transduzierten Zellen auf das Vektor-DNA-Material übertragen, sowie der Einfluss des Kapsids darauf sind bislang nur unzureichend erforscht. In einem langfristig angelegten Experiment in Mäusen konnte ich mittels Nanopore-Sequenzierung zeigen, dass die DNA-Methylierung auf den gebildeten Episomen vernachlässigbar ist. Dagegen wurden deutliche Unterschiede in der Deposition von Histonmodifikationen auf durch AAV2 bzw. AAV9 übertragenen Episomen beobachtet. Zu späteren Zeitpunkten gingen für die Genexpression förderliche Histon-Acetylierungen auf AAV9-vermittelten Transgenen verloren, was mit einer Repression der Transgenexpression einherging. AAV2 hingegen zeigte eine expressionsförderliche epigenetische Regulation, war jedoch durch eine verlangsamte Entpackung des Genoms limitiert.

Im zweiten Teil der Arbeit beschreibe ich die Entdeckung eines ausgeprägten, 15 Basenpaare umfassenden YY/RR-Dinukleotidmusters, das einzigartig für die Gattung *Dependoparvovirus* ist. Dieses Muster ähnelt dem 10 bp langen Histon-Positionierungscode eukaryotischer Genome, weist jedoch eine größere Periodenlänge auf, was auf einen anderen Bindungspartner hindeutet – möglicherweise zusammenhängend mit viraler Replikation oder Verpackung. Die biologische Relevanz dieses Musters zeigte sich in dessen Anreicherung während der Selektion rekombinierter AAV-*cap*-Sequenzen. Darüber

hinaus konnte ich nachweisen, dass rAAV-Vektoren mit Sequenzen, die die wildtypische Periodizität aufwiesen, in einem verkleinertem Produktionssystem höhere Titer erzielten. Das lässt auf eine funktionelle Rolle des Musters bei der Vektorproduktion schließen. Die vorliegenden Daten deuten darauf hin, dass eine verstärkte Interaktion der DNA mit Rep-Proteinen die Existenz dieses Musters erklären könnte. In Zukunft könnte dieses Muster genutzt werden, um rAAV-Transgenkassetten zu optimieren.

Contents

Abbreviations

AAV Adeno-associated virus

Ad human Adenovirus

AdH Adenohelper plasmid

bgh-pA Bovine growth hormone polyadenylation signal

bp base pairs

CMV Cytomegalovirus

CMVenh Cytomegalovirus enhancer

ddPCR droplet digital PCR

DMEM Dulbecco's modified eagle medium

DNA desoxyribonucleic acid

DNMT DNA methyltransferase

dsDNA double-stranded DNA

EtOH ethanol

eYFP enhanced yellow fluorescent protein

FBS fetal bovine serum

GdF Goodness of Fit

gDNA genomic DNA

GFAP glial fibrillary acidic protein

GFP green fluorescent protein

H3K27ac Histone 3 lysine 27 acetylation

H3K27me3 Histone 3 lysine 27 tri-methylation

HAT histone acetyltransferase

HDAC Histone deacetylase

ICTV International Committee on Taxonomy of Viruses

ITR inverted terminal repeat

kb kilobases

LB liquid broth

loess locally estimated scatterplot smoothing

LP1 liver promoter 1

MOI multiplicity of infection

NEB New England Bioscience

nt nucleotide

OBD Origin Binding Domain

ONT Oxford Nanopore Technologies

p.i. post-injection

pA Polyadenylation signal

PRC2 Polycomb-repressor complex 2

PTM post-translational modification

qPCR quantitative PCR

rAAV recombinant AAV

RBE Rep-binding element

ssDNA single-stranded DNA

VP1u VP1-unique

ZmBWV *Zophobas morio* black wasting virus

List of Figures

List of Tables

1. Introduction

1.1 Adeno-associated virus

Adeno-associated virus (AAV) were first described in the 1960s, initially as by-products of Adenovirus productions [1, 2]. In the early days of AAV research, it was often described as a 'defective' virus [3, 4, 5], which is somewhat misleading. AAV is dependent on the co-infection of other viruses to undergo its own replication. Additionally, it does not encode a polymerase and thus also relies on host cell factors for its replication. However, these characteristics do not render AAV defective but are significant traits within their niche as a satellite virus (section 1.1.1). Simultaneously, these characteristics are major contributors to its success as a gene-therapy vector (section 1.1.3).

1.1.1 Dependoparvoviruses and satellite viruses

AAV is classified as a member of the *Parvoviridae* family. Parvoviruses are small (22-28 nm; parvus *lat.*: small) in size and split into two subfamilies mainly differentiated by their host range: *Parvovirinae* and *Densovirinae*. AAV is a member of the *Parvovirinae* subfamily that contains parvoviruses that infect vertebrate species, such as humans, birds, and reptiles. The *Densovirinae* contain viruses that infect invertebrate hosts and are usually highly pathogenic [6]. All parvoviruses share a common genome architecture including a single-stranded DNA (ssDNA) genome of roughly 5 kb with one half of the genome encoding non-structural proteins (by convention the left half; Rep proteins in AAV) and the other half encoding the structural proteins making up the capsids (by convention the right half; VP proteins in AAV). An infamous and pathogenic member of the *Parvoviridae* is the human parvovirus B19, which causes 'fifth disease' (also called 'slapped face syndrome' or erythema infectiosum) in children [7].

The family *Parvoviridae* currently encompasses 13 genera. AAV is a member of the genus *Dependoparvovirus*. Dependoparvoviruses have been identified as infecting a wide array of vertebrates, including reptiles, birds, and mammals [8]. As the name suggests,

most viruses in the genus *Dependoparvovirus* share the dependency on the co-infection of a helper virus for efficient replication. Exceptions are the 'goose parvovirus' and 'muscovy duck parvovirus' that can replicate in embryonic cells as well as the 'bearded dragon dependoparvovirus' whose independence is still debated [9, 10]. The classical helper virus for AAV is the human Adenovirus (Ad), but Herpesviruses, Bocavirus, Papillomavirus, and Baculovirus are also capable of providing or replacing helper functions [11, 12, 13, 14, 15, 16]. The Ad helper factors aid the AAV life cycle on multiple levels. They are necessary for AAV transcription initiation (E1A), AAV protein stabilisation (VA RNA), and second-strand synthesis and replication (E2A, E1B55k, E4orf6) [11]. Vice versa, AAV is also capable of interfering with the Ad life cycle. Specifically, AAV downregulates adenoviral genes, ultimately inhibiting the Ad desoxyribonucleic acid (DNA) synthesis and leading to a 50-fold decrease in Ad particle production [17]. This provides a competitive advantage for AAV as it parasitises the helper virus together with the infected host cell. This interaction has likely evolved from self-replicating precursors but appears to benefit AAV greatly as demonstrated in the wide array of known host species [18].

The dependence on a helper virus and their complex relationship is a well-described phenomenon where the dependent viruses are called satellite viruses (sometimes virophages or cheating viruses) [19, 20]. The International Committee on Taxonomy of Viruses (ICTV) lists AAV as definite satellite viruses in their current release, although they are rarely called such in the literature [20]. To be classified as a satellite virus, the virus in question requires helper functions provided by a helper virus, as it lacks essential functions in its own genome. They also need their genome to be distinct from their helper virus so as not to be confused with 'defective interfering particles'. Although common, satellite viruses do not always have a parasitising relationship with their helper virus. Examples of this include Umbraviruses that provide a movement protein for their helper, which is essential for plant-infecting viruses [21, 22]. Satellite viruses have been mostly described in plants and their viruses. Currently, AAV and the Hepatitis D Virus are the only examples of satellite viruses infecting humans. Underlining their complex relationship, satellite and helper can undergo rapid co-evolution [23]. AAV itself has been reported to produce satellite subgenomic particles (defective interfering particles) that seem to be an evolved aspect that co-regulates the AAV life cycle rather than a by-product [24]. By fulfilling all characteristics of being a satellite virus, AAVs (and by extension most members of the genus *Dependoparvovirus*) set themselves apart from the family *Parvoviridae* and demonstrate their unique usefulness when engineered as a vector for gene therapy.

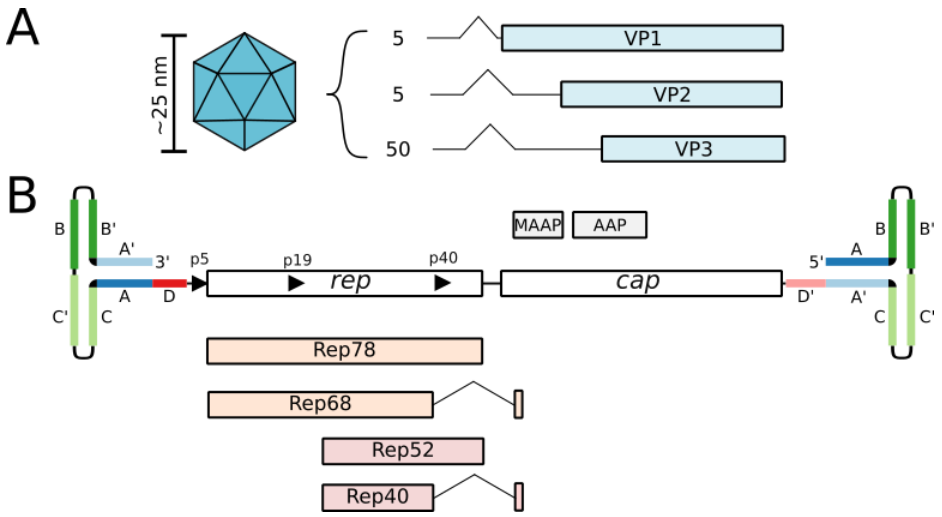

Figure 1.1. AAV genome organisation.
A) The circa 25 nm AAV capsid consists of the viral proteins VP1, VP2, and VP3 at a ratio of 1:1:10. The capsid is made up of 60 subunits, resulting in 5 VP1, 5 VP2, and 50 VP3 proteins. The VP proteins are expressed by the p40 promoter and their ratio is a result of splicing and leaky ribosomal scanning. B) The AAV genome is an about 4.7 kb single-stranded DNA molecule flanked by two ITR sequences, that form a characteristic T-shape on either side of the genome. The *cap* gene encodes the aforementioned VP proteins as well as frame-shifted the MAAP and AAP proteins. The *rep* gene encodes four different Rep proteins responsible for the replication and encapsidation of the genome into pre-formed capsids.

1.1.2 AAV capsid and genome organisation

The icosahedral AAV particle has a size between 22 and 26 nm, depending on the measurement method [25, 26]. It is made up of 60 subunits of the capsid proteins VP1, VP2, and VP3 at a ratio of about 1:1:10 (about 5, 5, and 50 copies, respectively) and a symmetry of T=1 (Figure 1.1 A). AAV is a relatively small virus with an equally short genome, which does not mean it is any less complex. It has evolved overlapping genes with multiple splice variants, resulting in an information-dense genome in which many sections have more than one function. The packaged AAV genome consists of a ssDNA molecule of about 4.7 kb that can be packaged in either the forward (+) or reverse (-) orientation in equal ratios [27]. The ends of the genome persist as two 145 bp inverted terminal repeat (ITR) sequences. They consist of four partially palindromic sequences, alphabetically named A, B, C, and D, which form characteristic T-shapes on either end of the AAV genome (Figure 1.1 B). They play an indispensable role in genome replication and are the only necessary signals for packaging flanked DNA into pre-formed capsids [28, 29]. This makes AAV an easily modifiable system and is a cornerstone of AAV-based gene therapy (more in section 1.1.3).

The two major genes in the AAV genome are the *rep* and *cap* genes. Starting with the latter, the transcription of the *cap* encoded VP proteins is driven by the p40 promoter (Figure 1.1 A). The expression at the proper ratio of 1:1:10 (VP1:VP2:VP3) is regulated by alternative

splicing and leaky ribosomal scanning [30, 31]. The minorly transcribed of the two *cap* gene splice variants contains the start codons of all VP proteins [32, 33]. The major splice variant is shorter and contains only start codons for VP2 and VP3 - ACG for VP2 and ATG for VP3. Because only the latter possesses the canonical start codon with an optimal Kozak sequence, it leads to the preference of VP3 over VP2 through leaky ribosomal scanning [31]. Together with the minor availability of the VP1-containing splice variant, this results in the observed VP ratio of 1:1:10. The common region between all VP proteins is termed the VP3-common region. VP3 is the only protein necessary (together with AAP) to form an icosahedral capsid [34, 16]. The VP1-unique region and the VP1-VP2-unique region provide functions necessary for post-uptake steps during infection (see section 1.1.4). Within the *cap* gene lie two more proteins termed the assembly-activating protein (AAP) and the membrane-associated accessory protein (MAAP). Both of these have a shifted reading frame of +1 in regard to the VP reading frame and are expressed via leaky ribosomal scanning. AAP is a chaperone protein necessary for the correct assembly of the AAV capsid of most serotypes. It functions by interacting with the hydrophobic C-terminal VP3-common region of all VP proteins, directing them to the nucleolus for assembly and stabilizing VP interactions [34, 35]. AAP itself is not a component of the mature capsid. MAAP is the most recently discovered AAV protein found by performing a scan of all possible point mutations within the *cap* gene [36]. Its major function is as a viral egress factor, but it is also implicated in capsid quality control and viral transcription regulation [37, 38].

The *rep* gene encodes four proteins named after their molecular weight: Rep78, Rep68, Rep52, and Rep40. Similar to the VP proteins, the Rep proteins are partially overlapping and share different regions amongst each other (Figure 1.1 B). All Reps have an ATP-dependent SF3 family DNA helicase with 3'-to-5' processivity [39, 40]. Unlike other SF3 helicases that form hexameric rings, the Rep proteins portray an irregular oligomerisation behaviour. The small Rep proteins (Rep52 and Rep40) lack a complete oligomerisation domain and act as monomers whereas the large Rep proteins (Rep78 and Rep68) form mono-, di-, hexa-, hepta-, or even octameres, depending on their concentration and the substrate [41, 42, 43]. The large Reps are transcribed by the p5 promoter and differentiate themselves from the small Reps by additionally possessing an Origin Binding Domain (OBD) necessary for terminal resolution. They are the only Rep proteins necessary for AAV DNA replication. AAV replication is based on a strand displacement mechanism in which the T-shaped ITR sequence acts as the primer for the replication. Rep68 binds the Rep binding element via its OBD and nicks the DNA at the terminal resolution site. Thus, the large Rep proteins resolve intermediate replication products, resulting in individual AAV genomes following end-repair of the ITR sequence [44]. Additionally, the large Rep proteins regulate the

expression of AAV genes depending on the presence of a helper virus and are necessary for the integration of AAV into the human locus termed AAVS1 on chromosome 19 [45, 46]. The small Rep proteins are transcribed by the p19 promoter and only contain the helicase domain and a zinc finger domain (only Rep40). They are required for the packaging of the ssDNA genome into preformed capsids in a 3' to 5' direction, probably through one of the capsid's fivefold pores, though the exact mechanism remains unknown [47, 48, 49].

1.1.3 Recombinant AAV and their use in Gene Therapy

Currently, there are over 200 completed and ongoing AAV-based trials [50] with nine approved AAV-based gene-therapy vectors for the treatment of various genetic disorders by the U.S. Food and Drug Administration (FDA) or European Medicines Agency (EMA): Beqvez, Elevidys, Glybera (withdrawn in 2017 due to high cost of treatment [51]), Hemgenix, Kebilidi, Luxturna, Roctavian, Upstaza, and Zolgensma [52, 50]).

Commonly cited advantages of AAV in their use as gene therapy vectors include their wide natural tropism, ease of manipulation, and apparent non-pathogenicity[1]. The advantages of AAV quickly captured the attention of researchers, leading to the beginning of its development into a gene delivery vector just 15 years after its accidental discovery. The entire AAV2 sequence was first cloned into a bacterial backbone in the early 1980s [57, 58], followed soon after by the first use of AAV as a mammalian cloning vector [59]. In this latter work, the *cap* gene of an AAV2-containing plasmid was replaced by a neomycin resistance gene and successfully used to transduce human D6 cells, granting them neomycin resistance. Fully recombinant AAV (rAAV) particles were first generated in the Lab of Nicholas Muzyczka in 1988 [60], by completely replacing the *rep* and *cap* genes between the ITR sequences. rAAVs are incapable of autonomous replication regardless of the presence of a helper virus and thus benefit from increased safety over lytic viruses. AAV2 remains the best-studied serotype of AAV and still supplies the ITR and *rep* sequences for the majority of rAAV productions to this day. However, the *cap* sequence is routinely pseudotyped by those from other AAV serotypes [61, 62, 63].

There are various methods of producing rAAVs from cell culture. The procedure that asserted itself as an industry standard is the triple transfection of HEK293 cells with three plasmids containing Adenovirus helper functions, the *rep/cap* plasmid, and a plasmid containing the ITR-flanked transgene [64, 65, 66]. This procedure does not produce any

[1] Recently, the non-pathogenicity of AAV has been brought into question when high levels of AAV2 alongside human adenovirus and herpesvirus were detected in children during an outbreak of acute hepatitis [53, 54, 55]. A complete causative relationship between AAV detection and hepatitis could not be confirmed as of the writing of this thesis [56].

contaminating helper Adenovirus or replicative AAV virus and is another development made to increase the safety of rAAV in clinical applications. The Adenohelper plasmid (AdH) replaces the helper virus components and contains the adenoviral genes E2A, E4, and the non-coding VA RNA. The remaining helper functions are supplied by the HEK293 cells that were originally immortalised by integration of the adenoviral genes E1A and E1B and subsequently express them [67]. The *rep/cap* plasmid does not contain ITR sequences and supplies the AAV-related genes *rep* and *cap* in trans, which are necessary for viral genome replication and encapsidation (section 1.1.2). The third and final plasmid contains a gene-of-interest that is cloned into an acceptor plasmid between two ITR sequences, which are recognised by the Rep proteins and ultimately packaged into an AAV capsid.

The use of rAAV as a gene therapy vector entered the clinic in the late 1990s with the injection of AAV2-based vector encoding the coagulation factor F.IX for the treatment of severe haemophilia [68, 69, 70]. These clinical trials could prove the concept of AAV-based gene therapy in humans, but were unable to alleviate the disease phenotype, as the expression was transient and too weak. The low efficiency of the first AAV trial was attributed to the adaptive immune system and cytotoxic action of CD8-positive cells towards transduced hepatocytes [71, 72]. Next to the cellular immune response, the innate immune system response [73] and persisting humoral immune response [74] all contribute to the low efficiencies and therefore high doses necessary for past and ongoing clinical trials.

Capsid engineering and DNA family shuffling

To overcome some of the drawbacks, much of the research in the past two decades has focused on the generation of engineered capsid variants. These variants can be generated by a plethora of methods, which can be roughly sorted into rational designs or random diversification. Rationally designed capsid variants might include the insertion of a known receptor-binding peptide into one of the nine variable loops of the AAV capsid to retarget the resulting variant [75]. This approach requires *a priori* knowledge about an eligible binding partner and is therefore limited by the known biology. The generation of diversified libraries and their subsequent selections based on desired properties do not require such knowledge. These methods include but are not limited to peptide display libraries [76, 77], mosaic capsids [78], error-prone PCR approaches [79, 80], or DNA family shuffling [81, 82, 83].

Despite the efficacy of alternative methods, here I will emphasise DNA family shuffling, given its role in the presented results. DNA family shuffling is based on random fragmentation of closely related DNA molecules and the subsequent re-assembly by PCR reactions [84]. Parental sequences are the *cap* sequences of closely related parental AAV serotypes, which are shuffled in this approach to create novel *cap* sequences. The shuffling

breaks and disrupts the coding sequences of the *cap* gene and reassembles the breakpoints into putatively functional and novel VP1, VP2, and VP3 coding sequences. The novel sequences ideally combine properties of the parental serotypes or even introduce new characteristics, such as a specific tropism or increased antibody evasion. Naturally, many shuffled sequences do not result in improved or even functioning capsids and need to be deselected from capsids with desired characteristics. This is done through the application of a suitable selection pressure, such as the shuffled capsid's ability to transduce cells of a preferred host [85] or selection in the presence of a pool of human antisera [81]. DNA family shuffling yielded widely appreciated engineered variants such as AAV-DJ [81], AAV-LK03 [85], or AAVMYO2 and 3 [86].

1.1.4 AAV infection and episome formation

AAV infection usually follows the steps: attachment, uptake, trafficking, endosomal escape, nuclear entry, and uncoating. The exact mechanism highly depends on the capsid serotype in question.

Cell attachment of AAV can be mediated by multiple molecules. Heparan sulfate proteoglycans were the first discovered cell-entry factors identified for AAV2 [87]. They are also important for AAV3 and AAV6, but binding of other serotypes can be facilitated by different molecules [88], such as sialic acids for AAV4/AAV5 [89] or N-linked galactose for AAV9 [90]. Binding stabilisation and internalisation are mediated by different membrane-bound co-receptor proteins. A prominent example is the so-named AAV receptor (AAVR) that has been shown to be necessary for the transduction of multiple serotypes [91]. Given the name, AAVR was initially believed to be the primary receptor necessary for AAV binding, but it has later been updated to be an entry factor, mostly necessary for post-attachment steps [92, 93]. The transduction of other AAV serotypes is dependent on other co-receptors to stabilise AAV cell binding, such as different integrins or laminins [94, 95]. Internalisation of AAV2 can occur through multiple pathways, as the clathrin-dependent pathway, pinocytosis, and clathrin-independent carriers all seem to be involved [96, 97, 98]. For successful transduction, AAV needs to travel to the Golgi apparatus through retrograde transport, probably with the help of the aforementioned entry factor AAVR. [99]. During its endosomal trafficking, the AAV capsid undergoes conformational changes that culminate in its release into the cytosol. The VP1- and VP2-unique regions (VP1u and VP2u) that are usually hidden on the lumen side of the capsid become exposed, triggered by a conformational change induced by the lowered pH in the endosome [100, 101]. The N-terminus of VP1 contains a phospholipase domain [102], whose externalisation leads to the disruption of the endosome and release of the AAV virions into the cytoplasm. It is

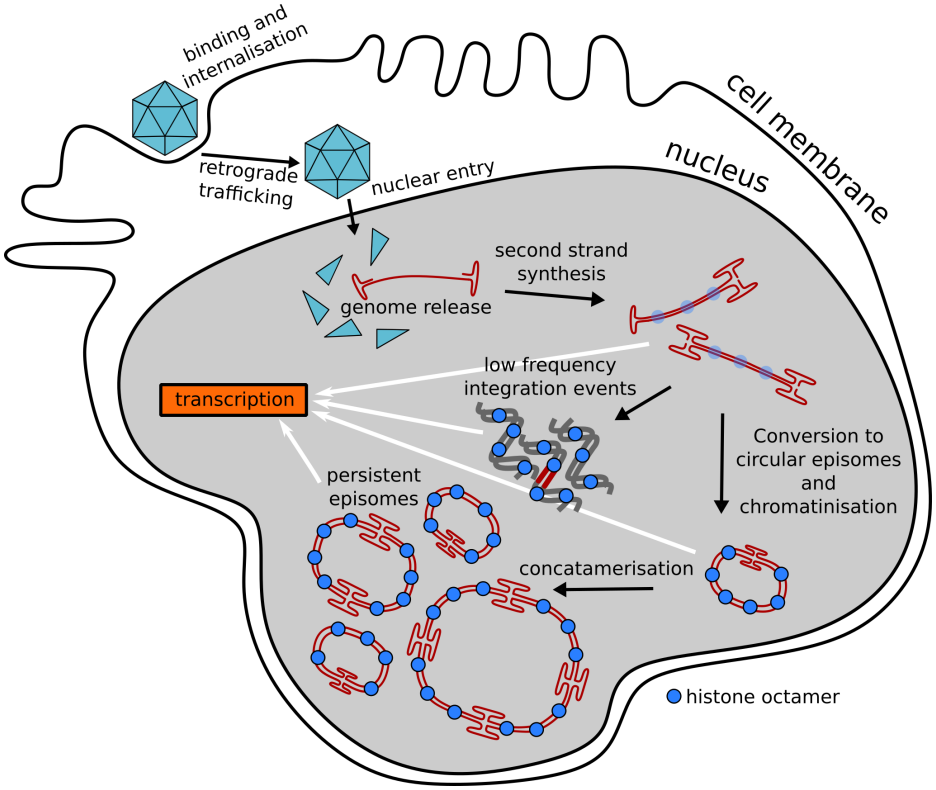

Figure 1.2. Schematic of rAAV episome formation.
The AAV-delivered DNA is shown in red, and the cell genomic DNA in dark grey. Histone octamers are displayed in blue. After genome release, the linear double-stranded episome is formed by second-strand synthesis. Circularisation occurs with the help of cell DNA damage response proteins. Circular episomes are associated with histones [104]. The association with linear episomes is not known but plausible, hence the implied histones in light blue. Histones and DNA are not drawn to scale. The delivered DNA can integrate into the genomic DNA at a low frequency and contributes to transcription. High-molecular-weight concatamers can be found years after transduction. Visualisation adapted from [105].

commonly assumed that the AAV particle enters the nucleus intact through the nuclear core complex [99]. Inside the nucleus, the genome is released, possibly caused by a genome ejection mechanism induced by another conformational change of the capsid [103].

Episomes are DNA molecules that exist alongside chromosomal DNA without being integrated into it. *In vivo*, persistent rAAV transduction is commonly thought to be conveyed by circularised double-stranded DNA episomes, which can also persist as high-molecular-weight molecules formed by recombination events. Before circularisation can occur, the ssDNA genome needs to be converted to a dsDNA molecule. The ITR sequences prime the synthesis of the second strand, which presents a rate-limiting step for AAV transduction [106]. The ITR sequences on both ends of the linear DNA are essential for the recombinations and subsequent circularisation, and also seem to stabilise resulting episomes [107, 108, 109]. The preferred episomal state seems to be head-to-tail monomeric episomes that form as a result of intermolecular recombination [109]. The presence of

circular AAV episome has been shown in primate muscles up to 5 years post-transduction with capsids AAV1 and AAV8 [104]. Persistent episomes extracted from primate liver were more heterogeneous and subject to multiple recombinations compared to muscle episomes [110]. Kinetics of episome formation (and of the preceding second-strand synthesis) can paradoxically be influenced by the used AAV capsid [111, 112], but the exact mechanism through which the episome is circularised is not fully solved. The circularisation involves the activity of polymerases α, β, and ϵ and the activity of topoisomerases [113, 108]. A follow-up study, interrogating DNA repair pathways, found that neither non-homologous end joining nor homology-directed repair was independently responsible for ITR recombination and circularisation [114]. In a screen in recombination-deficient cell lines, the authors could mainly identify members of the MRN complex, responsible for the sensing of double-strand breaks, and the DNA damage response downstream actor Ataxia Telangiectasia Mutated (ATM), as influencing the rate of circular episome formation. There is also evidence that the shape of the ITR hairpin at the ends of the episomal molecule (more than its sequence) affects which DNA damage response pathways are activated, enabling circularisation [115].

Integration versus episome formation

Wild-type AAV (wtAAV) has been shown to be able to perform a latent infection in the absence of a helper virus in which it integrates into a specific locus on chromosome 19 known as AAVS1 [116, 117, 118, 119]. AAVS1 (and the ITR sequences) contains a Rep recognition site that interacts with ITR sequences in the presence of Rep proteins and is probably the cause of integration of wtAAV [117]. There are multiple more Rep binding sites identifiable on the human genome, but AAVS1 is the preferred locus for integration [120]. However, only about 0.1-0.5% of infectious wtAAV particles perform an integration [121]. In the absence of helper viruses, the wtAAV genome can also exist episomally in infected cells. For example, in a screen of 175 tissue samples for wtAAV, only dsDNA circular episomes and no integration events could be observed [122].

In contrast, rAAV predominantly does not integrate into the genome, since it lacks the necessary Rep activity. However, low-frequency integration events of rAAV have long been known [123]. The integration sites are distributed semi-randomly over the genome. They occur preferentially at sites of dsDNA damage and transcriptionally active genes, such as ribosomal RNA repeats [124, 125, 126]. Recent results suggest that integration events are more frequent and more important to transduction persistence than previously believed [127]. In this study, although initially high, the transgene expression in transduced primate livers rapidly declined over two months to a lower but steady level. Detected integration events consisted mostly of highly concatemerized vector genomes that were

found in about 1% of liver cells. Additionally, the number of integration events correlated with the low long-term expression in liver cells up to two years post-injection. This raises safety concerns, even though in this study, the integration events were not close to genes associated with hepatocellular carcinoma.

1.1.5 Transgene expression persistence

The long-term expression of rAAV-delivered transgenes can vary significantly. Among other factors, this can depend on the delivery method, the applied dose, the receiving species, or even the treated individual. Even though rAAV is used to treat a plethora of monogenic diseases, most of the long-term data in the literature on AAV transductions are from trials treating different forms of haemophilia. As already briefly described above, the first clinical trial using rAAV in humans as a gene delivery vector resulted in relatively low levels of human factor IX expression that were completely abolished around 8 weeks after injection [70]. The decline was linked to T-cell-mediated destruction of transduced hepatocytes. A preceding study on the long-term expression of canine factor IX in hemophilic dogs (haemophilia B) was able to correct their disease phenotype up to 1.5 years after injection with a dosage in the same range ($1e12$ vg/kg in dogs versus $2e12$ vg/kg in humans) as in the clinical trial in humans [128]. More recently, all subjects in a 10-year study on haemophilia A in dogs had therapeutic levels of transgene expression at the conclusion of the study [129]. Interestingly, some dogs showed an increase in transgene expression years after the initial treatment that was linked to the clonal expansion of cells with integrated vector genomes. There can also be major differences between treated individuals. Human liver biopsies performed two to four years after rAAV treatment for haemophilia A (Roctavian) showed substantial differences between individual recipients [105]. Circular full-length AAV episomes were detected in all responders, but inter-individual differences in expression were also detected. They were directly associated with the differential expression of transcription regulators, but were also hypothesised to be caused by differing epigenetic modifications.

From these examples, it becomes clear that the long-term efficacy of AAV transduction is difficult to predict and is modulated by multiple factors [130]. To this date, the main research focus is on immunological factors, such as the innate response of Toll-like receptors against viral DNA [131, 132] or (as already mentioned) cytotoxic T-cell response against transduced cells [70, 133]. Epigenetic factors are recognised as potential modulators, but the direct analyses of such modifications on the delivered transgenes are rare.

1.1.6 Transgene CpG methylation

Cytosine methylations of the CpG motif are a staple of mammalian epigenetic modifications. They have a clearly defined role during development as they are indispensable for correct cell differentiation and tissue development [134, 135]. During early mammalian embryogenesis, the entire methylome is reset to prepare the embryonic cells for cell fate decisions [136]. Within the differentiated cells of mammals (and most other vertebrates), the vast majority of CpG motifs are methylated. The endosomal Toll-like receptor 9 takes advantage of this by recognising DNA with unmethylated CpGs as foreign and initiating an immune response, which can also be detrimental to rAAV transductions [137, 73]. CpG motifs appear non-randomly on the genome, mostly within so-called CpG islands, which predominantly reside in promoter regions. The methylations of CpG islands are generally associated with the inactivation of genes and heterochromatin formation, as they inhibit the binding of transcription factors and recruit repressive proteins [138, 139]. However, this correlation is not so simple as the methylation of genes can sometimes also increase transcription [140]. Artificial *de novo* methylations were associated with heterochromatin formation, but about 48% of methylated genes showed either no change or even higher expression [141].

AAV gene therapy is based on the delivery of therapeutic DNA, preferably to differentiated and quiescent cells, in which epigenetic methylation patterns are already established. It is therefore of interest how *de novo* methylations might affect the expression of a transgene. *De novo* methylations are generated by DNA methyltransferases 3A and 3B (DNMT3A, DNMT3B). The effects of methylation on foreign DNA have already been studied in the 1980s, where the authors found no noticeable methylation on a plasmid that was used to transfect mammalian cells in culture [142]. Similarly, *de novo* methylations played no role in the silencing of plasmids that were used to transduce mouse livers *in vivo* [143]. In contrast, when using a viral vector, methylations appear to be a bigger hurdle for efficient transcription. The Cytomegalovirus (CMV) promoter within an adenovirus vector was subject to substantial methylation starting only one day after transduction in mouse muscle tissue [144]. Gene therapy based on retroviruses is also hampered by interfering DNA methylations of the vector cargo [145].

Studies on the methylations of the AAV-delivered episomal transgene are sparse and in part contradictory. In wild-type AAV and rAAV, the packaged genomes are not methylated [146, 147]. Transgenes delivered into primate muscle tissue by an AAV1 capsid and driven by the Rous Sarcoma Virus (RSV) promoter did not show increased methylation even 37 months post-injection [148]. However, this work only interrogated the methylation of two CpG-island regions within the RSV promoter. In contrast to this result, transduction is

markedly increased when growing cells in the presence of the methyltransferase inhibitor azacitidine, indicating that *de novo* methylation has a repressive effect [149]. None of these studies interrogated the methylation of specifically episomal AAV-delivered DNA and the possible influence of the capsid.

Bisulfite and enzymatic methyl sequencing

The presence of 5-methyl cytosine (5mC) is classically detected through conversion techniques such as bisulfite conversion. In this approach, DNA is treated with sodium bisulfite, which reacts with cytosine (C) bases in the DNA and deaminates them, resulting in uracil (U) [150, 151]. Because of its pairing properties, U is converted into thymine (T) in subsequent PCR amplification steps. 5mC is protected from the deamination and subsequent conversion to T. In combination with next-generation sequencing (bisulfite-seq), the sequences are aligned to a reference with conversion-aware aligners, such as 'bwa-met' [152]. For the assessment of CpG methylation percentages, the number of converted T and unmodified C are compared at every reference CpG site. The treatment with sodium bisulfite can degrade the DNA substrate, which can induce a bias in the conversion of DNA and ultimately coverage rates [153]. An enzymatic variant of this technique has been developed as a modern alternative, called enzymatic methyl sequencing (EM-seq) [154]. It is fundamentally identical to bisulfite-seq, but uses APOBEC3A (Apolipoprotein B mRNA Editing Enzyme, Catalytic Subunit 3A) as a deaminase. In contrast to bisulfite, APOBEC can also deaminate 5mC. Therefore, additional protections of the methyl group must be undertaken. This is achieved by the joint action of TET2 (Tet Methylcytosine Dioxygenase 2) and T4-BGT (T4-phage Beta-Glucosyltransferase). TET2 oxidizes 5mC into 5-hydroxymethyl cytosine, which can then be glycosylated by T4-BGT, thus protecting it from deamination by APOBEC. This is followed by PCR amplification, sequencing, and data analysis as would be done for standard bisulfite sequencing. EM-seq is a milder alternative to bisulfite-seq that can outperform it in conversion rate, input needed, and DNA methylation accuracy (among other metrics) [154].

Nanopore sequencing

Nanopore sequencing is a powerful alternative as it can directly detect base modifications without the need for lengthy or harsh conversion protocols. Starting with the sample, it takes less than an eight-hour workday to obtain the first sequences. The greatest advantage of nanopore sequencing lies in its ability to sequence any length of DNA molecule, which has been recently put to use in the first telomere-to-telomere assembly of a human genome [155].

Nanopore sequencing is commercialised through the company Oxford Nanopore Technologies (ONT). In this sequencing method, a protein pore (nanopore) is embedded in a synthetic membrane along which a voltage is applied. This causes a flow of charged ions in solution to travel through the pores. During library preparation, a motor protein is ligated to the sample DNA molecules. For sequencing, the motor protein binds the nanopore, and a single strand of DNA is pushed through the pore in an ATP-dependent mechanism. The direction of the DNA translocation is opposed to the flow of ions through the pore. In this translocation, every base of the DNA molecule characteristically blocks the flow of ions and therefore also induces minuscule but measurable changes in the current. These changes in current (called the squiggle) can be translated back to a nucleotide sequence in a process termed basecalling. The chemistry used in ONT-sold flow cells was recently changed from the prior R9 chemistry to the R10 chemistry. The most notable improvement in this update is the introduction of a second 'reader' in the pores of the flowcell, which is essentially another bottleneck that the DNA molecule must pass through, which increases the accuracy of the produced squiggle [156]. The algorithms responsible for basecalling from the squiggle have similarly undergone substantial evolution, from Hidden-Markov-Models in the early versions to bi-directional recurrent neural networks in the latest programs 'guppy' (mostly used during the R9 chemistry lifespan; closed-source) and 'dorado' (latest program for calling; open-source [157]). The models used within the basecallers are trained on known, diverse DNA molecules to enable accurate basecalling. Because the shape of the nanopore is different between the R9 and R10 chemistry, the shape of the squiggle can also be fundamentally different. As a result, the basecaller models are largely not exchangeable between the different chemistries. Since 5mC is physically different to unmodified cytosine, the squiggle can also be used to measure the methylation state of individual cytosines in the sequenced DNA with good correlations to bisulfite sequenced controls [158]. There are multiple open-source callers for methylations from ONT sequencing data, most of which were established and trained on the R9 version [159]. The models used within 'dorado' are continuously being updated and are currently the state-of-the-art in R10 methylation calling [160].

1.1.7 Transgene histone modifications

The genomic DNA within eukaryotic nuclei is wrapped around histones and folded into a higher-order chromatin structure. The association of histones with the underlying DNA can lead to densely packaged DNA that is transcriptionally inactive, termed heterochromatin. More loosely packaged and accessible DNA is termed euchromatin. The organisation of the genome into hetero- and euchromatin is the key to differential gene regulation and

cell identity [161]. Histones are an octameric protein complex consisting of two H2A-H2B and two H3-H4 dimers. Around 147 bp of DNA are wrapped around a histone core, forming a nucleosome [162] (also see Figure 1.3 C). The N-termini (sometimes also the C-termini) of each individual histone protein are known as the histone tails. They protrude from the octamer core and are essential for the regulation of chromatin formation. By themselves, they function as internucleosomal linkers (H4 tail) [163] or stabilise the nucleosome (H3 tail) [164]. Because of their location close to the DNA entry and exit sites on the histone, the H3 tail has a disproportionately high impact on transcription regulation through DNA interactions compared to other histone tails [165, 166]. Histones unlock their full role in epigenetic regulation through multiple post-translational modifications (PTMs). More than ten different modifications occurring at over 60 amino acids on the histone tails have been described so far [167]. Among the most common ones are methylations and acetylations, which occur on lysine residues on the histone tails. Acetylations are associated with highly transcribed genes. They neutralise the charge of the lysine and thus increase electrostatic repulsion between the negatively charged DNA backbone and the DNA, leading to a more open chromatin state. This allows the binding of transcription factors or other members of the transcription machinery [168]. For example, the acetylations of the H3 tail at lysine 27 (H3K27ac) and at lysine 9 (H3K9ac) are associated with actively transcribed retrotransposons and are generally enriched at transcription start sites [169, 170]. Methylations of the histone tails can be associated with active or repressed transcription, depending on the location of the modified lysine. For example, whereas H3K4me3 (triple methylation of lysine 4 in the histone 3 tail) is found at actively transcribed promoters, H3K27me3 is heavily associated with silenced genes and heterochromatin [170]. Even though they are both associated with gene repression, CpG DNA methylation and H3K27me3 are mutually exclusive [171, 172, 173]. The presence of CpG methylation inhibits the action of polycomb repressive complex 2 (PRC2), which is responsible for the methylation of H3K27. Similarly, the presence of H3K27me3 can inhibit the access of DNMTs to DNA.

The circular AAV episome is bound by histones in transduced primate muscles [104] and presumably in other tissues as well. It is therefore plausible to assume that histone modifications can also influence the expression of the transgene. It has already been shown in the 1990s that silenced rAAV genomes can be reactivated by the Histone deacetylase (HDAC) inhibitors butyrate and trichostatin [174]. HDACs are responsible for the removal of acetyl groups from histone tails and act antagonistically to histone acetyltransferases (HATs). A similar increase in transduction was later observed with a different HDAC inhibitor in multiple cancer cell lines and additionally associated with increased association

of acetylated H3 with the AAV-delivered episomes [175]. The human silencing hub (HUSH) complex, through interaction with the DNA-binding protein NPP20, was shown to be directly involved in the epigenetic silencing of rAAV-delivered DNA [176]. HUSH directly deposits the repressive histone modification H3K9me3 also on DNA delivered by the murine leukemia virus. Furthermore, the authors of the same study are among the first to provide evidence that the used capsid influences the HUSH-mediated silencing, as the transduction through capsid AAVrh32.33 was especially improved by the knockdown of NPP20. Very recently, more evidence has been reported that suggests the AAV capsid itself to be able to alter the epigenetic modifications of the delivered DNA [177, 178]. Gonzalez-Sandoval et al. showed that the different transduction abilities of capsids AAV-LK03 and AAV-DJ are linked to decreased levels of the activating marks H3K27ac and H3K4me3 on the AAV-LK03-delivered genomes. They could furthermore rescue the phenotype with a single amino acid insertion into AAV-LK03 [177]. Loeb et al. transferred the N-terminal part of the VP1-unique region (VP1u) of AAV8 onto an avian AAV, which is ordinarily not able to transduce human cells. The transferral of the N-terminus rescued transduction in human cells and was correlated with increased levels of multiple histone PTMs on the delivered DNA [178].

Cut and Tag sequencing

The detection of histone modifications on DNA has been subject to major revolutions in the past decades. Chromatin immunoprecipitation and subsequent sequencing by next-generation short-read techniques (ChIP-seq) had established itself in the late 2000s [179]. It is based on the crosslinking and subsequent fragmentation of DNA together with its bound proteins. A specific antibody against the protein or PTM of choice (e.g. against H3K27ac) can be used to pull down pieces of DNA attached to it. Sequencing and alignment of the fragments to a reference genome reveal locations of binding along the reference. ChIP-seq is plagued by relatively high levels of background, which were improved upon by the introduction of techniques such as CUT&RUN (Cleavage Under Targets and Release Using Nuclease) or CUT&Tag (Cleavage under Target Tagmentation) [180, 181]. Both of these are based on the *in situ* binding of an antibody to a protein or PTM of interest within a permeabilized nucleus, without the need for crosslinking. CUT&Tag works by tethering a protein A-fused Tn5 transposase to the bound antibodies. The addition of magnesium activates the transposase that has been pre-loaded with sequencing adapters. The adapters are inserted in the vicinity of the protein or PTM in question and the resulting fragments can be subsequently amplified by a PCR reaction. CUT&Tag yields a higher resolution and coverage of individual peaks compared to ChIP sequencing, while requiring a significantly lower amount of overall reads [181].

1.2 Patterns in genomic DNA

1.2.1 Distinct repeating elements

The first draft of the human genome from the beginning of the millennium has revealed that only about 1% of the genome can be regarded as coding DNA [182]. The search for meaning in the non-coding part of the genome has resulted in the discovery of many repeating DNA sequences. About 50% of the human genome consists of such repeating sequences, whose functions remain enigmatic to this day [183]. Genomic DNA repeats are roughly classified into interspersed and tandem repeats.

Interspersed elements include short and long interspersed nuclear elements (SINE and LINE, respectively) and are found all across the human genome. They are mostly comprised of transposable elements (transposons) that can either directly or indirectly change their location within the genome. A well-studied example of a SINE is the primate Alu element and its various subfamilies that comprise about 10% of the human genome [184]. The Alu element is implicated in the regulation of multiple human genes and has recently been discovered to be involved in the generation of enhancer-promoter loops [185, 186]. Given their interspersed nature, these elements do not appear periodically.

Tandem repeats (also called satellite DNA) are repeating stretches of DNA that can range in size from 2 to 1000s of base pairs. They are the main components of centromeric and telomeric regions in eukaryotic chromosomes, in which they appear periodically one after another.

Centromeres are genomic regions within every chromosome that enable the formation of the kinetochore, an essential component of the mitotic spindle apparatus. The human centromeric sequences consist of periodically repeated tandem repeats of 171 bp monomers called alpha-satellites [187, 188]. Currently, the formation of the kinetochore is generally regarded to be epigenetically orchestrated and not sequence-dependent. The kinetochore is assembled with the help of nucleosomes containing a centromere-specific H3-variant that can create neo-centromeres independent of the underlying sequence [189, 190]. The centromeric tandem repeat sequence is speculated to be a by-product of unequal exchange between sister chromatids [191].

The telomeric tandem repeat in humans is made up of multiples of the hexanucleotide TTAGGG repeated to a length of several kilobases at the ends of every linear chromosome [192]. Telomeres are associated with the capping shelterin complex that protects the DNA ends from being recognised by the cellular damage response machinery [193]. Because of the nature of DNA replication and the linearity of eukaryotic chromosomes, the 3'-end

of a parental DNA molecule cannot be replicated entirely, leading to a shortening of the telomeres with every round of replication in telomerase-deficient cells. Once the telomeres become too short to bind enough capping protein, the DNA damage response is triggered and proliferation is inhibited [194]. Unlike the centromeric tandem repeat, the telomeric tandem repeat has been directly linked to a distinct structure. The G-rich forward strand of the telomeric tandem repeat sequence has been shown to form G-quadruplexes under molecular crowding conditions [195, 196]. G-quadruplexes are an alternative structure to the Watson-Crick DNA pairing, utilising Hoogsteen base pairing. As the name suggests, G-quadruplexes can only be observed between Guanine (G) bases in G-rich strands. One plane of a G-quadruplex consists of four Guanine bases stabilised by a centrally bound monovalent cation (commonly potassium). The formation of this DNA abnormality has been linked to transcriptional regulation as they are often found near promoter regions [197]. Less is known about the biological significance of this structure in telomeres, but it has been implicated in the correct bundling of homologous chromatids during meiosis and telomeric maintenance [198, 197].

1.2.2 Fuzzy periodic patterns

The repeating patterns described hitherto are noticeable with a macro-view of the genome. It is relatively apparent when a specific sequence is duplicated, e.g. an Alu element can be accurately assigned as such, no matter where on the genome it is located. However, there are more inconspicuous patterns that can emerge when looking at a genome in a more detailed scope. These patterns usually only become apparent when analysing and averaging over multiple kilobases of sequence information at the same time.

The 3-bp pattern of coding sequences

An example of this is the non-random nucleotide distribution within coding sequences. They exhibit a 3-bp periodicity of multiple nucleotides [199, 200]. This periodicity of coding sequences can be explained by the nature of codons being three base pairs long and a certain dominance of some synonymous codons over others. This dominance is proposed to be due to modern genomes containing vestiges of a primordial codon consisting of the bases RNY (R being purine, A or G; Y being pyrimidine, C or T; and N being any nucleotide) [201, 202, 203]. This RNY-based codon is simpler but can still encode eight amino acids, each encoded by two codons [204]. The amino acids that can be encoded by RNY are congruent with the chronological order of appearance [205]. It would have constituted the genomic information in the RNA world and was also proposed as a potential monitoring device during translation [206, 207]. Over time, this primordial codon is suggested to have mutated into

the codons we have in modern organisms, as remnants of this primordial pattern are still detectable. This characteristic DNA pattern can be exploited in the search for exons within coding sequences on newly assembled genomes. In this approach, the character string of a DNA sequence is converted into a numerical string that can be utilised for an analysis of its frequency domain [208, 209]. A protein coding sequence can be correctly classified as such if it portrays a high power density at the frequency of one-third corresponding to a period of 3.

The 10-bp dinucleotide periodicity

Next to the 3-bp periodicity, large-scale patterns were also discovered for repeating dinucleotides in genomic sequences [199, 210], of which the most prominent and researched is the 10.4-bp-periodicity of specific dinucleotides in eukaryotic DNA. The exact repeating dinucleotide motif can differ from genome to genome. For example, the strongest 10-bp periodicity in humans is the CG dinucleotide [211]. In *C. elegans* it is the YY and RR dinucleotides [212], in yeast the AA and TT dinucleotides [213], and in *Arabidopsis thaliana* the GG dinucleotide [214]. The fact that the detectable periodic pattern occurs for multiple sequences suggests that the underlying physical characteristic of the repeat, not the exact sequence, is key.

The three-dimensional structure of the DNA molecule is affected by its sequence. The classical shape of the double-stranded B-DNA double helix is defined by its components: the two strands of phosphate-sugar backbone and specific base pairings as well as base stackings in between (Figure 1.3 A). The shape of a DNA molecule is essential for its capability to interact with a protein and can thus be used to predict DNA-protein interactions [217]. Eluding to its importance for DNA shape, the physical properties of a base stack (two base pairs; a dinucleotide) are more important to DNA stability than are base pairs [218]. Additionally, they are key determinants of the non-planar deformability of the DNA molecule [219]. Two exemplary and important physical properties of a dinucleotide sequence are its Roll and its Helical Twist (Figure 1.3 B). For example, the TA dinucleotide has a significantly positive Roll angle, meaning at this base step the DNA molecule is characteristically bent towards the major groove [220]. If a dinucleotide appears periodically and thus the same Roll angle is present at a periodic interval, it can impart a large-scale bend into the DNA molecule [221]. This has been proposed as the 'wedge model' for DNA bending [222].

The bend resulting from the 10-bp-periodicity that is detectable in eukaryotic DNA is implicated in the facilitation of nucleosome formation [199]. This idea has been expanded into a potential DNA sequence code for the positioning of nucleosomes [216, 223]. With a

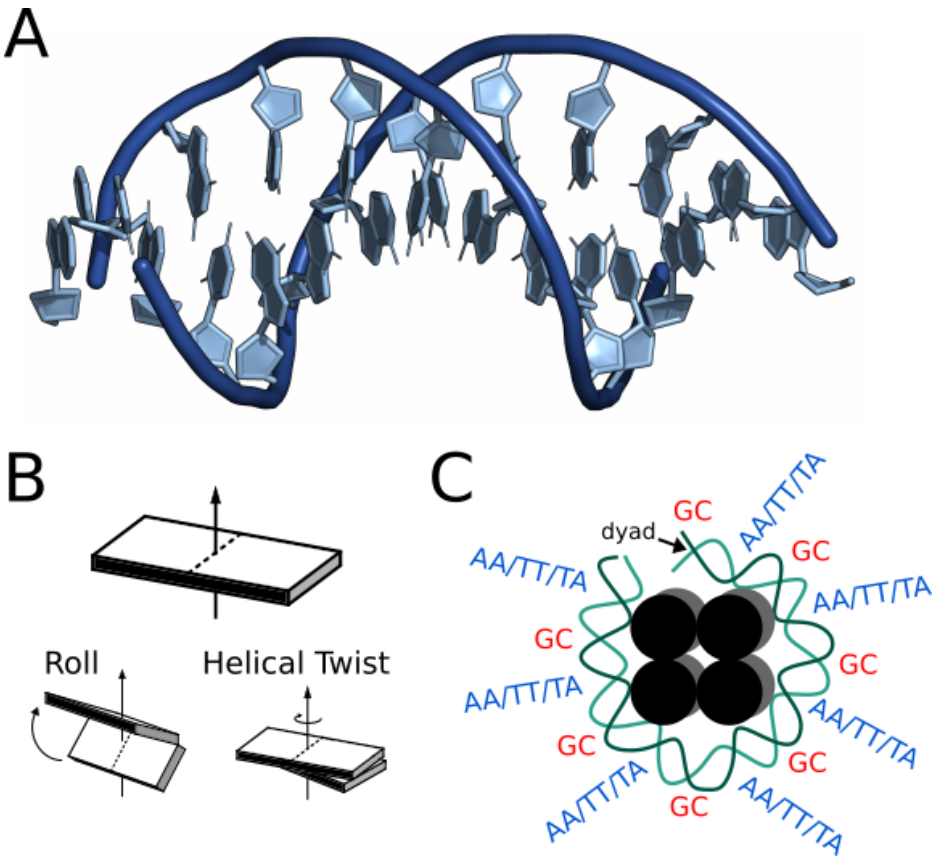

Figure 1.3. Influence of dinucleotides on DNA structure and DNA-protein interactions. A) Structure of PDB 1BNA. Phosphor backbone in dark blue. Deoxyribose and bases in light blue. B) Top: Representation of base-pair as a rectangular plane. Bases are paired roughly at the dotted line. Bottom: Examples of Roll and Helical Twist that can differ depending on the sequence of a dinucleotide. Representations are exaggerated. Adapted from [215]. C) The preference of chicken and yeast nucleosomal DNA for periodic sequences. DNA is wound around a histone octamer (black). GC and TT/AA/TA dinucleotides have a 10-bp periodicity that is offset with regard to each other by 5 bp. The dyad position is indicated with an arrow. Adapted from [216].

probabilistic model based on dinucleotide distributions, Segal and colleagues were able to accurately predict the sequence bias for nucleosome formation and state that around 50% of all nucleosomes *in vivo* are positioned by sequence-specific characteristics. In particular, chicken and yeast nucleosomal DNA have a 10-bp periodic pattern of the AA/TT/TA dinucleotides with a periodic 10-bp GC dinucleotide repeat shifted by 5 bp ([216], Figure 1.3 C). The sequence alone is insufficient to fully explain nucleosome positioning *in vivo*. Other models were proposed to complement the sequence-dependent positioning. They suggest that the first nucleosome after a transcription start site (+1 nucleosome) creates a barrier against which the following array of nucleosomes is packed [224, 225, 226]. However, the positioning of the -1/+1 nucleosomes is considerably influenced by the underlying sequence [224].

A similar 10- to 11-bp periodicity can also be detected in prokaryotic sequences [227]. This is intriguing as they have a widely different genome architecture that typically does not encode classical histone proteins. The pattern in prokaryotes has been related to the coiling of the circular prokaryotic genomes. Eubacteria usually have a periodicity of around 11 bp that is associated with negative supercoiling, whereas extremophiles portray a slightly shorter period around 10 bp that seems to facilitate positive supercoiling [228, 229]. This has been recently put into question in an analysis that could not correlate periodic length with optimal growth temperatures, even though the periodicity itself seems to be under selection [230]. The authors of this recent work instead suggest a processive moiety or other binding partner as a reason for the selection of prokaryotic periodicity, similar to histone binding in eukaryotic genomes.

1.2.3 Patterns in viral DNA

Fuzzy sequence patterns, as described so far, are understudied in viruses. The patterns are usually only clearly visible when multiple kb of sequence information are analysed simultaneously. Therefore, the relatively short viral genomes were largely ignored for such assessments, with a few exceptions.

3-bp coding sequence periodicity

The 3-bp periodicity detectable in multiple genomes (see section 1.2.2) can also be found in viral genomes. Howe and Song showed that an amino acid bias is insufficient to explain the detected periodicity and also presented a spectral analysis method of viral sequences as an alternative to retrace evolutionary steps [231].

A study on the genomic sequence of SARS-CoV-2 (severe acute respiratory syndrome

coronavirus 2) could also identify a 3-bp periodicity in its RNA genome. The author correlated its strength to the evolutionary fitness of the virus [232].

Larger scale periodicities greater than 3 bp

Interestingly, the genome of the simian virus 40 (SV40; a member of the *polyomaviridae* family) was among the first in which large-scale dinucleotide periodicities were discovered, as it portrays a relatively strong 10-bp-periodicity [199, 233, 234]. SV40 presents a unique case among viruses because its genome is associated with histones both in the nucleus, where the SV40 genome forms minichromosomes, as well as in the virion, where its association with histones remains. SV40 has long served as a model for chromatin organisation, among other processes, such as DNA replication and transcription regulation [235]. Therefore, it appears obvious that the periodicity on the viral genome serves the same purpose as on the eukaryotic genomes and SV40 likely has evolved a genome that promotes nucleosome formation [236, 233].

There are conflicting reports about patterns in the genomes of bacteriophages. One large-scale study suggests periodicities in phage genomes to be much weaker compared to prokaryotes [237]. This is somewhat surprising because phages are known to pack their genomes extremely tightly into the capsid [238]. Therefore, it seems likely that their DNA has also evolved structural features such as dinucleotide periodicities to facilitate the tight bending that is necessary for their packaging. Perhaps DNA bends, such as the ones inducible by periodic dinucleotide repeats, are not necessary for the almost crystalline structure that the packaged DNA needs to adopt in bacteriophages. Another study reports on the strong periodicity of *Erwinia* phages that also exhibit a circa 11-bp periodicity, similar to the periodicity in negatively supercoiling eubacteria [230]. This result is used to support the argument that such a periodicity is the result of the facilitation of interaction with an unknown DNA-binding protein and/or processive moiety.

Although no direct assessment of DNA sequence periodicity has been made, the nucleoprotein p6 of phage $\phi 29$ has been proposed to bind the phage genome based on the DNA's physical properties, such as its bendability [239]. A monomer of the p6 protein binds DNA every 12 bp and also shows a preference to bind DNA that can bend in these intervals [240].

In a study on the chromatin organisation of invading adenoviruses, a periodicity of about 5.4 bp for the SS and WW dinucleotides was uncovered in pVII-bound DNA stretches within the Adenovirus genome [241]. The pVII nucleoprotein is encoded by Adenovirus and coats the Adenovirus genome inside the capsid. Upon infection, adenoviral early transcribed

genes are more accessible, probably because they are less occupied by pVII proteins. The periodicity is suggested to guide the viral packaging and thus also the transcription regulation during infection.

1.3 Goal of this work

This thesis is split into two parts: (i) I wanted to elucidate the epigenetic modifications deposited on the AAV-delivered episome in time and thus performed capsid-dependent experiments in the mouse animal model, and (ii) I describe the discovery of a specific dinucleotide pattern in AAV genomic DNA and explored how it can influence different aspects of rAAVs.

Epigenetic regulation is a cornerstone of eukaryotes and dictates cell differentiation, development, and accompanying phenotypes of humans, animals, and plants alike. However, little information exists on the epigenetic marks on DNA delivered by the viral vector AAV and on the influence of the used capsid. In this part of the thesis, I explored how nanopore sequencing can be used to directly detect methylations on episomal DNA from transduced mouse livers and how it compares to more established techniques. Furthermore, I wanted to shine light on how epigenetic changes occur in AAV-delivered transgenes over a longer time frame of up to 12 weeks post-injection and how these are correlated with transgene expression. I employed nanopore sequencing to interrogate the changing methylation state of the episomal transgene and, additionally, perform CUT&Tag sequencing against the post-translational histone modifications H3K27ac and H3K27me3.

The second part explores a built-in sequence feature of *Dependoparvovirus* genomes. Many DNA motifs are detected by binding moieties not directly by sequence, but by physical properties that a DNA molecule with the motif possesses. An example of this is the dinucleotide periodicity that is observable in histone-binding DNA, which has a characteristic bend facilitating winding around the histone core. The genetic cargo of viruses needs to undergo severe compression in the capsids and also needs to efficiently bind its own and host proteins to undergo its lifecycle. Sequence patterns are severely understudied in the genomic sequences of viruses. In this work, I describe the discovery of a dinucleotide pattern unique to dependoparvoviral DNA, which is distinct from similar patterns in its repeat length. Furthermore, I wanted to characterise the function of this pattern concerning rAAV titer, transduction, and physical properties of the resulting vector.

2. Materials and Methods

2.1 Code availability

Throughout the Materials and Methods section, I will describe the function and purpose of multiple scripts and functions and will directly reference them by their file name. All mentioned files can be found under my **personal GitHub**[1], on zenodo [242], or within the group-share folder structure available from within the Grimm lab.

2.2 Cell culture

All cell culture work was performed in sterile conditions in laminar-flow cell culture hoods. Cell cultures were incubated at 37°C with 5% CO_2. HEK293T cells were cultured in 175cm² flasks (Greiner Bio-One) in Dulbecco's modified eagle medium (DMEM) GlutaMAX (Thermo Scientific) supplemented with 10% fetal bovine serum (FBS) and 100 U/mL Penicillin/Streptomycin (Pen/Strep). For cell harvesting, the medium was aspirated, and 0.25% Trypsin (Thermo Scientific) was added for 2 minutes or until cells were visibly dislodged. Trypsination was stopped by adding 10 mL DMEM (10% FBS, 100 U/mL Pen/Strep). Raji B-cells were grown in RPMI-1640 medium supplemented with 10% FBS and 100 U/mL Pen/Strep. Harvest was performed by centrifuging the suspension cell culture at 300xg for 5 minutes. Cells were counted with an automated Countess cell counter (Thermo Scientific).

2.3 AAV production

2.3.1 Transfection and Harvest

Large-scale AAV production was performed in 22 cm cell culture dishes seeded with HEK293T cells at $4e6$ cells in 22 mL DMEM (10% FBS, 100 U/mL Pen/Strep). Depending

[1] `https://github.com/ConradinBaumgartl/thesis_code_2025/tree/main`

on the amount of virus, either 1-5 dishes (small gradient) or 10-30 dishes (large gradient) were seeded with cells. After 2 day of incubation cells were transfected with 2 mL of transfection mix containing a final concentration of 300 mM NaCl, PEI MAX with an amine to phosphate ratio of 30, and a total of 44 µg of DNA with a 1:1:1 molar ratio of rep/cap plasmid, AdH, and transgene plasmid. The transfection mix was added drop-wise, and the dish was slightly agitated to mix. 3 days after transfection, the cells were harvested from the plate by using a cell lifter (Corning) and resuspended in their own medium. The cells were pelleted at 800 x g for 15 minutes and resuspended in 20 mL Benzonase buffer. Lysis of the cells was performed by 5 freeze-thaw cycles, after which Benzonase (Merck) was added to a final concentration of 50 U/mL. Digestion of unpackaged DNA was performed at 37°C for 1 hour while inverting the tube every 10 minutes. The digested lysates were then cleared from cell debris by 3 rounds of centrifugation at 4000 x g for 15 minutes. After every round, the pellet was discarded and the supernatant transferred to a new tube.

2.3.2 Iodixanol purification

AAVs in this work were purified by iodixanol purification. For small gradients using a glass long tipped pasteur pipette the cell lysates were transferred to centrifuge tubes (small gradient: re-seal polyallomer Seton Scientific 16 × 76 mm; large gradient: QuickSeal centrifuge tubes 25 × 89 mm Beckman Coulter) and consecutively sub-layered with 2/2/2/2 mL for a small and 7/5/4/4 mL for a large gradient of the 15%, 25%, 40%, and 60% iodixanol solutions respectively. Tubes were sealed using the Tube Sealer and meticulously balanced. Ultracentrifugation was performed in the OptimaTM L-90K ultracentrifuge (70.1 TI for small and 70 TI for large gradients, Beckman Coulter) at 50000 rpm at 4 °C for 2 h for small or 2.5h for large gradients. After centrifugation, using syringe needles, the tubes were punctured at the top and bottom for pressure and sample release, respectively. By controlling the pressure release on the top puncture, the lower phase was released dropwise from the tube bottom. For small gradients, the bottommost 1.5 mL were discarded while the following 0.8 mL were collected. For large gradients, the bottommost 3 mL were discarded while the following 1.5 mL were collected.

Fractionation

Fractionation was performed only for large gradients in this work. As described above, the lowest 3 mL were discarded, and the following 1.5 mL were collected as the filled capsid fraction (F1). Three more 0.5 mL fractions were taken in the same manner (F2-F3).

2.3.3 Size filtration - Iodixanol removal

For virus samples used in transducing mice and fractionated samples, a filtration to remove iodixanol from purified virus products was performed. The collected sample was filled up to 15 mL with DPBS and transferred to an Amicon Ultra-15 Centrifugal Filter Unit (MWCO 100000; Merck). Centrifugal filters were then centrifuged at 500 × g for 2–5 minutes or until circa 2 mL remained in the filter. The filter was again filled to 15 mL with DPBS and centrifuged as before, until reaching the desired volume of 200 – 400 µL. The filter was washed by pipetting the remaining volume repeatedly over it, a small sample taken for quantification, and the purified AAV sample stored at -80°C.

2.3.4 ddPCR for AAV titer quantification

Purified virus was diluted in water in a sterile hood with dilution factors (dil) from $1e5$ to $1e7$, depending on the number of plates used for production and the expected yield. A primer/probe mix of the Cytomegalovirus enhancer (CMVenh) primer/probe set and the ITR primer/probe set (for sequences see 2.8) was prepared by mixing 9 µL of each forward and reverse primer (100 µM) and 2.5 µL of each probe (100 µM) with 9 µL of water (50 µL total). One ddPCR reaction consisted of 11 µL ddPCR Supermix (Bio-Rad, 1863024), 1.1 µL primer/probe mix, 4.4 µL nuclease-free water, and 5.5 µL of diluted virus sample. 20 µL of the samples were used to generate droplets as per the manufacturer's protocol using a QX200 Droplet Generator. The generated droplets were transferred to a 96-well plate, and the plate was sealed. PCR was performed on a C1000 Touch Thermal Cycler according to the program in Table 2.1. Finally, the plate was measured on a QX200 Droplet Reader using the 'ABS' experiment.

$$vg/mL = 200 \times cp \times dil \times (1 - j) \tag{2.1}$$

The droplet reader returns the copies per 20 µL well of the measured primer/probe set as well as the number of droplets positive for either or both tested primer/probe sets. A droplet containing a complete genome should be positive for the transgene (CMVenh; FAM) and for the ITR (HEX). The QX200 Droplet Reader measures each droplet for the fluorescence of FAM (fluorescein) in Channel 1 (Ch1) and of HEX (Hexachlorofluorescein) in Channel 2 (Ch2). To get the fraction of partially filled capsids, the number of transgene-only-positive droplets (Ch1+Ch2-) is divided by the number of double-positive droplets (Ch1+Ch2+), resulting in the factor j. The final viral titer vg/mL is calculated using equation 2.1, where cp is the copy number (in the 20 µL well) of the transgene as assessed by the QX200 Droplet Reader Software (QuantaSoft Software), dil is the dilution factor of the viral sample,

Temp	Time	
94°C	10 minutes	
94°C 58°C	30 seconds 60 seconds	40 cycles
58°C	10 minutes	
4°C	hold	

Table 2.1. Cycling conditions for PCR reactions of droplets for ddPCR

and j the fraction of partially filled capsids. In ddPCR measurements without the ITR primer/probe set, j is assumed to be 0.

2.3.5 Small-scale 96-well production

The 96-well productions were performed using HEK293T cells seeded in DMEM substituted with 10% FBS and 1% Penicillin-Streptomycin. Cells were seeded in the 96-well plate at $1.25e4$ cells in 100 µL medium per well and incubated overnight. To ensure an equal amount of DNA for transfection, plasmid DNA was quantified using a Qubit Fluorometer. Using PEI-MAX, each well was transfected with 6 fmol each of AdH, rep2/cap2 helper plasmid, and ITR-flanked transgene plasmid. To maintain minimal deviation in the amount of added helper plasmids, a master mix of AdH and rep2/cap2 plasmid was prepared (per well: 1.633 µL 300 mM NaCl, 6 fmol rep2/cap2, and 6 fmol AdH). Similarly, a PEI-mastermix was prepared (per well: 1.633 µL 300 mM NaCl, 0.9 µL nuclease-free water, and 0.73 µL PEI-MAX). To address pipetting errors, every column of 8 wells of the 96-well plates was transfected with an individually prepared transfection mix. One transfection mix (enough for 12 wells) contained 20.4 µL of Helper mix and 72 fmol of transgene plasmid topped with water to a total volume of 39 µL. The transfection was finalised by adding 39 µL of PEI-mastermix, vortexing vigorously, and incubating for 10 minutes at room temperature. Each well was transfected with 6.5 µL of transfection mix, and the plate was slightly agitated to mix. The plate was incubated for 72 hours at 37°C.

Cells were harvested by pipetting the medium up and down to dislodge as many cells as possible and transferred to a new cup. The plate was then washed with 50 µL DPBS, which was also added to the same cup. The harvested cells were subjected to 5 rounds of freeze-thaw cycles and subsequently centrifuged for 10 minutes at 1000xg. 5 µL of the cell lysate was mixed with 37 µL water, 5 µL DNase I buffer, and 3 µL DNase I (2000 U/mL, NEB M0303L), mixed by pipetting, and incubated at 37°C for 15 hours. To inactivate the DNase, firstly, 6 µL of 50 mM EDTA was added to each cup and mixed by pipetting, secondly, the reaction was incubated at 75°C for 15 minutes. The DNase digestion was then diluted 1:1000

in nuclease-free water and quantified via ddPCR with primer/probe sets against the CMVenh (FAM) and beta-lactamase (bla; HEX). The concentration (titer) was calculated in the same way as in section 2.3.4, for both primer/probe sets. The beta-lactamase is measured to gather information on the remaining plasmid DNA left after DNase digestion. All transfected plasmids contain a beta-lactamase gene, but only one out of three also contains the CMVenh of the transgene. The final concentration is thus calculated according to equation 2.2, where $c_{corrected}$ is the bla-corrected viral concentration, c_{CMVenh} is the concentration of CMVenh, and c_{bla} is the concentration of bla.

The titers of the control transgene (Kana or lambda) were averaged, and all titers in the plots are displayed as log2 fold-change values against that average.

$$c_{corrected} = c_{CMVenh} - \frac{c_{bla}}{3} \tag{2.2}$$

2.4 Mouse work

2.4.1 Long-term mouse experiment

Mice were handled and sacrificed by Dr. Jonas Becker. Mouse experiments in this study were approved by German authorities (35-9185.81/G-26/20). A total of 27 five-week-old female C57BL/6J mice were tail-vein-injected with PBS or $5e11$ vg per mouse of AAV2 or AAV9 packaged CMV-eGFP transgene in 150 µL injection volume (filled with PBS). Virus was generated from large-scale productions from the plasmids with internal IDs #2510 (CMV-eGFP transgene), #183 (Rep2/Cap2), #189 (Rep2/Cap9), and #1111 (AdH). Virus was produced and filtered as described in section 2.3. Four mice were injected per timepoint and AAV serotype. Mice were sacrificed after 2, 6, and 12 weeks post-injection. An additional mouse per timepoint was injected with PBS. Liver, heart, kidney, musculus quadriceps, spleen, and brain were extracted from the mice and stored in RNAlater at 4°C until further processed. Additionally, sections of liver, musculus quadriceps, and heart were shock-frosted in liquid nitrogen for CUT&Tag sequencing experiments (section 2.5).

2.4.2 DNA extraction from stored murine tissue samples

The stored mouse tissues stem from work performed mainly by Claire Domeger and Jonas Becker [243]. The tissues were in storage at 4°C in RNAlater solution (Thermo Fisher Scientific; AM7021) for 14 months before this work began. DNA extraction was performed using the Qiagen DNeasy Blood and Tissue kit. About 10 mg of tissue was cut using a new scalpel blade and transferred to a cup containing 180 µL of ALT buffer and a steel bead. The

tissues were homogenised using the Tissue Lyser LT from Qiagen at 50 Hz for 45 seconds. If tissue chunks were still visible, the sample in question was vortexed and subjected to another round in the homogeniser. After that, 20 µL of Proteinase K (supplied by kit) was added and the samples incubated at 56°C for 90 minutes during which the samples were vortexed occasionally to aid tissue dissociation. 200 µL of AL buffer, together with 200 µL of 99% ethanol (EtOH), was added, and the samples centrifuged for 2 minutes at 10000 x g. 500 µL of the supernatant was transferred to a kit-provided spin column and centrifuged for 1 minute at 13000xg. The supernatant was discarded, and 500 µL of AW1 was added to the column, and the samples were centrifuged as before. The supernatant was discarded, and the column was placed in a new cup. The column was washed again with 500 µL of AW2 and centrifuged for 3 minutes at 13000 x g. After this, the DNA was eluted from the column with 200 µL of AE buffer, quantified with Nanodrop, and stored at -20°C.

2.4.3 DNA and RNA extraction from murine tissues generated in this work

DNA and RNA extractions from mouse tissues were performed using the Qiagen AllPrep DNA/RNA Minikits with minor modifications to the protocol. About 20 mg of tissue stored at 4°C in RNAlater (section 2.4.1) was sectioned using a new scalpel blade and added to a cup containing a steel bead and 600 µL RLT Plus buffer (provided by the kit) with 1% beta-mercaptoethanol. The tissue section was lysed as in 2.4.2, after which 20 µL of Proteinase K was added, incubated at 56°C for 1 hour, and finally centrifuged for 5 minutes at max speed. The supernatant was transferred to an AllPrep DNA spin column and the protocol followed until the first wash step with buffer RW1. Instead of 700 µL, I only added 350 µL of RW1, centrifuged, and then performed an on-column digest. To do that, 70 µL RDD buffer with 10 µL RNase-free DNase I stock solution (Qiagen 79254) was premixed, added to the column, and incubated at room temperature for 15 minutes. The DNase was washed out with the remaining 350 µL RW1 buffer, and the rest of the purification was performed as per the manufacturer's instructions. Initial quantifications of DNA and RNA were performed using Nanodrop.

cDNA synthesis

Before reverse transcription, the samples were digested with DNase I off-column to remove any residual amount of DNA from the RNA extractions. 424 ng of RNA was diluted in 35 µL of nuclease-free water and mixed with 4 µL of RDD buffer and 1 µL of RNase-free DNase I (Qiagen 79254). The reaction was incubated at room temperature for 30 minutes, and the DNase was heat-inactivated at 75°C for 10 minutes. The cDNA synthesis was performed

Temp	Time
25 °C	10 minutes
37 °C	120 minutes
85 °C	5 minutes
4°C	hold

Table 2.2. Thermocycler settings for cDNA synthesis

using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems 4368813). A reaction consists of 4 µL 10x RT buffer, 1.6 µL 25x dNTP Mix (100 mM), 4 µL 10x RT random primers, 2 µL MultiScribe Reverse Transcriptase, and 28.4 µL of the DNase-digested RNA sample. The reaction was incubated according to Table 2.2.

2.4.4 ddPCR for vg/dg and vt/hkt assessment

Measurement of vg/dg

The ddPCR for vector genomes per diploid genomes (vg/dg) was performed against the primer/probe sets CMVenh (FAM) and murine RPP30g6 (HEX). Primer/probe mix was created the same way as described in section 2.3.4. DNA extractions were diluted to 1 ng/µL each. Diluted HindIII was added to each ddPCR reaction mix and was created by mixing 75 µL of Diluent B with 25 µL of HindIII (both NEB). One ddPCR reaction consisted of 11 µL ddPCR Supermix (Bio-Rad, 1863024), 1.1 µL primer/probe mix, 3.3 µL nuclease-free water, 1.1 µL of diluted HindIII, and 5.5 µL of 1 ng/µL genomic DNA (gDNA). The rest of the ddPCR was performed as described in section 2.3.4. The vg/dg is calculated as the number of transgene-positive droplets (Ch1+Ch2+ plus Ch1+Ch2-) divided by half the number of muRPP30g6-positive droplets (Ch1+Ch2+ plus Ch1-Ch2+).

Measurement of vt/hkt

The ddPCR for the assessment of viral transcripts per housekeeper transcripts (vt/hkt) was done against the primer/probe sets of eGFP (HEX) and murine RPP30c9 (HEX). The ddPCR was performed as described in section 2.3.4, with 5.5 µL of undiluted cDNA from section 2.4.3. The vt/hkt is calculated by dividing the number of GFP-positive droplets (Ch1+Ch2+ plus Ch1+Ch2-) by the number of muRPP30c9-positive droplets (Ch1+Ch2+ plus Ch1-Ch2+).

2.5 CUT&Tag

2.5.1 CUT&Tag Library preparation

CUT&Tag Library preparation was conducted using the CUT&Tag Tissue Kit from Active Motif (53170). 10 mg of flash-frozen mouse liver (from section 2.4.1) was cut on a petri dish on ice. The cut was minced as much as possible with a scalpel, 1 mL of Tissue Lysis Buffer (supplied by the kit) was added and incubated for 2 minutes in the cooled petri dish. The tissue fragments with the buffer were transferred to a 15 mL Dounce Homogenizer (Roth) and homogenized with the tight pestle 30 times until no tissue chunks were remaining. The lysed tissue solution was strained through a 40 µM strainer (provided by the kit) and further processed according to the manufacturer's protocol. The number of nuclei in the lysed samples was quantified using a Luna Cell Counter (Logos Biosystems) with brightfield and the gating set for particles between 5 and 15 µM in size. $2.5e5$ nuclei were used for the remaining steps of the protocol, which were completed as specified in the manufacturer's instructions (53170, version A2). Primary antibody was either rabbit@H3K27ac (Active Motif, 39034) or rabbit@H3K27me3 (Active Motif, 39157). The secondary antibody was the kit-supplied guinea-pig@rabbit antibody.

The finished libraries were quantified using a Bioanalyzer with an Agilent DNA 1000 Kit (Agilent 5067-1504). On average, the libraries had a 3-peaked size distribution with the major peak being at 330 bp and a concentration of 1-5 ng/µL measured via the High Sensitivity DNA Qubit fluorometer. They were sequenced in multiple rounds on the NextSeq 2000 Illumina system. Samples were pooled with between 14 and 16 samples per flowcell with either the P1-SBS or P2-SBS reagent kit, resulting in between 12 and 53 million paired reads per sample (see Table 3.2).

2.5.2 CUT&Tag data analysis

Quantification of AAV transgene modifications

Analysis of cut and tag data was performed using the workflow manager snakemake [244] and found under 'cutandtag_analysis/snakefile'. Paired-end cut and tag sequencing reads were aligned to a custom genome containing the mouse genome (GCF_000001635.27) merged with the transgene sequence. Alignment was performed using bowtie2 [245] with parameters 'sensitive-local', 'no-mixed', and 'no-discordant'. Multiple filtering options for the alignments were taken into consideration to check for the robustness of the method. Aligned fragments were marked for PCR duplication using samtools' 'fixmate' and 'markdup' programs [246]. Filtering for primary alignments, no PCR duplicates, and a

minimum alignment quality of 30 was performed by samtools 'view' ('-f 3 -F 3328 -q 30'). This step is referred to in the results text as 'filtering'. Read-pairs with a minimum 120 and maximum 400 bp insert size were selected by using the program 'alignmentSieve' from the deeptools package [247] and are referred to as size filtered.

The modification rate on the transgene was quantified by equation 2.3, where PTM is the post-translational modification (either H3K27me3 or H3K27ac), $reads_{transgene}$ is the number of reads aligned to the transgene, $reads_{mouse}$ the number of reads aligned to the mouse assembly excluding the mitochondrial genome (NC_005089.1), and vg/dg is the number of viral genomes per diploid genome (see 2.4.4). Associated code found in 'cutandtag0x.analysis.Rmd'.

$$ratio_{PTM} = \frac{reads_{transgene}}{reads_{mouse} \times vg/dg} \tag{2.3}$$

The ratio of histone modifications was calculated as described in equation 2.4. This normalisation makes the vg/dg measurement obsolete.

$$ratio_{norm} = \frac{ratio_{H3K27me3}}{ratio_{H3K27ac}} \tag{2.4}$$

DEseq2 analysis

I used SEACR 1.3 [248] to identify peaks of histone modifications from the alignment using parameters '0.01 norm stringent', to return the top 1% of peaks. To make sure the AAV delivered transgenes are also analysed, I manually added a corresponding peak to the list of found peaks identified by SEACR. Fragment counts associated with every peak were directly extracted from the alignment bam file and a differential analysis performed using DESeq2 [249]. AAV2 was compared to AAV9 at timepoints 2 and 12 weeks post-infection. Associated code can be found in 'cutandtag0x.DEesq.Rmd'.

2.6 Episome sequencing

2.6.1 Exonuclease-based episomal enrichment

DNA samples were enriched for circular episomes by Exonuclease treatment. Regardless of concentration, 32 µL of DNA of interest were added to 5 µL Cutsmart, 10 µL 10 mM ATP, 2 µL RecBCD (Exonuclease V), and 1 µL of a restriction enzyme that does not have any recognition motifs within the transgene (EcoRI-HF for transgenes with CMV or GFAP, and

KpnI-HF for genomes with LP1). The restriction enzyme was added to fragment genomic DNA to increase the efficiency of the exonuclease treatment. RecBCD is considered a 'plasmid-safe' exonuclease as it does not digest nicked circular DNA. The reaction was incubated for 1 hour at 37°C, then cleaned with 50 µL AMPureXP beads (1:1 ratio), and eluted in 19 µL of nuclease-free water.

2.6.2 Enzymatic Methyl sequencing

Library generation for enzymatic methyl sequencing (EM-seq) was performed using the NEBNext Enzymatic Methyl-seq Kit (NEB E7120). Total DNA extracted from mouse livers was either enzymatically enriched for episomes (see 2.6.1) or directly subjected to fragmentation by sonication on the Covaris S2 Sonicator (5 Intensity, 10% duty cycle, 200 cycles per burst). The enzymatic conversion and library preparation were done according to the manufacturer's protocol. Because of a difference in concentration the enzymatically enriched samples were amplified by 12 cycles and the total DNA samples by 5 cycles of PCR. Sequencing was performed on the Illumina MiniSeq system with 42 bp paired-end reads, which yielded between 4 and 6 million paired reads per sample.

Reads were filtered by a minimum length of 35 bp with seqkit [250] and low quality bases, as well as adapter sequences (according to a file containing all commonly used Illumina adapters), were trimmed from the reads using trimmomatic [251]. Reads were aligned to the mouse genome (GCF_000001635.27) containing an extra entry with the AAV delivered transgene using bwa-meth [152]. Alignment was filtered using samtools [252], while retaining only primary alignments with a minimum quality of 30 (parameters: '-f 3 -F 3328 -q 30'). MethylDackel was used to aggregate and tabulate information on methylated CpG sites [253].

2.6.3 Nanopore Sequencing of episomal DNA

R9 sequencing of stored mouse tissues

R9 runs for the METEORE analysis were done on the episomally enriched samples (2.6.1) using the Rapid Barcoding kit from Oxford Nanopore Technologies (SQK-RBK004) and following the manufacturer's protocol. Sequencing was performed on a FLO-MIN106 and the MinION Mk1b sequencer.

R10 sequencing of long-term mouse samples

The DNA samples obtained from mouse livers from the long-term transduction experiment (section 2.4.1) were sequenced with the latest R10 chemistry obtained through ONT. For

the library preparation, the enriched DNA samples (section 2.6.1) were treated with either NotI-HF or KpnI-HF to produce a single cut between promoter and YFP and again purified with AMPureXP Beads at a 1:1 ratio. Sequencing adapters were ligated to the generated DNA ends using the ONT-provided barcoding and ligation kit (SQK-LSK114.24) together with the associated protocol from the manufacturer.

2.6.4 Analysis of Nanopore episome data

Transgene alignment

Alignment against the transgene was performed using minimap2 directly [254] (preset '-x map-ont') or dorado, which also allows for minimap2 alignment (preset 'lr:hq') against the corresponding reference obtained from the plasmid map. Samples obtained from the transposase-based rapid barcoding kit (here only R9 samples) were aligned to the 'native' reference of the transgene, starting and ending with the ITR sequence. The samples obtained from a single-cut and the ligation-based barcoding kit (here mostly R10 samples) were aligned to a 'linearized' reference starting and ending with the cut site between the promoter and YFP coding region. This was done to obtain adequate alignment of long reads across the ITR sequence, which led to noticeably lower coverage when using the 'native' reference.

Coverage for all nanopore sequencing runs as well as for the CUT&Tag sequencing runs was calculated according to equation 2.5 by dividing the sum of all aligned bases ($bp_{aligned}$) by the total length of the reference ($bp_{reference}$).

$$coverage = \frac{bp_{aligned}}{bp_{reference}} \tag{2.5}$$

METEORE

METEORE is a pipeline that combines multiple public basecallers for methylation calling [159]. For rebasecalling, I used the ONT-provided 'dna_r9.4.1_450bps_hac' configuration file with guppy. I used the calls from Guppy (ONT), Nanopolish [255], Megalodon [256], and Deepsignal [257]. The individual tools were all installed in separate conda environments as specified by the yaml files provided by the METEORE GitHub. An auxiliary extraction script is also required and was obtained from [258]. The pipeline was run as outlined in the METEORE instructions, and the outputs from the individual tools were combined. Only CpG sites with a minimum coverage of 10 reads were taken into consideration for plotting. The associated code for the pipeline execution and figure generation can be found in 'meteore.Rmd'/'meteore_newcontrols.Rmd' and 'meteore_visualisation.Rmd'

respectively.

Dorado methylation calling

The full code for the analysis and plotting in this section can be found in 'LTmouse_episome.readme.Rmd' and 'LTmouse_episome.align.Rmd'.

Dorado (version 0.7.4+59ab908) with the parameters '–min-qscore 8 –emit-moves –kit-name SQK-NBD114-24 -r sup,5mCG_5hmCG' was used for rebasecalling. The final parameter 'sup,5mCG_5hmCG' ensures the utilisation of the latest available 'super accurate' model while also calling methylation information (latest model at the time: 'dna_r10.4.1_e8.2_400bps_sup@v4.3.0_5mCG_5hmCG@v1'). Demultiplexing is performed by dorado automatically when it is supplied with the kit information. Rebasecalled reads were filtered for a minimum length of 2000 bp and aligned to the transgene reference using the dorado aligner program (which uses minimap2 [254]) and default parameters. The resulting bam file was split into individual barcodes with the dorado demux program and the '–no-classify' parameter. Every unmapped read and non-primary-alignment in the bam files was removed using samtools ('samtools view -F 2308'). The methylation information was tabulated by using the program modbam2bed [259] with the parameters '–aggregate –combine –cpg –extended'.

Pooling and subsampling were performed with scripts 'LTmouse_episome.pool.Rmd' and 'LTmouse_episome.subsample.Rmd'.

Plots were generated with 'LTmouse_episome.analysis.Rmd'.

ITR recombination analysis

Associated code can be found in 'LTmouse_episome.ITRalign.rmd'. ITR recombination analysis was performed on reads obtained from liver samples of mice transduced with AAV9-CMV-eYFP sequenced with the R10 chemistry and the ligation-based library generation protocol (all 3 time points combined). Nanopore reads can exist in either the forward or reverse orientation. For this analysis, all reads were re-oriented to the same orientation using samtools. Individual ITR annotations (A-sequence, B-loop, C-loop, D-sequence) were aligned to the reads using blastn [260] (parameters '-gapopen 0 -gapextend 2 -word_size 9 -reward 1 -penalty -2') and alignments with an e-value of above 0.05 discarded. The length of an ITR was determined by comparing the alignment coordinates of the first 2 D-sequence appearances within each read. Reads with ITRs of above 1000 bp in length were discarded.

2.7 AAV dinucleotide periodicity analysis methods

2.7.1 Dinucleotide auto-distance histograms and normalization thereof

The method to calculate the distribution of distances of a specific dinucleotide to each of its occurrences was adapted from the Bioconductor package 'periodicDNA' [261]. The associated functions are 'calculate_histogram' and 'normalize_histogram' within the script 'Periodicity6.R'. The script was partly co-authored by Sebastian Buddecke. To reduce working memory requirements, sequences of a length of above 10000 bp were split into sets of sequences with a maximum length of 10000 bp and the calculations were performed on the subsets. For the histograms, the location of every dinucleotide is recorded in a list on which a distance matrix is calculated. Only distances of below 300 bp were taken into consideration. The counts of the resulting histogram are processed in 2 steps. First, they are smoothed by averaging over a 3-bp window to reduce the influence of the 3-bp pattern in coding sequences. Finally, the background level is calculated by averaging over a 7-bp window and subsequently subtracted from the smoothed histogram to obtain the normalised counts. Since the subtraction of the background entails averaging over a 7-bp window, this leads to the first (and last) 3 bp not being normalised and remaining abnormally high. For clarity, these 3 bp were set to 0. Similarly, to clarify that position 0 contains the measured dinucleotide, that position was set to 1.3 times the value of the maximum in the first 50-bp of normalised counts. For visualising the convergence of multiple of such distance histograms, a ggplot2 'geom_smooth' object with 'method=loess' and 'span=0.3' was used [262].

2.7.2 Dinucleotide cross-distance histograms

To calculate the distribution of distances of a dinucleotide to the locations of another dinucleotide, a similar approach to section 2.7.1 was undertaken. The associated code can be found in the script 'Periodicity6_crossDistances.R'. Instead of calculating distances between all occurrences of a single dinucleotide, the script calculates the set of distances of an anchored dinucleotide to all occurrences of another supplied dinucleotide. The smoothing and normalisation are done in the same way as section 2.7.1.

2.7.3 Quantification of periodicity

To quantify the periodicity of a set of sequences, I measure the Goodness of Fit of a dampened sine function to the histogram as described in [230]. The associated function is called 'fit_sine' within the script 'Periodicity6.R'. A sine function as described in equation

2.6 is fit to the normalised counts distance histograms (distance < 80), where H is the half-life, λ the period, and A the amplitude of the fit.

$$f(x) = e^{-\frac{ln(2)}{H}} \times A \times sin(\frac{2\pi x}{\lambda} + \frac{\pi}{2}) \tag{2.6}$$

The 'nls' function of R with the Gauss-Newton algorithm, with a maximum of 10000 iterations, and with a minimum step size of $\frac{1}{2048}$ was used to perform the fit. Initiation values for every fit were chosen to be $A = 1e-3; H = 20; \lambda = 20$. Fitting was attempted 300 times, with initiation values randomized in a specific range ($A \in [5e-4, 1e-2]; H \in \{20..50\}; \lambda \in \{5..30\}$) and all successful fits retained. Goodness of Fit (GdF) is here defined as the standardized deviance of the fitted model, defined in equation 2.7 where $deviance$ is the deviance of the fit and σ is the standard deviation of the histogram.

$$GdF = \frac{deviance}{\sigma} \tag{2.7}$$

The fit with the lowest GdF value is considered the best fit and retained.

The strength of the periodicity is here defined as the reciprocal of the log of GdF over the amplitude of the sine fit, defined in equation 2.8 where GdF is the Goodness of Fit and A is the amplitude.

$$periodicity = log2(\frac{GdF}{A})^{-1} \tag{2.8}$$

2.7.4 Analysis of periodicity in other viral genera and model organisms

The associated code can be found in scripts 'download_sequences.r' and 'Periodicity2_sinecurve.r'. The former was in parts co-authored by Sebastian Buddecke. Sequences for every genus were downloaded using the R package 'rentrez' [263], with the query "genus[Organism] and 'complete genome'". Sequences below 1000 bp were discarded, and sequence duplicates were removed with seqkit [250]. This was done with all genera within the Kingdom of *Shotokuvirae* as listed in the International Committee on Taxonomy of Viruses (ICTV). Genera with less than one million bp of total unique sequence information were omitted from the analysis. In addition to the Kingdom of *Shotokuvirae*, also the genera in the family of *Adenoviridae* and *Orthoherpesviridae* were analysed since the most researched AAV helper viruses (Adeno- and Herpesvirus) belong to those families. Since AAV is a satellite virus, Hepatitis D sequences and sequences associated

with the terms *Lavidaviridae* and *Alphasatellitidae* were additionally analyzed. Since viral genomes predominantly consist of coding sequences, as a control, the consensus coding sequences of the human genome were downloaded from the Consensus Coding Sequence project[2] and processed as above. Similarly, all open reading frames (orf_genomic_all) were downloaded for the *Saccharomyces cerevisiae* S288C R64 assembly from the sgd-archive[3]. All coding sequences for *c. elegans* assembly WBcel235 were downloaded from ensembl[4]. The *e. coli* sequence was obtained from the accession number U00096.3.

To keep computation time to a minimum and also keep individual genera comparable to one another, the individual sequence sets from every genus were subsampled to achieve a total length of $1e5$ bp. This was done 10 times. These subsamples were then subjected to periodicity quantification as described in section 2.7.3 individually, after which the subsample calculations are re-combined to obtain the final values for any given genus. The period, GdF, and periodicity values are the median values from the 10 subsamples.

For the bias interrogation of the dependoparvovirus sequence set, I downloaded the archetypical sequence of the 27 dependoparvovirus species as listed by the International Committee on the Taxonomy of Viruses (ICTV) [8]. I performed the same analysis as described above without subsampling the dataset. The individual accession numbers are: U22967, MW046577, MT138326, KX583629, MT138328, AY186198, MN175614, MN794870, OQ198157, OQ198130, OQ198151, OQ198126, OQ198128, GU226971, MG745677, AF085716, MT138301, MT138277, JN420372, AF043303, OQ101836, MF416383, MF416384, MN242366, MN242367, AY349010, and KP733794.

2.7.5 Shuffling enrichment

AAV cap library and selections thereof

The capsid library consisted of the AAV capsid serotypes 1 through 9 and serotype rh10, and was originally created by Anne-Kathrin Herrmann. It was in storage for about 5 years before this work began. The plasmid library consisted of an ITR-flanked *rep* and a shuffled *cap*. Virus was produced in a dual transfection approach with AdH (internal ID #1111) at a molar ratio of 1:20 (library:AdH), purified by iodixanol gradient, and subsequently filtered as described in section 2.3. Quantification was performed as described in section 2.3.4.

For the transduction of Raji B-cells, $2e5$ cells in 2 mL RPMI-1640 medium were transferred to a 6-well plate. Wells were transduced with the virus library at a multiplicity

[2] `https://ftp.ncbi.nlm.nih.gov/pub/CCDS/current_human/`
[3] `http://sgd-archive.yeastgenome.org/sequence/S288C_reference/orf_dna/`
[4] `https://ftp.ensembl.org/pub/release-113/fasta/caenorhabditis_elegans/cds/`

Temp	Time	
98°C	2 minutes	
98°C	15 seconds	
62°C	15 seconds	35 cycles
72°C	75 seconds	
72°C	5 minutes	
4°C	hold	

Table 2.3. Cycling conditions for PCR reactions with the Phusion polymerase

of infection of $1e5$. Cells were incubated for 2 days and then harvested by centrifugation at 300xg for 5 minutes. The pellet was washed twice in PBS. To remove any AAV capsids sticking to the cell membranes, the pellet was incubated with 200 µL of 0.25% Trypsin for 10 minutes at 37°C. The pellets were washed twice in PBS as before. DNA was extracted from the cell pellet with the DNeasy Blood and Tissue Kit according to the manufacturer's protocol.

Rescue PCR from the virus production or the Raji-cell pellets was performed with the Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific, F548), with a final primer concentration of 400 nM for LseqFor and LseqRev (see Table 2.8), 1 µL template, and final volume of 50 µL in water (Table 2.3).

The resulting PCR amplicon was either processed by Nanopore Sequencing Library preparation (section 2.7.5) or re-cloned for another virus library generation step. To achieve the latter, 3 µg of PCR amplicon were digested with AscI and PacI overnight at 37°C and subsequently ligated with 500 ng of a similarly digested acceptor plasmid (internal ID #1608). The mix was ligated with T4 Ligase and incubated at room temperature overnight. The plasmid was transformed into electrocompetent cells by electroporation. Up to eleven electroporations were performed per ligation mix and mixed again afterwards. Electroporation was performed by adding 1 µL of ligation mix to 30 µL of E. cloni 10G Supreme cells (Lucigen) and pulsing on a Gene Pulser Xcell at 1.8 kV. Directly after electroporation, 970 µL SOB medium was added and the bacteria incubated at 37°C for one hour. A sample of all electroporations of one ligation reaction was diluted 1:100000 and plated on an LB-Ampicillin plate. Library diversity was estimated by counting colonies that grew overnight.

Nanopore Sequencing and Analysis

Libraries were generated using the Ligation Sequencing Kit from ONT (SQK-NBD114.24), and the associated protocol (15Sep2022) was performed with 200 fmol of PCR amplicons as

input per sample. Sequencing was performed on a Mk1B with an R10.4.1 flowcell. Reads were rebasecalled using Dorado with the latest SUP model and a quality threshold of 8. All reads were subsequently filtered by length (2250 < read_length < 2340) and Phred quality (read_quality > 12), which left between 55 and 206 million reads per sample.

From the total reads, I randomly subsampled down to 600 reads 10 times using 'seqkit sample -n 600' and removed exact sequence duplicates using 'seqkit rmdup -s' (0 in most cases) [250]. The 10 subsamples were individually assessed for their YY/RR periodicity in the same way as described in section 2.7.3.

2.7.6 Cloning of the periodic stuffer DNA plasmids

The ITR-backbone for all plasmids with periodic stuffer stems from a dual luciferase construct (internal ID #714). Unknown to me at the beginning of this work, whole-plasmid sequencing revealed that one of the ITR sequences in the backbone had a deletion in the B-loop sequence after a considerable part of this work had already been completed. Because only one of the two ITR sequences contained a mutation, the effect on viral titer was likely limited. Nevertheless, more recently created 4.8 kb constructs were generated with the backbone from plasmid pSSV9CLuWb (generously provided by Dr. Kleopatra Rapti), which contained intact ITR sequences confirmed by whole-plasmid sequencing. The backbones contain a bacterial origin of replication, an ampicillin resistance gene, and two ITR sequences. The insert was separated from the backbone by SphI and SpeI digestion and agarose gel cleanup. The CMV-eGFP-BGHpA cassette (1.8 kb) was amplified from plasmid #2054 and cloned with primer overhangs for SphI and separate complementary sequences for Type II restriction endonuclease cloning (BsaI; see Table 2.8). Different stuffer DNA was inserted downstream of the polyadenylation site (1.8 kb for 3.8 kb genomes; 2.7 kb for 4.8 kb genomes). The stuffer of the kana control plasmid (3.8 kb genome control) contains the kanamycin resistance gene from plasmid #1103 (internal ID) along with fragments of its origin of replication and its own stuffer DNA (see Table 2.8). The lambda control plasmid (4.8 kb genome control) contains a fragment (coordinates 17464-21785; see Table 2.8) of phage lambda DNA (NEB N3011S). The wt-fragments were amplified from the AAV2 sequence containing rep2/cap2 plasmid (WHC2; internal ID #0183) with primers as indicated in Table 2.8.

The 'YY' and 'YYRR' constructs (3.8 kb genomes) were generated *in silico*. The associated code can be found under 'modify_kana_period.ipynb', in particular see functions 'mod_sequence_YY' and 'mod_sequence_RR'. The kanamycin stuffer sequence was permuted to generate imperfect periodicities starting from existing dinucleotides. Another

dinucleotide of the same kind was inserted at the period of interest, depending on a specific skip probability. This skip probability was incorporated to mimic the imperfect periodicity of the wildtype AAV sequences. The YY stuffer was created with a skip probability of 0.8 and a period of 15. The YYRR stuffer was created by iteratively passing over the same sequence 64 times and (i) modifying the YY periodicity with a skip probability of 0.99 and period 15 and (ii) modifying the RR periodicity with a skip probability of 0.99 and a period of 21 on each pass. For these sequences, only the forward direction of the AAV2 sequence was taken into account for their design, which resulted in a YY periodicity of 15 and the RR periodicity of 21. The general 15 bp periodicity for both YY and RR dinucleotides was only later clarified in the analysis of the genus dependoparvovirus. For cloning, both stuffer sequences were split into 2 to 3 individual fragments each, because the IDT-calculated complexity was measured as too high for synthesis. The fragments were synthesized by IDT and assembled with the same Type II-based cloning strategy as before.

The fragments containing a modified lambda stuffer (M1, M2, and M3) were designed *in silico* with the python scripts 'WTF_mask.ipynb' (M1 and M2) and 'rigid_mask.ipynb' (M3). For M1 and M2, the YY/RR pattern from the wt-Fragment stuffer was imprinted on a sequence mask that was ultimately applied to the sequence of the lambda stuffer. The M1 mask contains all YY or RR dinucleotides from the wt-fragments stuffer. The M2 mask contains only the YY dinucleotides that are spaced exactly 15-bp apart from each other (same for the RR dinucleotides). The M3 stuffer contains a rigid mask of randomised YY and RR dinucleotides spaced exactly at 15 bp periods and offset regarding each other, agnostic of the wt-fragments stuffer sequence. Because the complexity of the fragments was again measured to be too high by the web tool from Twist, each sequence was split into 2 or 3 individual gene fragments with the same cloning strategy as before. Sites within the fragments with the cloning-relevant motifs were manually removed, and the fragments were synthesized by Twist Bioscience as gene fragments. After cloning, all plasmid sequences were confirmed by whole-plasmid sequencing.

2.8 Molecular methods

2.8.1 Bacterial methods

MAX Efficiency DH5α Competent Cells (Thermo Fisher Scientific) *e. coli* bacteria were used throughout this work for the cloning and amplification of plasmids. Transformations were performed by adding 1 to 5 µL of plasmid to a freshly thawed tube containing 50 µL of bacteria. The ligation mix was incubated on ice for 15 minutes. Heat-shock was performed by heating the tube to 42°C for 45 seconds and then quickly cooling the tube

back down on ice for 2 minutes. The bacteria were plated on LB-ampicillin (50 µg/mL Ampicillin) agar plates and incubated overnight at 37°C. Colonies were picked the next day, and an appropriate volume of LB-ampicillin was inoculated (LB medium with 50 µg/mL Ampicillin). Depending on the necessary amount of plasmid, they were either grown in 2 mL (Mini-prep), 50 mL (Midi-prep) or 400 mL (Maxi-prep) of LB-ampicillin. Usually, Mini-preps were used for cloning, and Midi- or Maxi-preps are used for HEK293T cell transfection and virus production. Liquid cultures were incubated overnight at 37°C with continuous agitation. On the next day, the bacteria were pelleted from the liquid cultures by centrifugation at 4000xg for 15 minutes. Supernatant was discarded, and the pellet was either frozen at -20°C or used in the corresponding plasmid purification kit protocols (see Table 2.9).

Glycerol stocks were created by adding 500 µL of autoclaved 50% glycerol to 500 µL *e. coli* that were grown overnight in LB medium (+Ampicillin). The glycerol stocks were stored at -80°C.

2.8.2 Agarose Gels

Agarose gels were created with Biozym LE Agarose in TAE buffer, usually at 1% (w/v) unless otherwise specified. Agarose was dissolved in TAE by heating it in a microwave. One mL of 6x Loading Dye (NEB) was supplemented with 10 µL of Gelred Nucleic Acid Gel Stain (10000x) and used to load DNA onto the gel. Gels were run at 100 V for 30 to 60 minutes.

2.8.3 Cloning by restriction and ligation

Cloning steps were performed by PCR amplification using Q5 High-Fidelity 2X Master Mix (NEB M0492S) and final primer concentrations of 500 µM each. Primers were designed to contain overlaps for Type I or Type II restriction enzymes. The prefix and suffix sequences for cloning with Type II restriction enzymes were adapted from the Barrick Lab website[5]. PCR fragments were cleaned with the QIAquick PCR Purification Kit. Backbone from a donor plasmid was cut by restriction enzymes, cleaned by cutting the corresponding band from an agarose gel, and finally purified with the QIAquick Gel Extraction Kit. Ligations were performed with T4 Ligase (NEB M0202L) with incubation times between 1 hour and overnight at room temperature.

[5] `https://barricklab.org/twiki/bin/view/Lab/ProtocolsBTKDesignANewPart`

2.8.4 Hirt DNA extraction and Southern Blotting

HEK293T cells were seeded in a 6-well plate with $2.5e5$ cells per well in 2 mL DMEM with 10% FBS and 1% Pen/Strep. 2 days later the cells were transfected with 120 fmol per plasmid of *rep/cap* helper, AdH, and transgene plasmid and PEI max. One transfection mix contained 120 fmol per plasmid, 20 µL PEI-max, 26 µL 1.5 M NaCl, filled up to a final volume of 130 µL with water. 120 µL of the transfection mix was finally used to transfect the HEK293T cells.

3 days post-transfection, cells were collected by pipetting and transferred to a 1.5 mL cup and centrifuged for 8 minutes at 1600 rpm. The pellet was washed once in 500 µL PBS and centrifuged the same way again. Supernatant was discarded, and the pellet was resuspended in a small amount left in the cup. One mL of Hirt Lysis buffer was added and incubated for three hours at 55°C and afterwards on ice for 10 minutes. 250 µL of 5 M NaCl was added dropwise to precipitate high molecular-weight DNA. The tube was mixed and left at 4°C overnight. The samples were then centrifuged at maximum speed for one hour and the supernatant transferred to a new tube. The supernatant was centrifuged again for 15 minutes and then 700 µL cleaned with 700 µL of Phenol/Chloroform/Isoamly alcohol (Roth). 600 µL of the aqueous phase was mixed with 420 µL of isopropanol and mixed by vortexing. After centrifuging for 30 minutes at full speed, the supernatant was discarded and the pellet washed once with 70% EtOH. After another centrifugation and removal of the supernatant, the DNA pellet was mixed with 4 µL Cutsmart, 1 µL DpnI, and 35 µL water. The DpnI digestion was incubated overnight at 37°C. The agarose gel was loaded with 10 µL of the DpnI-digested pellet and run at 100 V for about 2 hours. The agarose gel was rinsed in water and then incubated in 0.2 M HCl for 7 minutes at room temperature with slight agitation. The gel was rinsed again and incubated in gel denaturation buffer for 30 minutes at room temperature. After another quick rinse, the gel was incubated for 30 minutes in gel neutralisation buffer.

The Blot sandwich was assembled from bottom to the top in the following order: a plastic tray reservoir filled with 20xSSC, a bridge to keep the agarose out of the reservoir, a whatman paper wick that was soaked in 20xSSC long enough to touch the 20xSSC in the reservoir, the agarose gel with pockets downwards, a dry nylon membrane, dry thick piece of whatman paper, dry stack of paper towels, and finally a box of gloves as weight. The blot was left at room temperature overnight. The next day, the sandwich was disassembled, and the DNA side of the nylon membrane was marked. While still wet, the membrane was UV-crosslinked in a CL-1000 crosslinker set to 1000 µJ and run twice. The membrane was quickly rinsed in water and then pre-hybridized in 10 mL of hybridisation buffer (Roche 11796895001) for 30 minutes at 40°C. In the meantime, 3.5 mL per 100 cm^2

of hybridisation buffer was pre-warmed to 40°C. The DIG-labeled probe was generated with the PCR DIG Probe Synthesis Kit (Roche 11636090910) with the CMVenh forward and reverse primers according to the manufacturer's protocol. 2 µL of probe per mL of pre-warming hybridisation buffer in 100 µL of hybridisation buffer was denatured by heating to 95°C for 2 minutes and then directly cooling on ice. The denatured probe was added to the warm hybridisation buffer, the pre-hybridisation buffer was decanted, and the buffer with probe was added. The membrane with the hybridisation buffer was sealed in a plastic bag and incubated at 40°C overnight. The next day, the hybridisation buffer was decanted and the membrane washed once in 2xSSC at room temperature for 5 minutes and then in 0.5xSSC at 40°C for another 5 minutes.

The detection of the DIG-labeled probe was performed according to the protocols of the DIG Wash and Block Buffer Set (Roche 11585762001) and DIG Nucleic Acid Detection Kit (Roche 11175041910) and imaged on an Azure 400 imager.

2.8.5 Hypo- and Hyper-methylated control DNA for ONT

To assess the accuracy of nanopore-based methylation calling, I also created hypo- and hyper-methylated controls. The negative (native) controls are based on PCR amplicons of the plasmids with promoters CMV, liver promoter 1 (LP1), and GFAP (internal IDs #2510, #2512, and #2554, respectively; all originally generated by Claire Domenger), that were generated by Q5 polymerase PCR (as described in section 2.8.3). The hypermethylated (positive) DNA control was created by using 5 µL of PCR product (about 1 µg) and adding it to 1 µL SAM 32 mM, 5 µL NEB2, and 1 µL of M.SssI (all reagents from NEB M0226S) in a final reaction volume of 50 µL in water. The reaction was incubated for 4 hours at 37°C. Inactivation was performed for 20 minutes at 68°C. Finally, the methylated DNA was cleaned up with ProNex Beads at a 2x ratio of beads to sample (see section 2.8.10).

Methylation efficiency was tested with the restriction enzymes HpaII (NEB R0171S) and MspI (NEB R0106S), which are respectively methylation sensitive or insensitive to CpG methylations within their shared restriction motif of CCGG.

2.8.6 SDS-PAGE and Silver Stain

30 µL of an iodixanol-purified AAV sample was mixed with 9 µL 4x Laemmli Sample Buffer (Bio-Rad) and 1 µL β-mercaptoethanol (Roth). Samples were incubated at 95°C for 10 minutes. 15 µL of the prepared samples were loaded on mini-PROTEAN TGX Precast 7.5% gels in addition to 5 µL of PageRuler Plus Prestained Protein Ladder 10-250 kDa (Thermo 26619). The gels were run for 1 hour at 80 V for the first 20 minutes, later increased

Temp	Time	
95 °C	10 minutes	
95 °C 60 °C	20 seconds 60 seconds	40 cycles

Table 2.4. Thermocycler settings for probe-based qPCR reactions on StepOne Plus System

to 120 V. Silver stains on purified AAV productions were performed using the SilverQuest Silver Staining Kit (Thermo) according to the manufacturer's instructions.

2.8.7 Rolling Circle Amplification

Rolling Circle Amplification (RCA) was conducted using the lambda $\phi 29$ polymerase with the associated reaction buffer as obtained through NEB (M0269S). The sample was denatured by heating 3.5 µL of DNA, 1 µL exo-nuclease-resistant random hexamers, and 0.5 µL of $\phi 29$ reaction buffer at 95°C for 5 minutes and then directly putting it on ice until cooled. The reaction mix was generated by adding 2 µL 10 mM dNTP, 1.5 µL $\phi 29$ reaction buffer, 10.5 µL water, and 1 µL $\phi 29$ polymerase to the 5 µL denatured DNA. The reaction was incubated at 30°C overnight and heat-inactivated at 65°C for 10 minutes.

2.8.8 T5 exonuclease assay

The T5 exonuclease assay measures the accessibility of a DNA substrate to exonuclease activity. DNA samples were diluted to 5 ng/µL in 90 µL nuclease-free water and split into 2 tubes containing 44 µL each (digest and native). Both tubes were mixed with 5 µL of NEB buffer 4 and either 1 µL of water (undigested) or 1 µL of T5 exonuclease (NEB M0663). Reactions were incubated at 37°C for 1 hour and 85°C for 15 minutes. Reactions were performed in duplicates. Quantitative PCR was performed using the primer/probe set against CMVenh (see Table 2.8) and the 2x Sensimix probe kit set with ROX (Meridian, BIO-83005) with 1 reaction consisting of 12.5 µL 2x Sensimix, 0.5 µL ROX, 0.1 µL forward primer (10 µM), 0.1 µL reverse primer (10 µM), 0.025 µL probe (10 µM), 11.2 µL PCR grade water, and 1 µL of sample. The qPCR was conducted as outlined in Table 2.4. The exonuclease sensitivity is defined in equation 2.9, where FE_{exo} is the exonuclease sensitivity and Ct_{dig} and Ct_{nat} are the Ct values from the T5-digested and undigested (native) samples, respectively.

$$FE_{exo} = 2^{Ct_{dig} - Ct_{nat}} \times 100 \tag{2.9}$$

2.8.9 Transduction experiments in HEK293T and Huh7 cells

$1.5e6$ HEK293T or Huh7 cells per well were seeded on a clear-bottom 96-well plate (Corning) in 100 µL DMEM (10% FBS, 1% Pen/Strep). They were immediately transduced at a multiplicity of infection (MOI) of $1e4$, with three replicates per construct and per production (every construct was separately produced three times), resulting in nine total transduction replicates per construct. GFP emission was measured three days post-transduction. The medium was replaced by 100 µL PBS and every well was measured by a Spark Multimode Microplate Reader (Tecan).

2.8.10 Magnetic beads DNA clean up

In this work, I have used either AMPureXP beads or ProNex magnetic beads for cleaning up DNA. It is indicated which beads I used for which clean-up. The specified ratio of beads was added to the sample and left to incubate for 2-5 minutes at room temperature. Beads were then separated from the liquid using a magnetic stand. Supernatant was removed, and the beads were washed 2 times in freshly prepared 80% EtOH (AMpure XP) or ProNex bead wash buffer. After letting the beads dry until they obtained a matte shine, they were eluted with the specified amount of nuclease-free water. Beads were resuspended in the water and incubated at room temperature for 2-5 minutes. The beads are again separated from the liquid in the magnetic stand, and the water containing the purified DNA is transferred to a new cup.

2.9 Lists of Materials, Kits, and Reagents used in this work

2.9.1 Devices and Tools used in this work

Application	Device	Provider
Pipetting		
Pipetting	accu-jet® pro Pipette Controller	BrandTech Scientific (Essex, UK)
Pipetting	Research plus (2.5, 10, 20, 100, 200, 1000)	Eppendorf (Hamburg, Germany)
Pipetting	12-channel research plus (10, 100, 300)	Eppendorf (Hamburg, Germany)
Centrifugation		
Centrifugation	5415R Benchtop Centrifuge	Eppendorf (Hamburg, Germany)
Centrifugation	Allegra X-12	Beckman Coulter (Brea, USA)
Centrifugation	Avanti J-26 XP	Beckman Coulter (Brea, USA)

Ultracentrifugation	OptimaTM L-90K Ultracentrifuge	Beckman Coulter (Brea, USA)
Ultracentrifugation rotor	Rotor 70 TI/70.1 TI	Beckman Coulter (Brea, USA)
Centrifugation rotor	Rotor JA-10	Beckman Coulter (Brea, USA)
Cell culture		
Automated cell counter	Countess	Thermo Fisher Scientific (Waltham, USA)
Incubator for eukaryotic cells	Heracell 150 Incubator	Thermo Fisher Scientific (Waltham, USA)
Sterile cell culture working bench	Herasafe KS 12	Thermo Fisher Scientific (Waltham, USA)
Gel Electrophoresis and blotting		
Agarose Gel Chamber	Mupid One	Biozym (Hessisch Oldendorf, Germany)
SDS-PAGE chamber system	Mini-PROTEAN Tetra Cell	Bio-Rad (Hercules, USA)
Power device for SDS-PAGE	PowerPac HV	Bio-Rad (Hercules, USA)
Imaging for dot blot, silver stain, agarose gels	Azure 400 Visible Fluorescent Imager	Azure Biosystems (Dublin, USA)
ddPCR		
Droplet generation for ddPCR	QX200 Droplet Generator	Bio-Rad (Hercules, USA)
PCR in ddPCR	C1000 Touch Thermal Cycler	Bio-Rad (Hercules, USA)
ddPCR droplet analysis	QX200 Droplet Reader	Bio-Rad (Hercules, USA)
CUT&Tag		
Brightfield nuclei counting	Luna-FL	Logos Biosystems (Gyeonggi-do, South Korea)
Tissue dissociation	Dounce Homogenizer 15 mL	Carl ROTH (Karlsruhe, Germany)
Sequencing		
Nanopore Sequencing	Mk1B	Oxford Nanopore (Oxford, UK)
Miscellaneous		
Tissue lysis for DNA extraction	TissueLyser LT	Qiagen (Hilden, Germany)
Sterile hood for qPCR and PCR preparation	Captair Bio Smart PCR-Hood	Erlab (Val-de-Reuil, France)
PCR cycler	Nexus GSx1/GX2e	Eppendorf (Hamburg, Germany)
qRT-PCR cycler (96-well plates)	StepOne Plus	Applied Biosystems/Thermo Fisher Scientific (Waltham, USA)
Shaking incubator for bacteria	Multitron	INFORS HT (Basel, Switzerland)

DNA/RNA concentration measurment	NanoDrop 2000 Spectrophotometer	Thermo Fisher Scientific (Waltham, USA)
DNA quantification	Qubit 2.0 Fluorometer	Thermo Fisher Scientific (Waltham, USA)
DNA quality assessment	2100 Bioanalyzer	Agilent Technologies (Santa Clara, USA)

Table 2.5. List of devices, tools, gadgets and contraptions used in this work

2.9.2 In-house buffer recipes

Name	Ingredients
15% iodixanol	25% (v/v) OptiPrep™, 75% (v/v) PBS-MK-NaCl
20xSSC	3.0 M NaCl, 0.3 M Sodium Citrate, pH 7.0
25% iodixanol	41.56% (v/v) OptiPrep™, 58.19% (v/v) PBS-MK, 0.25% (v/v) Phenol red stock
40% iodixanol	66.67% (v/v) OptiPrep™, 33.33% (v/v) PBS-MK
60% iodixanol	99.75% (v/v) OptiPrep™, 0.25% (v/v) Phenol red stock
Benzonase buffer	150 mM NaCl, 50 mM TRIS-HCl (pH 8.0), 2 mM MgCl2
Gel Denaturation Buffer	1.5 M NaCl, 0.5 M NaOH
Gel Neutralisation Buffer	3 M NaCl, 0.5 M Tris-HCl
Hirt Lysis Buffer	10 mM Tris pH 8, 1% SDS, 10 mM EDTA, 250 µg/mL Proteinase K
LB agar	1.5% (w/v) Bacto agar, 1% (w/v) NaCl, 1% (w/v) Bacto tryptone, 0.5% (w/v) Bacto yeast extract
LB medium	1% (w/v) NaCl, 1% (w/v) Bacto tryptone, 0.5% (w/v) Bacto yeast extract
PBS-MK	PBS (1X), 2.5 mM KCl, 1 mM MgCl2
PBS-MK-NaCl	1 M NaCl in PBS-MK
Phenol red stock	Nuclease-free H2O, 0.5% Phenol red
TAE buffer	2 M TRIS, 1 M acetic acid, 50 mM EDTA
TBS-T	1.25 M NaCl, 250 mM Tris/HCl, pH 7.4; 0.05% (v/v) Tween20

Table 2.6. List of in-house buffer recipes used throughout this work.

2.9.3 Reagents and Primers

Reagent	Provider
0.25% Trypsin / EDTA	Gibco/Thermo Fisher Scientific (Waltham, USA)
1 × Dulbecco's phosphate buffered saline (PBS)	Gibco/Thermo Fisher Scientific (Waltham, USA)
1 kb Plus DNA ladder	Thermo Fisher Scientific (Waltham, USA)
10x Tris/Glycine/SDS Electrophoresis Buffer	Bio-Rad (Hercules, 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)
Biozym LE Agarose	Biozym Scientific (Hessisch Oldendorf, Germany)
CutSmart buffer	New England Biolabs (Ipswich, USA)
ddPCR Supermix for Probes (No dUTP)	Bio-Rad (Hercules, USA)
Deoxynucleotide (dNTP) Solution Mix (10 mM of each)	New England Biolabs (Ipswich, USA)
Diluent B	New England Biolabs (Ipswich, USA)
DMEM Glutamax	Gibco/Thermo Fisher Scientific (Waltham, USA)
Droplet Generation Oil for Probes	Bio-Rad (Hercules, USA)
EDTA	GRÜSSING GmbH (Filsum, Germany)
Ethanol absolute (EtOH)	Merck (Darmstadt, Germany)
Fetal bovine serum (FBS)	Capricorn Scientific (Ebsdorfergrund, Germany)
Gel Loading Dye, Purple (6X)	New England Biolabs (Ipswich, USA)
Gelred Nucleic Acid Gel Stain	Biotium (Fremont, USA)
Glycerol	VWR chemicals (Fenenay-sous-Bais, France)
Isopropanol	Merck (Darmstadt, Germany)
Laemmli Sample Buffer 4x	Bio-Rad (Hercules, USA)
NEBuffer 1.1, 2.1, 3.1	New England Biolabs (Ipswich, USA)
Nuclease-free H2O	Qiagen (Hilden, Germany)
OptiPrep™ (Iodixanol)	Progen (Heidelberg, Germany)
PageRuler Plus Prestained Protein Ladder	Thermo Fisher Scientific (Waltham, USA)
Penicillin / Streptomycin (P/S)	Gibco/Thermo Fisher Scientific (Waltham, USA)
Phenol red	Merck (Darmstadt, Germany)
Polyethylenimine (PEI MAX)	Polysciences Europe GmbH (Eppelheim, Germany)
RNAlater	Thermo Fisher Scientific (Waltham, USA)
SensiMix II Probe Kit	Meridian (Cincinnati, USA)
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	New England Biolabs (Ipswich, USA)
TE Buffer	Thermo Fisher Scientific (Waltham, USA)
TrickTrack DNA Loading dye (6x)	Thermo Fisher Scientific (Waltham, USA)
TRIS	Roth (Karlsruhe, Germany)
TRIS-HCl	Roth (Karlsruhe, Germany)
Triton X-100	Merck (Darmstadt, Germany)
Trypan Blue Solution, 0.4%	Thermo Fisher Scientific (Waltham, USA)
Tween20	Roth (Karlsruhe, Germany)

Table 2.7. List of commercial reagents used in this work

name	**Sequence 5' – 3'**	**Usage**	**Details**
wtf1_SphI_fwd	TCTGCATGCACGAAAAAGTTCG GCAAGAGG	Cloning	wt-fragments Gen 1 – Fragment 1.1

wtf1_BsaI_c1_rev	AGGTCTCACGTTGCTGCTGGTGTTCGAAGG	Cloning	wt-fragments Gen 1 – Fragment 1.1
wtf2_BsaI_c1_fwd	AGGTCTCAAACGGGAAATTGGCATTGCGATTCC	Cloning	wt-fragments Gen 1 – Fragment 1.2
wtf2_SpeI_rev	TCTACTAGTAGGAAGTGTACTGAATTTCGGG	Cloning	wt-fragments Gen 1 – Fragment 1.2
wtf1.2_SphI_fwd	TCTGCATGCATCGGGTGGCTCGTGGACAAGG	Cloning	wt-fragments Gen 2 – Fragment 2.1
wtf1.2_BsaI_c1_rev	AGGTCTCACGTTAACATCCGGTCTTGCAACGG	Cloning	wt-fragments Gen 2 – Fragment 2.1
wtf2.2_BsaI_c1_fwd	AGGTCTCAAACGTACAGGCAGTGGCGCACCAATGG	Cloning	wt-fragments Gen 2 – Fragment 2.2
wtf2.3_BsaI_c2_rev	AGGTCTCACATATAGCCACGGGATTGGTTGTCC	Cloning	wt-fragments Gen 2 – Fragment 2.2
wtf3.1_BsaI_c2_fwd	AGGTCTCATATGCACTTTTCACCACGTGACTGG	Cloning	wt-fragments Gen 2 – Fragment 2.3
wtf2.2_SpeI_rev	TCTACTAGTATAGCCACGGGATTGGTTGTCC	Cloning	wt-fragments Gen 2 – Fragment 2.3
KanaStuffer_fwd_BsaI_CS1	TGGTCTCAAACGCCGTCCTCCCTCACCGCTG	Cloning	Kanamycin stuffer
KanaStuffer_rev_BsaI_SpeI	TGGTCTCACTAGTACCGCTGCGCCTTATCCGGTAACTATC	Cloning	Kanamycin stuffer
CMVeGFP_fwd_BsaI_BglII	TGGTCTCAGATCTCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGAC	Cloning	eGFP cassette
CMVeGFP_rev_BsaI_CS1	TGGTCTCACGTTCTCCCCAGCATGCCTGCTATTGTCTTC	Cloning	eGFP cassette
CB_epPCR_lambda2_SphI_fwd	AAGCATGCGAGTATCCGTGAGAACGACG	Cloning	lambdaStuffer
CB_lambda_ctrl-Lambda_SpeI_rev	ATACTAGTGGAAAACCGTCAACCTGCAGG	Cloning	lambdaStuffer
CMVenhancer probe	FAM-CGGTAAACTGCCCACTTGGCAGT-BHQ1	ddPCR or qPCR	CMVenh
CMVenhancer fwd	AACGCCAATAGGGACTTTCC	ddPCR or qPCR	CMVenh
CMVenhancer rev	GGGCGTACTTGGCATATGAT	ddPCR or qPCR	CMVenh
bla probe	HEX-CAGTGCTGCCATAACCATGAGTGA-3IABKFQ	ddPCR or qPCR	bla
bla fwd	GCATCTTACGGATGGCATGA	ddPCR or qPCR	bla
bla rev	GTCCTCCGATCGTTGTCAGAA	ddPCR or qPCR	bla
ITR probe	HEX-CACTCCCTCTCTGCGCGCTCG-BHQ1	ddPCR or qPCR	ddPCR ITR
ITR fwd	GGAACCCCTAGTGATGGAGTT	ddPCR or qPCR	ddPCR ITR
ITR rev	CGGCCTCAGTGAGCGA	ddPCR or qPCR	ddPCR ITR
rep probe	FAM-TGATCGTCACCTCCAACA-BHQ1	ddPCR or qPCR	library quantification

rep fwd	AAGTCCTCGGCCCAGATAGAC	ddPCR or qPCR	library quantification
rep rev	CAATCACGGCGCACATGT	ddPCR or qPCR	library quantification
muRPP30_g6 probe	HEX-CCCTGTTAATGTGGTAAGTATTGCTCT-BHQ1	ddPCR or qPCR	gDNA housekeeper
muRPP30_g6 fwd	GATGTGGATTTAGTCTGTATAACTG	ddPCR or qPCR	gDNA housekeeper
muRPP30_g6 rev	CAGATCCAAGTTCAGAACACAA	ddPCR or qPCR	gDNA housekeeper
muRPP30_c9 probe	HEX-TGCCACGTCATATGGTCCTCTTATTTCC-BHQ1	ddPCR or qPCR	cDNA housekeeper
muRPP30_c9 fwd	TGTCCAGTGCTGCAGAAAG	ddPCR or qPCR	cDNA housekeeper
muRPP30_c9 rev	GCCCAAACAGCAGTCCTAA	ddPCR or qPCR	cDNA housekeeper
eGFP probe	FAM-ACGACGGCAACTACA-BHQ1	ddPCR or qPCR	cDNA target
eGFP fwd	GAGCGCACCATCTTCTTCAAG	ddPCR or qPCR	cDNA target
eGFP rev	TGTCGCCCTCGAACTTCAC	ddPCR or qPCR	cDNA target
CB_JCV_fwd	CTCTAGAGCAACTAGCGACATTG	ONT neg. ctrl.	amplification of 2510
#1923	CAATTAAGCTTGGGCAAATAATATCGGTGGC	ONT neg. ctrl.	amplification of 2510, 2512
002_CMVeGFP_rev_BsaI_CS1	TGGTCTCACGTTCTCCCCAGCATGCCTGCTATTGTCTTC	ONT neg. ctrl.	amplification of 2554
JB_GFAP_fwd	AGAGCAACTAGTCCCACCTCCCTCTCTGTGCTG	ONT neg. ctrl.	amplification of 2554
CB_LP1_fwd	GCAACTAGTCCCTAAAATGGGCAAACATTGCAAGC	ONT neg. ctrl.	amplification of 2512
#822_LseqFor	GATCTGGTCAATGTGGATTTG	shuffled enrichment	amplification of library
#823_LseqRev	GACCGCAGCCTTTCGAATGTC	shuffled enrichment	amplification of library

Table 2.8. List of primers and probes used in this work

2.9.4 Kits used in this work

Application	Kit	Provider
DNA extraction from tcells and tissues	DNeasy Blood & Tissue Kit	Qiagen (Hilden, Germany)
DNA and RNA extraction from cells and tissues	AllPrep Mini Kit	Qiagen (Hilden, Germany)
DNA extraction from agarose gels	QIAquick Gel Extraction Kit	Qiagen (Hilden, Germany)

DNA purification	QIAquick PCR Purification Kit	Qiagen (Hilden, Germany)
Plasmid Purification (small-scale)	QIAprep Spin Miniprep Kit	Qiagen (Hilden, Germany)
Plasmid Purification (mid-scale)	PureYield Plasmid Midiprep System	Promega (Madison, USA)
Plasmid Purification (large-scale)	NucleoBond PC 500	Macherey-Nagel (Düren, Germany)
Qubit dsDNA quantification	Qubit dsDNA HS Assay Kit	Thermo Fisher Scientific (Waltham, USA)
Magnetic bead purification	ProNex Size-Selective Purification System	Promega (Madison, USA)
Silver staining	SilverQuest Silver Staining Kit	Invitrogen/Thermo Fisher Scientific (Waltham, USA)
On-column digest of genomic DNA for RNA extraction	RNase-Free DNase Set	Qiagen (Hilden, Germany)
cDNA synthesis	High-Capacity cDNA Reverse Transcription Kit	Applied Biosystems/Thermo Fisher Scientific (Waltham, USA)
ddPCR	ddPCR Supermix for Probes (No dUTP)	Bio-Rad (Hercules, USA)
qPCR (probe-based)	Sensimix II Probe Kit	Bioline (London, UK)
enzymatic methyl conversion and library generation	E7120	NEB
nanopore library prep; ligation-based barcoding	SQK-LSK109 (R9), SQK-LSK114.24 (R10)	Oxford Nanopore (Oxford, UK)
nanopore library prep; transposase-based barcoding	SQK-RBK004	Oxford Nanopore (Oxford, UK)
nanopore flowcell washing	EXP-WSH004	Oxford Nanopore (Oxford, UK)
nanopore flowcell flushing	EXP-FLP002 (R9), EXP-FLP004 (R10)	Oxford Nanopore (Oxford, UK)
Cut and Tag	53170	Active Motif (Carlsbad, California, USA)
Southern Blot	PCR DIG Probe Synthesis Kit (Roche 11 636 090 910)	Roche
Southern Blot	DIG Wash and Block Buffer Set (Roche 11 585 762 001)	Roche
Southern Blot	DIG Nucleic Acid Detection Kit (Roche 11 175 041 910)	Roche
Southern Blot	DNA Molecular Weight Marker VII, DIG-labelled (Roche 11 669 940 910)	Roche
Southern Blot	DIG Easy Hyb™ Granulat (Roche 11 796 895 001)	Roche

Table 2.9. List of all used kits with providers

2.10 Software used in this work

Software	Version	reference
Python		
scipy	1.9.2	[264]
matplotlib	3.8.4	[265]
numpy	1.24.3	[266]
pandas	1.1.5	[267]
seaborn	0.13.2	[268]
Biopython	1.79	[269]
R		
R	4.4.2	[270]
Rstudio	2024.12.0	[271]
tidyverse	2.0.0	[272]
ggpubr	0.6.0	[273]
dplyr	1.1.4	[274]
forcats	1.0.0	[275]
ggplot2	3.5.1	[262]
purrr	1.0.2	[276]
readr	2.1.5	[277]
stringr	1.5.1	[278]
tibble	3.2.1	[279]
tidyr	1.3.1	[280]
Biostrings	2.72.1	[281]
GenomicRanges	1.56.1	[282]
DEseq2	1.44.0	[249]
chromVar	1.26.0	[283]
General		
ImageJ	1.53k	[284]
conda	23.3.1	[285]
SEACR	1.3	[248]
bowtie2	2.3.5.1	[245]
seqkit	2.5.1	[250]
fastqc	0.11.9	[286]
multiqc	1.12	[287]
qualimap	2.2.2	[288]
deeptools	3.3.2	[247]
snakemake	7.26	[244]
bwa-met	0.2.7	[152]
MethylDackel	0.3.0 (using HTSlib version 1.2.1)	[253]
samtools	1.10	[252]
modbam2bed	v0.10.0	[259]
dorado	0.7.4+59ab908	[157]
guppy	6.1.5	[289]
nanopolish	0.13.2	[255]
megalodon	2.5.0	[256]
deepsignal	v0.2.0	[257]

Table 2.10. List of all relevant software with version number used in this work

2.10.1 Other software

AI tools

Throughout this thesis, I have used the services of perplexity [290] for literature research and (in a minor role) chatGPT [291] for rewording cumbersome sentences. AI was not used to generate text longer than one sentence. Unless otherwise stated, the data analysis and data visualisation were done independently by me.

Biorender schematics

Some schematics were generated using the online service provided by Biorender. The links to the individual biorender figures are provided as citations as required by biorender's Terms of Service.

3. Results

3.1 CpG methylations and histone modifications on AAV delivered transgenes

3.1.1 AAV derived episomes can be directly sequenced using nanopore sequencing

Even though DNA methylation is a widely studied subject and its impact on gene expression is well known, there is little literature about the methylation state of the AAV-delivered transgene. Because of its availability and lack of tedious and lengthy library preparation protocols, I had initially opted to analyse these DNA methylations using nanopore sequencing supplied by Oxford Nanopore Technologies (ONT). In comparison to the host genomic DNA, the AAV-delivered transgene only makes up a fraction of the total DNA that would be obtained from an extraction. Therefore, traditional bulk sequencing approaches would create a significant amount of 'useless' information by mainly sequencing host DNA. Therefore, I sought to develop a protocol with which to enrich a DNA sample for episomal AAV transgenes that can be directly sequenced by nanopore sequencing. Amplification by PCR is not an option for library preparation, since the base modifications would be lost in the process. In an effort to sequence episomal AAV-delivered transgenes, I had transduced HEK293T cells with an AAV2 packaged transgene as a preliminary experiment. I considered using multiple features of the episomal transgene as leverage for enrichment. It is considered to persist as a mono-, or multimeric circular episome that is smaller than the chromosomal DNA of the host. However, sequencing libraries (i) created from Hirt DNA (size-enrichment), (ii) generated by pull-down using biotinylated oligomers and streptavidin-beads (sequence-enrichment), or (iii) enriched by Cas9 restrictions on dephosphorylated total DNA [292] (sequence-enrichment) yielded next to no reads attributable to the transgene (data not shown here). At the time, I attributed the lack of transgenic reads from cell culture to the unknown state that the AAV2-derived episome has in HEK293T cells 2 days post-transduction (more in section 3.1.7).

Therefore, I decided to pursue mouse samples from previous experiments conducted by Jonas Becker and Claire Domenger in our group. These mice were transduced via a tail vein injection with $1e12$ vg/mouse of an AAV9 encoding a single-stranded transgene containing multiple promoter constructs (one per vector) driving a *yfp* transgene cassette. They were sacrificed 2 weeks post-injection, after which multiple tissues were extracted and stored in RNAlater at 4°C for about 1 year before the here described work began. As interesting candidates with putative differing epigenetic modifications, I decided to focus on the samples from mice transduced with the cassettes containing the promoters CMV, LP1, and GFAP. CMV is the ubiquitously expressed cytomegalovirus promoter, which is widely used in AAV constructs both in the laboratory and in clinical studies. However, in the clinic, the CMV promoter has been reported to undergo silencing at later time points [144]. LP1 is a liver-specific promoter which consists of multiple liver-specific enhancer/promoter fragments [293]. The GFAP promoter endogenously expresses the murine glial fibrillary acidic protein and has recently been shown by our group to express unexpectedly well in mouse livers in an AAV context [243].

To confirm the presence of episomal DNA in total DNA extractions, I began by examining total DNA extracted from mouse livers. The liver is usually among the best AAV-transduced targets in the mouse model organism. To confirm the presence of circular episomes, I used rolling circle amplification (RCA) together with a single-cutter within the transgene cassette. RCA exponentially amplifies circular DNA using random hexameric primers, while the single-cutter restriction enzyme is used to fragment the amplicons, which is expected to form a banding pattern of multiples of the AAV transgene length (3.9 kb, CMV promoter-containing cassette in this case, Figure 3.1 A). To enrich total DNA extractions for circular episomal DNA, I used an exonuclease (RecBCD) to digest and deplete chromosomal DNA. Without performing an RCA reaction, no remaining DNA was detectable on a 1% agarose gel after exonuclease digestion. Episomal DNA could be confirmed when performing the RCA on an exonuclease-enriched sample, visualised by the expected banding of the RCA product after digestion by a single-cutter (Figure 3.1 A; monomer, dimer, trimer). This showed that episomes were present in the mouse liver DNA samples and are still detectable after exonuclease treatment. Since RCA with random hexameric primers is extremely sensitive to all circular DNA, the water control in all replicates (only one shown) also contained amplified DNA with a different banding pattern, which suggested that this is most likely contaminating and unrelated plasmid DNA (Figure 3.1 A; two rightmost lanes). Ligation-based library preparation (see section 2.6.3) followed by nanopore sequencing of the exonuclease-enriched total DNA from mouse livers, revealed a distinct transgene-sized peak at 3.9 kb in the read size distribution histogram (Figure 3.1 B). Together with

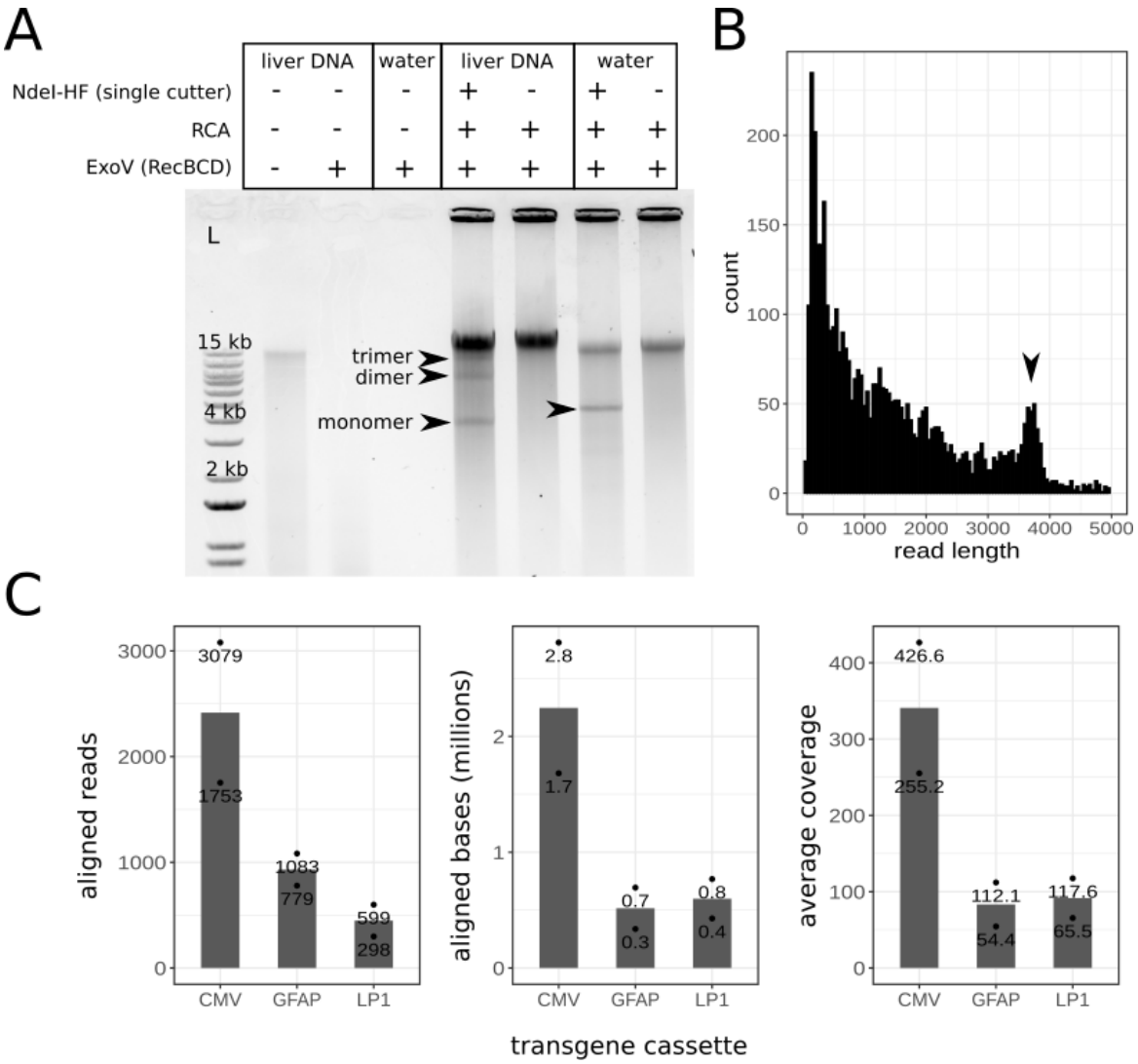

Figure 3.1. Transgenic episomal AAV DNA can directly be sequenced using nanopore sequencing following exonuclease enrichment from the mouse liver.
A) 1% agarose gel of episomal DNA detected in total mouse liver DNA extractions. Chromosomal DNA was depleted using exonuclease digestion (RecBCD), followed by a rolling circle amplification (RCA) step. The amplified DNA was broken back down to single molecules of AAV transgene size by a single-cutter restriction enzyme (NdeI-HF). The arrows indicate mono-, di-, and trimers resolved from the RCA product by the restriction digestion. The arrow for the water control shows a band resolved from the RCA product attributable to contaminating plasmid DNA. Ladder (L) is a 1kb Plus DNA Ladder (Thermo). B) Read size histogram from DNA enriched by exonuclease digestion, directly sequenced by nanopore sequencing. The arrow indicates a peak at the expected size of the transgene (3.9 kb). This sample was prepared using another single-cutter to prepare the episome for ligation-based library preparation. C) Aligned reads, aligned bases, and resulting coverage on the references of the used transgene cassettes using enriched total DNA from mouse livers. These samples were prepared using a transposase-based library preparation protocol.

alignments to the transgene cassette (not shown), this showed that the AAV-derived episome can directly be sequenced from the total DNA of the transduced mouse liver.

Next, I extracted DNA from livers, brain, and muscle (*musculus quadriceps*) of the same mouse samples transduced with the CMV, GFAP, and LP1 transgene cassettes packaged in AAV9 from two mice each. The extracted DNA was enriched in the same manner using an exonuclease to deplete genomic DNA as much as possible before library preparation. The libraries for this experiment were sequenced using an R9 flowcell and the rapid transposase-based preparation kit (see section 2.6.3). The roll-back to a previous generation of ONT chemistry (R9) was a conscious decision, as most of the developed methylation callers that exist were trained on data obtained using R9 chemistry. Aligning to the individual AAV transgene references revealed that only liver samples yielded enough reads for a meaningful analysis (muscle and brain samples had five to ten transgene-aligned reads). Sequencing the liver samples produced between 0.3 and 2.8 million aligned bases (300 to 3000 reads), leading to a coverage between 54x to 427x (Figure 3.1 C). The number of aligned reads varied by a factor of ten, yet a usable coverage was obtained with all transgene cassettes from the liver, thanks to the lack of length restrictions of ONT reads.

3.1.2 Low methylation of the transgene cassette measured by ONT and EM-seq 2 weeks post injection

Having shown that direct nanopore sequencing of the AAV transgene is possible, I set out to analyse the nanopore reads for the methylation state of the transgene in the mouse liver samples. For calling methylation rates of CpG sites from the raw voltage data, I employed the METEORE pipeline [159], which supplies a framework that conveniently combines multiple publicly available callers with the ONT-developed caller guppy. To ensure that these callers could accurately distinguish between methylated and native CpG sites, I additionally generated positive and negative controls. The negative control is a PCR-amplified transgene from every used cassette, excluding the ITR sequences. The positive control was the same PCR amplicon hyper-methylated by the enzyme M.SssI. The hyper-methylation was confirmed by digestion using restriction enzymes HpaII (methylation sensitive) and MspI (methylation insensitive), both targeting the CCGG motif (data not shown). All controls were also sequenced by nanopore. From the available callers in the METEORE pipeline, I chose guppy, nanopolish, megalodon, and deepsignal to estimate the CpG methylation frequency on the enriched liver DNA samples as well as the controls (Figure 3.2 A). To keep the different transgenes comparable to each other, I restricted the analysis to CpG motifs within the individual promoter and the eYFP coding regions (exact same sequence in all cassettes). The promoter was analysed because methylations there are

thought to have the highest impact on transcription. The coding region was analysed as it has the same sequence in all used transgenes, unlike the promoter. All tools managed to separate the negative and positive controls from each other; however, nanopolish and deepsignal appeared to make more accurate calls regarding the controls. The used tools call methylations on the transgenic DNA to a similar frequency as the native DNA controls - both in the promoter and in the protein-coding region. According to nanopore, this suggests there is no methylation present on the transgene.

To obtain a second opinion from a more established technique, I employed enzymatic methyl-sequencing, a method that works similarly to bisulfite sequencing (see section 1.1.6). One sample of each mouse DNA liver extraction was first enzymatically enriched and then subjected to conversion and library preparation. Before the library preparation, native lambda DNA and hyper-methylated pUC19 DNA are spiked in to assess the conversion and protection efficiency. The controls showed acceptable conversion rates of the native control between 99.5% and 99.8% and of the negative control between 3.3% and 4.2% (not shown). The coverage on the transgene was between 750x and 1800x, i.e., even higher than the ONT sequencing runs, which allows me to make meaningful conclusions about the methylation state. It is also worthwhile to mention that in the EM-seq library preparation, next to the samples enzymatically enriched for episomal DNA, I had also processed samples that were taken directly from the mouse liver DNA extraction without any tampering. Those samples only achieved a coverage on the transgenic DNA of 4x to 6x, with some regions not being covered at all. The exonuclease enrichment led to a transgene coverage increase of between 180 and 450 times in the EM-seq experiments, which underscores its importance for efficiently sequencing episomal AAV transgenes.

Across the promoter and coding regions of the transgene, the methylation rate was measured by EM-seq to be close to 0%, whereas the median nanopore calls were around 5% averaged over all used tools (3.2 B). When assessing the methylation rate across the genome, a similar picture emerged. EM-seq measured methylations were close to 0% throughout the promoter, coding region, and stuffer (Figure 3.2 C). The ONT measured methylations were slightly higher yet stayed below 10%. Importantly, they hardly differed from the native control average (Figure 3.2 C, dashed yellow line). The nanopore methylation rate on the transgene seemed to be mainly noise, which was also demonstrated in the low correlation values between nanopore and EM-seq CpG methylation rates (Figure 3.2 D). Interestingly, the methylation rate in the ITR annotations was around 20% for EM-seq, whereas nanopore displayed the same deviation around the negative control as all other regions (Figure 3.2 C). This phenomenon will be discussed further in section 3.1.5. In conclusion, methylation estimation calling by nanopore sequencing worked well on the episomal transgene but was

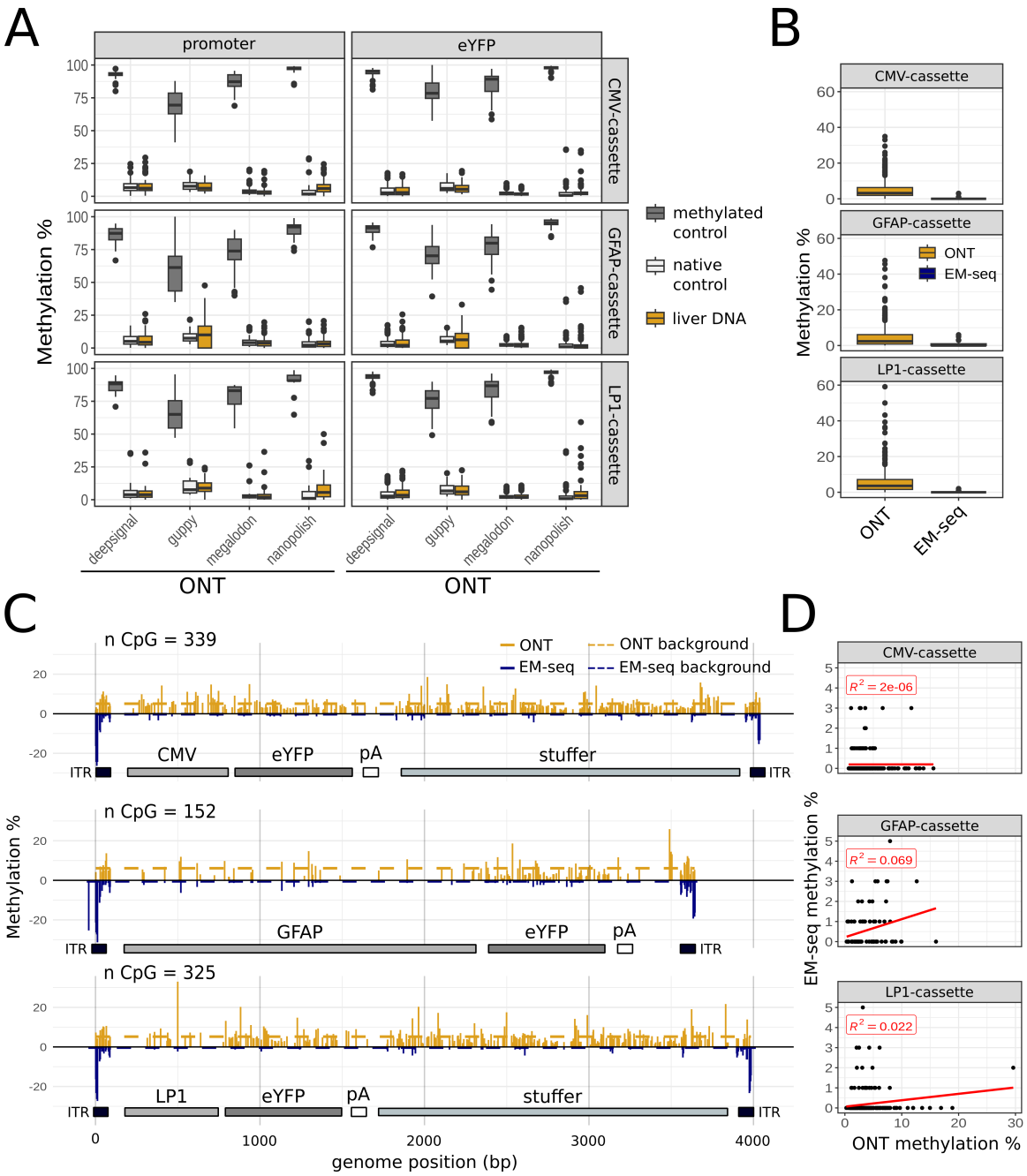

Figure 3.2. CpG methylation rates on transgenic episomal DNA as measured by ONT and EM-seq from mouse livers transduced by CMV, GFAP, and LP1 promoter-driven cassettes.

A) Methylation rates as called by different callers from ONT data. Only CpG sites within the promoter region and the eYFP annotation are shown. Replicates were combined. Promoters differ in the number of CpGs: CMV = 33, GFAP = 36, LP1 = 14. B) Comparison of the ONT methylation averaged over all callers with the rate observed by EM-seq across the promoter and eYFP annotations. C) Transgene-wide methylation rates over the different cassettes. Positive values are the ONT methylation calls averaged over all used tools in A. Negative values are methylation frequencies measured by EM-seq. The dashed lines represent the background signal as measured by an unmethylated (native) DNA control with the individual technology. Annotations for every cassette are indicated below. The number of CpG motifs on every transgene is also indicated. D) Pearson correlation between ONT and EM-seq of the individual methylation frequencies per CpG.

plagued by a relatively high rate of false-positive methylation predictions. However, both nanopore and EM-seq agreed on neglible methylation rates across the promoter and coding regions of the AAV transgene 2 weeks post-injection. Being aware of its drawbacks, the low cost and ease-of-use of nanopore sequencing make it a powerful alternative to established methods.

3.1.3 12-week study of epigenetic modifications on the AAV transgene in mice

After 2 weeks, there was no noteworthy methylation detectable in any of the coding or regulatory annotations of the AAV9-packaged transgene in the liver samples of mice. To obtain a better understanding of the trajectory of epigenetic transgene regulation over time, I devised an expanded experiment analysing the murine liver at later time points post-injection. Since there was no discernible difference in the methylation rate of different promoters, I decided to focus on the well-characterised and widely used CMV promoter. In this experiment, I decided to compare different capsids packaging the same transgene cassette. The capsid can direct its encoded transgene down widely differing expression trajectories, which have been causally linked to differences in epigenetic regulation [178, 177]. Four C57BL/6 mice for every condition were injected with a CMV promoter-driven eYFP cassette (same CMV cassette as in section 3.1.2) packaged into either the AAV2 or the AAV9 capsid. AAV9 widely transduces tissues in the mouse, whereas AAV2 mainly transduces the liver [294]. Mice were sacrificed after 2, 6, or 12 weeks, and livers were extracted. Next to methylation, I also analysed different histone modifications present on the delivered transgenes at the chosen time points (Figure 3.3 A). To assess the transduction capabilities of the different vectors, viral genomes per diploid genomes (vg/dg) were measured using digital-droplet PCR (ddPCR; Figure 3.3 B). Both vectors were present at similar levels 2 weeks post-injection, and only slightly diverged thereafter. The AAV9-delivered transgene vg/dg slightly increased, whereas the AAV2-delivered transgene vg/dg slightly decreased over time, with the biggest difference measured at the 12-week time point. A bigger discrepancy could be seen in the transcription of the used vectors (Figure 3.3 C). Transcription measurements were taken from liver RNA extracts, also using ddPCR, assessing viral transcripts (*yfp*) against housekeeper transcripts (murine *RPP30*; vt/hkt). Here, the largest difference was observed after 2 weeks, where the AAV9-delivered transgene produces about 6x more transcripts compared to the AAV2-delivered transgene. Over the time points, the transcription of the AAV9 transgene decreased, yet stayed above that of the AAV2-delivered one at all measured points. This was an intriguing observation suggesting that the markedly different behaviours of AAV2 and AAV9 might be caused

	AAV2			AAV9		
replicate	2 weeks	6 weeks	12 weeks	2 weeks	6 weeks	12 weeks
1	10 (9x)	4 (3x)	16 (11x)	261 (230x)	200 (169x)	81 (72x)
2	52 (32x)	26 (21x)	8 (7x)	361 (310x)	154 (132x)	56 (50x)
3	10 (9x)	34 (24x)	4 (3x)	213 (182x)	217 (191x)	76 (66x)
4	5 (5x)	3 (3x)	8 (7x)	407 (350x)	260 (221x)	110 (96x)

Table 3.1. Reads and resulting coverage in brackets of all four mice per time point on the transgene reference.

by differential epigenetic regulation of the delivered transgenes, originally triggered by the used capsid.

3.1.4 CpG methylation marginally increases over time

In the preliminary experiments on stored mouse tissues (section 3.1.2), I found that I can directly sequence DNA from AAV-derived transgenes, but also that there was no methylation present on the transgene two weeks post-injection (p.i.). Nanopore sequencing using the R9 chemistry and an ensemble of methylation caller tools was able to distinguish accurately between negative and positive controls. For this experiment, I decided to use the R10 ONT chemistry, because (i) using a variety of different methylation calling tools yielded no substantial benefit, and because (ii) the updated R10 chemistry has been shown to more accurately identify methylations [296]. DNA extractions were enriched for episomal DNA using an exonuclease as before (see 2.6.1 and 3.1.1). Instead of using a transposase-based library preparation, here, the enriched DNA was digested with a single-cutter to allow for the ligation of sequencing adapters and subsequent library preparation. This step was changed to obtain more complete reads covering the entire transgene, including the ITR-recombination site (see section 3.1.5). The methylation analysis is unaffected by this step.

Using this approach and aligning only sufficiently long reads to the transgene reference yielded a coverage between 50x and 350x for the AAV9-transduced samples. However, the coverages of the AAV2-transduced samples were mostly in the range of single digits (Table 3.1). The low coverage of the individual samples prevented me from making meaningful observations about the methylation status of the transgene while accurately comparing AAV2 with AAV9. Therefore, I pooled all replicates from the same AAV capsid and time point, resulting in a minimum of 30 aligned reads per condition. To maintain comparability, all pooled samples were subsampled to 30 reads each. Dorado and the latest 'super accurate' model ('sup', version 4.3) were used for base and methylation calling. At this level of coverage, the median methylation across the entire transgene (including ITR and stuffer

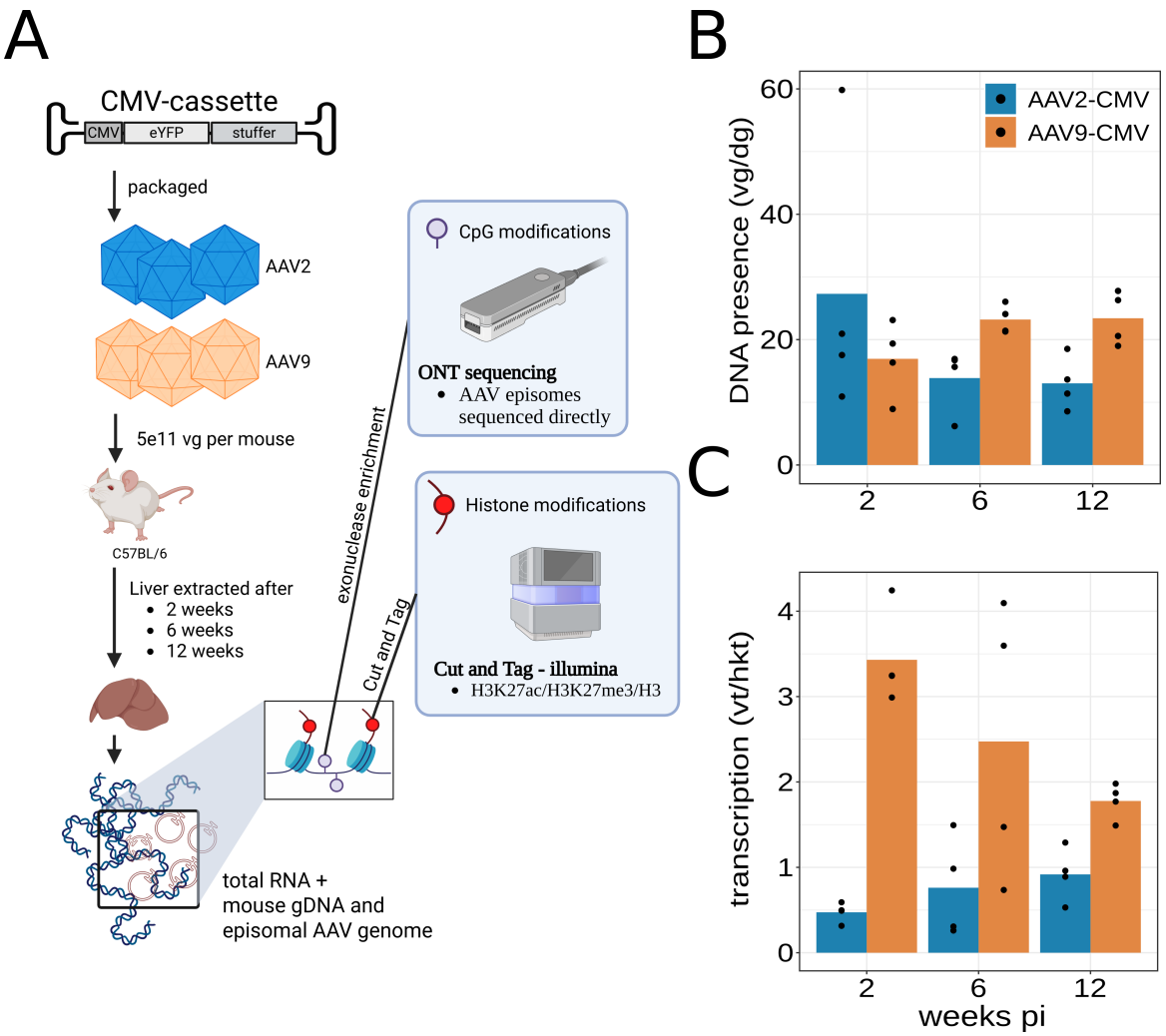

Figure 3.3. Long-term mouse transduction experiment setup and initial efficiency data. A) Experimental scheme. C57BL/6 mice were tail vein-injected with $1e5$ vector genomes of either AAV2 or AAV9 packaging a CMV-*yfp*-transgene-cassette. Mice were sacrificed after 2, 6, and 12 weeks and livers extracted. From the livers, RNA was extracted to assess transcription, and DNA was extracted for measuring transduction and performing episome sequencing. Additionally, the same liver tissue was used for CUT&Tag library preparation. Figure created with Biorender [295]. B) Viral vector genomes per diploid genomes (vg/dg) measured by ddPCR. C) Transgene transcription measured by ddPCR as viral transcripts over housekeeper transcripts (vt/hkt).

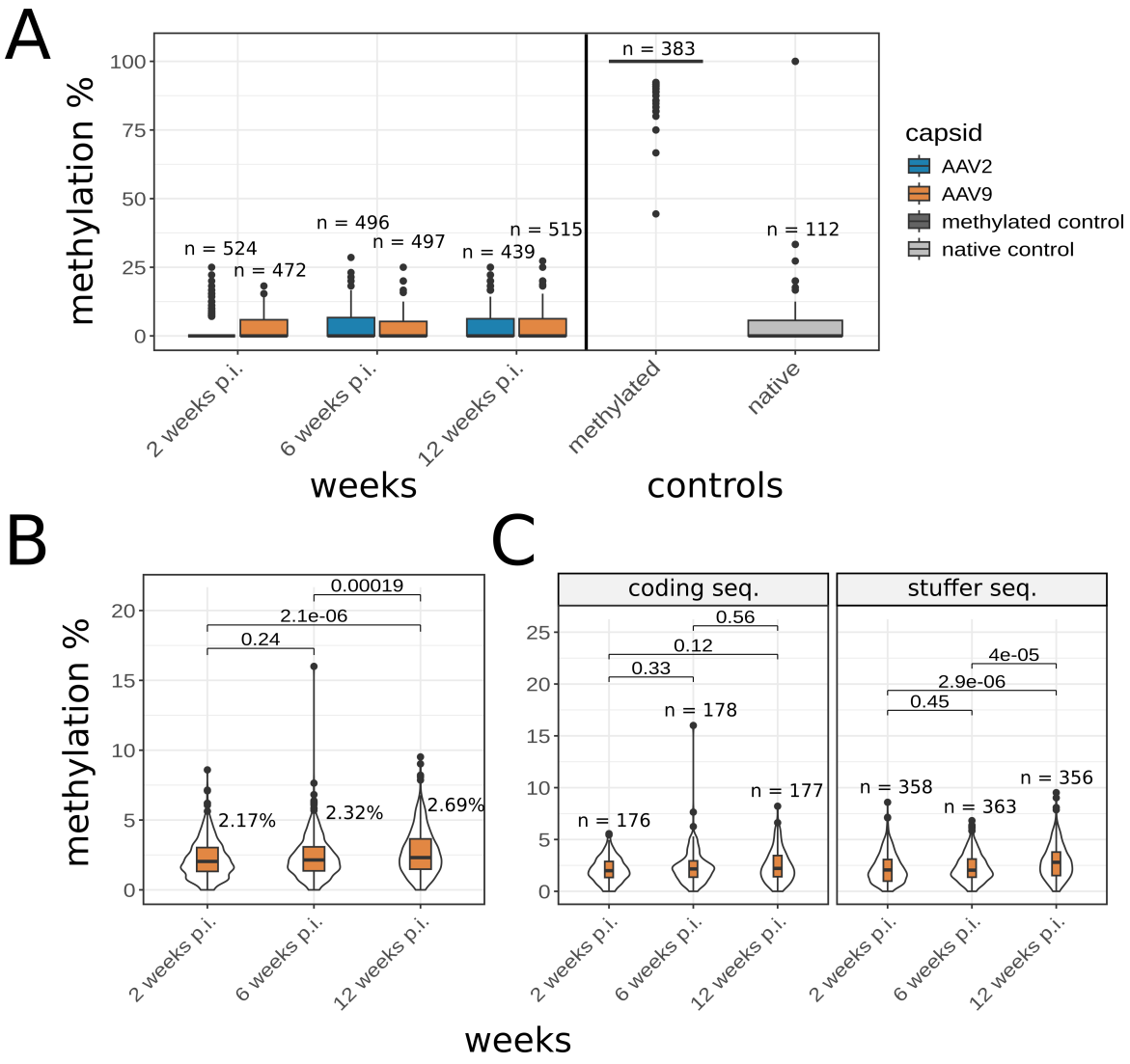

Figure 3.4. Long-term changes in transgene methylations measured by nanopore sequencing.

A) Methylation rate of AAV2 and AAV9 delivered transgenes in mouse livers determined using dorado and R10 chemistry. Positive (methylated) and negative (native) controls are the raw data from section 3.1.2 processed with the latest models and dorado. Replicates were pooled and subsampled to 30 reads each. The number of CpG sites remaining after stringency filters is indicated. B) Methylation rate across the entire transgene in AAV9 samples. Samples were pooled and subsampled to 300 reads each. The percentage indicates the average methylation rate at every time point. The Wilcoxon-Mann-Whitney test was used to test for statistical significance, and p-values are indicated. C) Methylation rates of the AAV9 transgene cassette, split by functionality. The coding part entails the CMV-*yfp*-pA (1.6 kb), and the stuffer part entails the remaining lacZ stuffer sequence (2 kb). The number of analysed CpG sites for every sample is indicated. The Wilcoxon-Mann-Whitney test was used to test for statistical significance, and p-values are indicated.

sequences) remains close to 0% at all tested time points and indistinguishable from the native control (Figure 3.4 A). The methylated and native control raw data are the same as in section 3.1.2, but processed with the latest models, also using dorado and also subsampled to 30 reads each. Because of the applied stringency filtering, the number of retained CpG sites moderately differed between the samples. The native control having fewer CpG sites remaining after filtering probably stemmed from two factors. Firstly, the PCR done to create the controls did not span the entire stuffer region of the transgene cassette, and secondly, the R9 chemistry likely led to more ambiguous calls that were filtered during data analysis (see 2.6.4).

Because the coverage of the AAV9 transduced samples was considerably higher, I was able to analyse them in more detail. I continued with pooled samples, but this time subsampled to 300 reads each, i.e, ten times higher than in the comparison of AAV2 with AAV9. Again, the entire transgene cassette, including the ITR sequences, was used for analysis. At this depth, a slight, but statistically significant increase in methylation rate could be observed across the time points (Figure 3.4 B). The average methylation rate marginally increased from 2.17% to 2.69%. This is, however, unlikely to explain the vastly diminished transcription I saw in the AAV9-delivered transgenes at later time points. To figure out if a specific region of the transgene was responsible for the increase in methylation, I split the transgene into two halves. The first half is termed the coding half and contains the regulatory and coding sequences (CMV-*yfp*-pA, 1.6 kb). The second half is made of stuffer DNA and contains the lacZ stuffer (2 kb) used to expand the transgene cassette to a total length of 4 kb, necessary for efficient packaging. The increase in methylation was only statistically significant in the stuffer half of the transgene cassette (Figure 3.4 C). This indicates that the methylation mainly occurs in DNA, which should not - at least in theory - affect the transcription of the transgene. However, the number of CpGs in the stuffer DNA was also greater than in the coding half, which could also contribute to a higher statistical significance.

In conclusion, the methylation of the episomal form in mouse livers of the tested CMV cassette was very low at all tested time points. The capsid used for the transgene delivery also did not influence methylation on the transgene. The AAV9-delivered transgenes marginally increased over time, to an extent that is unlikely to be biologically significant.

3.1.5 Methylation rates and recombinations of the ITR sequences

The R9 transposase-based library preparation conducted in section 3.1.2 also yields reads of the ITR sequence, but the information on read length is lost due to the fragmentation. The R10 nanopore sequencing experiment done for section 3.1.4 was performed using a

ligation-based library preparation approach. This allowed for the retrieval of long reads that also stretch across the ITR sequence in the reference. Using this data, I wanted to address two questions focused on the episomal ITR sequences. Firstly, are the ITR-sequences more highly methylated as the EM-seq results suggest? Secondly, how does the recombined ITR sequence in an episomal AAV in the murine liver look like?

Addressing the first question, I aligned all reads from the experiment on long-term methylations against a reference containing only both ITR sequences as annotated in the plasmid map. To address this question with adequate read coverage, only reads from the AAV9 samples were taken into consideration without subsampling them. I could obtain adequate coverage on all of the 16 CpG motifs present on the used ITR reference. The majority of the CpGs were covered by 10 or more reads at every time point, allowing me to make a meaningful observation on the methylation state. Mirroring the data from the complete transgene (Figure 3.4), the methylation rates on the ITR sequences remained low, with only a marginal and non-significant increase detectable (Figure 3.5 A). To showcase the difference between the ONT-measured ITR methylation and the EM-seq data, I continued by comparing the methylation rate on the ITR sequences to that of the negative control in a log fold-enrichment approach. This resulted in EM-seq portraying CpG methylation on the ITR sequences to be four times higher than on unmethylated lambda DNA (Figure 3.5 B). In contrast, the mean of the R9 sequenced data was very similar to that of its negative control. The R10 sequenced data even appeared to have a lower methylation rate than the negative control, but this difference can be pinned on the sequencing chemistry, considering that the negative control was R9 sequenced. The ITR sequence was found to be overly methylated only when using the EM-seq approach, but not with nanopore sequencing.

These data raised the suspicion that EM-seq produces false-positive artefacts in the ITR sequences. In the genomes of mammals, cytosine methylation is constrained to the motif CpG. In plants, cytosine methylations can also be observed in other motifs, such as CHH or CHG, where H stands for C, A, or T. The observation of methylations in these motif contexts within my EM-seq data hints at an incomplete conversion of cytosines, since they should not be methylated in the mouse. To test this, I used the EM-seq data and extracted the methylation rates of different motifs using MethylDackel. The methylation rates of these motifs mirror that of the CpG motif, with the ITRs being elevated over the rest of the transgene (Figure 3.5 C).

To address the second question about the nature of ITR recombinations in the episomes, I performed a series of blast alignments on the AAV9-transduced samples. I preferred AAV9 samples over the AAV2-delivered samples because of their higher read numbers. Every

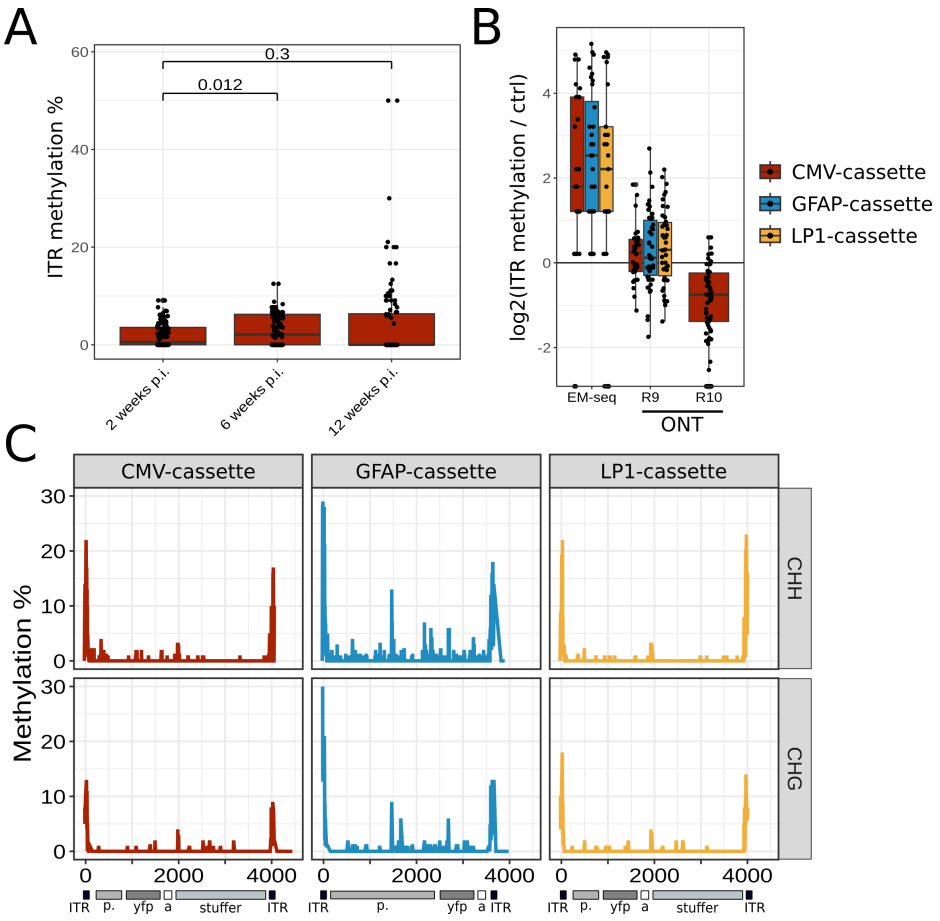

Figure 3.5. Methylation rate within the ITR sequences of the episomal AAV transgene. A) Methylation rate on CpG motifs (black dots) across the ITR sequences of the transgene of AAV9 transduced mice over three time points post-injection as measured with the ligation-based library preparation and R10 nanopore sequencing. The Wilcoxon-Mann-Whitney test was used to test for statistical significance, and p-values are indicated. B) Log fold-change of the ITR CpG methylation rate over the rate on the individual negative controls for EM-seq and nanopore sequencing with R9 and R10 chemistry. Every dot corresponds to one individual CpG site. Only the 2-week time point is taken into consideration for the long-term samples using R10 chemistry. C) Methylation of motifs CHH and CHG according to EM-seq. A line is drawn through every motif occurrence. Only motifs with a coverage of over 10 were taken into consideration. The transgene-cassette annotation is shown below, with 'p.' being the promoter, 'yfp' the eYFP coding sequence, 'a' the polyadenylation signal, 'stuffer' the stuffer DNA, and 'ITR' the ITR sequences.

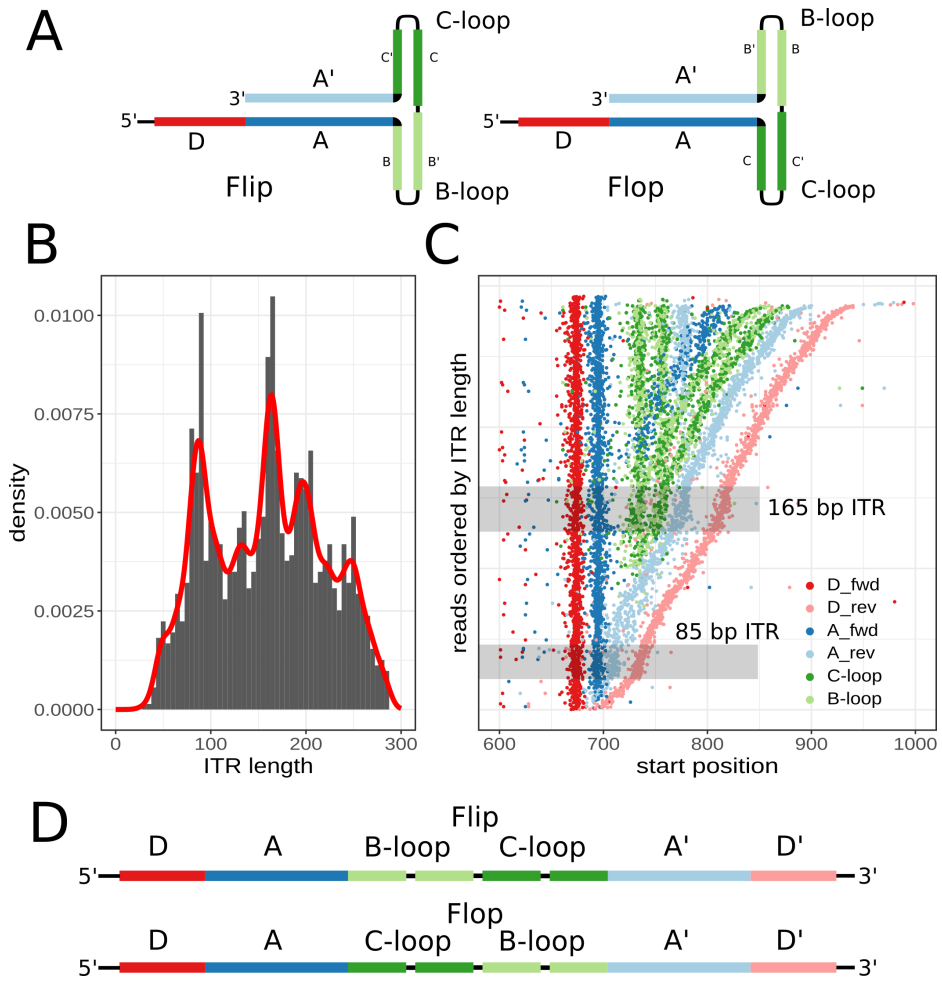

Figure 3.6. Recombination of the ITR sequences of the AAV9-delivered episomal transgene.

A) Schematic of ITR sequences with individual sections in Flip and Flop orientation. B) Histogram of the ITR lengths obtained from blast analysis of every individual read. A kernel-density estimation is shown in red. C) The order of alignments of individual ITR sections. Reads are ordered by ITR size on the Y-axis, with larger sizes at the top. The X-axis represents the start of the ITR section alignment. Every read has a varying number of dots (ITR sections) assigned to it, all of which are plotted in a horizontal line at the same Y-axis value. The grey boxes indicate the two major ITR-length variants of 165 and 85 bp. D) Schematic of the most prominent recombined ITR sequence showing both the Flip and Flop variation.

read was used as the query of a blastn alignment against every individual section of the ITR sequence (A-sequence, B-loop, C-loop, D-sequence) as subject (Figure 3.6 A). Only hits with an e-value lower than 0.05 were retained. The length of the ITR sequence within every read was determined by comparing the location of the first two D-sequences aligned to the read. In total, 1453 out of 2396 sequences (61%) from the AAV9 samples had an identifiable ITR sequence of below 1000 bp in length. The length distribution of the ITR-sequences showed them to vary substantially, but still displayed 3 major peaks with distinct sizes (Figure 3.6 B). The major peak was at around 165 bp, which is the theoretical length of an intact ITR sequence with two D-sequences in the order: D-A-B/C-B/C-A-D (also illustrated in Figure 3.6 D). The second-largest peak was at around 85 bp, which, upon closer analysis, turned out to be ITR sequences devoid of B- or C-loops (D-A-A-D). To gain a better understanding of all possible rearrangements within the ITR sequences, I plotted the start position of every ITR section of every read and arranged the reads according to the length of the ITR (Figure 3.6 C). Owing to my definition, all reads started and ended with a D-sequence, which in most cases was followed by an A-sequence. Shorter ITR sequences below 100 bp continued with another set of reverse A- and D-sequences. The dominant recombination form maintained all ITR sections (Figure 3.6 D). They (and longer ITR sequences greater than 180 bp) continued with either the B- or C-loops, depending on whether the Flip or Flop configuration of the ITR was subject to the recombination. Interestingly, ITR sequences with a length of above 180 bp appeared to have another set of forward and reverse A-sequences in between another set of B- and C-loops.

As anticipated, ITR recombinations can be quite complex, yet they are not completely random, with some forms being more abundantly observed than others. The B- and C-loop sequences often fell victim to recombination and deletion events, probably owing to their strong secondary structure.

3.1.6 Accumulation of H3K27me3 in comparison to H3k27ac in AAV9-delivered transgenes

Next to the methylation by nanopore sequencing, I also analysed the accumulation of different histone modifications on the AAV transgene. Livers from three mice of every condition were subjected to CUT&Tag sequencing targeting the histone modification H3K27me3 and H3K27ac. Sequencing yielded 14 to 53 million reads per sample, which resulted in acceptable alignment rates and coverages (Table 3.2). These modifications were chosen in particular, as they have already been shown to differ when comparing transgenes delivered by different capsids in cell culture [177]. Briefly, H3K27me3 is associated with the down-regulation of nearby genes and is enriched in heterochromatic regions of

the genome. H3K27ac is regarded as an activating modification, often found around the transcription start sites of actively transcribed genes [170]. CUT&Tag is a library generation method, in which an antibody against a target protein (histone modification in my case) is used to direct a protein A-fused transposase into the vicinity of the target [181] (also see section 1.1.7). The transposase has been preloaded with sequencing adapters that are inserted in proximity to the target. The resulting fragments are PCR-amplified and paired-end sequenced with short-read sequencing (Illumina). The alignments produced by these libraries to the reference can give insights into the location of a target protein on the reference genome.

sample	reads	alignment %	coverage	Run
K27ac_2-6-1	2.90E+07	98.46%	1.2693X	Run 1
K27ac_2-6-2	2.64E+07	98.15%	1.1522X	
K27ac_9-6-1	3.34E+07	98.06%	1.4551X	
K27ac_9-6-2	3.39E+07	97.94%	1.4789X	
K27ac_2-12-1	1.80E+07	95.08%	0.7507X	Run 2
K27ac_2-12-2	2.31E+07	96.64%	0.9892X	
K27ac_2-12-3	4.59E+07	96.92%	1.9705X	
K27ac_2-2-1	5.31E+07	97.08%	2.2805X	
K27ac_2-2-2	3.66E+07	97.66%	1.5854X	
K27ac_2-2-3	4.55E+07	96.97%	1.9507X	
K27ac_2-6-3	4.76E+07	97.44%	2.0575X	
K27ac_9-12-1	4.84E+07	96.27%	2.06X	
K27ac_9-12-2	2.16E+07	95.64%	0.909X	
K27ac_9-12-3	3.69E+07	97.43%	1.5936X	
K27ac_9-2-1	4.26E+07	96.70%	1.8171X	
K27ac_9-2-2	4.47E+07	97.87%	1.9424X	
K27ac_9-2-3	4.26E+07	97.54%	1.8464X	
K27ac_9-6-3	2.63E+07	93.81%	1.095X	
K27me3_2-12-1	2.32E+07	98.04%	1.0097X	Run 3
K27me3_2-12-2	2.89E+07	98.59%	1.2673X	
K27me3_2-12-3	3.15E+07	98.49%	1.384X	
K27me3_2-2-1	2.60E+07	98.28%	1.1353X	
K27me3_2-2-2	2.36E+07	98.48%	1.0342X	
K27me3_2-2-3	2.94E+07	98.41%	1.2905X	
K27me3_2-6-1	1.38E+07	98.58%	0.6059X	
K27me3_2-6-2	1.23E+07	98.61%	0.5423X	
K27me3_9-12-1	2.73E+07	97.85%	1.1926X	
K27me3_9-12-2	2.42E+07	98.75%	1.0634X	
K27me3_9-12-3	2.45E+07	98.66%	1.0759X	
K27me3_9-2-1	2.24E+07	98.44%	0.9783X	
K27me3_9-2-2	2.84E+07	98.35%	1.244X	
K27me3_9-2-3	3.17E+07	98.51%	1.391X	
K27me3_9-6-1	5.20E+07	98.06%	2.2691X	
K27me3_9-6-2	2.81E+07	98.23%	1.231X	

Table 3.2. Number of paired CUT&Tag sequencing information from mouse liver as obtained from three runs. Samples are named after the tested modification and the number of the mouse. The mouse number naming follows this convention: 'capsidSerotype-weeks-animalNo.' Alignment rate and coverage are gathered from bowtie2 alignments against the mouse genome, together with the transgene sequence.

CUT&Tag sequencing produces robust data on the transgene cassettes

To demonstrate that the CUT&Tag experiment functions as expected in my hands, I studied different aspects of the aligned fragments. For simplicity's sake, in this section, I only show data from mice two weeks post-transduction, but samples from other time points showed comparable results. Reads were aligned to a genome consisting of the GRCm39 mouse assembly fused with the CMV-transgene-cassette as an additional entry. The associated alignment percentage and resulting coverage on the mouse reference genome can be found in Table 3.2. The fragment length histogram on the mouse genome displayed a classical sawtooth pattern in all samples (Figure 3.7 A). This pattern likely originates from the fact that the fused transposase is not completely random in its insertion site. As the transposase is transfixed to the protein of interest, it preferentially inserts its adapter sequences into the transposase-facing side of the DNA double-helix rather than the transposase-far side of the molecule, creating a 10-bp-sawtooth pattern of fragment sizes. The experiments against H3K27me3 and H3K27ac both produced a major peak at a size of about 180-200 bp (Figure 3.7 A). One nucleosome encompasses about 147 bp of DNA, making the major peak close to a mononucleosomal size. The peak structure for the H3K27ac experiments differed, as it portrayed a relatively large sub-nucleosomal peak of about 80 to 90 bp. Expectedly, the alignments of the H3K27ac and H3K27me3 samples showed widely different profiles on the mouse genome, exemplified by the genes Smad3 and Prdm12 (Figure 3.7 B). Smad3 is associated with the TGFβ/SMAD-pathway and shown to be hyper-acetylated in other tissues [297]. Prdm12 is a gene pivotal for neurogenesis and thus repressed in liver cell types. Smad3 showed a high accumulation of the activating H3K27ac mark in its promoter/enhancer region, but comparatively very few H3K27me3 marks. Vice versa, Prdm12 showed a sizeable accumulation of H3K27me3 in its promoter/enhancer, but also in its coding region, whereas there were few to no reads of the H3K27ac modification. This result reaffirmed that the CUT&Tag experiment worked well in my hands and could produce the expected results.

Next, I analysed the fragment alignments from the different experiments on the transgene reference (Figure 3.7 C). There were more reads aligned to the AAV9-delivered transgenes compared to the AAV2-delivered ones for both tested modifications. The alignment

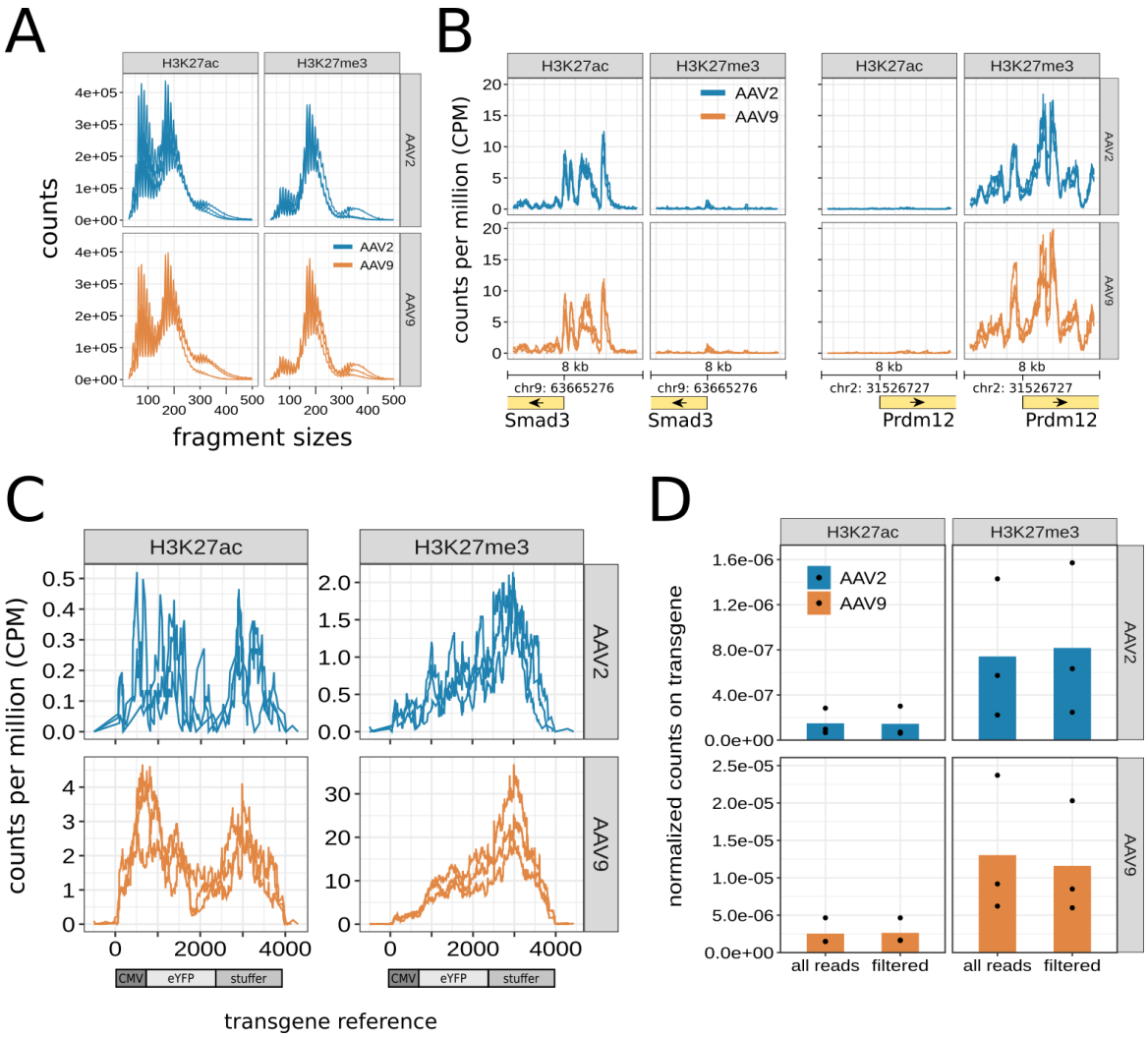

Figure 3.7. Robustness of CUT&Tag against H3K27ac and H3K27me3 exemplified in mice two weeks post AAV vector injection.
A) Size histogram of fragments aligned to the mouse genome showing the typical sawtooth pattern for CUT&Tag obtained libraries. Every line represents one mouse from three total mice per condition. B) Coverage distribution in an 8 kb window along genes Smad3 and Prdm12 for both post-translational histone modifications. Read coverage was normalised to counts per million aligned reads (CPM). Every line represents one mouse from three total mice per condition. C) Coverage distribution of reads aligned to the transgene. Read coverage was normalised to counts per million aligned reads (CPM). Note that all displayed samples were scaled independently. Every line represents one mouse from three total mice per condition. D) Normalised number of reads aligned to the transgene for the different post-translational histone modifications. Reads were either not filtered or filtered for PCR duplicates and nucleosomal-sized fragments (120-400 bp).

distribution also differed slightly. Fragments from the H3K27ac experiment covered the transgene more or less uniformly, while the fragments from the H3K27me3 experiment were mostly found on the stuffer DNA. Most importantly, all experiments produced fragments that could be uniquely aligned to the transgenic reference. In order to quantify the level of histone modifications on the transgenes, I normalised the number of reads on the transgene reference by the number of reads aligned to the mouse genome and by the present vg/dg assessed by ddPCR (see section 3.1.3). Additionally, I also applied stringency filters to the alignment to study whether (especially on the transgene reference) abnormally short reads or PCR duplicates could have had an impact on the attained number. The filtered alignments thusly encompassed only nucleosomally sized fragments (140-200 bp) and fragments that were not flagged as PCR duplicates. Both of these are factors that can affect the number of aligned fragments. However, the filtering approach did not greatly alter the normalised number of reads on the transgene in any of the conditions (Figure 3.7 D). This indicated that the obtained CUT&Tag data and the derived quantification of the transgenic reads were highly robust and thus meaningful conclusions could be drawn from them.

AAV9-delivered transgenes lose H3K27ac over time

Having shown that the CUT&Tag data gives robust results, I continued with comparing the accumulation of histone modifications on the transgenes at different time points. As mentioned briefly already, the number of reads on the transgene was highly dependent on the tested modification and the used capsid (Figure 3.8 A). The samples from AAV9-transduced mice produce more alignments to the transgene than the samples from AAV2-transduced mice. Likewise, the experiment against H3K27me3 yielded a higher number of alignments than the H3K27ac one, regardless of the used capsid. Interestingly, the time point also influenced the amount of observed reads on the transgenes, as there is an opposing trend visible for AAV2- and AAV9-transduced samples. H3K27ac on the AAV9-delivered transgene portrayed a decreasing trend, whereas the ones on the AAV2-delivered transgene had more reads at later time points. This is a pattern that was also observed in the transcription, where AAV9 showed a decrease over time and AAV2 a slight increase. To put the histone modification number in relation to the expression, I performed a correlation analysis of those numbers in every mouse sample, which showed mainly weak and non-significant correlations (Figure 3.8 B).

Since vg/dg appears to moderately correlate with the amount of histone modifications, it might constitute a confounding factor. Additionally, the vg/dg measurements were taken with a different method and thus could introduce unwanted noise. This makes the interpretation of the computed value difficult. To arrive at useful conclusions, I therefore

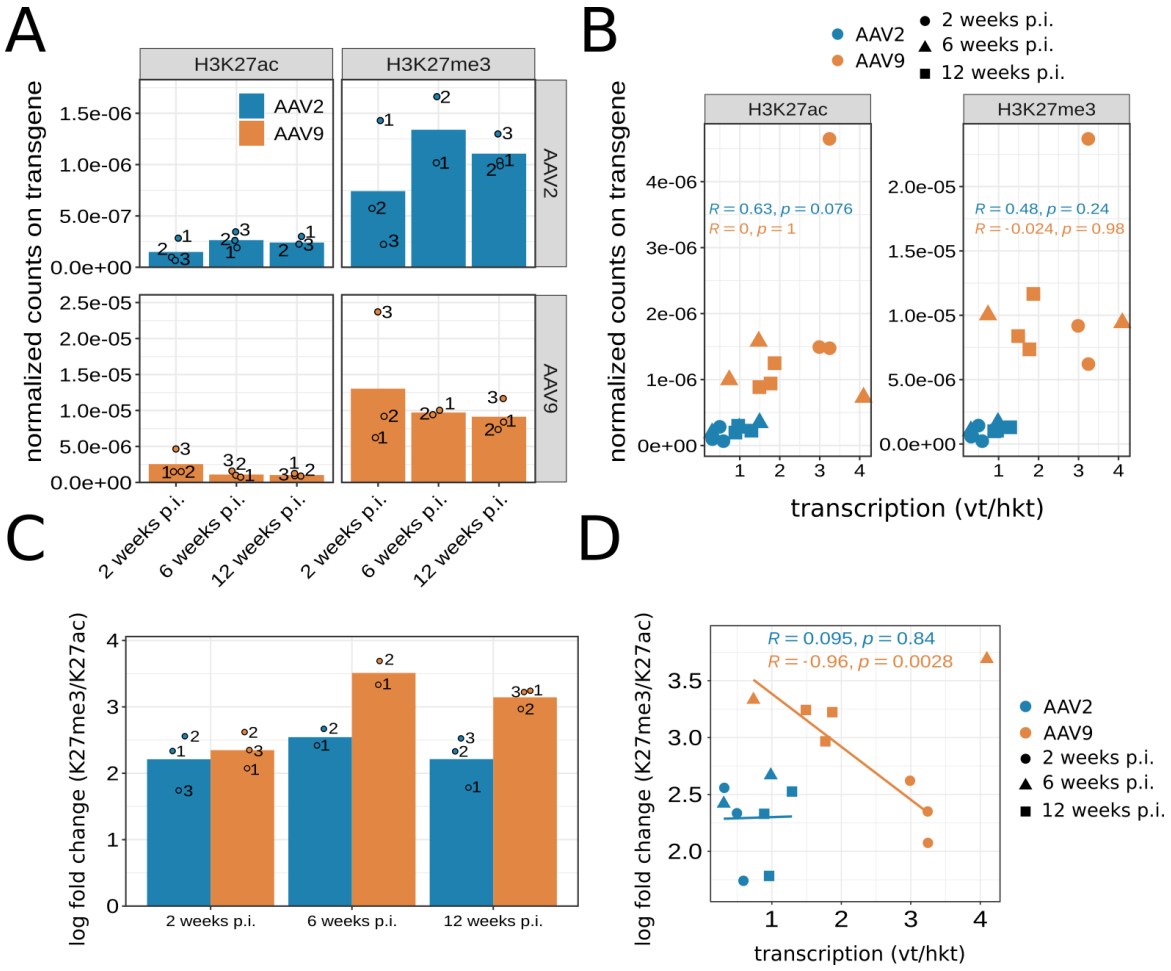

Figure 3.8. Quantification of histone modifications on the AAV delivered transgene. A) Normalised read counts on the transgene for all samples. Every point and associated number represent one mouse sample. The bar represents the mean of the replicates of the same condition. B) Dot plot of the normalised read counts against transcription values obtained from the same samples. A Spearman correlation was conducted, and the p-value and coefficient are indicated. The dot shape indicates the time point, and the colour indicates the injected capsid. C) Fold-change of H3K27me3 over H3K27ac in all samples. Every dot and associated number represent one mouse sample. The bar represents the mean of all replicates in the same condition. D) Dot plot of the histone modification ratio with transcription values obtained from the same samples. A Spearman correlation was conducted, and the p-value and coefficient are indicated. One point from the AAV9 6-week samples (top right) was omitted for the correlation analysis. The dot shape indicates the time point, and the colour indicates the injected capsid.

computed the ratio of H3K27me3 to H3K27ac at all time points. By the nature of this calculation, the vg/dg factor is eliminated. Comparing multiple histone modifications in this manner gives an estimate of the epigenetic changes on the transgene of an activating mark (H3K27ac) in relation to an inactivating mark (H3K27me3). For every mouse sample, I divided the normalised count for H3K27me3 by the normalised count for H3K27ac on the transgenes to obtain a ratio of H3K27me3 over H3K27ac (Figure 3.8 C). At all time points, there were at least four times more H3K27me3-derived reads on the transgenes than there were H3K27ac-derived reads. The ratio of the AAV2-delivered transgenes did not alter much with time, but AAV9 displayed an increased ratio at later time points. The modification ratio of the AAV2-delivered transgenes did not correlate with transcription, but it did for the AAV9-delivered transgene. When removing one outlier, a correlation analysis resulted in a significant p-value for the correlation of the modification ratio with transcription for AAV9-delivered samples (Figure 3.8 D). The modification ratio of the AAV2-delivered samples, however, did not correlate with transcription.

In conclusion, these results indicate that both tested histone modifications deposited on the transgene are subject to change over time. However, in comparing the levels of H3K27ac to H3K27me3, I observed that the AAV9-delivered transgenes become depleted of the activating histone modification H3K27ac sometime after the 2-week time point, which correlates with the decrease in transcription observed in the same transgene. This does, however, not explain why the advantageous modification ratio of AAV2-delivered transgenes does not translate into an equally high transcription.

3.1.7 Episomal state of the delivered AAV genomes

So far in the long-term mouse study, I have continuously compared AAV2- against AAV9-transduced mice. However, vectors derived from these two different capsids differ substantially in their biology. Most prominently, they differ in their biodistribution, but also in their post-uptake processing, as the AAV2 capsid has been shown to be much slower in the formation of episomes compared to AAV8, a close relative of AAV9 [111]. The slow uncoating of AAV2 also negatively impacts the onset of transcription.

To assess the differences in episome formation between AAV2 and AAV9 at different time points, I used a T5 exonuclease assay [298]. Briefly, circular episomes are resistant to exonuclease digestion. Determining the level of exonuclease sensitivity indicates whether a transgene has successfully formed a circular episome. A circular episome can only form once the virus has successfully uncoated and the second strand has been synthesised. Digested and undigested samples were compared using qPCR, which resulted in the

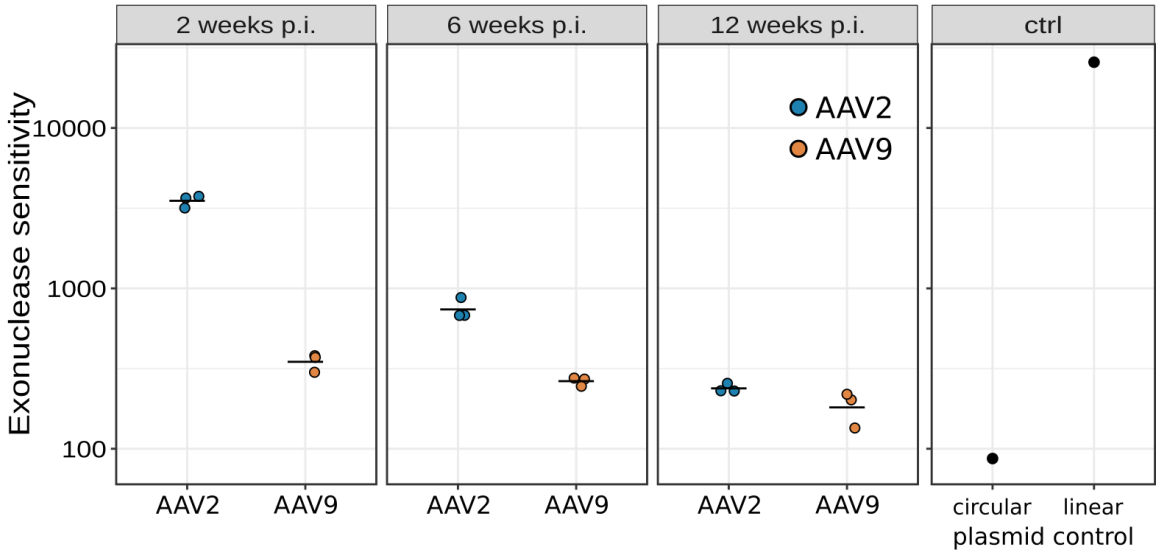

Figure 3.9. T5 exonuclease sensitivity of AAV delivered DNA.
Every dot is one DNA sample obtained from a transduced mouse liver. The horizontal black line represents the mean of the tested samples. The same mice were analysed as in the CUT&Tag sequencing experiments. Controls consist of plasmid DNA either in its circular form or linearised by a single-cutter restriction enzyme.

exonuclease sensitivity measurement (see section 2.8.8). The transgenes delivered by both capsids became more resistant to exonuclease over time, indicating that they either continued to form circular episomes or their linear forms were being degraded over time (Figure 3.9). At all time points, AAV2-delivered transgenes showed a higher sensitivity towards T5 exonuclease compared to AAV9-delivered ones. However, the difference between them became smaller at later time points. The AAV2-delivered transgene appeared to undergo a much more drastic change towards exonuclease-resistant episomes. Keeping in mind that DNA presence did not widely differ between the AAV2- and AAV9-delivered transgenes (Figure 3.3 B), this result suggests that the AAV2-delivered transgene takes much longer to form a circular episome compared to the AAV9-delivered transgene.

3.1.8 Differential analysis of the effect of AAV2 and AAV9 transduction on histone modification peaks on the mouse genome

The analyses in the previous sections were mainly focused on the difference in the transgenes at different time points, but I deemed it also interesting to see what influence the AAV vector had on the histone modifications on the host genome. To achieve this, I compared the peaks of the AAV9- with the AAV2-injected samples for both tested histone modifications. Peaks of histone modifications were found using SEACR [248], and a differential analysis was

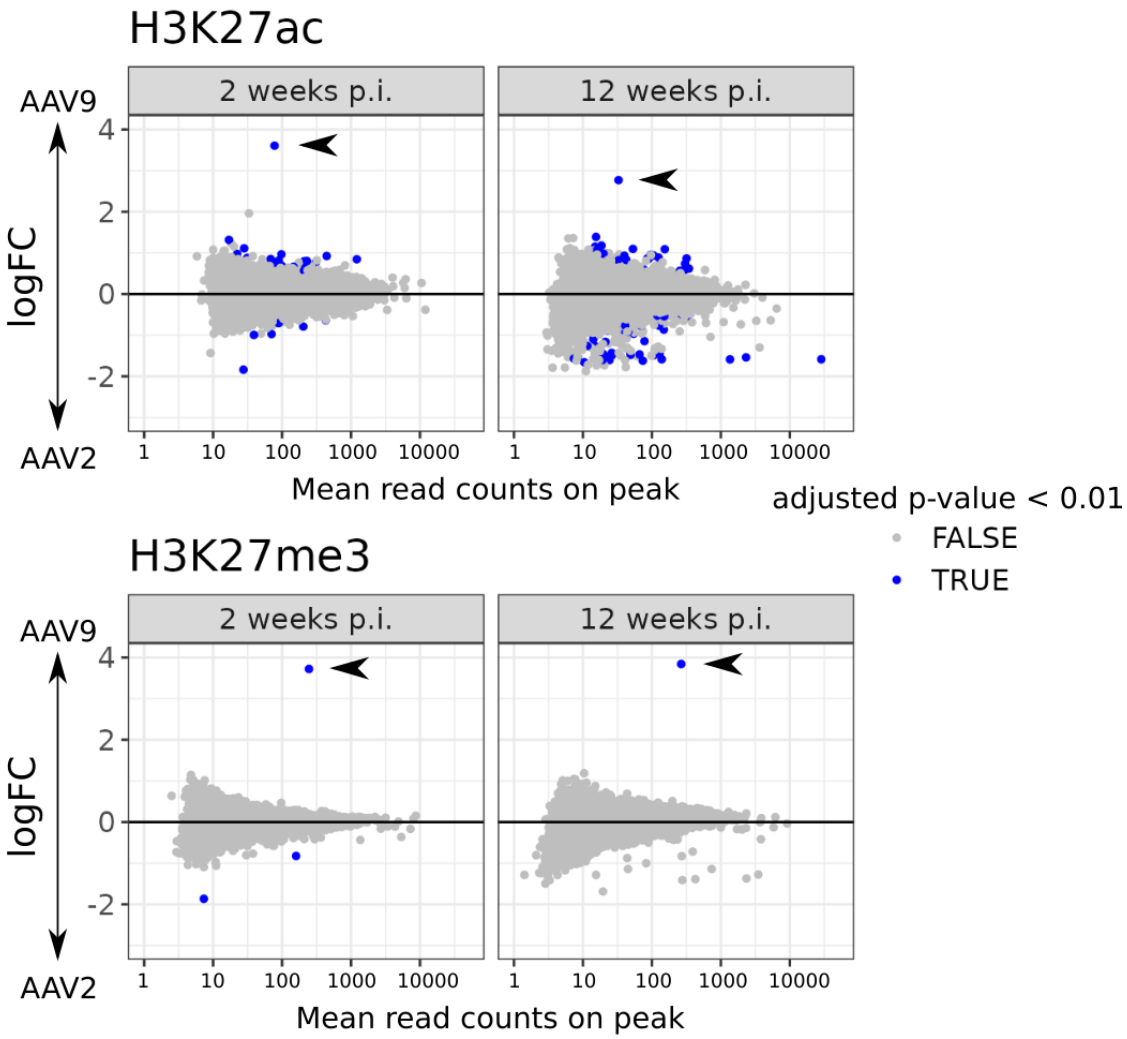

Figure 3.10. Differential analysis of peak sizes of H3K27ac and H3K27me3 comparing AAV2- and AAV9-transduced samples at 2 weeks and 12 weeks.
Every dot represents one peak on the mouse genome plus transgene reference. The Y-axis represents the log fold-change of AAV9 over AAV2. The X-axis represents the mean number of counts in the peak over all replicates. The arrows indicate the peak on the transgene. The Wald test was performed in the DEseq2 framework, and Benjamini-Hochberg was used to correct for multiple testing. All blue dots have an adjusted p-value below 0.01.

conducted using DESeq [249]. The majority of histone modification peaks remained similar between the AAV2 and AAV9 injected mice (Figure 3.10). For H3K27me3, there were only three or one significantly different peaks for the 2- and 12-weeks samples, respectively, one of which is always the transgene as indicated by an arrow. There were more significantly different genes for the H3K27ac modification, for which a gene ontology enrichment was performed additionally. Significant peaks were associated with the closest gene reference, which was possible for 21 peaks in the 2 weeks p.i. sample and 65 peaks in the 12 weeks p.i. sample. The set of genes in the 2 weeks p.i. samples did not have any significantly enriched terms (neither biological process nor molecular function). The set of genes in the 12 weeks p.i. sample had 'cell motility' as the only significantly enriched biological process (no significantly enriched molecular process). In conclusion, this analysis could show that the AAV capsid has a relatively small effect on the epigenetic landscape of the mouse liver, at least regarding the tested histone modifications and extreme time points.

3.2 The genomic di-nucleotide pattern of Adeno-associated viruses

The DNA molecule has a defined shape of a winding double helix comprising the phosphate backbone, base pairings, and base stacks. The shape of the DNA molecule, however, needs to undergo permutations for it to be able to interact with other molecules (mainly proteins) to allow the processes that make DNA the information carrier of an organism. The canonical DNA bases A, C, G, and T are distinct from each other. The primary difference is that they are either pyrimidine- (Y) or purine- (R) based. A base pair always consists of one of each pairing with the other. Given the different sizes of the pyrimidine and purine bases, base stacks of either one base type can cause slight permutations in the double helix structure of the DNA molecule. A bend can be introduced into the helix if the same base stack (and thus the same permutation) exists in regularly spaced periodic distances (see section 1.2.2). This bend can facilitate the association of the DNA molecule with interacting proteins. This is best demonstrated in the 10 bp dinucleotide periodicity of eukaryotic DNA that facilitates histone association [216, 230]. A similar pattern has been found on the viral DNA of adenoviruses, which is implicated in the binding of adenoviral packaging proteins [241].

Here, I am interested in possible dinucleotide patterns in the genomes of AAV and the implications they might have for viral packaging, replication, and ultimately recombinant vector design.

3.2.1 The YY/RR dinucleotide pattern in the 13 primate AAV serotypes

I began my analysis with the wild-type genomes of the 13 classical primate AAV serotypes, because of their prevalence in research and applications. I mainly focused on the YY and RR dinucleotide combinations, which consist of all pyrimidine (YY: TT, CC, CT, TC) or all purine (RR: AA, GG, GA, AG) dinucleotides, respectively (Figure 3.11 A; red and blue marked dinucleotides). A special feature of the YY dinucleotide is that its reverse complement is always an RR dinucleotide and vice versa. The analysis of dinucleotide periodicity first entails determining all positions of a specific dinucleotide. From the dinucleotide coordinates, a distance matrix was calculated, which was then processed into a histogram (exemplified in Figure 3.11 A; middle panel). The counts from the resulting histogram were processed in two steps to obtain the normalized counts: (i) a smoothing window of 3 bp was applied to minimise the omnipresent 3 bp DNA periodicity, and (ii) an averaged background window of 7-bp was subtracted from the smoothed histogram for normalisation (Figure 3.11 A; rightmost panel).

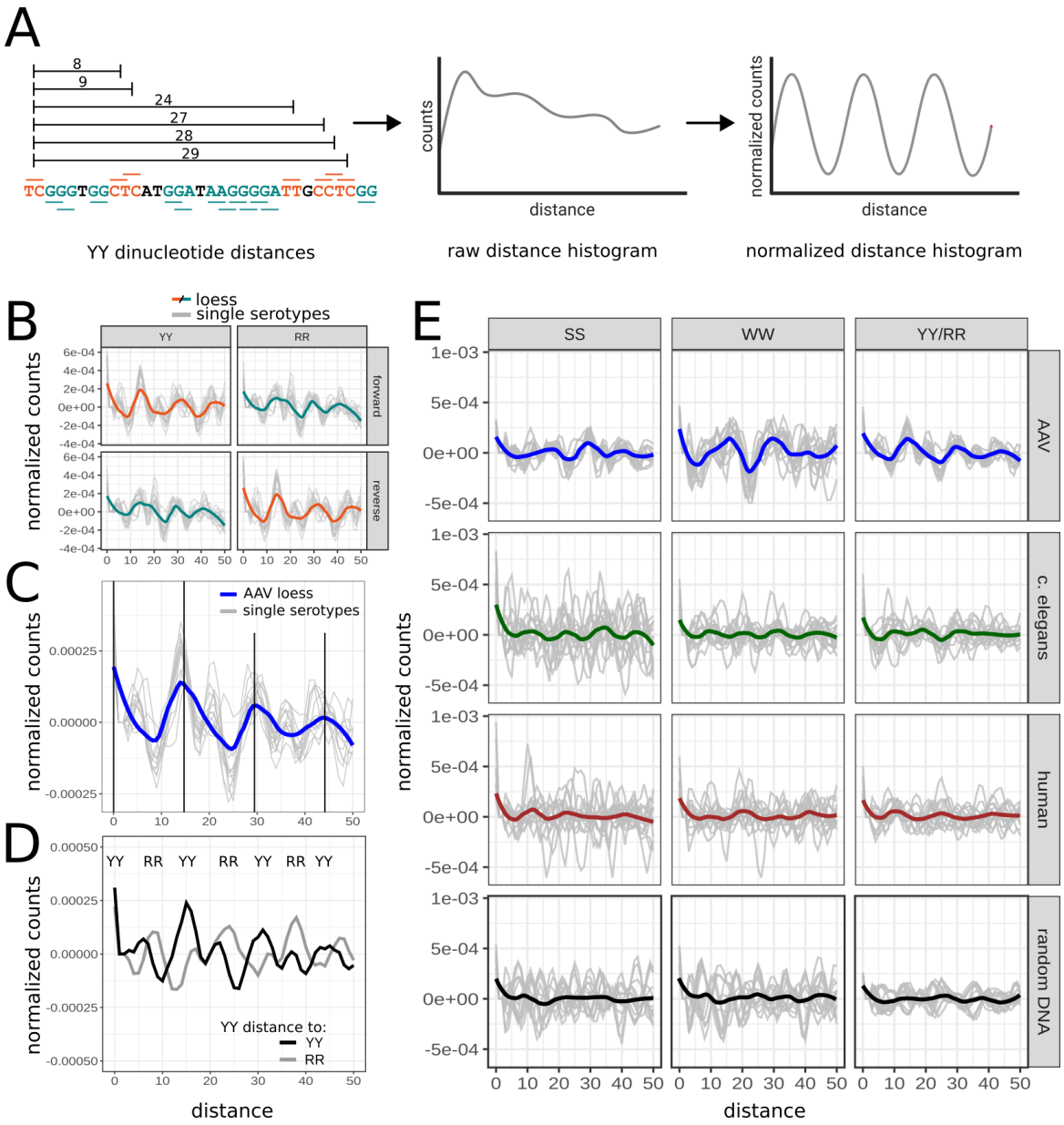

Figure 3.11. The dinucleotide pattern on the 13 primate AAV serotypes.
A) Exemplary representation of the YY dinucleotide periodicity calculation. The distances of every dinucleotide to every other occurrence are measured (YY dinucleotides shown in red). The distances from the first YY to all others are indicated. RR dinucleotides are YY dinucleotides on the opposing strands and are shown here in blue. A histogram is calculated from the assessed distances. The histogram is finally smoothed and normalised to obtain the normalised counts. Assets from biorender [299]. B) Normalised histograms of YY and RR dinucleotides are shown separately for the forward and reverse strands of the 13 primate AAV serotypes. Each grey line represents a single AAV serotype. The coloured line represents the loess of all individual counts. C) Combined distance histogram of YY and RR on both strands of the 13 primate AAV serotypes with equally spaced vertical lines with a period of 14.8 bp. Each grey line represents a single AAV serotype. The blue line represents the loess of all individual counts. D) Distance histogram of the YY dinucleotides to other YY and RR dinucleotides showing how a YY and RR pattern can exist on the same strand by being shifted regarding each other. E) Histograms for different dinucleotides in sequence sets from AAV, model organisms, and randomized DNA. 13 random 5-kb-long sequences were chosen from within the model organism genomes. Each grey line represents a single sequence and the coloured line represents the loess of all sequences.

Initially, I analysed multiple dinucleotides on every individual sequence separately for both strands. The most significant pattern on both strands was observed for the YY and RR dinucleotide combinations (Figure 3.11 B). Both of these showed a periodicity of the dinucleotides repeating every 15 bp when averaging (loess) over all 13 AAV serotypes. Smaller peaks, that were not present in all sequences, present themselves in between the 15 bp peaks with a periodicity of 7-8 bp (also smoothed out by the loess). Other dinucleotides and combinations thereof showed either non-periodic patterns or patterns that were only apparent on one strand at a time (not shown). Since the YY/RR pattern appears on both strands in a similar manner and those dinucleotides are always complementary to the one on the opposing strand, from here on I analysed both strands simultaneously (Figure 3.11 C). Given that both the YY and RR patterns exist on the same strand simultaneously, the two patterns must be somewhat shifted with regard to each other. To analyse this, I calculated the cross-correlation of the YY dinucleotides not only with themselves but also with RR dinucleotides. In other words, from the perspective of each YY I recorded the distances to all other YY but also all RR dinucleotides. I did this for the 13 aforementioned serotypes in a pooled approach and can clearly observe that the RR pattern seems to be shifted about 8 bp in regard to the YY pattern, which allows both 15 bp patterns to co-exist on the same strand (Figure 3.11 D).

Attempts at localising the pattern on the AAV genomes did not yield any particular region with increased periodicity, which is likely due to the anyway short size of the AAV genome (data not shown). Despite their small individual size, the YY/RR pattern observable on the 13 primate wild-type AAV genomes (4.1 - 4.8 kb in size) is remarkably pronounced. To obtain a comparative view, I also performed a similar analysis for the YY/RR, SS, and WW dinucleotides in 13 random stretches (5 kb in size) from coding sequences of *C. elegans*, human, and from randomised DNA (Figure 3.11 E). Only in *C. elegans* could the 10 bp eukaryotic pattern be clearly discerned, notably though in dinucleotides SS and WW as opposed to YY/RR. In human DNA, in the SS and YY/RR dinucleotide histograms, only the first peak at around 10 bp can be seen, after which the pattern begins to fade. The fact that the well-described eukaryotic pattern is not always visible in similarly limited stretches of their genomes demonstrates the clarity of the pattern in the analysed AAV serotypes. The existence of such a clear, unique, and emphasised pattern strongly suggests its association with a biological function.

3.2.2 The dinucleotide pattern is conserved only in the genus *Dependoparvovirus*

To gain an understanding of the prevalence of this YY/RR dinucleotide pattern, I expanded the analysis to the kingdom of *Shotokuvirae*, AAV helper viruses, satellite viruses, and control genome sequences.

The *Shotokuvirae* kingdom encompasses primarily ssDNA as well as some dsDNA viruses. Since AAV itself is a ssDNA virus and the measured genomic pattern deviates from the well-described 10 bp repeat seen in eukaryotes, I set out to analyse other ssDNA viruses in the *Shotokuvirae* to ascertain whether it is a fundamental characteristic of their genome organisation. AAV is dependent on many factors provided by helper viruses, from which I also included genomes in the analysis. The YY/RR dinucleotide pattern might be a common feature among the AAV and helper virus genomes, as it might facilitate the association of the DNA with a specific helper factor. I also deemed it possible that this sequence feature might be present in other helper-dependent viruses, which are often called satellite viruses in the literature. Therefore, I also included viral sequences associated with the term 'satellite virus' in the analysis. More on the specific included sequences can be found in section 2.7.4.

Because of the small size of viral genomes, I opted to analyse all sequences from the same genus as one combined sequence set. I only took genera with at least $1e5$ bp of combined sequence length into consideration, which resulted in more than 25000 unique sequences from 35 genera within the kingdom *Shotokuvirae*, as well as satellite viruses and AAV helper viruses. To ensure adequate comparison, I adapted a previously published quantification strategy [230] that utilises the fit of a damped sine function to model the periodicity of the distance histogram (Figure 3.12 A). I provided two examples showing both a histogram and a fit sine wave for the genus *Dependoparvovirus* and the genome of *C. elegans* for the YY dinucleotide. The analysis of *C. elegans* produces a better fit that is reflected in the Goodness of Fit (GdF), amplitude and the resulting periodicity values ($periodicity = log2(\frac{GdF}{amplitude})^{-1}$). The fit is most apparent in the first 30-50 bp distance, as then the pattern begins to fade and disappear into the background.

For every tested genus or sequence set, I then plotted the period of the best fit against the periodicity (Figure 3.12 B). The majority of measured periods for all tested dinucleotides lay between 10 and 11 bp, which is especially visible at the top of the periodicity measurements (Figure 3.12 B; right panel). The genus *Dependoparvovirus* (containing all publicly available AAV sequences) ranked among the top hits for its YY periodicity (Figure 3.12 C). Other highly periodic sequence sets included Papilloma- and Polyomaviruses (details in Table 3.3). Remarkably, the 15 bp period stood out and was not detected in other genera

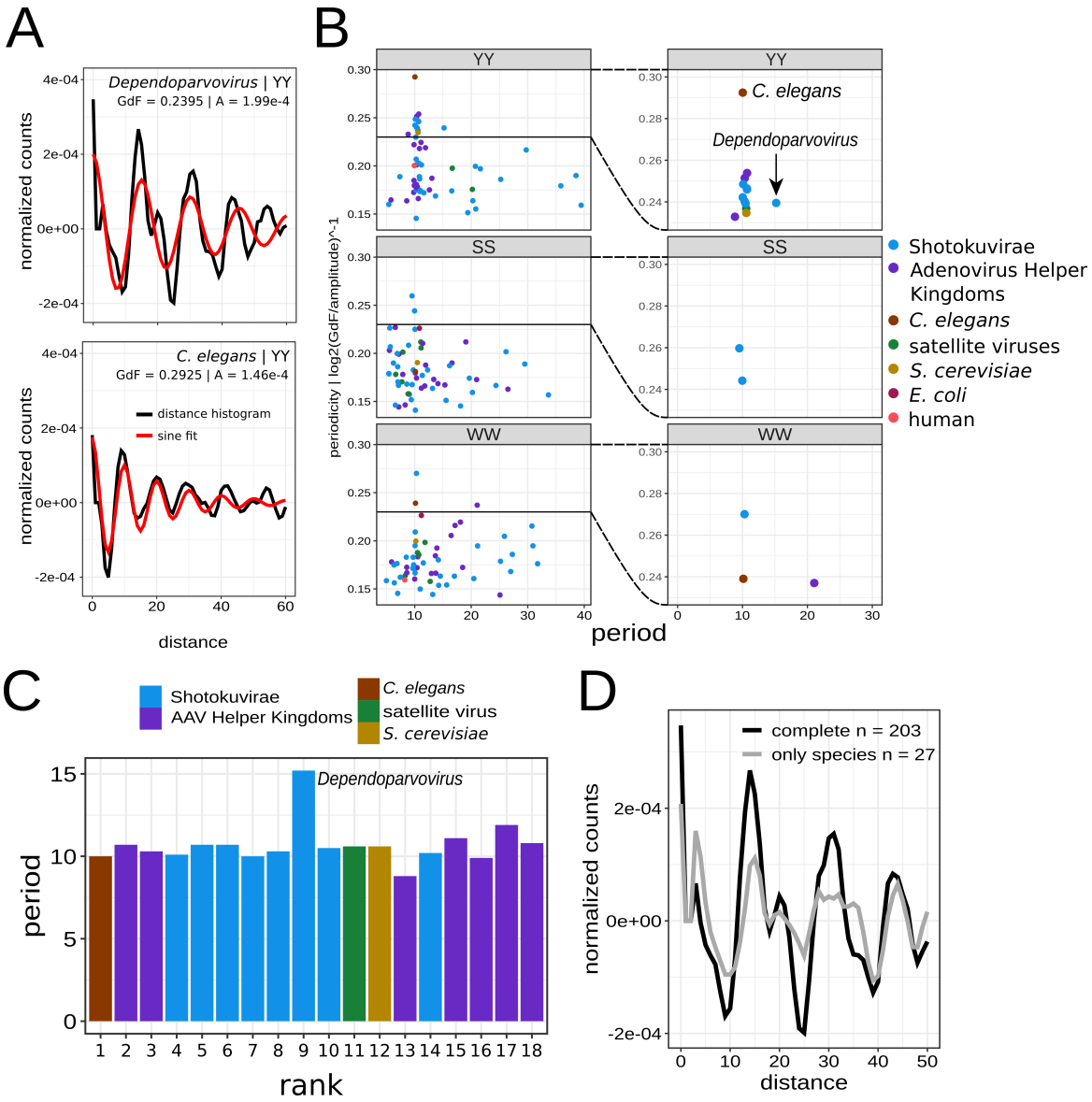

Figure 3.12. Periodicity quantification in different genera of *Shotokuvirae* and other reference genomes.

A) Normalised histogram of the YY/RR dinucleotide distances (black) and associated sine fit (red) for the genus *Dependoparvovirus* and the entire genome of *C. elegans*. The Goodness of Fit (GdF) and fit amplitude are indicated. B) Periodicity strength against the fit period for multiple sequence sets for dinucleotides YY, SS, and WW. Each dot is a sequence set stemming from one genus of the virus kingdom *Shotokuvirae*, AAV helper viruses, or satellite viruses. The dots for the reference genomes *E. coli*, *C. elegans*, yeast, and human were derived from all coding sequences of the respective model organisms. The left panel represents the total data, with the horizontal line being the cutoff for the top hits defined at 0.23. The right panel represents the top hits with *C. elegans* and *Dependoparvovirus* labelled specifically. C) YY periodicity top hits shown as a bar graph with the rank on the x-axis and the period on the y-axis. The genus *Dependoparvovirus* is indicated. D) Normalised histogram of the YY/RR dinucleotide distances of the complete *Dependoparvovirus* set (black) and of a set containing the 27 species of *Dependoparvovirus* (grey).

Group	Genus	Period	GdF	Amplitude	Periodicity
C. elegans	*C. elegans*	10	1.58E-03	1.46E-04	2.92E-01
Helper Kingdom	*Bossavirus*	10.7	1.91E-03	1.06E-04	2.54E-01
Helper Kingdom	*Iltovirus*	10.3	1.15E-03	7.91E-05	2.51E-01
Shotokuvirae	*Deltapolyomavirus*	10.1	1.92E-03	1.18E-04	2.49E-01
Shotokuvirae	*Betapapillomavirus*	10.7	2.19E-03	1.34E-04	2.46E-01
Shotokuvirae	*Gammapapillomavirus*	10.7	1.89E-03	1.16E-04	2.46E-01
Shotokuvirae	*Porprismacovirus*	10	2.98E-03	1.63E-04	2.42E-01
Shotokuvirae	*Gammapolyomavirus*	10.3	2.04E-03	1.11E-04	2.40E-01
Shotokuvirae	*Dependoparvovirus*	15.2	3.59E-03	1.99E-04	2.40E-01
Shotokuvirae	*Chipapillomavirus*	10.5	2.01E-03	1.10E-04	2.39E-01
Satellite virus	*Lavidaviridae*	10.6	1.49E-03	7.48E-05	2.37E-01
Yeast	S288C	10.6	1.37E-03	7.15E-05	2.35E-01
Helper Kingdom	*Simplexvirus*	8.8	1.79E-03	9.43E-05	2.33E-01
Shotokuvirae	*Alphatorquevirus*	10.2	2.40E-03	1.15E-04	2.30E-01
Helper Kingdom	*Varicellovirus*	11.1	1.36E-03	5.17E-05	2.24E-01
Helper Kingdom	*Rhadinovirus*	9.9	1.55E-03	6.08E-05	2.22E-01
Helper Kingdom	*Atadenovirus*	11.9	1.44E-03	6.31E-05	2.19E-01
Helper Kingdom	*Quwivirus*	10.8	1.43E-03	5.96E-05	2.18E-01

Table 3.3. The top YY periodic sequence sets based on the periodicity of their fits. Periodicity is calculated from the Goodness of Fit (GdF) and the Amplitude of a sine function fit to the distance histogram of YY dinucleotides, all of which are indicated in the table. The fit also has a period length, which is indicated in the Table. Same data points as in Figure 3.12 C.

of *Shotokuvirae* nor in other satellite viruses, which makes it a unique feature of the genus *Dependoparvovirus*.

There are 203 sequences in the sequence set for *Dependoparvovirus*. However, they have a strong bias towards human/primate and duck/goose (anseriform) infecting species, with 82 and 94 annotated species each. For comparison, there are only two sequences each for reptile and rodent-infecting AAVs. 17 sequences have no annotated host at all. To test whether my *Dependoparvovirus* dataset is heavily biased because of the high number of primate and bird infecting sequences, I analysed the 27 individual species of the genus *Dependoparvovirus* as listed by the ICTV [8] (see section 2.7.4). The *Dependoparvovirus* species are mostly differentiated by host species, which include bats, rodents, birds, dogs, and cats. I generated the normalised histogram in the same manner as before for the YY/RR dinucleotides and compared it to the one obtained from the analysis of the complete *Dependoparvovirus* set (Figure 3.12 D). Both histograms showed a comparable pattern in periodicity of about 15 bp. The histogram from the set only containing sequences from the 27 species had a less pronounced histogram. This can be explained by the set containing about ten times fewer sequences than the complete one. Fewer sequences leave less room for the sequence pattern to stand out from the background. These results demonstrate that the observed periodicity is a marker of the genus *Dependoparvovirus* regardless of the host species.

3.2.3 The dinucleotide pattern is being selected for from a shuffled *cap* library

The data presented in this section (3.2.3) was jointly created with Sebastian Heß and is also presented as part of his Bachelor's Thesis conducted in the Grimm lab. The experimental design and computational analysis have been performed by me.

The presence of this unique periodic pattern on AAV genomic DNA indicates that it likely provides an advantage for AAV. I designed an experiment to interrogate whether sequences with higher periodicity will be selected over multiple rounds of selection. The basis for this experiment is the DNA family shuffling of closely related AAV capsid sequences and subsequent selection steps. This method is ordinarily used to identify AAV capsid variants with antibody-evading abilities and other sought-after properties. Here, I used this procedure to partially disrupt the inherent periodic pattern on the viral DNA and analysed which sequences are enriched upon selection. The parental *cap* genes of AAV serotypes 1 through 9 as well as serotype rh10 were used to generate the initial capsid plasmid library, which was originally created by Anne-Kathrin Herrmann in the Grimm lab. This plasmid library was used to generate a virus library with every ensuing viral capsid ideally packaging the

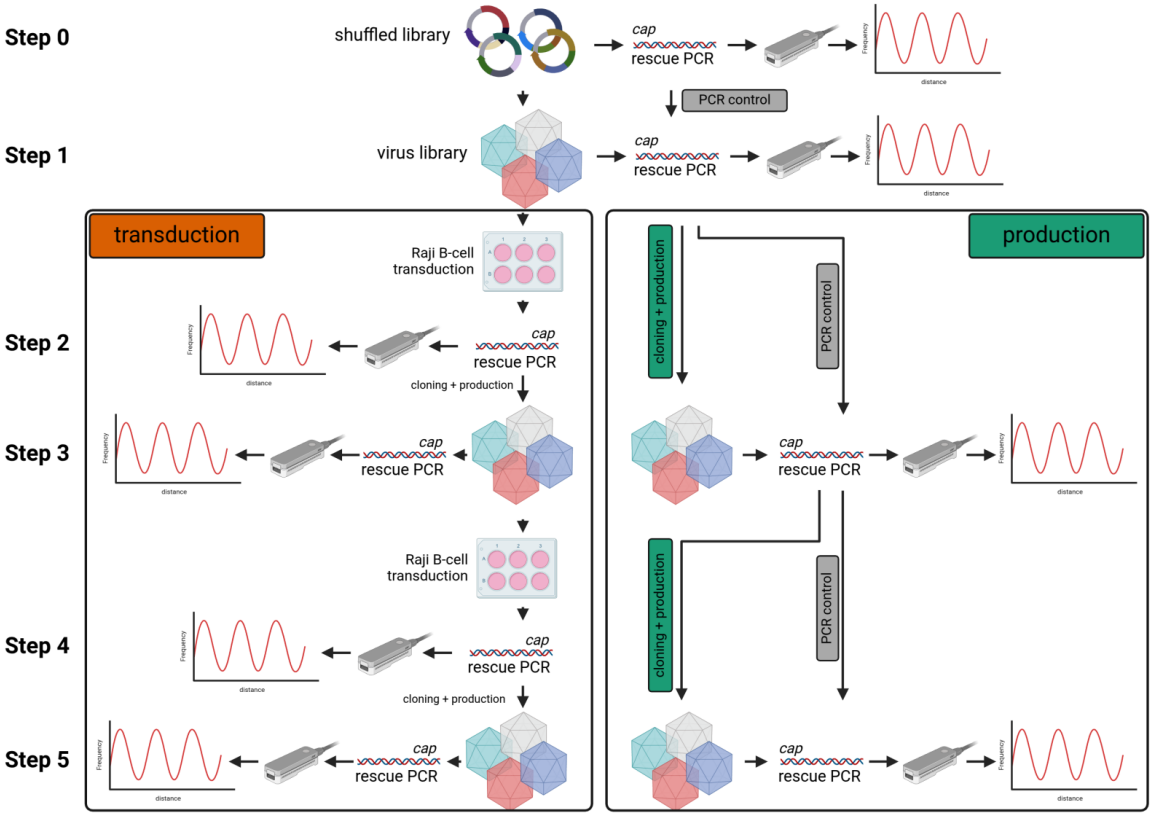

Figure 3.13. Schematic of the shuffling enrichment experiment.
Stepwise illustration of the experimental setup. A DNA-family shuffled *cap* gene library was used to produce a virus library, which was then either used to transduce Raji cells, followed by amplification from the cells (transduction) or directly amplified without transduction (production). This amplified *cap* library was then used to produce another round of virus library that was processed once as before. At every step, some of the amplified *cap* gene was retained for nanopore sequencing and subsequent quantification of periodicity. Next to the transduction and production halves, a PCR control was included, consisting of repeated PCR amplifications of the initial shuffled *cap* gene library. Figure was created in biorender [300].

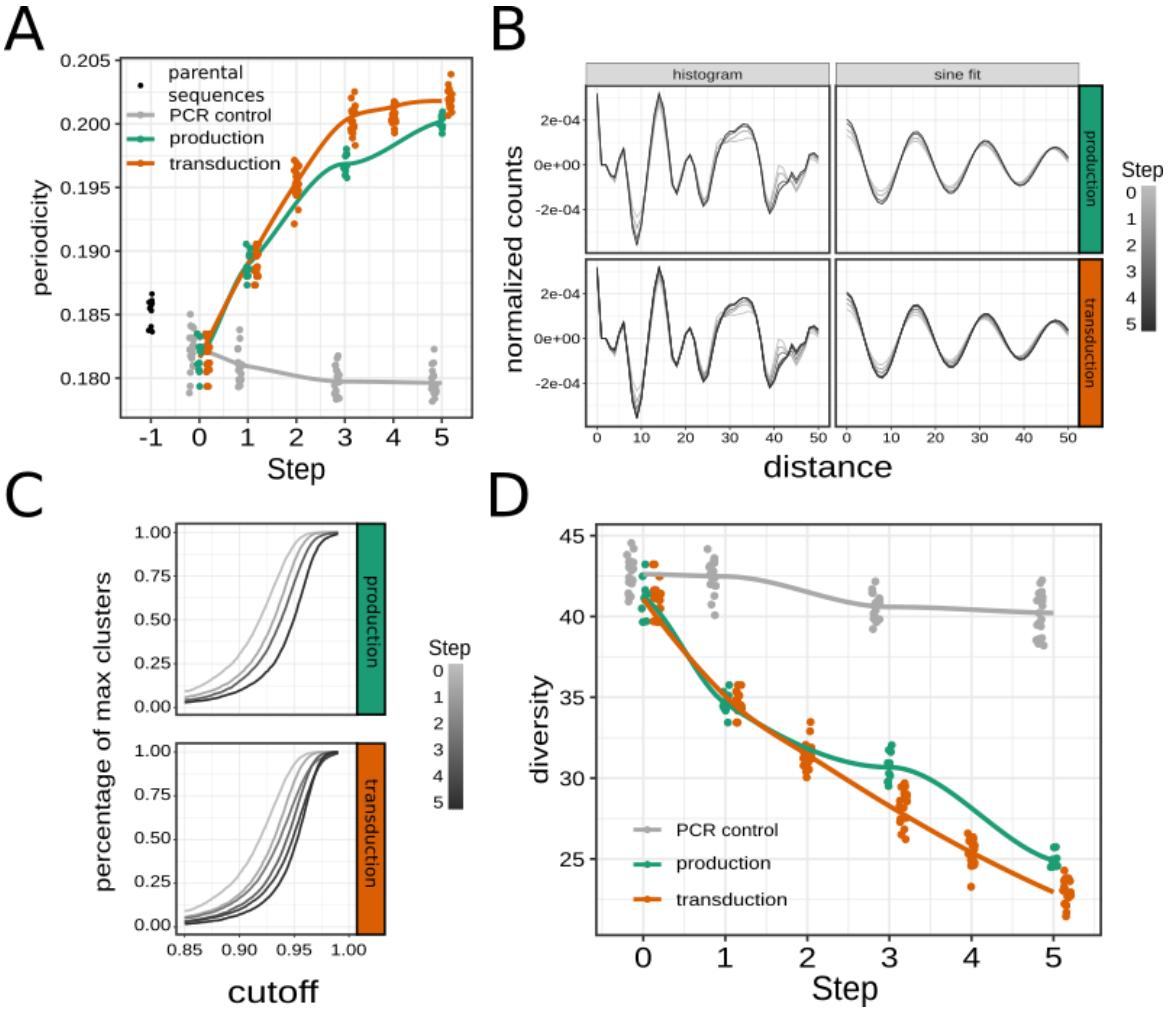

Figure 3.14. Periodicity and diversity across selection steps in the shuffling enrichment experiment.

The data for this figure was jointly created with Sebastian Hess. A) Periodicity over the selection steps. The parental sequences (AAV1-AAV9 and AAVrh10) were analysed to simulate step -1 (black dots). Each point represents a random subsample of 600 sequences from the sequenced *cap* genes and deviates slightly along the x-axis for clarity. The line represents a loess along the individual group of dots. B) YY/RR dinucleotide histograms and associated sine fit for the production and transduction half of the experiment. Experiment steps are shown in increasing grey-scale values. C) Diversity estimation using a clustering approach. The 600 subsampled sequences were clustered with an iteratively increasing sequence-similarity cutoff for cluster assignment. The resulting curve is plotted again with increasing grey-scale values for experimental steps. D) Diversity is estimated by measuring the area under the curve in C. Every dot represents one random subsample of 600 *cap* genes from the total reads, jittered slightly along the x-axis for clarity. The line represents the loess along the individual group of dots.

viral genome that encodes for it (Figure 3.13, Step 0 to Step 1). The virus library was then either directly subjected to PCR amplification of the *cap* gene or used to transduce Raji cells, followed by amplification from cell culture (Figure 3.13; production and transduction side, respectively). The amplified *cap* sequences were then used to generate another virus library. These steps were repeated once to the final number of two full rounds. Alongside the production and transduction experiments, I included a control in which only the *cap* sequences were amplified 3 times without subjecting them to the selective pressures of transduction and/or production. At every step, nanopore sequencing was performed to assess the periodicity of the amplified *cap* genes. The nanopore reads were filtered by a minimum quality of 12 and the expected *cap* gene length of about 2.2 kb, which left between 30000 and 90000 *cap* sequences per sample available for analysis. To keep computation to a reasonable time, I subsampled 600 reads from the total sequenced reads ten times, and assessed the YY/RR periodicity in the same way as in section 3.2.2.

The periodicity of parental sequences was indeed disrupted by the shuffling (Figure 3.14 A; Steps -1 to 0). Over the selection steps, the periodicity of the capsid sequences continually increased for both the transduction and production half of the experiment, which was not the case for the PCR controls. A more detailed analysis at every selection step revealed that the increase in periodicity coincided with a change in the distance histograms and associated sine fits (Figure 3.14 B). The YY/RR dinucleotide pattern of sequences from later selection rounds was characterised by a deepening of the first valley at around 9 bp and the appearance of a clearer second peak at around 30 bp, which was mirrored by the sine fit gaining a larger amplitude. This was true for the production and transduction halves of the experiment. Throughout, the periodicity of the YY/RR pattern remained unchanged at 15. To assess whether sequences were in fact being selected, I quantified the diversity of the sequences at every step in a clustering-based approach. Briefly, I clustered all 600 subsampled sequences while iterating through a sequence similarity threshold. At higher thresholds, sequences were placed into more and more clusters until every cluster was only comprised of one sequence (Figure 3.14 C). A sequence set of low diversity has more similar sequences in each cluster, for which it takes a higher threshold for it to be split apart. The approximated area under the ensuing curve gives an estimate of diversity. The diversity of both the production and transduction sequences was markedly decreasing, whereas the PCR controls retained a similar level throughout all steps (Figure 3.14 D). Capsid sequences were being selected and became less diverse, while their YY/RR periodicity was increasing concurrently, regardless of the selective pressure.

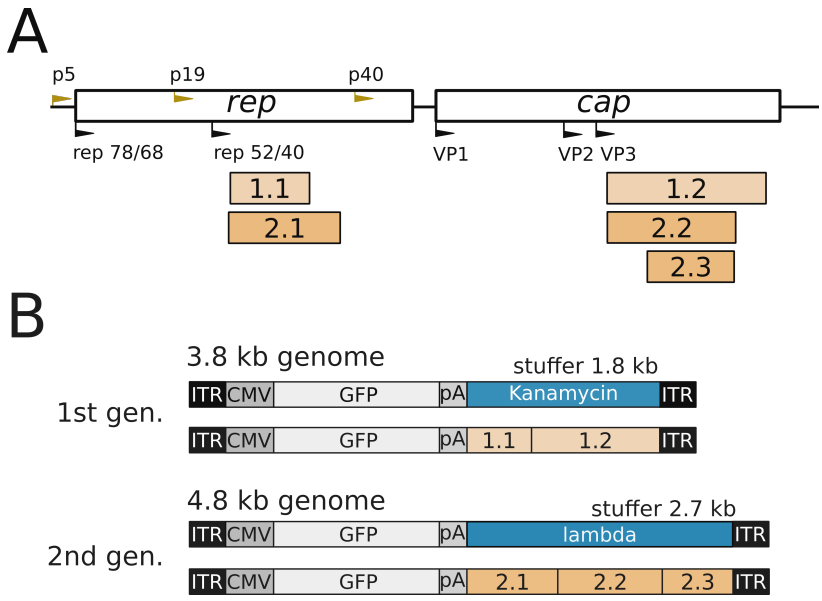

Figure 3.15. AAV2 wild-type genome and wt-fragment containing transgene cassettes. A) AAV2 wild-type genome annotations. The black arrows represent the start codons of the major protein products of AAV (excluding MAAP and AAP). The yellow arrows indicate the position of viral promoters. The bars underneath the wild-type genome indicate the fragments from which the 3.8 kb and the 4.8 kb wt-fragments cassettes were generated. B) Detailed schematic of the 3.8 kb and 4.8 kb control and wt-fragment transgene cassettes. They are distinguished by a different stuffer DNA, as all contain an AAV2 ITR sequence, as well as a CMV promoter-driven *gfp* gene.

3.2.4 Inserting YY/RR periodicity into transgene cassettes

I have shown that AAV genomes and genomes from the entire genus of *Dependoparvovirus*es possess a strong periodic repeat of the YY/RR dinucleotides, which was selected for when interrogating a shuffled *cap* library, but the question of why this pattern exists remains. To begin assessing this, I generated multiple stuffer DNA stretches with different methodologies, which I inserted into a cassette comprising a CMV promoter driving a *gfp* reporter gene. The insertion of stuffer DNA into transgene cassettes to expand their length is a common procedure, as shorter transgenes can impede correct viral packaging [301]. Starting from the wild-type genome of AAV2, I combined multiple fragments of the *rep* and *cap* genes to generate inert pieces of stuffer DNA that possessed the inherent AAV dinucleotide pattern without carrying over any known AAV protein products or regulatory sequences. The fragments did not contain any known start or stop codon nor any known promoter sequence (Figure 3.15 A). Thus, I created two sets of control and wt-fragment-containing (wt-fragments) cassettes of 3.8 and 4.8 kb in length (Figure 3.15 B). The control construct was either a bacterial kanamycin resistance gene for the 3.8 kb cassette or a stretch of lambda phage DNA for the 4.8 kb cassette. The two sets represent improvements made in two generations of tested transgene cassettes.

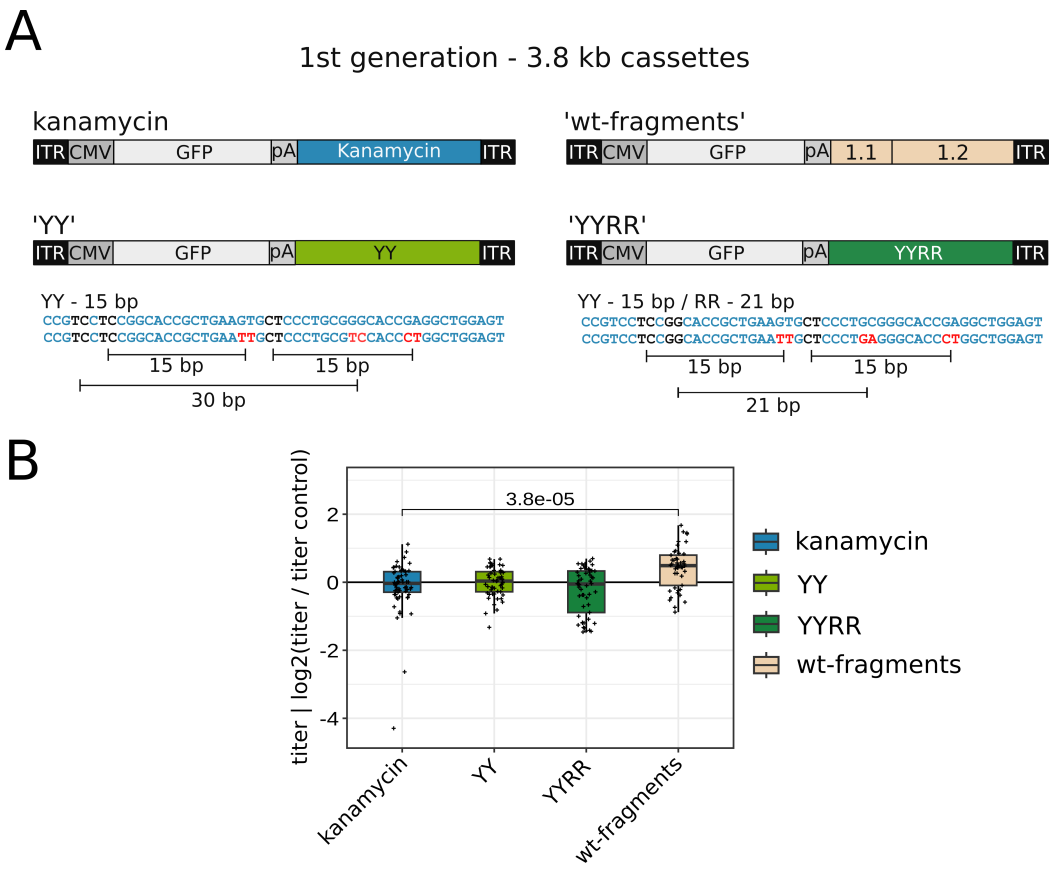

Figure 3.16. First generation dinucleotide pattern imitation attempts.
A) Schematic representation of the first-generation transgene cassettes 'YY' and 'YYRR'. In the 'YY' transgene, starting from a YY dinucleotide existing in the kanamycin stuffer sequence (shown in black), other YY dinucleotides were randomly inserted at 15, 30, or 45 bp from the starting point. The 'YYRR' transgene was generated similarly. In addition to the YY inclusion, also an RR dinucleotide was inserted randomly at 21 bp from an RR starting point (black). B) Small-scale viral titer estimations of all 3.8 kb first-generation transgene cassettes compared to the kanamycin control. The Wilcoxon-Mann-Whitney test was used to test for statistical significance, and only significant p-values (< 0.01) are indicated.

First-generation imitation

In an effort to emulate the wild-type AAV2 pattern on inert stuffer DNA, I computationally inserted YY or YY/RR dinucleotides into the kanamycin resistance gene to ultimately test for a difference in viral production titers. The changes here described were only performed for the 3.8 kb genomes, which are termed here the first generation imitations (Figure 3.16 A). For imitation, I only tampered with the forward sequence of the kanamycin control stuffer sequence. The first transgene, termed 'YY', was generated by randomly mutating other dinucleotides into YY dinucleotides 15 bp downstream of already existing YYs in the kanamycin stuffer sequence. This increased the YY dinucleotide proportion from 25% in the wild-type sequences to 47% in the 'YY' stuffer. The second transgene, termed 'YYRR', in addition to the YY dinucleotides, also had randomly inserted RR dinucleotides (for more information, see section 2.7.6). This design is presented anachronistically here

as it represents an older understanding of mine that only took the forward strand of the AAV2 wild-type genomic dinucleotide patterns into consideration. I have since updated my perspective to include both strands of all closely related AAVs and arrived at the 15 bp periodicity for both YY and RR, which I have presented in previous sections. Nevertheless, the inclusion of the RR at 21 bp intervals still pushed the YY dinucleotide proportions down to 30% and thus closer to the wild-type AAV2 proportions.

I measured the viral titer in a small-scale production system, in which HEK293T cells were triple-transfected in a 96-well format with the designed cassettes, AdH, and the AAV2 *rep* and *cap* genes. Three days after transfection, the cells from every well were lysed and DNase-digested to remove unpackaged DNA, after which the viral titer was measured using ddPCR. This experiment included between 45 and 61 biological replicates, split across multiple plates measured at different times. The wt-fragment stuffer produced a significantly higher titer compared to the kanamycin control cassette (Figure 3.16 B). However, both the titers of the 'YY' and 'YYRR' transgene cassettes did not differ from the control cassette.

Second-generation imitation

The first-generation 3.8 kb genomes were far shorter in length than the 4.8 kb genomes within wild-type AAV2 virus particles. Because the length and thus pressure of the packaged DNA might cause a difference in a putative interaction of dinucleotides with the capsid interior, I created a second generation of transgenes and wild-type pattern imitations with a length of 4.8 kb (Figure 3.17 A). The 4.8 kb wt-fragments cassette also produced a statistically significantly higher titer in the 96-well small-scale production format compared to its control, similar to the 3.8 kb genomes (Figure 3.17 B). This shows that the increase in viral titer of the wt-fragments is irrespective of the exact control sequence or the packaged genome size. Thus, I conclude the wt-fragments stuffer must contain a sequence-based signal that boosts viral titer.

Unlike the inert wt-fragments, the first-generation imitations failed to yield a higher titer compared to the control. Therefore, for the second generation, I stayed closer to the pattern of these fragments and directly copied the dinucleotide patterns existing on the wt-fragments onto stuffer DNA, which I termed M1, M2, and M3. In these designs, I mapped the wt-fragments dinucleotide pattern onto the lambda control sequence using different masks (Figure 3.17 A). M1 possesses a mask of all YY or RR dinucleotides present on the wt-fragments. Since this mutates 80% of the lambda sequence, I also designed M2 in which the mask only contained YY or RR dinucleotides that are exactly 15 bp apart from another YY or RR, respectively. This approach only mutated a little less than 50% of the lambda DNA. The mask for sequence M3 contains - agnostic to the wt-fragments sequence - a

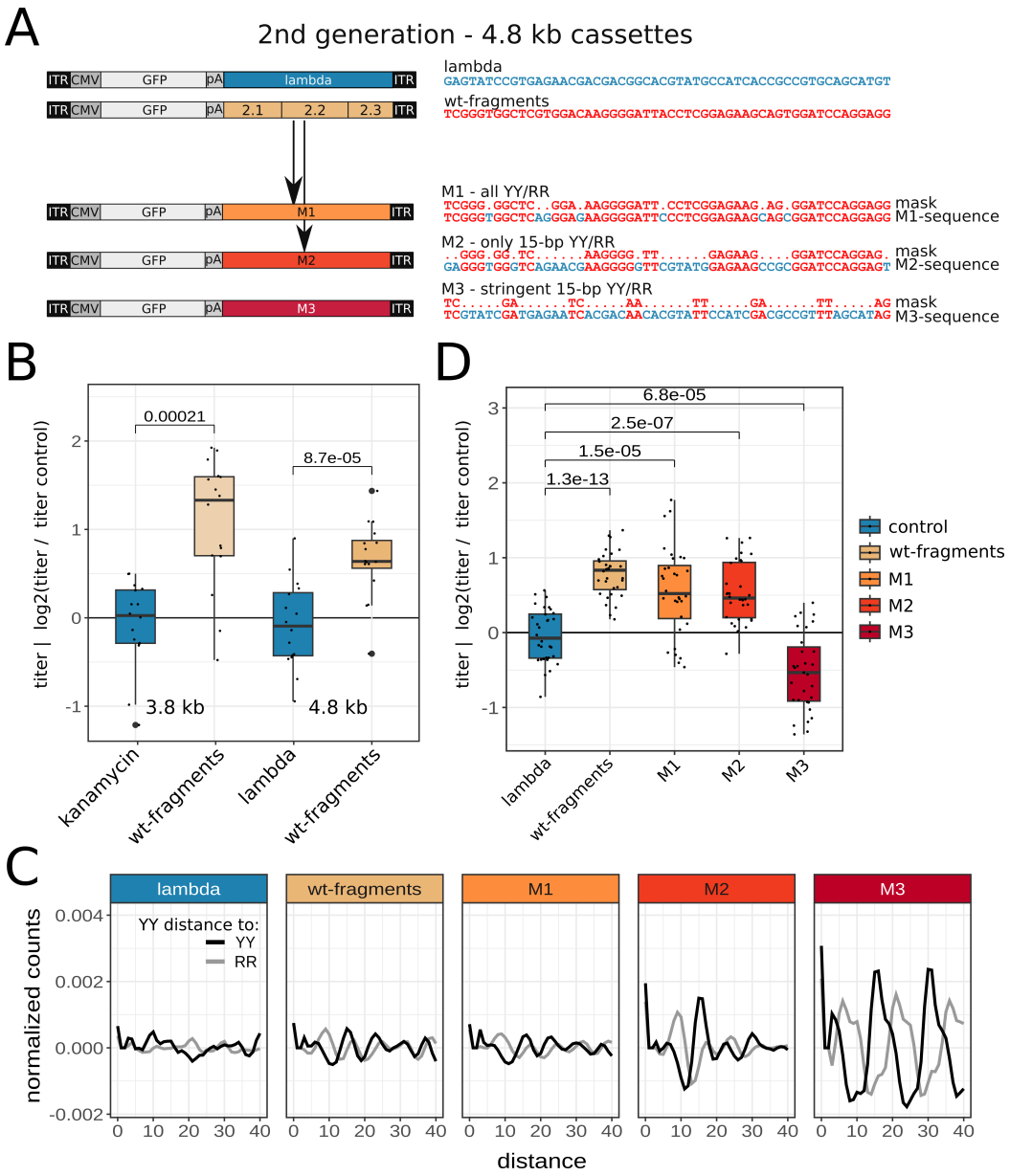

Figure 3.17. Second-generation dinucleotide pattern imitation attempts.
A) Schematic representation of the second-generation 4.8 kb sized genomes. All transgene cassettes contain an AAV2 ITR as well as a CMV promoter-driven *gfp* next to a 2.7 kb stuffer sequence. The control stuffer sequence was amplified from lambda phage DNA. The wt-fragments are comprised of three inert sequence fragments from the AAV2 wild-type genome. M1, M2, and M3 are all imitation attempts of the dinucleotide pattern of the wt-fragments containing transgene cassette. M1 was generated by applying a mask containing all YY/RR dinucleotides from the wt-fragments to the lambda stuffer DNA. M2 was generated by applying a mask only containing YY/RR dinucleotides from the wt-fragments to the lambda DNA stuffer that are 15 bp apart from one another. The M3 transgene cassette was generated by directly applying a mask of offset 15 bp YY and RR dinucleotides to the lambda stuffer DNA. B) Comparison of the wt-fragments transgene cassettes of both lengths. The control cassette contains a kanamycin resistance in the 3.8 kb genomes and a lambda DNA fragment in the 4.8 kb genomes. The Wilcoxon-Mann-Whitney test was used to test for statistical significance, and significant p-values are indicated. C) YY dinucleotide auto- and cross-correlation to RR for all tested transgene cassettes in this section. Similar depiction as Figure 3.11 D. D) Small-scale production viral titer estimations of all tested transgene cassettes. All titers are depicted as log2 fold-changes to the lambda DNA control. The Wilcoxon-Mann-Whitney test was used to test for statistical significance, and significant p-values are indicated.

perfectly 15 bp spaced repeat of YY and RR dinucleotides shifted by 8 bp in regards to each other as observed in section 3.2.1 and Figure 3.11 D. The exact dinucleotide pair for M3 was randomly chosen at every position. To assess whether my imitation was successful, I performed auto- and cross-correlations of the YY and RR dinucleotide patterns (similar to what was performed in Figure 3.11 D). These showed that indeed the lambda control DNA, unlike the wt-fragments cassette, does not have an observable dinucleotide pattern (Figure 3.17 C). M1, having 80% of the sequence of wt-fragments, also stays very close to the wt-fragments pattern. M2 showed an exaggerated peak at 15 bp for YY to YY auto-correlation and 9 bp for the YY to RR cross-correlation. Unsurprisingly, M3 possesses a highly exaggerated and regular distance histogram for YY and RR dinucleotides.

Testing these cassettes in the 96-well small-scale production system and 32 individual biological replicates each, revealed that M1 and M2 were behaving very similarly to wt-fragments as they yielded a statistically significantly greater viral titer compared to the control (Figure 3.17 D). However, M3 did not follow the same trend and fell short of the control, indicating that this attempt at rigorously imitating the pattern did not yield a transgene with an increased viral titer.

No difference in replication intermediates but improved transduction of M2

The lab work and part of the experimental design in this section were performed by Dr. Jonas Becker and Emma Gerstmann in the Grimm lab. The entire data analysis and part of the experimental design and work were performed by me.

The mechanism that leads to an increase in titer for the wt-fragment and M1/M2 transgenes remains unclear. A possible mechanism is an increased Rep-mediated replication of the transgenes in question. To test this, we performed a triple-transfection of the second-generation transgene plasmids with AdH and *rep/cap* helper plasmids on HEK293T cells in a 6-well format. Low molecular-weight Hirt DNA was extracted three days post-transfection and DpnI-digested before loading on an agarose gel and subsequently subjected to Southern blotting (Figure 3.18 A). The used probe was specific for the CMVenh sequence, present in every transgene cassette. The Southern blot showed multiple bands that were specific to the DpnI digestion of the plasmid (marked with white asterisks), as shown in Figure 3.18 B. All five transgenes showed the expected replication mono- and dimer intermediates at about 5 and 10 kb, respectively. The intensity was comparable in all transgenes in question. The wt-F and M1 plasmids additionally showed a lower band for the replication monomer that could be attributed to a putative recombination event. Fragments 2.2 and 2.3 in wt-F consist of partially overlapping sequences (see Figure 3.15), which might lead to unexpected rearrangement during the replication of the ITR-flanked genomes (possibly a deletion of

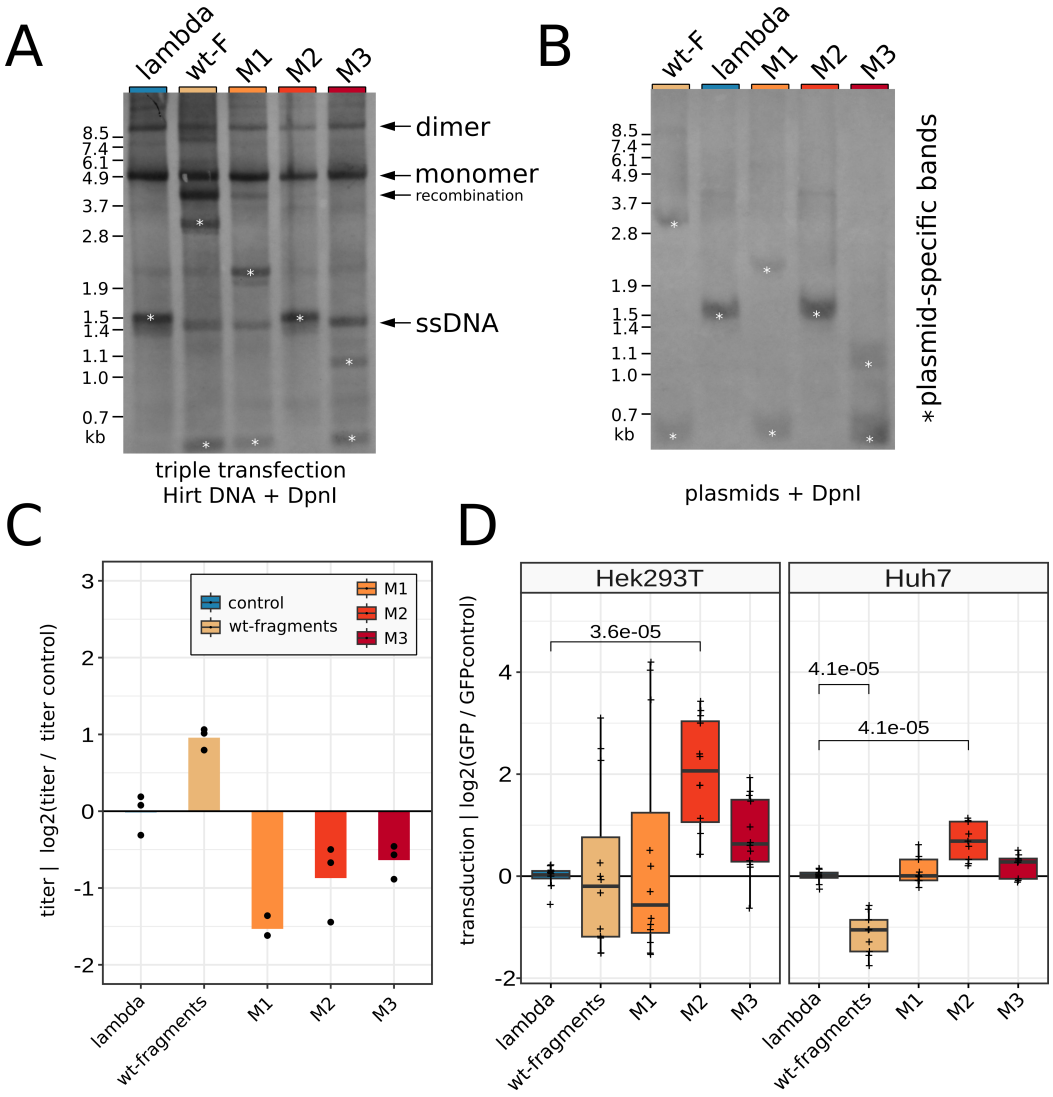

Figure 3.18. Assessment of replication intermediates and transduction behaviour of the second-generation transgene cassettes.
A) Southern blot of HEK293T cells that were triple-transfected with the second-generation transgene cassettes, AdH, and *rep/cap* plasmid. Hirt DNA was extracted three days post-transfection and digested with DpnI before loading. The probe against the CMVenh sequence was DIG-labelled. Mono- and dimer replication intermediates, ssDNA, and a putative recombination band are indicated. Plasmid-specific bands are marked with a white asterisk based on the bands in B. B) Southern blot of transgene cassette plasmids digested with DpnI and visualised with the same probe as in A. Plasmid-specific bands are marked with a white asterisk. C) Viral titer estimated from a large-scale production. Values are log2 fold-changes in comparison to the lambda control DNA transgene cassette. D) GFP fluorescence of HEK293T or Huh7 cells three days post-transduction with the dinucleotide imitation transgene cassettes. Values are log2 fold-changes regarding the lambda DNA control transgene. The Wilcoxon-Mann-Whitney test was used to test for statistical significance, and significant p-values are indicated.

2.2). This might also have led to recombination in M1, since it inherits 80% of the wt-F sequence. The band at 1.5 kb is attributed to 4.7 kb of ssDNA (based on results from [109]) and can not be quantitatively assessed because of overlapping plasmid-specific bands.

For subsequent transduction experiments, we produced the second-generation transgenes in a large-scale dish production. The increase in viral titer of M1 and M2 from the 96-well productions was not transferrable to the large-scale dish production (Figure 3.18 C). However, the wt-containing transgene continued to outperform the control transgene as it displayed a two-fold greater titer. The lower titer of transgenes M1-M3 in large-scale dish productions may be due to the necessary purification by density gradient centrifugation.

To interrogate whether the created stuffer DNA designs also behave differently in their transduction, HEK293T and Huh7 cells were transduced at a multiplicity of infection of 10000 vector genomes per cell. Medium was replaced by PBS and GFP emission was measured three days post-transduction. In both HEK293T and Huh7 cells, the M2 transgene cassette outperformed all others and showed about 50 to 100% more GFP signal compared to the control (Figure 3.18 D). Interestingly, the wt-fragments cassette did not perform better than the lambda control cassette and even performed significantly worse in Huh7 cells. These results indicate that a significant change in the properties of the resulting AAV vector can be induced by merely manipulating an inert part of the transgene sequence outside of the ITR or coding regions. The YY/RR dinucleotide seems to influence the vector properties on multiple levels, as it can also affect transduction.

Mass photometry measurements of wt-fragments

There was no evident difference in replication intermediates between the second-generation transgene cassettes. Therefore, I employed mass photometry measurements in an effort to find a potential biophysical difference between the control and wt-fragments containing particles that might explain the contrasting viral titers. Lambda and wt-F were produced in 44 dishes each in a large-scale production. After density gradient centrifugation, I pulled four fractions from the iodixanol gradient and subsequently purified them by size filtration (Figure 3.19 A). The first fraction (F1) is commonly drawn to obtain filled AAV capsids (1.5 mL). AAV productions are known to produce a variable number of partially filled and empty capsids, which would travel further up in the gradient. The following fractions (F2-F4; 0.5 mL each) were taken to interrogate any differences in partly filled or empty capsids between the tested transgene cassettes. A silver stain of the drawn fractions revealed F1 of both cassettes containing expected proportions and sizes of the VP1, VP2, and VP3 proteins (1:1:10; 84.1 kDa, 68.9 kDa and 62.2 kDa; Figure 3.19 B). While F2 and F3 seemed mostly empty, F4 again showed the same viral capsid protein pattern as F1 does, albeit at lower

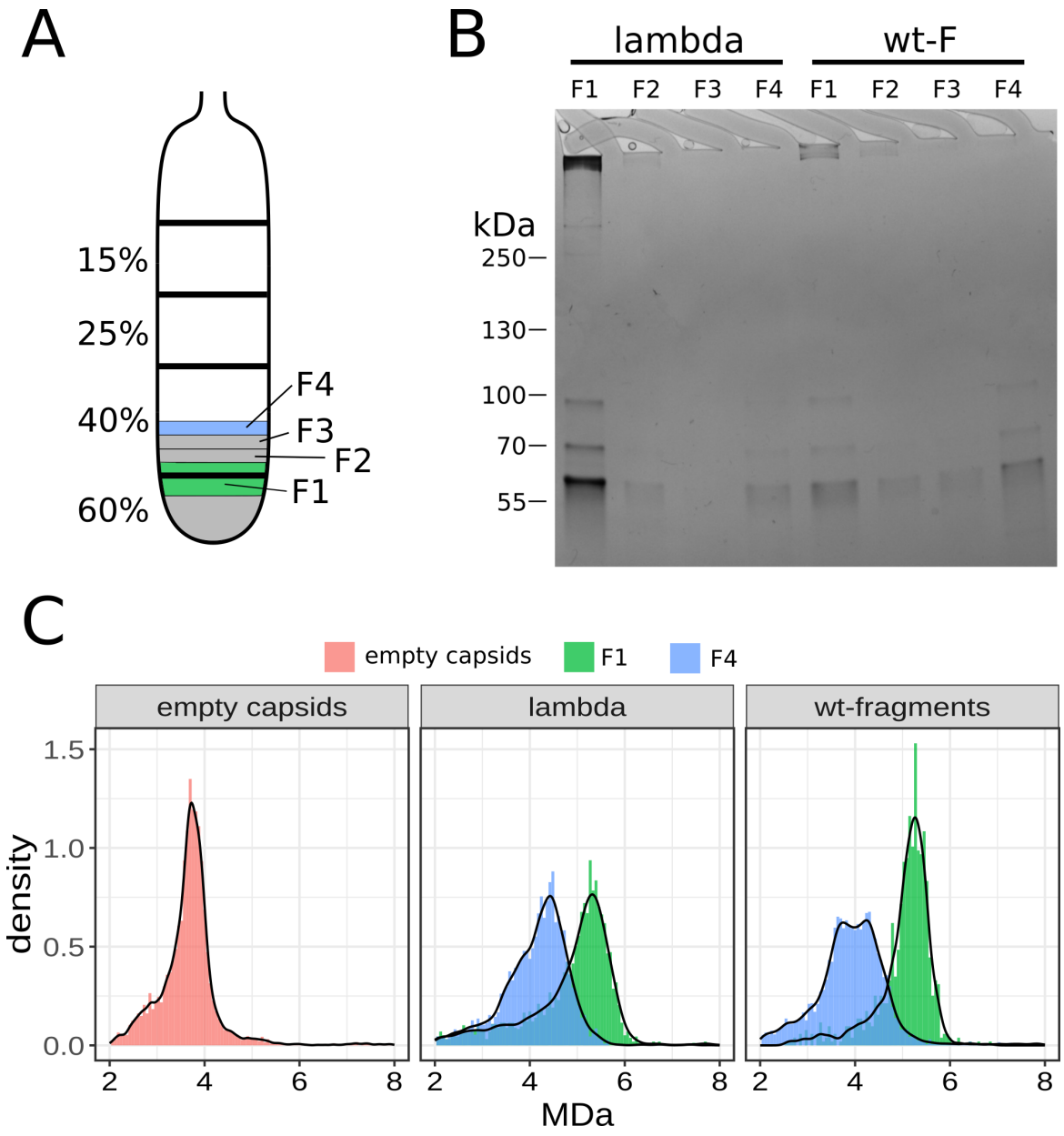

Figure 3.19. Silver stain and mass photometry measurements comparing of different iodixanol fractions from the wt-fragments transgene cassette.
A) Schematic of the different interrogated fractions pulled from the iodixanol gradient. The crude lysate is sub-layered by a 60%, 40%, 25%, and 15% iodixanol solution in that sequence. The first fraction (F1) is the top-most section of the 60% solution and the next 0.5 mL of the 40% iodixanol solution. The subsequent fractions (F2-F4) are the following: 0.5 mL of the 40% iodixanol solution. B) Silver stain of the different iodixanol fractions. The molecular weight marker is indicated in kDa. Each lane in the gel represents a single fraction. C) Histograms from the particle count obtained by mass photometry measurements. The black line represents a kernel density estimation of the histogram. Only measured particles of a relevant size between 2 and 8 MDa are shown.

intensity. Thus, F4 probably contained partially filled or empty particles. The F1 of both productions showed a sizeable impurity close to the pocket of the gel.

Mass photometry measurements leverage the light scattering of particles to ascertain their mass. Empty AAV particles have a mass of about 3.9 MDa (control empty particles were generously gifted by Julie Garcia Gonzalez-Calero). Together with the 4.8 kb single-stranded DNA genome, this sums up to a mass of about 5.2 MDa for completely filled capsids. Partially filled capsids would be found between 3.9 and 5.2 MDa. Mass photometry was performed at the EMBL Heidelberg Protein Core Facility together with Dr. Karine Lapouge. The measurements revealed distinct peaks for the F1 and empty particles at 5.2 and 3.9 MDa, respectively (Figure 3.19 C). A difference could be observed in the F4 fraction. Both possessed a differentially pronounced double-peak structure, where one shoulder was at about 3.9 MDa (empty capsids) and the other was slightly larger at around 4.2 MDa (partially filled). The proportion of these two peaks regarding each other differed between the lambda control and the wt-fragments cassette. The lambda control F4 distribution seemed to favour the partially filled capsids over the empty capsids. Similarly, the filled capsids peak (F1) was sharper in the wt-fragments, indicating that in this production, fewer data points were far from the mean, and thus more were at the theoretical mass of 5.2 MDa.

4. Discussion

In this work, I have explored two related aspects of AAV biology. Epigenetic regulation is pivotal for the regulation of gene expression, however, its influence on DNA delivered by AAV gene therapy vectors is severely understudied. Therefore, AAV-delivered episomes were interrogated in terms of DNA CpG methylations and the presence of the histone modifications H3K27me3 and H3K27ac. In the second part of this thesis, a unique dinucleotide pattern on the wild-type genomes of AAV is described and characterised. The pattern is selected from a pool of shuffled *cap sequences* and its presence appears to be beneficial for rAAV vector production.

4.1 Epigenetic regulation of the rAAV-delivered transgenes

There is conflicting evidence in the literature regarding the methylation state of the AAV episome. On the one hand, direct measurements of the transgenes several months after transduction did not show meaningful methylation rates in primate muscle tissue [148]. Additionally, the insertion of CpG-free ITRs into a vector has yielded comparable transduction results to a wild-type ITR-containing vector, suggesting that the CpG methylation of the vector is presumably not a limiting factor of transduction [302]. On the other hand, there is evidence that methylation inhibitors can increase the AAV transduction rate [149]. The methylation inhibitor would alter the entire methylome of the cell culture, and thus could influence the transgene expression indirectly. Therefore, the available evidence collectively favours the conclusion that methylation makes little to no difference in AAV-based transgene expression. Nevertheless, because of the importance of DNA methylation in epigenetic regulation, it remains imperative to consider novel methods for assessing it in the AAV gene therapy context.

4.1.1 Nanopore methylation calling on the transgene is comparable to EM-seq

From a methodological standpoint, this work underscores both the potential and limitations of using nanopore sequencing to measure methylations in the CpG context within circular AAV transgenes.

While nanopore sequencing enables direct methylation detection without chemical conversion, it suffers from relatively high false-positive estimations. Initially, I had used the R9 chemistry of nanopore sequencing together with several well-established and comparable tools to invoke the 'wisdom of the crowd' and ultimately improve accuracy. However, the negative control was never measured to be exactly at 0%. This is a sequence-independent phenomenon, since it was detectable in the promoter region (different sequence and length in every transgene-cassette) and in the eYFP coding region (same sequence in every cassette). This false-positive rate is in a comparable range to the METEORE benchmark study for the used tools [159]. In the same benchmark study, guppy is labelled as an underpredictor and nanopolish as an overpredictor, respectively predicting a lower and higher methylation rate compared to the ground truth. The underprediction of guppy can also be seen in my data, in which the positive hypermethylated control is only predicted to be between 60 and 80% methylated. The high false-positive rate appears to be linked to nanopore sequencing, as it is also seen in the later R10 nanopore sequencing runs. Ultimately, this suggests that nanopore sequencing is likely not able to perform sensitive methylation calling. Nevertheless, by comparing the nanopore-sequenced tissue sample to the negative control, a generally low methylation rate in the AAV transgene can be concluded.

EM-seq was used as a control technique to analyse methylations on the AAV episome. It uses an enzymatic approach to convert unmethylated cytosines eventually to thymines (see section 1.1.6). This technique found methylation levels close to 0% across the entire transgene, congruent with the nanopore result not deviating from its negative control. The ITR sequences were measured to be methylated by EM-seq at a rate of around 20%, which is likely to be an artefact (more in section 4.1.2). Correlations of nanopore and EM-seq methylation frequencies in the rest of the transgene (excluding ITR) were poor, again suggesting the nanopore methylations were mainly caused by the inaccuracy of the calling. Because of the noisy methylation measurements of unmethylated DNA from nanopore sequencing, EM-seq presents a superior alternative for assessing DNA methylation of a sample where a low methylation rate is expected. However, nanopore sequencing offers several advantages over EM-seq. The first is its ease of use: a sequencing run can be initiated within a single 8-hour workday, starting from extracted DNA without needing to rely on the

capacity of a sequencing core facility. Additionally, the sequencing of long reads achievable by nanopore can give insights into the per-molecule methylation and also recombinatorial state of the episome (latter discussed in section 4.2).

4.1.2 Methylation is not affecting AAV episomal transgene expression in the liver of mice

In the tested samples two weeks post-injection (p.i.) of CMV, GFAP, and LP1 promoter-driven cassettes packaged in AAV9, the nanopore-predicted methylation rates were not divergent from the negative control. All of the used methylation callers for R9 nanopore sequencing data agreed, suggesting a methylation rate close to 0% in the entire transgene, including the ITR sequence. Both the promoter and eYFP coding regions in all cassettes showed a similarly low methylation rate, suggesting that all tools could call methylations independent of the underlying sequence. Similarly, measuring methylation with EM-seq, showed methylation rates at/or close to 0% across the entire transgenes, except for the ITR sequences (see section 4.1.2).

In a subsequent experiment, I looked into the CpG methylation on a CMV promoter-driven cassette in the livers of mice sacrificed at different time points after the injection. The transgene expression was highly time-dependent and increased slightly for AAV2-transduced samples but decreased for AAV9-transduced samples. Because the ensemble approach from different tools in R9 sequencing did not appear to be necessary, here only R10 sequencing and the latest dorado methylation caller were employed. Similar to the previous data at two weeks p.i., methylations on the transgene cassette were indistinguishable from the negative control at all time points. There was also no difference between samples treated with capsid AAV2 or AAV9, as both displayed a similarly low CpG methylation rate. The yield of episomal reads was comparably low for AAV2-transduced samples and only allowed analysis at a coverage of about 30x. AAV9-transduced samples were re-analysed with an increased coverage of around 300x. This analysis revealed that over time, there was indeed a minor, but statistically significant, increase in methylation detectable with nanopore sequencing. This increase was minuscule, as the mean methylation frequency only increased from 2.17% to 2.69%. This is unlikely to play a part in explaining the time-dependent changes in transgene expression, but it proves that methylation is occurring, albeit at a very slow rate. This increase in methylation is unlikely to stem from integrated transgenes accumulating CpG methylations because of the exonuclease enrichment step that was used in this sequencing run.

At least until the 12-week timepoint, CpG methylation does not majorly affect transgene

expression in the mouse liver, regardless of the capsid or promoter used. This data is congruent with other research that found the methylation of plasmid or of the delivered AAV transgene to be inconsequential for its expression [143, 148].

In contrast to AAV, in the Adenovirus vector system, the CMV promoter was subject to significant methylation after *in vivo* transduction in the muscle of rats with a first-generation Adenovirus vector [144]. It is unclear what leads to the increased methylation of the CMV promoter in Adenovirus vector DNA in this context, but the *de novo* methyltransferases DNMT3A and DNMT3B are likely to be involved. The transduced tissue could also influence the epigenetic regulation of a delivered transgene. Since I only interrogated the mouse liver, a direct comparison with Adenovirus in rat muscle is not possible. However, also in the Adenovirus vector space, the evidence for methylation being detrimental is contradictory. Another study suggests that epigenetic silencing is performed by altered histone modifications and not CpG methylation [303]. This study, however, mainly interrogated whether the prokaryotic origin of the stuffer sequence affected transgene regulation. A possible resolution of this discrepancy is that the histone-deacetylase HDAC1 can interact with DNMT3A, which could induce methylations at sites that are already being targeted for histone deacetylation [304].

Why the delivered AAV transgene is not affected by CpG methylations remains unclear. There is one piece of evidence that shows the methylation inhibitor azacitidine to be beneficial for AAV transduction [149]. However, this is only an indirect proof. Outside of an AAV context, the silencing of transfected DNA was also not correlated with DNA methylation, but is associated with histone modifications [143, 305]. The majority of evidence points towards methylation not being an issue for AAV-delivered transgene expression, even multiple years post-treatment.

In fact, the non-existent methylation of the transgene could be considered worrisome. Expressed endogenous genes display a distinct methylation pattern derived from the general hypermethylation of CpG motifs within the genomes of higher animals. In highly expressed genes, CpGs close to the promoter are unmethylated, whereas CpGs within the gene body are hypermethylated [306]. The fact that the AAV-delivered transgene does not follow this methylation pattern suggests that it has not been fully incorporated into the chromatinized DNA within the nucleus and might face scrutiny as a result.

The ITR sequence is likely not methylated

The ITR sequences are an integral part of every transgene, as they are the only necessary packaging signal that also enables vector replication during production. They enable the

circularisation of the episome after transduction [107]. Naturally, substantial research has been devoted to studying the effects the ITR sequence can enact on the transgene expression [108, 115, 307, 302, 28]. However, as with all epigenetic work on the AAV episome, information on the ITR sequence is sparse. At a congress presentation, the ITR sequences of AAV episomes formed *in vivo* were reported to be overly methylated when interrogated with bisulfite sequencing [308]. In the EM-seq data from AAV-delivered episomes two weeks p.i., I also observed elevated levels of methylations in the ITR sequences. However, the nanopore sequencing of the same sequence in multiple samples and additional timepoints after 6 and 12 weeks did not show methylations that were deviating from the negative control. EM-seq measured methylations were also increased in motifs that are usually not methylated in mammals, like CHH and CHG. The most likely explanation for the increased methylation rates in ITR sequences according to bisulfite or EM-seq is that the strong secondary structure of the ITR sequences counteracts the conversion. The ITR sequences adopt highly ordered T-shaped secondary structures, which are infamously problematic when subjecting them to Sanger sequencing [309].

The base-pairing of double-stranded DNA can heavily inhibit the conversion rate of an unmethylated cytosine to uracil in bisulfite-based conversion [310]. Similarly, APOBEC almost exclusively accepts single-stranded molecules as its substrate [311]. Nanopore sequencing does not seem to be affected by strong secondary structures. The motor protein that is necessary for translocating DNA through the nanopore during nanopore sequencing needs a strong helicase activity for this purpose [312]. Therefore, it is likely able to resolve and sequence the ITR regardless of its strong secondary structure. Secondary structure does not seem to impact the squiggle and subsequent basecalling, since no aberrant behaviour was seen in the ITR regions compared to others.

4.1.3 Histone modifications and transgene expression - AAV9 is bad but AAV2 is worse

It became increasingly obvious from my results and those from the literature that methylation of the CpG motif within the AAV episome was not a driving factor for epigenetic silencing of the transgene. There is growing evidence that histone modifications are the major players in the regulation of transgenic DNA. Therefore, I analysed the presence and quantity of the histone modifications on the AAV-delivered episomes. This was conducted on the livers of mice transduced with AAV2 and AAV9 and sacrificed after 2, 6, and 12 weeks. I performed CUT&Tag sequencing on the mouse livers against H3K27me3 and H3K27ac. Briefly, CUT&Tag uses antibodies to direct a transposase loaded with sequencing adapters into the vicinity of the target histone modifications and thus enables the amplification of

DNA segments enriched with the modification of interest (more in section 1.1.7). This technique enabled me to assess the histone modifications in question in the mouse genome and simultaneously on the AAV transgene. At all time points and regardless of the used capsid, the samples yielded reads for both H3K27me3 and H3K27ac. This proves that in the liver, the transgene is associated with histones that are at least modified on their H3 tail.

There are three major observations from the CUT&Tag experiment. Firstly, in both transgenes, H3K27me3 accumulated more than H3K27ac. Secondly, only for AAV9-delivered samples, the ratio of H3K27me3 to H3K27ac shifted towards favouring H3K27me3 at later time points. Finally, there are always significantly fewer reads from the AAV2-delivered transgene compared to AAV9.

High levels of histone modifications might contribute to low CpG methylation

H3K27me3 appears to gather at higher levels on the transgene than H3K27ac does. Sensing invading DNA is a cornerstone of the innate immune system that can also elicit repressive epigenetic regulation of said DNA [313, 314]. For example, similarly high depositions of H3K27me3 were found in the early stages of infection on the genomes of Herpes simplex virus 1 (HSV-1) [315]. This silencing has been hypothesised to be particularly pronounced in hepatocytes, since they are subject to repeated onslaught by viruses from the gut virome [127].

A side effect of the high levels of H3K27me3 on the transgene might be the absence of methylated CpG sites. H3K27me3 and DNA methylations are considered to be mutually exclusive [171, 316]. The exact interplay is not known, but disruption of polycomb-repressor complex 2 (PRC2), which is responsible for depositing H3K27me3, leads to aberrant DNA methylations at loci that are usually repressed by H3K27me3 [171].

Vice versa, unmethylated DNA might also be prone to association with the inactivating mark H3K27me3, as PRC2 is inhibited by DNA methylations [172]. It might be worthwhile to devise an experiment in which vector transgene DNA is hypermethylated before transduction. This could be used to gather information on differences in epigenetic regulation when one of the markers for the innate immune system (unmethylated CpGs) is missing from the invading DNA.

AAV9-delivered transgenes lose H3K27ac over time

Two weeks post-injection, compared to the later time points, the AAV9-transduced samples had a ratio that was more in favour of H3K27ac, which later shifted even more towards H3K27me3. This can be interpreted as a loss of K27 acetylations over time. This shift

correlated well with the decreasing transcription of the AAV9-transduced liver samples, except for one liver sample from the 6-week time point. Other histone modifications are likely to play an additional part in the regulation, but were not analysed in this work.

It has to be noted that CUT&Tag sequencing is a composite approach. Hence, only the mean of all episomes and liver cells can be observed using this method. All results have to be evaluated in this light. There might have been a population of highly transcribing transgenes that was continuously being repressed over time. Similarly, the association of opposing histone modifications might occur on the same transgene competitively. Only later and for unknown reasons, H3K27ac is starting to become lost.

As already briefly mentioned above, the silencing of invading DNA is a crucial part of the innate immune response that can occur in a matter of hours, for example, in an HSV-1 infection [315]. If invading DNA is so readily silenced, why do the AAV9-transduced samples demonstrate (compared to later time points) high levels of H3K27ac and active transcription two weeks after transduction?

It is possible that by injecting vast amounts of virus into the bloodstream of the mice, the innate immune response of transduced cells becomes overwhelmed. Alternatively, the innate immune response's reaction is not as strong towards the AAV capsid as it would be for a virus such as HSV-1. In any case, the transgene seems to become chromatinised, initially without being extensively silenced. I hypothesise that the chromatinisation occurs through free histones from the nucleus upon second-strand synthesis of the transgene. These histones need to be expressed by the host cell. Because of their sheer abundance within the liver, I assume that mainly hepatocytes are transduced and that they are the main transduced cell type analysed in this work. Even though hepatocytes are quiescent in the healthy liver, they are still likely to express histones during the endoreplication of their genomes [317, 318]. Newly synthesised histones are often quickly modified with post-translational modifications, before they are deposited on DNA [319]. Both methylations and acetylations can be found on nascent histones [320, 321]. The newly chromatinised transgene could therefore be in a relatively naive regulatory state that allows ample transgene expression. The silencing of the transgene may then occur at a later stage, driven by the progressive deacetylation of histone tails associated with the transgene. This is congruent with HDAC inhibitors being advantageous for AAV transduction [174, 175]. NPP220 presumably plays an important role in this by possibly recruiting histone deacetylases to the transgene [176, 322]. It remains unclear, though, why this happens only after a week-long delay.

The capsid influences the levels of modifications on the episomal transgene

The explanation attempt above does not cover the better ratio of histone modifications on the AAV2-delivered transgene. This is potentially an effect of the capsid itself. Such downstream functions of the capsid are rare but not unheard of. The HIV capsid proteins were shown to have a function exceeding protection and delivery, as they form a subviral complex with their genome with a putative role in genome integration [323].

In a novel discovery, the AAV capsid was similarly implicated in having a downstream influence on the epigenetic regulation of the transgene [177, 178]. Both of these studies show an altered epigenetic regulation depending on changes made to the capsid, but the part of the capsid responsible for this interaction differs. According to [177], a missing glycine or threonine in a variable section of the VP3-common region (amino acid 264-267 according to the AAV2 sequence numbering) of the *cap* gene is responsible for the non-permissive epigenetic regulation. However, according to [178], the N-terminal part of the gene is responsible for the advantageous epigenetic regulation in their study. The AAV2 capsid sequence is also missing the aforementioned glycine or threonine, as indicated by the Kay lab [177]. Therefore, it seems more likely that the N-terminal VP1u sequence is allowing its permissive regulation. This would be easily tested in a swap of the N-terminal sequence. It might also be the case that neither of these two aspects is responsible, and a completely different part of the AAV2 capsid is the reason for my observations.

4.1.4 The limitation in comparing AAV2 and AAV9

The major limitation of comparing AAV2 with AAV9 is their different unpacking kinetics. They were initially chosen as subjects for this interrogation of differential epigenetic regulation because of their different behaviour within the mouse. Whereas AAV9 is a good transducer that also forms the basis of many capsid engineering efforts for systemic transductions [324, 325], AAV2 is an underperformer. This discrepancy has been subject to previous research. The conclusion was that AAV2 takes considerably longer to unpack and form a transcription-ready transgene compared to well-transducing serotypes, like AAV8 [111]. In my data, a similar picture emerged. In an exonuclease assay, I could see that the AAV2-delivered DNA was more sensitive towards the exonuclease than the AAV9-delivered one was. This suggests that AAV2 is not as readily circularised and remains in a linear form, maybe even still packaged in the capsid. The AAV9-delivered transgene, too, seems to undergo increased circularisation over time, which is not as drastic. Histones, together with their modifications, can presumably only form on DNA that at the very least has undergone second-strand synthesis and persists as a double-stranded molecule. The slower transduction of AAV2 serves as an explanation for the low number

of reads obtained from AAV2-transduced samples, both in the nanopore sequencing and the CUT&Tag experiments. Compared to the beneficial epigenetic regulation allowed by the AAV2 capsid, its slow second-strand synthesis apparently has more weight in the expression of the transgene. It is also possible that, due to the slow unpacking speed of AAV2, the epigenetic regulation of the host cells has not yet had the time to establish itself as it has for AAV9. Future experiments should compare capsid variants that do not differ in their transduction kinetics.

4.1.5 Outlook on the importance of AAV episome epigenetics

Current gene therapy in humans using rAAV vectors can attain high expression of a therapeutic gene only for a limited time, which can diminish to subtherapeutic levels at later time points. Even though vector genome loss through the action of the immune system might well be the major factor for this transience, epigenetic factors can demonstrably play a role too. Efforts in understanding the epigenetic regulation of the AAV-delivered transgene have only recently been brought back into the focus of AAV researchers. This renewed interest is undeniably influenced by advancements made in methods such as EM-seq and CUT&Tag sequencing. Nanopore sequencing is also continuing to grow its toolbox with, for example, a novel method for simultaneously screening histone modifications and CpG methylations [326].

The AAV9 capsid manages to enable high expression on a short-term basis, which is soon silenced via loss of histone tail acetylation. AAV2 does not seem to fall victim to this loss of acetylation, but is plagued by a slow unpacking rate, which is even more detrimental to transgene expression. Recent evidence points towards the capsid itself playing a role in the epigenetic regulation. In this work, I have only analysed two different histone modifications on the N-terminus of H3. Other H3 modifications, such as H3K4me3 or H3K9me3 are implicated in epigenetic regulation of the AAV transgene [177] and are thus interesting targets for future research. Future experiments could also aim at elucidating capsid features that allow an efficient incorporation of the transgene into the host chromatin, such as the N-terminus of the VP1u-region [178]. Insights into these features are potentially translatable to other gene therapy approaches and the delivery of foreign DNA for therapeutic purposes.

4.2 Direct nanopore sequencing of the transgene is possible and gives insight into the recombinatorial state of the episome

In this work, I used Oxford nanopore sequencing to directly sequence episomal AAV-delivered transgenes extracted from the livers of transduced mice. To enrich for circular episomal reads, the sequencing is preceded by an enzymatic enrichment step, which degrades linear DNA, like linear AAV genomes or the genomic DNA. In the well-transduced liver, this allows for enough reads from AAV9 delivered episomes to undertake meaningful analyses. The reads recovered from AAV2-delivered episomes were generally much lower than in AAV9-delivered ones, which is why the analysis of transgene recombination here is only performed on AAV9-treated samples. More on this discrepancy is discussed in section 4.1.4.

I was able to utilise this protocol to directly assess the methylation rate of AAV-delivered DNA (see discussion section 4.1.2). Additionally, I initially studied at ITR recombinations as they appear *in vivo* in the mouse liver. The most prominent form of ITR recombination sequences in a circular episome was reported to be a 'double-D' sequence, starting and ending with a D-sequence and containing the other ITR-subsequences in between [109, 327, 107]. I utilised this fact to identify ITR sequences within each individual nanopore read, by locating all occurrences of the D-sequences on the read and defining everything in between two such sequences as the recombined ITR. Using this approach, I could identify an ITR sequence in 61% of the total sequences that were obtained from nanopore sequencing of the mouse livers transduced with AAV9. The remaining 39% of sequences are likely also to possess an ITR sequence, which was not identified using my method. The identification of an ITR relied on the alignment of a D-sequence using the nucleotide blast algorithm [260]. Since nanopore reads have a lower accuracy compared to other sequencing methods and the ITR sequence is subject to recombinations, my method was likely not sensitive enough for the comprehensive detection of all ITR sequences.

Nevertheless, this analysis is supposed to yield an initial glimpse at the ITR recombinations within the episomes. The recombinations were found to be highly diverse in length. The most prominent recombined ITR sequence had a length of 165 bp and closely resembles the previously reported 'double-D' sequence. The B- and C-loops were similarly likely to be present in either the Flip or Flop orientation. The circularisation yielding this and other forms was shown not to depend solely on non-homologous end joining (NHEJ) or on homology-directed repair (HDR) [114]. However, the damage response protein ATM

and the DNA-break sensor complex MRN were both implicated. The high diversity of ITR lengths in my data is likely to be a result of deletions occurring during recombination of the ITR sequences, likely occurring during repair by NHEJ. The second-most abundant recombination form is about 85 bp in size and lacks both the B- and the C-loop. A similar form has also been reported before [327], but the high amount of this shorter form in my data was somewhat unexpected. Other experiments (see section 2.7.6) and colleagues' anecdotes imply that the B- and C-loops are prone to being deleted from plasmids during amplification in *E. coli*. In this case, it is unclear if the high occurrence of the 85 bp ITR form is due to the plasmid used for virus production having an erroneous ITR or if this recombination is a result of only intermolecular recombination.

With this investigation on the ITR recombinations, I have merely scratched the surface of the analyses of the circular episomes that are possible with the deployed method. The shape of an ITR hairpin within the AAV transgene affects which host proteins associate with it [115]. The shape of the recombined hairpin might also enable or inhibit binding of host proteins to the episome, which could be important for the long-term persistence of transgene expression in a transduced tissue. Therefore, understanding the recombinations of the episome depending on the used ITR sequence, the used capsid, the transgene, or the time after injection, might all prove pivotal for allowing long-term expression of the transgene. In my data, there was no discernible relationship between the time point after injection and the recombinatorial state of the transgene (not shown). However, interrogating time points closer to the injection might provide interesting insights into the kinetics of episome formation and associated ITR recombination. Another unexplored possibility with this technique is the quality control of episomal transgenes. One might use this protocol to investigate single-nucleotide polymorphisms or recombination events affecting the transgene-cassette and how these can affect long-term expression.

4.3 Periodic DNA patterns on the AAV genome

In this work, I described a YY or RR dinucleotide pattern on the genomes within the genus *Dependoparvovirus*. The pattern is found on both strands of the genome and has a periodic length of about 15 bp. I had initially discovered the pattern on the genomes of primate AAV serotypes, but it can also be discerned in a sequence set from the entire genus of *Dependoparvovirus*. Moreover, I also found hints towards a potential function of the pattern, as it (i) is enriched during the selection of shuffled capsid genes and (ii) stuffer sequences with various forms of the pattern outperform control cassettes in a small-scale production format.

4.3.1 The 15 bp YY/RR repeat is a marker for the genus *Dependoparvovirus*

The YY/RR pattern is remarkably pronounced and its repeat length at 15 bp is unique to the genus *Dependoparvovirus*. In a comparison of similarly sized sequences from *C. elegans* and from the human genome, the amplitude of the YY/RR distance histogram of AAV sequences was greater (Figure 3.11 E). An assessment of this pattern on the genus level in the kingdom *Shotokuvirae* and control genomes showed the genus *Dependoparvovirus* (of which AAV is a member) to be among the most periodic ones. The 15 bp repeat is also highly unique for the genus *Dependoparvovirus*, with all other highly periodic YY repeats being around 10 bp (Figure 3.12). Next to most eukaryotic genomes displaying a 10 bp periodicity, remarkably, many viral genera also displayed 10 bp periodicity. Some of the highly periodic genera were from the families of *Papillomaviridae* and *Polyomaviridae*. Viruses from these families associate their genomic dsDNA with histones from the host. Here, the 10 bp periodicity is easily justified, since it likely has the same purpose as in eukaryotic genomes. For other highly periodic genera, like the genus *Porprismacovirus* from the family *Smacoviridae*, there is no easy explanation for their strong periodicity. Histone association is not the only reason for dinucleotide periodicity, as it is also seen in prokaryotes. There, it is associated with the positive and negative supercoiling of genomes. The genomes of viruses from the families of *Papillomaviridae* and *Polyomaviridae* are supercoiled [328], which is possibly also aided by the YY/RR pattern. The circular genomes of *Smacoviridae* could profit from the same characteristic, even though their genomes are very short (around 2.5 kb) and are composed of ssDNA. Another explanation for the prevalence of the 10 bp periodicity in viruses would be that virus genomes almost exclusively consist of protein-coding sequences. The alternation of hydrophilic and hydrophobic amino acids in alpha-helices in these protein coding sequences can contribute to a 10 bp periodicity [329].

The comparison with other genera demonstrates the uniqueness of the 15 bp periodicity found in the genus *Dependoparvovirus*. It could potentially be used to identify *Dependoparvovirus* sequences in the analysis of metagenomic data, similar to how the 3-bp periodicity can be used for discovering protein-coding genes [208, 209]. Modern gene prediction pipelines such as BRAKER [330] still analyse the 3-bp periodicity within sequences to identify 'seed genes' to train subsequent, more accurate gene predictions. Similarly, using the frequency spectrum to narrow down the search space for alignment-based discovery of novel AAV capsid sequences might yield a considerable improvement in computation time. There is growing interest in the search for novel AAV capsid serotypes with potentially useful properties for gene therapy. This would enhance the capsid variety at the disposal of researchers seeking novel gene therapy vectors. These discoveries could then

be combined with a recent discovery, which enables *trans*-species transduction of AAVs by switching N-terminal VP1u sequences [178].

4.3.2 The YY/RR pattern potentially facilitates protein interaction

The function of the YY/RR pattern is not entirely clear, even though in this work colleagues and I have begun to gather some evidence. I hypothesise that this pattern represents DNA characteristics that do not directly result from the underlying protein-coding sequence. Such a characteristic might be the physical properties that the periodicity imposes on the DNA molecule. A comparable and prominent example of this is a dinucleotide periodicity found in the genomes of many eukaryotes (introduced in section 1.2.2). This dinucleotide pattern in eukaryotes facilitates the winding of the DNA double-helix around a histone octamer core and thus aids in the formation of chromatin [216, 223]. Similarly, in the context of AAV, the dinucleotide periodicity could likely facilitate the binding or processivity of a proteinaceous moiety.

In small-scale production experiments, I could observe an increased titer of transgenes having a periodic pattern that was close to the wild-type DNA sequence of the AAV2 genome over a control with no observable periodicity (see section 3.2.4). Why the titer is increased is not quite clear, but it could be traced back to an increased interaction of a protein with the transgenic DNA. Possible reasons why this increase does not translate to large-scale productions are explored further below.

Helper proteins or satellite virus feature

For its replication, AAV needs to utilise many proteins from helper viruses and the host cell. For example, the DNA-binding protein (DBP) of Adenovirus is encoded by the adenoviral E2A transcriptional unit, which is an essential factor for AAV replication [11]. Initially, I had regarded the periodicity of AAV genomes as a competitive advantage of AAV over its helper viruses. It could be an evolved aspect of AAV that enables it to bind specific replication-imperative proteins with a higher affinity than a helper virus, which would add another facet to the intriguing interactions of a satellite virus with its hosts. However, all sequences within the genus *Dependoparvovirus* contain the pattern. As the genus *Dependoparvovirus* also has independently replicating viruses as members, this generalisation can not be made.

Still, this does not mean that this periodicity has nothing to do with helper factors. As previously shown, the DNA of Adenovirus also portrays a periodicity of YY/RR dinucleotides with a repeat length of around 6 bp [241, 331]. However, this periodicity is likely associated with the binding of the adenoviral protein pVII, which is not a necessary

factor for AAV replication. Neither the Adenovirus nor the Herpesvirus families have a 15 bp periodicity pattern, which contradicts the hypothesis of this pattern being necessary for the binding of a helper virus protein.

Additionally, the periodicity might have been a general feature of satellite viruses. Along with many other analysed sequence sets, the sequences associated with the family of *Lavidaviridae*[1] also have a periodicity of around 10 bp (see section 3.2.2 and Table 3.3). This sequence set was the only one with a considerably high periodicity among the sets from satellite viruses. Therefore, the 15 bp YY/RR periodicity does not seem to be a general feature of satellite viruses.

In conclusion, it is unlikely that the 15 bp periodicity is a result of interactions with helper virus proteins or a general feature of satellite viruses.

Histones and other host proteins

The M2 variant of dinucleotide periodicity had an increased transduction in Huh7 and HEK293T cells (Figure 3.18 D). This is also a potentially intriguing discovery, which is admittedly underexplored in this work. The M2 variant is a derivation of M1 and the wt-fragments sequence. Therefore, the increased transduction of only M2 defies simple explanation. Because all sampled transgene cassettes were packaged in an AAV2 capsid, this might suggest a benefit that only becomes apparent after the transgene has been unpacked, such as increased affinity for a necessary host protein. Another hint at an advantage only triggering in the host cell is the enrichment of periodic sequences during the selection from a shuffled library (Figure 3.14). The sequences were subjected to two different selection pressures, namely, only production or production plus transduction. Sequences recovered from transduced Raji cells were more periodic compared to those used for the transduction.

I have already explored the association of transgenes with histones and their modifications in another section of this work (discussed in section 4.1.3). The available literature mostly discusses 10 to 11 bp periodicity correlating with increased association of DNA to histones. It is unknown how a 15 bp periodic sequence might affect binding. Some viruses, like Influenza, use histone mimicry to their advantage. Histone mimicry is based on virus-encoded histone-like proteins, which can interfere with the antiviral immune response [332]. Neither AAV nor the tested transgenes encode for such a protein, but being able to efficiently bind host histones might give the virus a pivotal advantage.

[1] In 2024, this family has been promoted to the order of *Lavidavirales* in the taxonomy browser of the International Committee on Taxonomy of Viruses (ICTV). Technically, therefore, the family no longer exists. However, there are still plenty of sequences that were uploaded to the NCBI nucleotide database under this family tag.

Cap interaction

Having a YY/RR motif at specific distances might either facilitate DNA folding inside the virion or enable specific interaction of the ssDNA genome with the lumen side of the capsid proteins. Therefore, I performed a mass photometry experiment, comparing different fractions of a density gradient of AAV capsids packaging two different transgene cassettes. The two cassettes differed in their stuffer DNA, which either contained periodic AAV2 genomic DNA or a non-periodic section of phage λ DNA. The mass of the filled capsid did not significantly differ in either of the recombinant viruses, but it did in a fraction, which comprised partially filled and empty capsids. The control had equally high peaks for empty and partially filled capsids, whereas the periodic transgene peak profile was skewed towards partially filled capsids. This might be an indication of the DNA preferentially being packaged into the capsids of AAV2. The obvious limitation of this experiment is the lack of replicates. Pulling fractions from an iodixanol gradient can suffer from inaccuracy. Therefore, this experiment should be independently repeated to increase the validity of the conclusions.

Extensive genome capsid interactions are mostly known for viruses that assemble their capsid around their genome, as, for example, in simple spherical RNA viruses such as those from the family *Bromoviridae* [333]. They are also known for ssDNA viruses such as the beak and feather disease virus from the family *Circoviridae*, which has an extensive interaction of its genome with its capsid [334]. The presence of the ssDNA genome allows the formation of an intact capsid, which, in the absence of the genome, forms a complex with a different capsid protein stoichiometry. This interaction is also believed to be the result of this virus forming its capsid around the genome. This makes comparisons to AAV difficult, which is believed to package its genome into preformed capsids [47, 48, 49].

Capsid genome interactions occurring on the lumen side of the capsid have also been demonstrated in Parvoviruses. The canine parvovirus (CPV) has relatively non-specific DNA-binding sites on the lumen side of its capsid. The CPV genome binds to these and folds in a fairly structured manner within the capsid [335]. Similarly, the packaging efficiency of minute virus of mice (another parvovirus) is significantly reduced with foreign DNA, suggesting a preference for its own DNA in a sequence-dependent manner [336]. A study on the Black Wasting disease affecting insect cultures for food production has very recently identified the parvovirus *Zophobas morio* black wasting virus (ZmBWV) as its cause [337]. High-resolution Cryo-EM experiments on ZmBWV have revealed a surprisingly high number of ordered nucleotides, which can be interpreted as the ssDNA genome that interacts with the capsid in multiple DNA-binding pockets.

There is also evidence that the ssDNA genomes of AAV interact with the inside of their capsid, yet here it is not as obvious as for other viruses. In an earlier work, VP1 and VP2 N-termini were implicated in DNA interaction on the lumen of the capsid [338]. In Cryo-EM maps of AAV serotypes AAV4, AAV8, AAVrh32.33, and AAVrh.8, DNA-binding pockets were discovered within the 3-fold axis of the capsid complex [339, 340, 341, 342]. The residues forming the DNA-binding pocket are conserved in all AAV isolates. Only a single nucleotide or dinucleotide (A or AC) was resolved in the binding pocket, suggesting a relatively minor association of the DNA with the capsid. Because the binding pocket is also occupied in recombinant AAV with a non-wild-type genome, it was concluded that this binding is likely not sequence-specific [340, 341]. It was also revealed that the nucleotide is not ordered at lower pH, indicating that the binding might be weakened within an acidifying endosome [340].

Most of this evidence would suggest that the binding of a ssDNA genome with AAV capsid occurs independently of the sequence. However, these arguments are made with specific binding motifs in mind, but not fuzzy patterns such as those described in this thesis. The presence of the YY/RR dinucleotide pattern may facilitate the binding to the aforementioned binding pockets. Future work could aim at elucidating the binding affinity of separate capsid proteins with different stretches of ssDNA with varying periodicities of dinucleotide patterns. This could be achieved by simple electrophoretic mobility shift assays (EMSA).

This explanation attempt does not encompass why the repeat length of the YY/RR pattern in the genus *Dependoparvovirus* deviates from the 10 bp patterns seen in most other genomes.

Rep interaction

Another possibility is the interaction of *Dependoparvovirus* genomes with the Rep protein, which is also necessary for their replication. The pattern of AAV being unique also hints at the binding moiety being unique to AAV. Rep is classified as an SF3 Helicase, but also possesses uncommon properties, such as its unconventional oligomerisation behaviour [41, 42, 43]. Naturally, all Rep proteins have some DNA-binding affinity through their SF3 Helicase activity. Additionally, the binding activity of the large Rep proteins is bolstered by their Origin Binding Domain (OBD). In a Cryo-EM experiment, the heptameric complex of Rep68 proteins bound to a dsDNA substrate was shown to encompass roughly 21 bp of the substrate; however, only 15 bp of which had clearly defined densities [43]. On further sharpening of the map, the final model of the bound DNA had a length of 14 bp, suggesting that a Rep68 heptamer binds tightly to 14 bp of dsDNA. This fits the YY/RR

pattern repeat length of dependoparvoviral DNA at 15 bp. It is possible that the periodicity of *Dependoparvovirus* sequences can aid in the binding of DNA to Rep68 heptamers. The large Rep proteins (Rep68 and Rep78) are necessary for AAV replication as they melt dsDNA and catalyse the terminal resolution within the ITR sequences [343]. Therefore, an increased affinity of the Rep proteins to DNA substrates might have an impact on the formation of replication intermediates or replication efficiency.

In a Southern blot targeting replication intermediates of transgenes with different intensities of periodicities (Figure 3.18 A), a difference in replication intermediates could not be observed. However, an increased interaction with Rep to periodic DNA can not be excluded, since Southern blots are only semi-quantitative.

Another possible hint at Rep interactions would be the increased replication of periodic transgenes in small-scale productions that were not translatable to dish productions (Figure 3.18 C). This might be the result of the more elaborate purification performed for large-scale productions. In the small-scale production, the crude cell lysate is treated with DNase to digest unpackaged DNA. The DNase digestion might not be able to remove DNA as efficiently as the more elaborate purification of the large-scale productions could. Large-scale productions are treated with Benzonase and then purified on an iodixanol gradient. This purification is likely to get rid of a larger portion of unpackaged DNA. An increased affinity for Rep proteins of the transgenic DNA may lead to more efficient replication, but not better packaging.

Finally, the enrichment of periodic sequences upon selection rounds of a shuffled *cap* library (see Figure 3.14) is potentially explicable by a preferred Rep interaction. A sequence with a greater affinity to Rep might be replicated more readily than others and thus be enriched during the replication of the viral genomes. Indeed, we could observe that sequences were becoming more periodic upon sequential production rounds. This does not exclude another effect that would benefit a more periodic sequence during production, such as increased binding affinity with the Cap proteins. However, based on the other evidence discussed here, an involvement of Rep appears to be more likely.

4.3.3 Limitations

Dinucleotides on ssDNA and a possibly more complex pattern

In the analysis of the pattern on the sequences of *Dependoparvovirus*, I have focused on a dinucleotide pattern in the sequence. Based on the wedge model for DNA helix bending [222], dinucleotides can have a great impact on the shape of a DNA molecule, without

affecting much of the sequence. There are two major limitations in this approach to be considered. First, the genomes of viruses in the family of *Parvoviridae* and the genus of *Dependoparvovirus* consist of ssDNA. The hypotheses on the dinucleotide periodicities in pro- and eukaryotic genomes are made with the presumption of a DNA double helix. How a nucleotide pattern might influence the structure or folding of a ssDNA molecule is unknown, although there is evidence of viral single-stranded nucleic acid genomes taking the shape of the capsid interior [344]. Secondly, limiting the scope to only dinucleotides could have masked the presence of other, more expansive patterns. The described dinucleotide pattern might be a result of a more complicated pattern. This became especially apparent in the non-improvement of the M3 cassette (Figure 3.17), which had a 'perfect' YY/RR pattern designed agnostically to an AAV reference sequence. How the pattern was implemented on transgene cassettes seems to make no difference in titer or transduction. Therefore, the true sequence pattern is still not clear. However, the true pattern might have been unknowingly incorporated into cassettes M1 and M2, explaining their persisting advantages. Unravelling the true pattern could be undertaken by exploring longer sequence patterns. For example, longer stretches of homopolymeric A or T sequences are significantly enriched in nucleosome-free regions, as they probably inhibit the winding of the DNA around the histone octamer [345, 346]. These DNA sequences are likely more rigid because of a strong association of water molecules or cations within their minor groove. Similar homopolymeric stretches could also be observed in the masks that were used for the generation of M1 and M2 (Figure 3.17). Future research should analyse the presence and importance of these sequence stretches on the genomes of AAV and the genus *Dependoparvovirus*, which were also present in the cassettes M1 and M2.

The similarity of wt-fragments to the AAV2 coding regions

I found periodic transgene cassettes to produce a higher titer in small-scale productions (section 3.2.4). The transgene cassettes wt-fragments, M1, and M2 are all based on inert sequences originating from the AAV2 *rep* and *cap* genes. It is possible that the similarity of these cassettes to the Rep/Cap plasmid used for production could have led to recombination events between them. Even though unlikely, such recombinations might have led to the creation of semi-replicative cassettes that might have restored some of the Rep or Cap functions. Going forward, this could be tested by direct nanopore sequencing of the virus production to analyse the quality of the packaged transgenic DNA.

4.3.4 An outlook on the sequence patterns in AAV

The here discovered YY/RR dinucleotide pattern is potentially an immensely interesting factor for gene therapy. A major drawback of contemporary gene therapy drugs is their massive price tags, which are (among other factors) caused by the elaborate procedures necessary for their production [347]. Altering the stuffer sequences of transgenes could lower production costs, which would be in the interest of manufacturers and ultimately patients. In the work presented in this thesis, my colleagues and I have gathered strong evidence for the dinucleotide periodicity having a biologically relevant function. This pattern might be worked into the coding regions of therapeutic transgenes through codon optimisation or more easily imprinted on stuffer DNA.

Bibliography

[1] R. W. Atchison, B. C. Casto, and W. M. Hammon. “ADENOVIRUS-ASSOCIATED DEFECTIVE VIRUS PARTICLES”. In: *Science* 149.3685 (Aug. 13, 1965), pp. 754–756. ISSN: 0036-8075. DOI: 10.1126/science.149.3685.754.

[2] M. D. Hoggan, N. R. Blacklow, and W. P. Rowe. “Studies of small DNA viruses found in various adenovirus preparations: physical, biological, and immunological characteristics.” In: *Proc Natl Acad Sci U S A* 55.6 (June 1966), pp. 1467–1474. ISSN: 0027-8424.

[3] C. J. Marcus, C. A. Laughlin, and B. J. Carter. “Adeno-associated virus RNA transcription in vivo”. In: *Eur J Biochem* 121.1 (Dec. 1981), pp. 147–154. ISSN: 0014-2956. DOI: 10.1111/j.1432-1033.1981.tb06443.x.

[4] U. Bantel-Schaal and H. zur Hausen. “Characterization of the DNA of a defective human parvovirus isolated from a genital site”. In: *Virology* 134.1 (Apr. 15, 1984), pp. 52–63. ISSN: 0042-6822. DOI: 10.1016/0042-6822(84)90271-x.

[5] M. A. Labow, P. L. Hermonat, and K. I. Berns. “Positive and negative autoregulation of the adeno-associated virus type 2 genome”. In: *J Virol* 60.1 (Oct. 1986), pp. 251–258. ISSN: 0022-538X. DOI: 10.1128/JVI.60.1.251-258.1986.

[6] S. F. Cotmore et al. “ICTV Virus Taxonomy Profile: Parvoviridae”. In: *Journal of General Virology* 100.3 (2019), pp. 367–368. ISSN: 1465-2099. DOI: 10.1099/jgv.0.001212.

[7] E. D. Heegaard and K. E. Brown. “Human parvovirus B19”. In: *Clin Microbiol Rev* 15.3 (July 2002), pp. 485–505. ISSN: 0893-8512. DOI: 10.1128/CMR.15.3.485-505.2002.

[8] *Genus: Dependoparvovirus | ICTV*. URL: https://ictv.global/report/chapter/parvoviridae/parvoviridae/dependoparvovirus (visited on 04/07/2025).

[9] Z. Zadori, J. Erdei, J. Nagy, and J. Kisary. “Characteristics of the genome of goose parvovirus”. In: *Avian Pathol* 23.2 (June 1994), pp. 359–364. ISSN: 0307-9457. DOI: 10.1080/03079459408419004.

[10] J. J. Pénzes, H. T. Pham, M. Benkö, and P. Tijssen. “Novel parvoviruses in reptiles and genome sequence of a lizard parvovirus shed light on Dependoparvovirus genus evolution”. In: *J Gen Virol* 96.9 (Sept. 2015), pp. 2769–2779. ISSN: 1465-2099. DOI: 10.1099/vir.0.000215.

[11] A. F. Meier, C. Fraefel, and M. Seyffert. “The Interplay between Adeno-Associated Virus and Its Helper Viruses”. In: *Viruses* 12.6 (June 19, 2020), p. 662. ISSN: 1999-4915. DOI: 10.3390/v12060662.

[12] Z. Wang et al. “Human Bocavirus 1 Is a Novel Helper for Adeno-associated Virus Replication”. In: *Journal of Virology* 91.18 (Aug. 24, 2017), 10.1128/jvi.00710–17. DOI: 10.1128/jvi.00710-17.

[13] M. Cao et al. “HPV-16 E1, E2 and E6 each complement the Ad5 helper gene set, increasing rAAV2 and wt AAV2 production”. In: *Gene Ther* 19.4 (Apr. 2012), pp. 418–424. ISSN: 1476-5462. DOI: 10.1038/gt.2011.115.

[14] M. Urabe, C. Ding, and R. M. Kotin. “Insect Cells as a Factory to Produce Adeno-Associated Virus Type 2 Vectors”. In: *Human Gene Therapy* 13.16 (Nov. 2002), pp. 1935–1943. ISSN: 1043-0342. DOI: 10.1089/10430340260355347.

[15] R. H. Smith, J. R. Levy, and R. M. Kotin. “A Simplified Baculovirus-AAV Expression Vector System Coupled With One-step Affinity Purification Yields High-titer rAAV Stocks From Insect Cells”. In: *Molecular Therapy* 17.11 (Nov. 1, 2009), pp. 1888–1896. ISSN: 1525-0016. DOI: 10.1038/mt.2009.128.

[16] S. Grosse et al. “Relevance of Assembly-Activating Protein for Adeno-associated Virus Vector Production and Capsid Protein Stability in Mammalian and Insect Cells”. In: *J Virol* 91.20 (Oct. 15, 2017), e01198–17. ISSN: 1098-5514. DOI: 10.1128/JVI.01198-17.

[17] J. M. Timpe, K. C. Verrill, and J. P. Trempe. “Effects of Adeno-Associated Virus on Adenovirus Replication and Gene Expression during Coinfection”. In: *J Virol* 80.16 (Aug. 2006), pp. 7807–7815. ISSN: 0022-538X. DOI: 10.1128/JVI.00198-06.

[18] E. Hildebrandt et al. “Evolution of dependoparvoviruses across geological timescales—implications for design of AAV-based gene therapy vectors”. In: *Virus Evolution* 6.2 (July 1, 2020), veaa043. ISSN: 2057-1577. DOI: 10.1093/ve/veaa043.

[19] A. Leeks, S. A. West, and M. Ghoul. “The evolution of cheating in viruses”. In: *Nat Commun* 12 (Nov. 26, 2021), p. 6928. ISSN: 2041-1723. DOI: 10.1038/s41467-021-27293-6.

[20] *Satellites and Other Virus-dependent Nucleic Acids | ICTV*. URL: https://ictv.global/report_9th/subviral/Satellites-introduction (visited on 04/07/2025).

[21] E. V. Ryabov et al. “Umbravirus gene expression helps potato leafroll virus to invade mesophyll tissues and to be transmitted mechanically between plants”. In: *Virology* 286.2 (Aug. 1, 2001), pp. 363–372. ISSN: 0042-6822. DOI: 10.1006/viro.2001.0982.

[22] M. Taliansky, L. Torrance, and N. O. Kalinina. “Role of plant virus movement proteins”. In: *Methods Mol Biol* 451 (2008), pp. 33–54. ISSN: 1064-3745. DOI: 10.1007/978-1-59745-102-4_3.

[23] B. Frígols et al. “Virus Satellites Drive Viral Evolution and Ecology”. In: *PLoS Genet* 11.10 (Oct. 23, 2015), e1005609. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1005609.

[24] J. Zhang et al. “Satellite Subgenomic Particles Are Key Regulators of Adeno-Associated Virus Life Cycle”. In: *Viruses* 13.6 (June 21, 2021), p. 1185. ISSN: 1999-4915. DOI: 10.3390/v13061185.

[25] Y. R. Nam et al. “Distinguishing between DNA-Loaded Full and Empty Capsids of Adeno-Associated Virus with Atomic Force Microscopy Imaging”. In: *Langmuir* 39.19 (May 16, 2023), pp. 6740–6747. ISSN: 0743-7463. DOI: 10.1021/acs.langmuir.3c00241.

[26] A. Bennett, M. Mietzsch, and M. Agbandje-McKenna. “Understanding capsid assembly and genome packaging for adeno-associated viruses”. In: *Future Virology* 12.6 (June 2017), pp. 283–297. ISSN: 1746-0794. DOI: 10.2217/fvl-2017-0011.

[27] K. I. Berns and S. Adler. “Separation of two types of adeno-associated virus particles containing complementary polynucleotide chains”. In: *J Virol* 9.2 (Feb. 1972), pp. 394–396. ISSN: 0022-538X. DOI: 10.1128/JVI.9.2.394-396.1972.

[28] Y. Chen 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”. In: *Microorganisms* 12.2 (Feb. 1, 2024), p. 310. ISSN: 2076-2607. DOI: 10.3390/microorganisms12020310.

[29] E. M. Shitik, I. K. Shalik, and D. V. Yudkin. “AAV- based vector improvements unrelated to capsid protein modification”. In: *Front Med (Lausanne)* 10 (2023), p. 1106085. ISSN: 2296-858X. DOI: 10.3389/fmed.2023.1106085.

[30] F. B. Johnson, H. L. Ozer, and M. D. Hoggan. “Structural Proteins of Adenovirus-Associated Virus Type 3”. In: *Journal of Virology* 8.6 (Dec. 1971), pp. 860–863. DOI: 10.1128/jvi.8.6.860-863.1971.

[31] H. Oyama et al. “Characterization of Adeno-Associated Virus Capsid Proteins with Two Types of VP3-Related Components by Capillary Gel Electrophoresis and Mass Spectrometry”. In: *Hum Gene Ther* 32.21 (Nov. 1, 2021), pp. 1403–1416. ISSN: 1043-0342. DOI: 10.1089/hum.2021.009.

[32] J. P. Trempe and B. J. Carter. “Alternate mRNA splicing is required for synthesis of adeno-associated virus VP1 capsid protein”. In: *Journal of Virology* 62.9 (Sept. 1988), pp. 3356–3363. DOI: 10.1128/jvi.62.9.3356-3363.1988.

[33] P. Cassinotti, M. Weitz, and J. D. Tratschin. “Organization of the adeno-associated virus (AAV) capsid gene: mapping of a minor spliced mRNA coding for virus capsid protein 1”. In: *Virology* 167.1 (Nov. 1988), pp. 176–184. ISSN: 0042-6822.

[34] F. Sonntag et al. “The Assembly-Activating Protein Promotes Capsid Assembly of Different Adeno-Associated Virus Serotypes□”. In: *J Virol* 85.23 (Dec. 2011), pp. 12686–12697. ISSN: 0022-538X. DOI: 10.1128/JVI.05359-11.

[35] L. F. Earley et al. “Adeno-associated Virus (AAV) Assembly-Activating Protein Is Not an Essential Requirement for Capsid Assembly of AAV Serotypes 4, 5, and 11”. In: *Journal of Virology* 91.3 (Jan. 18, 2017), 10.1128/jvi.01980–16. DOI: 10.1128/jvi.01980-16.

[36] P. J. Ogden, E. D. Kelsic, S. Sinai, and G. M. Church. “Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design”. In: *Science* 366.6469 (Nov. 29, 2019), pp. 1139–1143. ISSN: 1095-9203. DOI: 10.1126/science.aaw2900.

[37] L. Galibert et al. “Functional roles of the membrane-associated AAV protein MAAP”. In: *Sci Rep* 11 (Nov. 4, 2021). ISSN: 2045-2322. DOI: 10.1038/s41598-021-01220-7.

[38] C. Aksu Kuz et al. “Role of the membrane-associated accessory protein (MAAP) in adeno-associated virus (AAV) infection”. In: *J Virol* 98.6 (June 13, 2024), e0063324. ISSN: 1098-5514. DOI: 10.1128/jvi.00633-24.

[39] D. S. Im and N. Muzyczka. “The AAV origin binding protein Rep68 is an ATP-dependent site-specific endonuclease with DNA helicase activity”. In: *Cell* 61.3 (May 4, 1990), pp. 447–457. ISSN: 0092-8674. DOI: 10.1016/0092-8674(90)90526-k.

[40] A. B. Hickman and F. Dyda. “Binding and unwinding: SF3 viral helicases”. In: *Current Opinion in Structural Biology*. Folding and binding / Protein-nucleic acid interactions 15.1 (Feb. 1, 2005), pp. 77–85. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2004.12.001.

[41] F. Zarate-Perez 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”. In: *PLoS Pathog* 8.6 (June 14, 2012), e1002764. ISSN: 1553-7366. DOI: 10.1371/journal.ppat.1002764.

[42] F. Zarate-Perez et al. “Oligomeric Properties of Adeno-Associated Virus Rep68 Reflect Its Multifunctionality”. In: *J Virol* 87.2 (Jan. 2013), pp. 1232–1241. ISSN: 0022-538X. DOI: 10.1128/JVI.02441-12.

[43] R. Jaiswal et al. “Cryo-EM structure of AAV2 Rep68 bound to integration site AAVS1: insights into the mechanism of DNA melting”. In: *Nucleic Acids Res* 53.3 (Jan. 24, 2025), gkaf033. ISSN: 1362-4962. DOI: 10.1093/nar/gkaf033.

[44] K. I. Berns. “Parvovirus replication”. In: *Microbiological Reviews* 54.3 (Sept. 1990), pp. 316–329. DOI: 10.1128/mr.54.3.316-329.1990.

[45] D. J. Pereira, D. M. McCarty, and N. Muzyczka. “The adeno-associated virus (AAV) Rep protein acts as both a repressor and an activator to regulate AAV transcription during a productive infection”. In: *J Virol* 71.2 (Feb. 1997), pp. 1079–1088. ISSN: 0022-538X. DOI: 10.1128/JVI.71.2.1079-1088.1997.

[46] A. Recchia and F. Mavilio. “Site-specific integration by the adeno-associated virus rep protein”. In: *Curr Gene Ther* 11.5 (Oct. 2011), pp. 399–405. ISSN: 1875-5631. DOI: 10.2174/156652311797415809.

[47] J. A. King, R. Dubielzig, D. Grimm, and J. A. Kleinschmidt. “DNA helicase-mediated packaging of adeno-associated virus type 2 genomes into preformed capsids”. In: *EMBO J* 20.12 (June 15, 2001), pp. 3282–3291. ISSN: 0261-4189. DOI: 10.1093/emboj/20.12.3282.

[48] M. Yoon-Robarts et al. “Residues within the B’ motif are critical for DNA binding by the superfamily 3 helicase Rep40 of adeno-associated virus type 2”. In: *J Biol Chem* 279.48 (Nov. 26, 2004), pp. 50472–50481. ISSN: 0021-9258. DOI: 10.1074/jbc.M403900200.

[49] S. Bleker, F. Sonntag, and J. A. Kleinschmidt. “Mutational analysis of narrow pores at the fivefold symmetry axes of adeno-associated virus type 2 capsids reveals a dual role in genome packaging and activation of phospholipase A2 activity”. In: *J Virol* 79.4 (Feb. 2005), pp. 2528–2540. ISSN: 0022-538X. DOI: 10.1128/JVI.79.4.2528-2540.2005.

[50] J.-H. Wang et al. “Adeno-associated virus as a delivery vector for gene therapy of human diseases”. In: *Sig Transduct Target Ther* 9.1 (Apr. 3, 2024), pp. 1–33. ISSN: 2059-3635. DOI: 10.1038/s41392-024-01780-w.

[51] M. Senior. “After Glybera’s withdrawal, what’s next for gene therapy?” In: *Nature Biotechnology* 35.6 (June 1, 2017), pp. 491–492. ISSN: 1546-1696. DOI: 10.1038/nbt0617-491.

[52] C. f. B. E. a. Research. *Approved Cellular and Gene Therapy Products*. June 3, 2025. URL: https://www.fda.gov/vaccines-blood-biologics/cellular-gene-therapy-products/approved-cellular-and-gene-therapy-products (visited on 04/02/2025).

[53] S. Morfopoulou et al. “Genomic investigations of unexplained acute hepatitis in children”. In: *Nature* 617.7961 (May 2023), pp. 564–573. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06003-w.

[54] A. Ho et al. “Adeno-associated virus 2 infection in children with non-A–E hepatitis”. In: *Nature* 617.7961 (May 2023), pp. 555–563. ISSN: 1476-4687. DOI: 10.1038/s41586-023-05948-2.

[55] V. Servellita et al. “Adeno-associated virus type 2 in US children with acute severe hepatitis”. In: *Nature* 617.7961 (May 2023), pp. 574–580. ISSN: 1476-4687. DOI: 10.1038/s41586-023-05949-1.

[56] K.-i. Iwata et al. “Association between adeno-associated virus 2 and severe acute hepatitis of unknown etiology in Japanese children”. In: *Journal of Infection and Chemotherapy* 31.1 (Jan. 1, 2025), p. 102462. ISSN: 1341-321X. DOI: 10.1016/j.jiac.2024.07.002.

[57] R. J. Samulski, K. I. Berns, M. Tan, and N. Muzyczka. “Cloning of adeno-associated virus into pBR322: rescue of intact virus from the recombinant plasmid in human cells.” In: *Proc Natl Acad Sci U S A* 79.6 (Mar. 1982), pp. 2077–2081. ISSN: 0027-8424.

[58] C. A. Laughlin, J. D. Tratschin, H. Coon, and B. J. Carter. “Cloning of infectious adeno-associated virus genomes in bacterial plasmids”. In: *Gene* 23.1 (July 1983), pp. 65–73. ISSN: 0378-1119. DOI: 10.1016/0378-1119(83)90217-2.

[59] P. L. Hermonat and N. Muzyczka. “Use of adeno-associated virus as a mammalian DNA cloning vector: transduction of neomycin resistance into mammalian tissue culture cells”. In: *Proc Natl Acad Sci U S A* 81.20 (Oct. 1984), pp. 6466–6470. ISSN: 0027-8424. DOI: 10.1073/pnas.81.20.6466.

[60] S. K. McLaughlin, P. Collis, P. L. Hermonat, and N. Muzyczka. “Adeno-associated virus general transduction vectors: analysis of proviral structures”. In: *J Virol* 62.6 (June 1988), pp. 1963–1973. ISSN: 0022-538X. DOI: 10.1128/JVI.62.6.1963-1973.1988.

[61] W. Xiao et al. “Gene therapy vectors based on adeno-associated virus type 1”. In: *J Virol* 73.5 (May 1999), pp. 3994–4003. ISSN: 0022-538X. DOI: 10.1128/JVI.73.5.3994-4003.1999.

[62] G.-P. Gao et al. “Novel adeno-associated viruses from rhesus monkeys as vectors for human gene therapy”. In: *Proc Natl Acad Sci U S A* 99.18 (Sept. 3, 2002), pp. 11854–11859. ISSN: 0027-8424. DOI: 10.1073/pnas.182412299.

[63] C. Burger 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”. In: *Mol Ther* 10.2 (Aug. 2004), pp. 302–317. ISSN: 1525-0016. DOI: 10.1016/j.ymthe.2004.05.024.

[64] X. Xiao, J. Li, and R. J. Samulski. “Production of high-titer recombinant adeno-associated virus vectors in the absence of helper adenovirus”. In: *J Virol* 72.3 (Mar. 1998), pp. 2224–2232. ISSN: 0022-538X. DOI: 10.1128/JVI.72.3.2224-2232.1998.

[65] T. Matsushita et al. “Adeno-associated virus vectors can be efficiently produced without helper virus”. In: *Gene Ther* 5.7 (July 1998), pp. 938–945. ISSN: 0969-7128. DOI: 10.1038/sj.gt.3300680.

[66] R. J. Samulski, L. S. Chang, and T. Shenk. “Helper-free stocks of recombinant adeno-associated viruses: normal integration does not require viral gene expression”. In: *J Virol* 63.9 (Sept. 1989), pp. 3822–3828. ISSN: 0022-538X. DOI: 10.1128/JVI.63.9.3822-3828.1989.

[67] F. L. Graham, J. Smiley, W. C. Russell, and R. Nairn. “Characteristics of a human cell line transformed by DNA from human adenovirus type 5”. In: *J Gen Virol* 36.1 (July 1977), pp. 59–74. ISSN: 0022-1317. DOI: 10.1099/0022-1317-36-1-59.

[68] M. A. Kay et al. “Evidence for gene transfer and expression of factor IX in haemophilia B patients treated with an AAV vector”. In: *Nat Genet* 24.3 (Mar. 2000), pp. 257–261. ISSN: 1061-4036. DOI: 10.1038/73464.

[69] C. S. Manno et al. “AAV-mediated factor IX gene transfer to skeletal muscle in patients with severe hemophilia B”. In: *Blood* 101.8 (Apr. 15, 2003), pp. 2963–2972. ISSN: 0006-4971. DOI: 10.1182/blood-2002-10-3296.

[70] C. S. Manno et al. “Successful transduction of liver in hemophilia by AAV-Factor IX and limitations imposed by the host immune response”. In: *Nature Medicine* 12.3 (Mar. 2006), pp. 342–347. ISSN: 1546-170X. DOI: 10.1038/nm1358.

[71] F. Mingozzi et al. “Induction of immune tolerance to coagulation factor IX antigen by in vivo hepatic gene transfer”. In: *J Clin Invest* 111.9 (May 2003), pp. 1347–1356. ISSN: 0021-9738. DOI: 10.1172/JCI16887.

[72] F. Mingozzi et al. “CD8(+) T-cell responses to adeno-associated virus capsid in humans”. In: *Nat Med* 13.4 (Apr. 2007), pp. 419–422. ISSN: 1078-8956. DOI: 10.1038/nm1549.

[73] K. Rapti and D. Grimm. “Adeno-Associated Viruses (AAV) and Host Immunity - A Race Between the Hare and the Hedgehog”. In: *Front Immunol* 12 (2021), p. 753467. ISSN: 1664-3224. DOI: 10.3389/fimmu.2021.753467.

[74] T. Weber. “Anti-AAV Antibodies in AAV Gene Therapy: Current Challenges and Possible Solutions”. In: *Front Immunol* 12 (2021), p. 658399. ISSN: 1664-3224. DOI: 10.3389/fimmu.2021.658399.

[75] A. Girod et al. “Genetic capsid modifications allow efficient re-targeting of adeno-associated virus type 2”. In: *Nat Med* 5.9 (Sept. 1999), pp. 1052–1056. ISSN: 1078-8956. DOI: 10.1038/12491.

[76] L. Perabo et al. “In vitro selection of viral vectors with modified tropism: the adeno-associated virus display”. In: *Molecular Therapy* 8.1 (July 1, 2003), pp. 151–157. ISSN: 1525-0016. DOI: 10.1016/S1525-0016(03)00123-0.

[77] O. J. Müller et al. “Random peptide libraries displayed on adeno-associated virus to select for targeted gene therapy vectors”. In: *Nat Biotechnol* 21.9 (Sept. 2003), pp. 1040–1046. ISSN: 1546-1696. DOI: 10.1038/nbt856.

[78] J. E. Rabinowitz et al. “Cross-Dressing the Virion: the Transcapsidation of Adeno-Associated Virus Serotypes Functionally Defines Subgroups”. In: *J Virol* 78.9 (May 2004), pp. 4421–4432. ISSN: 0022-538X. DOI: 10.1128/JVI.78.9.4421-4432.2004.

[79] N. Maheshri, J. T. Koerber, B. K. Kaspar, and D. V. Schaffer. “Directed evolution of adeno-associated virus yields enhanced gene delivery vectors”. In: *Nat Biotechnol* 24.2 (Feb. 2006), pp. 198–204. ISSN: 1546-1696. DOI: 10.1038/nbt1182.

[80] K. J. D. A. Excoffon et al. “Directed evolution of adeno-associated virus to an infectious respiratory virus”. In: *Proc Natl Acad Sci U S A* 106.10 (Mar. 10, 2009), pp. 3865–3870. ISSN: 0027-8424. DOI: 10.1073/pnas.0813365106.

[81] D. Grimm et al. “In Vitro and In Vivo Gene Therapy Vector Evolution via Multispecies Interbreeding and Retargeting of Adeno-Associated Viruses”. In: *Journal of Virology* 82.12 (June 15, 2008), pp. 5887–5911. DOI: 10.1128/JVI.00254-08.

[82] J. T. Koerber, J.-H. Jang, and D. V. Schaffer. “DNA shuffling of adeno-associated virus yields functionally diverse viral progeny”. In: *Mol Ther* 16.10 (Oct. 2008), pp. 1703–1709. ISSN: 1525-0024. DOI: 10.1038/mt.2008.167.

[83] W. Li et al. “Engineering and Selection of Shuffled AAV Genomes: A New Strategy for Producing Targeted Biological Nanoparticles”. In: *Mol Ther* 16.7 (July 2008), pp. 1252–1260. ISSN: 1525-0016. DOI: 10.1038/mt.2008.100.

[84] W. P. Stemmer. “Rapid evolution of a protein in vitro by DNA shuffling”. In: *Nature* 370.6488 (Aug. 4, 1994), pp. 389–391. ISSN: 0028-0836. DOI: 10.1038/370389a0.

[85] L. Lisowski et al. “Selection and evaluation of clinically relevant AAV variants in a xenograft liver model”. In: *Nature* 506.7488 (Feb. 20, 2014), pp. 382–386. ISSN: 0028-0836. DOI: 10.1038/nature12875.

[86] J. El Andari et al. “Semirational bioengineering of AAV vectors with increased potency and specificity for systemic gene therapy of muscle disorders”. In: *Sci Adv* 8.38 (Sept. 23, 2022), eabn4704. ISSN: 2375-2548. DOI: 10.1126/sciadv.abn4704.

[87] C. Summerford and R. J. Samulski. “Membrane-Associated Heparan Sulfate Proteoglycan Is a Receptor for Adeno-Associated Virus Type 2 Virions”. In: *J Virol* 72.2 (Feb. 1998), pp. 1438–1445. ISSN: 0022-538X.

[88] B. P. Dhungel, C. G. Bailey, and J. E. J. Rasko. “Journey to the Center of the Cell: Tracing the Path of AAV Transduction”. In: *Trends in Molecular Medicine* 27.2 (Feb. 1, 2021), pp. 172–184. ISSN: 1471-4914, 1471-499X. DOI: 10.1016/j.molmed.2020.09.010.

[89] N. Kaludov et al. “Adeno-associated virus serotype 4 (AAV4) and AAV5 both require sialic acid binding for hemagglutination and efficient transduction but differ in sialic acid linkage specificity”. In: *J Virol* 75.15 (Aug. 2001), pp. 6884–6893. ISSN: 0022-538X. DOI: 10.1128/JVI.75.15.6884-6893.2001.

[90] S. Shen et al. “Terminal N-linked galactose is the primary receptor for adeno-associated virus 9”. In: *J Biol Chem* 286.15 (Apr. 15, 2011), pp. 13532–13540. ISSN: 1083-351X. DOI: 10.1074/jbc.M110.210922.

[91] S. Pillay et al. “An essential receptor for adeno-associated virus infection”. In: *Nature* 530.7588 (Feb. 4, 2016), pp. 108–112. ISSN: 0028-0836. DOI: 10.1038/nature16465.

[92] A. M. Dudek et al. “An Alternate Route for Adeno-associated Virus (AAV) Entry Independent of AAV Receptor”. In: *J Virol* 92.7 (Apr. 1, 2018), e02213–17. ISSN: 1098-5514. DOI: 10.1128/JVI.02213-17.

[93] A. M. Dudek et al. “GPR108 Is a Highly Conserved AAV Entry Factor”. In: *Mol Ther* 28.2 (Feb. 5, 2020), pp. 367–381. ISSN: 1525-0024. DOI: 10.1016/j.ymthe.2019.11.005.

[94] A. Asokan et al. “Adeno-associated virus type 2 contains an integrin alpha5beta1 binding domain essential for viral cell entry”. In: *J Virol* 80.18 (Sept. 2006), pp. 8961–8969. ISSN: 0022-538X. DOI: 10.1128/JVI.00843-06.

[95] B. Akache et al. “The 37/67-kilodalton laminin receptor is a receptor for adeno-associated virus serotypes 8, 2, 3, and 9”. In: *J Virol* 80.19 (Oct. 2006), pp. 9831–9836. ISSN: 0022-538X. DOI: 10.1128/JVI.00878-06.

[96] D. Duan et al. “Dynamin Is Required for Recombinant Adeno-Associated Virus Type 2 Infection”. In: *J Virol* 73.12 (Dec. 1999), pp. 10371–10376. ISSN: 0022-538X.

[97] A. D. Sanlioglu et al. “Novel approaches to augment adeno-associated virus type-2 endocytosis and transduction”. In: *Virus Research* 104.1 (Aug. 1, 2004), pp. 51–59. ISSN: 0168-1702. DOI: 10.1016/j.virusres.2004.03.002.

[98] M. Nonnenmacher and T. Weber. “Adeno-associated virus 2 infection requires endocytosis through the CLIC/GEEC pathway”. In: *Cell Host Microbe* 10.6 (Dec. 15, 2011), pp. 563–576. ISSN: 1934-6069. DOI: 10.1016/j.chom.2011.10.014.

[99] J. M. Riyad and T. Weber. “Intracellular trafficking of adeno-associated virus (AAV) vectors: challenges and future directions”. In: *Gene Ther* (Mar. 3, 2021), pp. 1–14. ISSN: 1476-5462. DOI: 10.1038/s41434-021-00243-z.

[100] F. Sonntag et al. “Adeno-Associated Virus Type 2 Capsids with Externalized VP1/VP2 Trafficking Domains Are Generated prior to Passage through the Cytoplasm and Are Maintained until Uncoating Occurs in the Nucleus”. In: *Journal of Virology* 80.22 (Nov. 2006), pp. 11040–11054. DOI: 10.1128/jvi.01056-06.

[101] B. Venkatakrishnan et al. “Structure and Dynamics of Adeno-Associated Virus Serotype 1 VP1-Unique N-Terminal Domain and Its Role in Capsid Trafficking”. In: *Journal of Virology* 87.9 (May 2013), pp. 4974–4984. DOI: 10.1128/jvi.02524-12.

[102] A. Girod et al. “The VP1 capsid protein of adeno-associated virus type 2 is carrying a phospholipase A2 domain required for virus infectivity”. In: *J Gen Virol* 83 (Pt 5 May 2002), pp. 973–978. ISSN: 0022-1317. DOI: 10.1099/0022-1317-83-5-973.

[103] K. Gliwa et al. “Biophysical and structural insights into AAV genome ejection”. In: *Journal of Virology* 0.0 (Feb. 5, 2025), e00899–24. DOI: 10.1128/jvi.00899-24.

[104] M. Penaud-Budloo et al. “Adeno-Associated Virus Vector Genomes Persist as Episomal Chromatin in Primate Muscle”. In: *Journal of Virology* (Aug. 2008). DOI: 10.1128/JVI.00649-08.

[105] S. Fong et al. “Interindividual variability in transgene mRNA and protein production following adeno-associated virus gene therapy for hemophilia A”. In: *Nat Med* (Apr. 11, 2022). ISSN: 1546-170X. DOI: 10.1038/s41591-022-01751-0.

[106] F. K. Ferrari, T. Samulski, T. Shenk, and R. J. Samulski. “Second-strand synthesis is a rate-limiting step for efficient transduction by recombinant adeno-associated virus vectors.” In: *J Virol* 70.5 (May 1996), pp. 3227–3234. ISSN: 0022-538X.

[107] Z. Yan, R. Zak, Y. Zhang, and J. F. Engelhardt. “Inverted terminal repeat sequences are important for intermolecular recombination and circularization of adeno-associated virus genomes”. In: *J Virol* 79.1 (Jan. 2005), pp. 364–379. ISSN: 0022-538X. DOI: 10.1128/JVI.79.1.364-379.2005.

[108] V. W. Choi, R. J. Samulski, and D. M. McCarty. “Effects of Adeno-Associated Virus DNA Hairpin Structure on Recombination”. In: *Journal of Virology* (June 1, 2005). DOI: 10.1128/JVI.79.11.6801-6807.2005.

[109] D. Duan et al. “Circular intermediates of recombinant adeno-associated virus have defined structural characteristics responsible for long-term episomal persistence in muscle tissue”. In: *J Virol* 72.11 (Nov. 1998), pp. 8568–8577. ISSN: 0022-538X. DOI: 10.1128/JVI.72.11.8568-8577.1998.

[110] X. Sun et al. “Molecular analysis of vector genome structures after liver transduction by conventional and self-complementary adeno-associated viral serotype vectors in murine and nonhuman primate models”. In: *Hum Gene Ther* 21.6 (June 2010), pp. 750–761. ISSN: 1557-7422. DOI: 10.1089/hum.2009.214.

[111] C. E. Thomas, T. A. Storm, Z. Huang, and M. A. Kay. “Rapid Uncoating of Vector Genomes Is the Key to Efficient Liver Transduction with Pseudotyped Adeno-Associated Virus Vectors”. In: *J Virol* 78.6 (Mar. 2004), pp. 3110–3122. ISSN: 0022-538X. DOI: 10.1128/JVI.78.6.3110-3122.2004.

[112] J.-J. Kim et al. “AAV capsid prioritization in normal and steatotic human livers maintained by machine perfusion”. In: *Nat Biotechnol* (Jan. 29, 2025), pp. 1–13. ISSN: 1546-1696. DOI: 10.1038/s41587-024-02523-6.

[113] J. Yang et al. “Concatamerization of Adeno-Associated Virus Circular Genomes Occurs through Intermolecular Recombination”. In: *J Virol* 73.11 (Nov. 1999), pp. 9468–9477. ISSN: 0022-538X.

[114] V. W. Choi, D. M. McCarty, and R. J. Samulski. “Host Cell DNA Repair Pathways in Adeno-Associated Viral Genome Processing”. In: *J Virol* 80.21 (Nov. 2006), pp. 10346–10356. ISSN: 0022-538X. DOI: 10.1128/JVI.00841-06.

[115] M. P. Cataldi and D. M. McCarty. “Hairpin-end conformation of adeno-associated virus genome determines interactions with DNA-repair pathways”. In: *Gene Ther* 20.6 (June 2013), pp. 686–693. ISSN: 1476-5462. DOI: 10.1038/gt.2012.86.

[116] R. Kotin, R. Linden, and K. Berns. “Characterization of a preferred site on human chromosome 19q for integration of adeno□associated virus DNA by non□homologous recombination.” In: *The EMBO Journal* 11.13 (Dec. 1992), pp. 5071–5078. ISSN: 0261-4189. DOI: 10.1002/j.1460-2075.1992.tb05614.x.

[117] M. D. Weitzman, S. R. Kyöstiö, R. M. Kotin, and R. A. Owens. "Adeno-associated virus (AAV) Rep proteins mediate complex formation between AAV DNA and its integration site in human DNA". In: *PNAS* 91.13 (June 21, 1994), pp. 5808–5812. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.91.13.5808.

[118] D. Hüser, S. Weger, and R. Heilbronn. "Kinetics and frequency of adeno-associated virus site-specific integration into human chromosome 19 monitored by quantitative real-time PCR". In: *J Virol* 76.15 (Aug. 2002), pp. 7554–7559. ISSN: 0022-538X. DOI: 10.1128/jvi.76.15.7554-7559.2002.

[119] H. Hamilton, J. Gomos, K. I. Berns, and E. Falck-Pedersen. "Adeno-associated virus site-specific integration and AAVS1 disruption". In: *J Virol* 78.15 (Aug. 2004), pp. 7874–7882. ISSN: 0022-538X. DOI: 10.1128/JVI.78.15.7874-7882.2004.

[120] R. S. Wonderling and R. A. Owens. "Binding sites for adeno-associated virus Rep proteins within the human genome." In: *J Virol* 71.3 (Mar. 1997), pp. 2528–2534. ISSN: 0022-538X.

[121] D. M. McCarty, S. M. Young, and R. J. Samulski. "Integration of adeno-associated virus (AAV) and recombinant AAV vectors". In: *Annu Rev Genet* 38 (2004), pp. 819–845. ISSN: 0066-4197. DOI: 10.1146/annurev.genet.37.110801.143717.

[122] B. C. Schnepp et al. "Characterization of Adeno-Associated Virus Genomes Isolated from Human Tissues". In: *J Virol* 79.23 (Dec. 2005), pp. 14793–14803. ISSN: 0022-538X. DOI: 10.1128/JVI.79.23.14793-14803.2005.

[123] H. Nakai et al. "Extrachromosomal Recombinant Adeno-Associated Virus Vector Genomes Are Primarily Responsible for Stable Liver Transduction In Vivo". In: *J Virol* 75.15 (Aug. 2001), pp. 6969–6976. ISSN: 0022-538X. DOI: 10.1128/JVI.75.15.6969-6976.2001.

[124] D. G. Miller, L. M. Petek, and D. W. Russell. "Adeno-associated virus vectors integrate at chromosome breakage sites". In: *Nat Genet* 36.7 (July 2004), pp. 767–773. ISSN: 1061-4036. DOI: 10.1038/ng1380.

[125] K. S. Hanlon et al. "High levels of AAV vector integration into CRISPR-induced DNA breaks". In: *Nat Commun* 10.1 (Sept. 30, 2019), p. 4439. ISSN: 2041-1723. DOI: 10.1038/s41467-019-12449-2.

[126] K. M. Martins et al. "Prevalent and Disseminated Recombinant and Wild-Type Adeno-Associated Virus Integration in Macaques and Humans". In: *Hum Gene Ther* 34.21 (Nov. 2023), pp. 1081–1094. ISSN: 1557-7422. DOI: 10.1089/hum.2023.134.

[127] J. A. Greig et al. "Integrated vector genomes may contribute to long-term expression in primate liver after AAV administration". In: *Nat Biotechnol* (Nov. 6, 2023). ISSN: 1546-1696. DOI: 10.1038/s41587-023-01974-7.

[128] J. D. Mount et al. "Sustained phenotypic correction of hemophilia B dogs with a factor IX null mutation by liver-directed gene therapy". In: *Blood* 99.8 (Apr. 15, 2002), pp. 2670–2676. ISSN: 0006-4971. DOI: 10.1182/blood.V99.8.2670.

[129] G. N. Nguyen et al. "A long-term study of AAV gene therapy in dogs with hemophilia A identifies clonal expansions of transduced liver cells". In: *Nat Biotechnol* 39.1 (Jan. 2021), pp. 47–55. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0741-7.

[130] M. Muhuri et al. "Durability of transgene expression after rAAV gene therapy". In: *Mol Ther* 30.4 (Apr. 6, 2022), pp. 1364–1380. ISSN: 1525-0024. DOI: 10.1016/j.ymthe.2022.03.004.

[131] A. T. Martino et al. "The genome of self-complementary adeno-associated viral vectors increases Toll-like receptor 9–dependent innate immune responses in the liver". In: *Blood* 117.24 (June 16, 2011), pp. 6459–6468. ISSN: 0006-4971. DOI: 10.1182/blood-2010-10-314518.

[132] S. M. Faust et al. "CpG-depleted adeno-associated virus vectors evade immune detection". In: *J Clin Invest* 123.7 (July 2013), pp. 2994–3001. ISSN: 1558-8238. DOI: 10.1172/JCI68205.

[133] G. C. Pien et al. "Capsid antigen presentation flags human hepatocytes for destruction after transduction by adeno-associated viral vectors". In: *J Clin Invest* 119.6 (June 2009), pp. 1688–1695. ISSN: 1558-8238. DOI: 10.1172/JCI36891.

[134] Z. D. Smith and A. Meissner. "DNA methylation: roles in mammalian development". In: *Nat Rev Genet* 14.3 (Mar. 2013), pp. 204–220. ISSN: 1471-0064. DOI: 10.1038/nrg3354.

[135] L. Laurent et al. "Dynamic changes in the human methylome during differentiation". In: *Genome Res.* 20.3 (Jan. 3, 2010), pp. 320–331. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.101907.109.

[136] C. Li et al. "DNA methylation reprogramming of functional elements during mammalian embryonic development". In: *Cell Discov* 4.1 (Aug. 7, 2018), pp. 1–12. ISSN: 2056-5968. DOI: 10.1038/s41421-018-0039-9.

[137] Z. Dongye, J. Li, and Y. Wu. "Toll-like receptor 9 agonists and combination therapies: strategies to modulate the tumour immune microenvironment for systemic anti-tumour immunity". In: *Br J Cancer* 127.9 (Nov. 2022), pp. 1584–1594. ISSN: 1532-1827. DOI: 10.1038/s41416-022-01876-6.

[138] L. D. Moore, T. Le, and G. Fan. “DNA Methylation and Its Basic Function”. In: *Neuropsychopharmacol* 38.1 (Jan. 2013), pp. 23–38. ISSN: 1740-634X. DOI: 10 . 1038 / npp.2012.112.

[139] S. I. S. Grewal. “The molecular basis of heterochromatin assembly and epigenetic inheritance”. In: *Molecular Cell* 83.11 (June 1, 2023), pp. 1767–1785. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2023.04.020.

[140] J. Smith et al. “Promoter DNA Hypermethylation and Paradoxical Gene Activation”. In: *Trends Cancer* 6.5 (May 2020), pp. 392–406. ISSN: 2405-8025. DOI: 10.1016/j.trecan.2020.02.007.

[141] A. de Mendoza et al. “Large-scale manipulation of promoter DNA methylation reveals context-specific transcriptional responses and stability”. In: *Genome Biology* 23.1 (July 26, 2022), p. 163. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02728-5.

[142] Y. Pollack, R. Stein, A. Razin, and H. Cedar. “Methylation of foreign DNA sequences in eukaryotic cells.” In: *Proceedings of the National Academy of Sciences* 77.11 (Nov. 1980), pp. 6463–6467. DOI: 10.1073/pnas.77.11.6463.

[143] Z.-Y. Chen et al. “Silencing of episomal transgene expression in liver by plasmid bacterial backbone DNA is independent of CpG methylation”. In: *Mol Ther* 16.3 (Mar. 2008), pp. 548–556. ISSN: 1525-0024. DOI: 10.1038/sj.mt.6300399.

[144] A. R. Brooks et al. “Transcriptional silencing is associated with extensive methylation of the CMV promoter following adenoviral gene delivery to muscle”. In: *The Journal of Gene Medicine* 6.4 (2004), pp. 395–404. ISSN: 1521-2254. DOI: 10.1002/jgm.516.

[145] J. Ellis. “Silencing and variegation of gammaretrovirus and lentivirus vectors”. In: *Hum Gene Ther* 16.11 (Nov. 2005), pp. 1241–1246. ISSN: 1043-0342. DOI: 10.1089/hum.2005.16.1241.

[146] R. Tóth et al. “Methylation Status of the Adeno-Associated Virus Type 2 (AAV2)”. In: *Viruses* 11.1 (Jan. 2019), p. 38. DOI: 10.3390/v11010038.

[147] M. T. Radukic et al. “Nanopore sequencing of native adeno-associated virus single-stranded DNA using a transposase-based rapid protocol”. In: *bioRxiv* (Dec. 28, 2019), p. 2019.12.27.885319. DOI: 10.1101/2019.12.27.885319.

[148] A. Léger et al. “Adeno-Associated Viral Vector-Mediated Transgene Expression Is Independent of DNA Methylation in Primate Liver and Skeletal Muscle”. In: *PLOS ONE* 6.6 (June 8, 2011), e20881. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0020881.

[149] D. Chanda et al. “Effects of Cellular Methylation on Transgene Expression and Site-Specific Integration of Adeno-Associated Virus”. In: *Genes (Basel)* 8.9 (Sept. 18, 2017). ISSN: 2073-4425. DOI: 10.3390/genes8090232.

[150] K. Miyata et al. “Bisulfite Sequencing for DNA Methylation Analysis of Primary Muscle Stem Cells”. In: *Methods Mol Biol* 1668 (2017), pp. 3–13. ISSN: 1064-3745. DOI: 10.1007/978-1-4939-7283-8_1.

[151] A. Chatterjee, P. A. Stockwell, E. J. Rodger, and I. M. Morison. “Comparison of alignment software for genome-wide bisulphite sequence data”. In: *Nucleic Acids Res* 40.10 (May 2012), e79. ISSN: 0305-1048. DOI: 10.1093/nar/gks150.

[152] B. S. Pedersen et al. *Fast and accurate alignment of long bisulfite-seq reads*. May 13, 2014. DOI: 10.48550/arXiv.1401.1129. arXiv: 1401.1129[q-bio].

[153] C. Grunau, S. J. Clark, and A. Rosenthal. “Bisulfite genomic sequencing: systematic investigation of critical experimental parameters”. In: *Nucleic Acids Res* 29.13 (July 1, 2001), e65. ISSN: 0305-1048.

[154] R. Vaisvila et al. “Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA”. In: *Genome Res* 31.7 (July 2021), pp. 1280–1289. ISSN: 1549-5469. DOI: 10.1101/gr.266551.120.

[155] S. Nurk et al. “The complete sequence of a human genome”. In: *Science* 376.6588 (Apr. 2022), pp. 44–53. DOI: 10.1126/science.abj6987.

[156] S. E. Van der Verren et al. “A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity”. In: *Nat Biotechnol* 38.12 (Dec. 2020), pp. 1415–1420. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0570-8.

[157] *nanoporetech/dorado*. Apr. 8, 2025. URL: https://github.com/nanoporetech/dorado (visited on 04/08/2025).

[158] Y. Liu et al. “DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation”. In: *Genome Biol* 22.1 (Oct. 18, 2021), p. 295. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02510-z.

[159] Z. W.-S. Yuen et al. “Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing”. In: *Nat Commun* 12.1 (June 8, 2021), p. 3438. ISSN: 2041-1723. DOI: 10.1038/s41467-021-23778-6.

[160] M. U. Ahsan et al. “A signal processing and deep learning framework for methylation detection using Oxford Nanopore sequencing”. In: *Nat Commun* 15.1 (Feb. 16, 2024), p. 1448. ISSN: 2041-1723. DOI: 10.1038/s41467-024-45778-y.

[161] M. R. Hübner, M. A. Eckersley-Maslin, and D. L. Spector. “Chromatin Organization and Transcriptional Regulation”. In: *Curr Opin Genet Dev* 23.2 (Apr. 2013), pp. 89–95. ISSN: 0959-437X. DOI: 10.1016/j.gde.2012.11.006.

[162] S. Martire and L. A. Banaszynski. “The roles of histone variants in fine-tuning chromatin organization and function”. In: *Nat Rev Mol Cell Biol* 21.9 (Sept. 2020), pp. 522–541. ISSN: 1471-0072. DOI: 10.1038/s41580-020-0262-8.

[163] G. Arya and T. Schlick. “Role of histone tails in chromatin folding revealed by a mesoscopic oligonucleosome model”. In: *Proceedings of the National Academy of Sciences* 103.44 (Oct. 31, 2006), pp. 16236–16241. DOI: 10.1073/pnas.0604817103.

[164] M. Ghoneim, H. A. Fuchs, and C. A. Musselman. “Histone tail conformations: A fuzzy affair with DNA”. In: *Trends Biochem Sci* 46.7 (July 2021), pp. 564–578. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2020.12.012.

[165] H. Santos-Rosa et al. “Histone H3 tail clipping regulates gene expression”. In: *Nat Struct Mol Biol* 16.1 (Jan. 2009), pp. 17–22. ISSN: 1545-9985. DOI: 10.1038/nsmb.1534.

[166] N. P. Nurse, I. Jimenez-Useche, I. T. Smith, and C. Yuan. “Clipping of Flexible Tails of Histones H3 and H4 Affects the Structure and Dynamics of the Nucleosome”. In: *Biophys J* 104.5 (Mar. 5, 2013), pp. 1081–1088. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2013.01.019.

[167] X. Duan et al. “The role of histone post-translational modifications in cancer and cancer immunity: functions, mechanisms and therapeutic implications”. In: *Front. Immunol.* 15 (Nov. 15, 2024), p. 1495221. ISSN: 1664-3224. DOI: 10.3389/fimmu.2024.1495221.

[168] A. J. Bannister and T. Kouzarides. “Regulation of chromatin by histone modifications”. In: *Cell Res* 21.3 (Mar. 2011), pp. 381–395. ISSN: 1748-7838. DOI: 10.1038/cr.2011.22.

[169] K. Wu et al. “Dynamics of histone acetylation during human early embryogenesis”. In: *Cell Discov* 9 (Mar. 14, 2023), p. 29. ISSN: 2056-5968. DOI: 10.1038/s41421-022-00514-y.

[170] Z. Wang et al. “Combinatorial patterns of histone acetylations and methylations in the human genome”. In: *Nat Genet* 40.7 (July 2008), pp. 897–903. ISSN: 1546-1718. DOI: 10.1038/ng.154.

[171] A. M. Lindroth et al. “Antagonism between DNA and H3K27 Methylation at the Imprinted Rasgrf1 Locus”. In: *PLoS Genet* 4.8 (Aug. 1, 2008), e1000145. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1000145.

[172] Y. Li, X. Chen, and C. Lu. “The interplay between DNA and histone methylation: molecular mechanisms and disease implications”. In: *EMBO reports* 22.5 (May 5, 2021), e51803. ISSN: 1469-221X. DOI: 10.15252/embr.202051803.

[173] J. Richard Albert et al. “DNA methylation shapes the Polycomb landscape during the exit from naive pluripotency”. In: *Nat Struct Mol Biol* 32.2 (Feb. 2025), pp. 346–357. ISSN: 1545-9985. DOI: 10.1038/s41594-024-01405-4.

[174] W. Y. Chen et al. “Reactivation of silenced, virally transduced genes by inhibitors of histone deacetylase”. In: *Proc Natl Acad Sci U S A* 94.11 (May 27, 1997), pp. 5798–5803. ISSN: 0027-8424. DOI: 10.1073/pnas.94.11.5798.

[175] T. Okada et al. “A Histone Deacetylase Inhibitor Enhances Recombinant Adeno-associated Virus-Mediated Gene Expression in Tumor Cells”. In: *Molecular Therapy* 13.4 (Apr. 1, 2006), pp. 738–746. ISSN: 1525-0016. DOI: 10.1016/j.ymthe.2005.11.010.

[176] A. Das et al. “Epigenetic silencing of recombinant AAV genomes by NP220 and the HUSH complex”. In: *J Virol* (Dec. 8, 2021), JVI0203921. ISSN: 1098-5514. DOI: 10.1128/JVI.02039-21.

[177] A. Gonzalez-Sandoval et al. “The AAV capsid can influence the epigenetic marking of rAAV delivered episomal genomes in a species dependent manner”. In: *Nat Commun* 14.1 (Apr. 28, 2023), p. 2448. ISSN: 2041-1723. DOI: 10.1038/s41467-023-38106-3.

[178] E. J. Loeb et al. “Capsid-mediated control of adeno-associated viral transcription determines host range”. In: *Cell Rep* 43.3 (Mar. 26, 2024), p. 113902. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2024.113902.

[179] T. S. Furey. “ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions”. In: *Nat Rev Genet* 13.12 (Dec. 2012), pp. 840–852. ISSN: 1471-0064. DOI: 10.1038/nrg3306.

[180] P. J. Skene and S. Henikoff. “An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites”. In: *eLife* 6 (Jan. 12, 2017). Ed. by D. Reinberg, e21856. ISSN: 2050-084X. DOI: 10.7554/eLife.21856.

[181] H. S. Kaya-Okur et al. “CUT&Tag for efficient epigenomic profiling of small samples and single cells”. In: *Nat Commun* 10.1 (Apr. 29, 2019), p. 1930. ISSN: 2041-1723. DOI: 10.1038/s41467-019-09982-5.

[182] E. S. Lander et al. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (Feb. 2001), pp. 860–921. ISSN: 1476-4687. DOI: 10.1038/35057062.

[183] K.-C. Liang, J. T. Tseng, S.-J. Tsai, and H. S. Sun. “Characterization and distribution of repetitive elements in association with genes in the human genome”. In: *Computational Biology and Chemistry*. 13th Asia Pacifi c Bioinformatics Conference, HsinChu, Taiwan, 21-23 January 2015 57 (Aug. 1, 2015), pp. 29–38. ISSN: 1476-9271. DOI: 10.1016/j.compbiolchem.2015.02.007.

[184] A. M. Roy-Engel et al. “Alu insertion polymorphisms for the study of human genomic diversity.” In: *Genetics* 159.1 (Sept. 2001), pp. 279–290. ISSN: 0016-6731.

[185] R. J. Britten. “DNA sequence insertion and evolutionary variation in gene regulation.” In: *Proc Natl Acad Sci U S A* 93.18 (Sept. 3, 1996), pp. 9374–9377. ISSN: 0027-8424.

[186] L. Liang et al. “Complementary Alu sequences mediate enhancer-promoter selectivity”. In: *Nature* 619.7971 (July 2023), pp. 868–875. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06323-x.

[187] H. F. Willard and J. S. Waye. “Hierarchical order in chromosome-specific human alpha satellite DNA”. In: *Trends in Genetics* 3 (Jan. 1, 1987), pp. 192–198. ISSN: 0168-9525. DOI: 10.1016/0168-9525(87)90232-0.

[188] N. Altemose et al. “Complete genomic and epigenetic maps of human centromeres”. In: *Science* 376.6588 (Apr. 2022), eabl4178. ISSN: 0036-8075. DOI: 10.1126/science.abl4178.

[189] W. C. Earnshaw and N. Rothfield. “Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma”. In: *Chromosoma* 91.3 (1985), pp. 313–321. ISSN: 0009-5915. DOI: 10.1007/BF00328227.

[190] G. H. Karpen and R. C. Allshire. “The case for epigenetic effects on centromere identity and function”. In: *Trends Genet* 13.12 (Dec. 1997), pp. 489–496. ISSN: 0168-9525. DOI: 10.1016/s0168-9525(97)01298-5.

[191] P. B. Talbert and S. Henikoff. “The genetics and epigenetics of satellite centromeres”. In: *Genome Res* 32.4 (Apr. 2022), pp. 608–615. ISSN: 1088-9051. DOI: 10.1101/gr.275351.121.

[192] T. de Lange et al. “Structure and variability of human chromosome ends”. In: *Mol Cell Biol* 10.2 (Feb. 1990), pp. 518–527. ISSN: 0270-7306. DOI: 10.1128/mcb.10.2.518-527.1990.

[193] T. de Lange. “Shelterin-Mediated Telomere Protection”. In: *Annu Rev Genet* 52 (Nov. 23, 2018), pp. 223–247. ISSN: 1545-2948. DOI: 10.1146/annurev-genet-032918-021921.

[194] B. Schumacher, J. Pothof, J. Vijg, and J. H. J. Hoeijmakers. “The central role of DNA damage in the ageing process”. In: *Nature* 592.7856 (Apr. 2021), pp. 695–703. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03307-7.

[195] Z.-y. Kan et al. “G-quadruplex formation in human telomeric (TTAGGG)4 sequence with complementary strand in close vicinity under molecularly crowded condition”. In: *Nucleic Acids Res* 35.11 (June 2007), pp. 3646–3653. ISSN: 0305-1048. DOI: 10.1093/nar/gkm203.

[196] J. L. Huppert. “Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes”. In: *Chem. Soc. Rev.* 37.7 (June 23, 2008), pp. 1375–1384. ISSN: 1460-4744. DOI: 10.1039/B702491F.

[197] J. E. Johnson, J. S. Smith, M. L. Kozak, and F. B. Johnson. “*In vivo veritas:* Using yeast to probe the biological functions of G-quadruplexes”. In: *Biochimie*. Targeting DNA Part II 90.8 (Aug. 1, 2008), pp. 1250–1263. ISSN: 0300-9084. DOI: 10.1016/j.biochi.2008.02.013.

[198] D. Sen and W. Gilbert. “Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis”. In: *Nature* 334.6180 (July 1988), pp. 364–366. ISSN: 1476-4687. DOI: 10.1038/334364a0.

[199] E. N. Trifonov and J. L. Sussman. “The pitch of chromatin DNA is reflected in its nucleotide sequence.” In: *Proc. Natl. Acad. Sci. U.S.A.* 77.7 (July 1980), pp. 3816–3820. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.77.7.3816.

[200] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis. “Periodicity in DNA coding sequences: Implications in gene evolution”. In: *Journal of Theoretical Biology* 151.3 (Aug. 7, 1991), pp. 323–331. ISSN: 0022-5193. DOI: 10.1016/S0022-5193(05)80381-9.

[201] M. Eigen and P. Schuster. “The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle”. In: *Naturwissenschaften* 64.11 (Nov. 1977), pp. 541–565. ISSN: 0028-1042. DOI: 10.1007/BF00450633.

[202] J. C. Shepherd. “Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification”. In: *Proc Natl Acad Sci U S A* 78.3 (Mar. 1981), pp. 1596–1600. ISSN: 0027-8424. DOI: 10.1073/pnas.78.3.1596.

[203] T. H. Jukes. “On the prevalence of certain codons (”RNY”) in genes for proteins”. In: *J Mol Evol* 42.4 (Apr. 1996), pp. 377–381. ISSN: 0022-2844. DOI: 10.1007/BF02498631.

[204] G. S. Zamudio and M. V. José. “On the Uniqueness of the Standard Genetic Code”. In: *Life (Basel)* 7.1 (Feb. 13, 2017), p. 7. ISSN: 2075-1729. DOI: 10.3390/life7010007.

[205] E. N. Trifonov. “Consensus temporal order of amino acids and evolution of the triplet code”. In: *Gene* 261.1 (Dec. 30, 2000), pp. 139–151. ISSN: 0378-1119. DOI: 10.1016/s0378-1119(00)00476-5.

[206] J. Lehmann. “Amplification of the sequences displaying the pattern RNY in the RNA world: the translation –> translation/replication hypothesis”. In: *J Theor Biol* 219.4 (Dec. 21, 2002), pp. 521–537. ISSN: 0022-5193. DOI: 10.1006/jtbi.2002.3142.

[207] E. N. Trifonov. “Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences”. In: *J Mol Biol* 194.4 (Apr. 20, 1987), pp. 643–652. ISSN: 0022-2836. DOI: 10.1016/0022-2836(87)90241-5.

[208] D. Anastassiou. “Frequency-domain analysis of biomolecular sequences”. In: *Bioinformatics* 16.12 (Dec. 2000), pp. 1073–1081. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/16.12.1073.

[209] L. Wang and L. D. Stein. “Localizing triplet periodicity in DNA and cDNA sequences”. In: *BMC Bioinformatics* 11.1 (Nov. 8, 2010), p. 550. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-550.

[210] E. Trifonov. “Sequence-dependent deformational anisotropy of chromatin DNA”. In: *Nucleic Acids Research* 8.17 (Sept. 11, 1980), pp. 4041–4054. ISSN: 0305-1048. DOI: 10.1093/nar/8.17.4041.

[211] T. Bettecken, Z. M. Frenkel, and E. N. Trifonov. “Human nucleosomes: special role of CG dinucleotides and Alu-nucleosomes”. In: *BMC Genomics* 12.1 (May 31, 2011), p. 273. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-273.

[212] F. Salih, B. Salih, and E. N. Trifonov. “Sequence Structure of Hidden 10.4-base Repeat in the Nucleosomes of C. elegans”. In: *Journal of Biomolecular Structure and Dynamics* 26.3 (Dec. 1, 2008), pp. 273–281. ISSN: 0739-1102. DOI: 10.1080/07391102.2008.10531241.

[213] A. B. Cohanim, Y. Kashi, and E. N. Trifonov. “Yeast nucleosome DNA pattern: deconvolution from genome sequences of S. cerevisiae”. In: *J Biomol Struct Dyn* 22.6 (June 2005), pp. 687–694. ISSN: 0739-1102. DOI: 10.1080/07391102.2005.10507035.

[214] T. Bettecken and E. N. Trifonov. “Repertoires of the Nucleosome-Positioning Dinucleotides”. In: *PLoS One* 4.11 (Nov. 2, 2009). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0007654.

[215] P. D. Dans et al. “Exploring polymorphisms in B-DNA helical conformations”. In: *Nucleic Acids Research* 40.21 (Nov. 1, 2012), pp. 10668–10678. ISSN: 0305-1048. DOI: 10.1093/nar/gks884.

[216] E. Segal et al. “A genomic code for nucleosome positioning”. In: *Nature* 442.7104 (Aug. 2006), pp. 772–778. ISSN: 1476-4687. DOI: 10.1038/nature04979.

[217] M. Andrabi et al. “Predicting conformational ensembles and genome-wide transcription factor binding sites from DNA sequences”. In: *Sci Rep* 7.1 (June 22, 2017), p. 4071. ISSN: 2045-2322. DOI: 10.1038/s41598-017-03199-6.

[218] P. S. Ho, M. Carter, P. S. Ho, and M. Carter. “DNA Structure: Alphabet Soup for the Cellular Soul”. In: *DNA Replication - Current Advances*. IntechOpen, Aug. 1, 2011. ISBN: 978-953-307-593-8. DOI: 10.5772/18536.

[219] W. K. Olson et al. “DNA sequence-dependent deformability deduced from protein-DNA crystal complexes”. In: *Proc Natl Acad Sci U S A* 95.19 (Sept. 15, 1998), pp. 11163–11168. ISSN: 0027-8424. DOI: 10.1073/pnas.95.19.11163.

[220] R. Stefl et al. “DNA A-tract bending in three dimensions: Solving the dA4T4 vs. dT4A4 conundrum”. In: *Proceedings of the National Academy of Sciences* 101.5 (Feb. 3, 2004), pp. 1177–1182. DOI: 10.1073/pnas.0308143100.

[221] C. R. Calladine, H. R. Drew, and M. J. McCall. “The intrinsic curvature of DNA in solution”. In: *J Mol Biol* 201.1 (May 5, 1988), pp. 127–137. ISSN: 0022-2836. DOI: 10.1016/0022-2836(88)90444-5.

[222] A. Bolshoy, P. McNamara, R. E. Harrington, and E. N. Trifonov. “Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles.” In: *Proc Natl Acad Sci U S A* 88.6 (Mar. 15, 1991), pp. 2312–2316. ISSN: 0027-8424.

[223] N. Kaplan et al. “The DNA-encoded nucleosome organization of a eukaryotic genome”. In: *Nature* 458.7236 (Mar. 19, 2009), pp. 362–366. ISSN: 0028-0836. DOI: 10.1038/nature07667.

[224] T. N. Mavrich et al. “A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome”. In: *Genome Res* 18.7 (July 2008), pp. 1073–1083. ISSN: 1088-9051. DOI: 10.1101/gr.078261.108.

[225] Y. Zhang et al. “Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo”. In: *Nat Struct Mol Biol* 16.8 (Aug. 2009), pp. 847–852. ISSN: 1545-9985. DOI: 10.1038/nsmb.1636.

[226] Z. Zhang et al. “A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome”. In: *Science* 332.6032 (May 20, 2011), pp. 977–980. ISSN: 1095-9203. DOI: 10.1126/science.1200508.

[227] H. Herzel, O. Weiss, and E. N. Trifonov. “10-11 bp periodicities in complete genomes reflect protein structure and DNA folding”. In: *Bioinformatics* 15.3 (Mar. 1999), pp. 187–193. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/15.3.187.

[228] P. Schieg and H. Herzel. “Periodicities of 10–11 bp as Indicators of the Supercoiled State of Genomic DNA”. In: *Journal of Molecular Biology* 343.4 (Oct. 29, 2004), pp. 891–901. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2004.08.068.

[229] J. Mrázek. “Comparative Analysis of Sequence Periodicity among Prokaryotic Genomes Points to Differences in Nucleoid Structure and a Relationship to Gene Expression”. In: *J Bacteriol* 192.14 (July 2010), pp. 3763–3772. ISSN: 0021-9193. DOI: 10.1128/JB.00149-10.

[230] A. Atzinger and J. G. Lawrence. “Selection for ancient periodic motifs that do not impart DNA bending”. In: *PLOS Genetics* 16.10 (Oct. 6, 2020), e1009042. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1009042.

[231] E. D. Howe and J. S. Song. “Categorical spectral analysis of periodicity in human and viral genomes”. In: *Nucleic Acids Research* 41.3 (Feb. 1, 2013), pp. 1395–1405. ISSN: 0305-1048. DOI: 10.1093/nar/gks1261.

[232] C. Yin. “Latent periodicity-2 in coronavirus SARS-CoV-2 genome: Evolutionary implications”. In: *J Theor Biol* 515 (Apr. 21, 2021), p. 110604. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2021.110604.

[233] M. Bina. “Periodicity of dinucleotides in nucleosomes derived from simian virus 40 chromatin”. In: *Journal of Molecular Biology* 235.1 (Jan. 7, 1994), pp. 198–208. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80026-9.

[234] A. Stein and M. Bina. “A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment”. In: *Nucleic Acids Research* 27.3 (Feb. 1, 1999), pp. 848–853. ISSN: 0305-1048. DOI: 10.1093/nar/27.3.848.

[235] C. Gruss and R. Knippers. “[7] The SV40 minichromosome”. In: *Methods in Molecular Genetics*. Ed. by K. W. Adolph. Vol. 7. Viral Gene Techniques. Academic Press, Jan. 1, 1995, pp. 101–113. DOI: 10.1016/S1067-2389(06)80039-7.

[236] M. F. Clarke, P. C. FitzGerald, J. M. Brubaker, and R. T. Simpson. “Sequence-specific interaction of histones with the simian virus 40 enhancer region in vitro.” In: *Journal of Biological Chemistry* 260.23 (Oct. 15, 1985), pp. 12394–12397. ISSN: 0021-9258. DOI: 10.1016/S0021-9258(17)38885-3.

[237] J. Abel and J. Mrázek. “Differences in DNA curvature-related sequence periodicity between prokaryotic chromosomes and phages, and relationship to chromosomal prophage content”. In: *BMC Genomics* 13.1 (May 15, 2012), p. 188. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-188.

[238] B. Prevo and W. C. Earnshaw. “DNA packaging by molecular motors: from bacteriophage to human chromosomes”. In: *Nat Rev Genet* 25.11 (Nov. 2024), pp. 785–802. ISSN: 1471-0064. DOI: 10.1038/s41576-024-00740-y.

[239] M. Serrano et al. “Signals at the bacteriophage phi 29 DNA replication origins required for protein p6 binding and activity.” In: *The EMBO Journal* 8.6 (June 1989), pp. 1879–1885. ISSN: 0261-4189. DOI: 10.1002/j.1460-2075.1989.tb03584.x.

[240] M. Alcorlo et al. “Analytical ultracentrifugation studies of phage phi29 protein p6 binding to DNA”. In: *J Mol Biol* 385.5 (Feb. 6, 2009), pp. 1616–1629. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2008.11.044.

[241] U. Schwartz et al. “Changes in adenoviral chromatin organization precede early gene activation upon infection”. In: *The EMBO Journal* n/a (n/a Aug. 29, 2023), e114162. ISSN: 0261-4189. DOI: 10.15252/embj.2023114162.

[242] ConradinBaumgartl. *ConradinBaumgartl/thesis_code_2025: v1.0.0*. Version PhD. May 2025. DOI: 10.5281/zenodo.15356216.

[243] J. Becker et al. “Identification of a robust promoter in mouse and human hepatocytes by in vivo biopanning of a barcoded AAV library”. In: *Mol Ther* (Apr. 21, 2025), S1525–0016(25)00301–6. ISSN: 1525-0024. DOI: 10.1016/j.ymthe.2025.04.027.

[244] J. Köster and S. Rahmann. “Snakemake—a scalable bioinformatics workflow engine”. In: *Bioinformatics* 28.19 (Oct. 1, 2012), pp. 2520–2522. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts480.

[245] B. Langmead and S. L. Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4 (Apr. 2012), pp. 357–359. ISSN: 1548-7105. DOI: 10.1038/nmeth.1923.

[246] P. Danecek et al. “Twelve years of SAMtools and BCFtools”. In: *GigaScience* 10.2 (Feb. 1, 2021), giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008.

[247] F. Ramírez et al. “deepTools2: a next generation web server for deep-sequencing data analysis”. In: *Nucleic Acids Res* 44 (W1 2016), W160–165. ISSN: 1362-4962. DOI: 10.1093/nar/gkw257.

[248] M. P. Meers, D. Tenenbaum, and S. Henikoff. “Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling”. In: *Epigenetics & Chromatin* 12.1 (July 12, 2019), p. 42. ISSN: 1756-8935. DOI: 10.1186/s13072-019-0287-4.

[249] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (Dec. 5, 2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

[250] W. Shen, B. Sipos, and L. Zhao. “SeqKit2: A Swiss army knife for sequence and alignment processing”. In: *iMeta* 3.3 (2024), e191. ISSN: 2770-596X. DOI: 10.1002/imt2.191.

[251] A. M. Bolger, M. Lohse, and B. Usadel. “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15 (Aug. 1, 2014), pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170.

[252] H. Li et al. “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16 (Aug. 15, 2009), pp. 2078–2079. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352.

[253] D. Ryan. *MethylDackel: A (mostly) universal methylation extractor for BS-seq experiments.* URL: https://github.com/dpryan79/MethylDackel.

[254] H. Li. “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18 (Sept. 15, 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191.

[255] J. T. Simpson et al. “Detecting DNA cytosine methylation using nanopore sequencing”. In: *Nat Methods* 14.4 (Apr. 2017), pp. 407–410. ISSN: 1548-7105. DOI: 10.1038/nmeth.4184.

[256] *megalodon*. Feb. 5, 2025. URL: https://github.com/nanoporetech/megalodon (visited on 02/19/2025).

[257] P. Ni et al. “DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning”. In: *Bioinformatics* 35.22 (Nov. 1, 2019), pp. 4586–4595. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz276.

[258] K. Palin. *GpC methylation from fast5:s to reference anchored calls (gcf52ref)*. May 25, 2021. URL: https://github.com/kpalin/gcf52ref (visited on 03/17/2023).

[259] *epi2me-labs/modbam2bed*. Nov. 15, 2024. URL: https://github.com/epi2me-labs/modbam2bed (visited on 02/26/2025).

[260] C. Camacho et al. “BLAST+: architecture and applications”. In: *BMC Bioinformatics* 10 (Dec. 15, 2009), p. 421. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-421.

[261] J. Serizay. “periodicDNA: Set of tools to identify periodic occurrences of k-mers in DNA sequences.” In: *bioconductor* (R package version 1.16.0 2024). DOI: 10.18129/B9.bioc.periodicDNA.

[262] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4.

[263] D. J. Winter. “rentrez: An R package for the NCBI eUtils API”. In: *The R Journal* 9.2 (2017), pp. 520–526. ISSN: 2073-4859.

[264] P. Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[265] J. D. Hunter. “Matplotlib: A 2D Graphics Environment”. In: *Computing in Science & Engineering* 9.3 (May 2007), pp. 90–95. ISSN: 1558-366X. DOI: 10.1109/MCSE.2007.55.

[266] C. R. Harris et al. “Array programming with NumPy”. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2.

[267] T. pandas development team. *pandas-dev/pandas: Pandas*. Version 1.1.5. Feb. 2020. DOI: 10.5281/zenodo.3509134.

[268] M. L. Waskom. “seaborn: statistical data visualization”. In: *Journal of Open Source Software* 6.60 (Apr. 6, 2021), p. 3021. ISSN: 2475-9066. DOI: 10.21105/joss.03021.

[269] P. J. A. Cock et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (June 1, 2009), pp. 1422–1423. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp163.

[270] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. 2021. URL: https://www.R-project.org/.

[271] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC. 2020. URL: http://www.rstudio.com/.

[272] H. Wickham et al. “Welcome to the Tidyverse”. In: *Journal of Open Source Software* 4.43 (Nov. 21, 2019), p. 1686. ISSN: 2475-9066. DOI: 10.21105/joss.01686.

[273] A. Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*. 2023. URL: https://rpkgs.datanovia.com/ggpubr/.

[274] H. Wickham et al. *dplyr: A Grammar of Data Manipulation*. 2023. URL: https://dplyr.tidyverse.org.

[275] H. Wickham. *forcats: Tools for Working with Categorical Variables (Factors)*. 2023. URL: https://forcats.tidyverse.org/.

[276] H. Wickham and L. Henry. *purrr: Functional Programming Tools*. 2025. URL: https://purrr.tidyverse.org/.

[277] H. Wickham, J. Hester, and J. Bryan. *readr: Read Rectangular Text Data*. 2024. URL: https://readr.tidyverse.org.

[278] H. Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. 2023. URL: https://stringr.tidyverse.org.

[279] K. Müller and H. Wickham. *tibble: Simple Data Frames*. 2024. URL: https://tibble.tidyverse.org/.

[280] H. Wickham, D. Vaughan, and M. Girlich. *tidyr: Tidy Messy Data*. 2024. URL: https://tidyr.tidyverse.org.

[281] H. Pagès, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: Efficient manipulation of biological strings*. 2024. DOI: 10.18129/B9.bioc.Biostrings. URL: https://bioconductor.org/packages/Biostrings.

[282] M. Lawrence et al. “Software for Computing and Annotating Genomic Ranges”. In: *PLoS Computational Biology* 9 (8 2013). DOI: 10.1371/journal.pcbi.1003118.

[283] A. N. Schep, B. Wu, J. D. Buenrostro, and W. J. Greenleaf. “chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data”. In: *Nat Methods* 14.10 (Oct. 2017), pp. 975–978. ISSN: 1548-7105. DOI: 10.1038/nmeth.4401.

[284] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri. “NIH Image to ImageJ: 25 years of image analysis”. In: *Nat Methods* 9.7 (July 2012), pp. 671–675. ISSN: 1548-7105. DOI: 10.1038/nmeth.2089.

[285] *Anaconda Software Distribution*. Version Vers. 2-2.4.0. 2020. URL: https://docs.anaconda.com/.

[286] *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (visited on 10/22/2020).

[287] P. Ewels, M. Magnusson, S. Lundin, and M. Käller. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (Oct. 1, 2016), pp. 3047–3048. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw354.

[288] K. Okonechnikov, A. Conesa, and F. García-Alcalde. “Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data”. In: *Bioinformatics* 32.2 (Jan. 15, 2016), pp. 292–294. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv566.

[289] *Welcome to Oxford Nanopore Technologies*. Oxford Nanopore Technologies. URL: https://nanoporetech.com/ (visited on 04/10/2025).

[290] Perplexity. *Perplexity.ai (AI Chatbot) [Large language model]*. 2023. URL: https://www.perplexity.ai/.

[291] OpenAI. *ChatGPT (Mar 14 version) [Large language model]*. 2023. URL: https://chat.openai.com/chat.

[292] T. Gilpatrick et al. “Targeted nanopore sequencing with Cas9-guided adapter ligation”. In: *Nat Biotechnol* 38.4 (Apr. 2020), pp. 433–438. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0407-5.

[293] A. C. Nathwani et al. “Self-complementary adeno-associated virus vectors containing a novel liver-specific human factor IX expression cassette enable highly efficient transduction of murine and

nonhuman primate liver". In: *Blood* 107.7 (Apr. 1, 2006), pp. 2653–2661. ISSN: 0006-4971. DOI: 10.1182/blood-2005-10-4035.

[294] C. Zincarelli, S. Soltys, G. Rengo, and J. E. Rabinowitz. "Analysis of AAV Serotypes 1–9 Mediated Gene Expression and Tropism in Mice After Systemic Injection". In: *Molecular Therapy* 16.6 (June 1, 2008), pp. 1073–1080. ISSN: 1525-0016. DOI: 10.1038/mt.2008.76.

[295] Biorender. *Created in Biorender*. 2025. URL: https://BioRender.com/oozcz9p.

[296] B. D. Sigurpalsdottir et al. "A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes". In: *Genome Biology* 25.1 (Mar. 11, 2024), p. 69. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03207-9.

[297] J. Pei et al. "H3K27ac acetylome signatures reveal the epigenomic reorganization in remodeled non-failing human hearts". In: *Clinical Epigenetics* 12.1 (July 14, 2020), p. 106. ISSN: 1868-7083. DOI: 10.1186/s13148-020-00895-5.

[298] A. Rossi et al. "Vector uncoating limits adeno-associated viral vector-mediated transduction of human dendritic cells and vector immunogenicity". In: *Sci Rep* 9.1 (Mar. 6, 2019), p. 3631. ISSN: 2045-2322. DOI: 10.1038/s41598-019-40071-1.

[299] Biorender. *Created in Biorender*. 2025. URL: https://BioRender.com/r79sqlc.

[300] Biorender. *Created in Biorender*. 2025. URL: https://BioRender.com/gk5m23t.

[301] J. Y. Dong, P. D. Fan, and R. A. Frizzell. "Quantitative analysis of the packaging capacity of recombinant adeno-associated virus". In: *Hum Gene Ther* 7.17 (Nov. 10, 1996), pp. 2101–2112. ISSN: 1043-0342. DOI: 10.1089/hum.1996.7.17-2101.

[302] X. Pan et al. "Rational engineering of a functional CpG-free ITR for AAV gene therapy". In: *Gene Ther* (Oct. 6, 2021), pp. 1–13. ISSN: 1476-5462. DOI: 10.1038/s41434-021-00296-0.

[303] P. J. Ross, M. A. Kennedy, and R. J. Parks. "Host cell detection of noncoding stuffer DNA contained in helper-dependent adenovirus vectors leads to epigenetic repression of transgene expression". In: *J Virol* 83.17 (Sept. 2009), pp. 8409–8417. ISSN: 1098-5514. DOI: 10.1128/JVI.00796-09.

[304] F. Fuks et al. "Dnmt3a binds deacetylases and is recruited by a sequence-specific repressor to silence transcription". In: *EMBO J* 20.10 (May 15, 2001), pp. 2536–2544. ISSN: 0261-4189. DOI: 10.1093/emboj/20.10.2536.

[305] E. Riu et al. "Histone Modifications are Associated with the Persistence or Silencing of Vector-mediated Transgene Expression In Vivo". In: *Molecular Therapy* 15.7 (July 1, 2007), pp. 1348–1355. ISSN: 1525-0016. DOI: 10.1038/sj.mt.6300177.

[306] D. Jjingo et al. "On the presence and role of human gene-body DNA methylation". In: *Oncotarget* 3.4 (May 9, 2012), pp. 462–474. ISSN: 1949-2553.

[307] L. F. Earley et al. "Adeno-Associated Virus Serotype-Specific Inverted Terminal Repeat Sequence Role in Vector Transgene Expression". In: *Hum Gene Ther* 31.3 (Feb. 1, 2020), pp. 151–162. ISSN: 1043-0342. DOI: 10.1089/hum.2019.274.

[308] K. O. Moss, M. Vora, and C. Russell. "Methylation of rAAV vector DNA is not a mechanism for differential transgene RNA expression from rAAV gene therapy in mouse and NHP liver". Edinburgh, Nov. 10, 2022.

[309] M. B. S. Al-Shuhaib and H. O. Hashim. "Mastering DNA chromatogram analysis in Sanger sequencing for reliable clinical analysis". In: *J Genet Eng Biotechnol* 21 (Nov. 13, 2023), p. 115. ISSN: 1687-157X. DOI: 10.1186/s43141-023-00587-6.

[310] R. Shapiro, B. Braverman, J. B. Louis, and R. E. Servis. "Nucleic acid reactivity and conformation. II. Reaction of cytosine and uracil with sodium bisulfite". In: *J Biol Chem* 248.11 (June 10, 1973), pp. 4060–4064. ISSN: 0021-9258.

[311] L. Budzko, A. Mierzwa, and M. Figlerowicz. "AID/APOBEC: an expanding repertoire of targets and functions". In: *Trends in Biochemical Sciences* 0.0 (Mar. 24, 2025). ISSN: 0968-0004. DOI: 10.1016/j.tibs.2025.02.006.

[312] Y. Wang et al. "Nanopore sequencing technology, bioinformatics and applications". In: *Nat Biotechnol* 39.11 (Nov. 2021), pp. 1348–1365. ISSN: 1087-0156. DOI: 10.1038/s41587-021-01108-x.

[313] Y. K. Chan and M. U. Gack. "Viral evasion of intracellular DNA and RNA sensing". In: *Nat Rev Microbiol* 14.6 (2016), pp. 360–373. ISSN: 1740-1526. DOI: 10.1038/nrmicro.2016.45.

[314] X. Wang et al. "Epigenetic regulation in antiviral innate immunity". In: *Eur J Immunol* 51.7 (July 2021), pp. 1641–1651. ISSN: 1521-4141. DOI: 10.1002/eji.202048975.

[315] C. Gao et al. "The epigenetic landscapes of histone modifications on HSV-1 genome in human THP-1 cells". In: *Antiviral Research* 176 (Apr. 1, 2020), p. 104730. ISSN: 0166-3542. DOI: 10.1016/j.antiviral.2020.104730.

[316] K. Fu, G. Bonora, and M. Pellegrini. “Interactions between core histone marks and DNA methyltransferases predict DNA methylation patterns observed in human cells and tissues”. In: *Epigenetics* 15.3 (Sept. 17, 2019), pp. 272–282. ISSN: 1559-2294. DOI: 10.1080/15592294.2019.1666649.

[317] T. Meshi, K. I. Taoka, and M. Iwabuchi. “Regulation of histone gene expression during the cell cycle”. In: *Plant Mol Biol* 43.5 (Aug. 2000), pp. 643–657. ISSN: 0167-4412. DOI: 10.1023/a:1006421821964.

[318] Z. Shu, S. Row, and W.-M. Deng. “Endoreplication: The Good, the Bad, and the Ugly”. In: *Trends Cell Biol* 28.6 (June 2018), pp. 465–474. ISSN: 0962-8924. DOI: 10.1016/j.tcb.2018.02.006.

[319] L. López-Hernández, P. Toolan-Kerr, A. J. Bannister, and G. Millán-Zambrano. “Dynamic histone modification patterns coordinating DNA processes”. In: *Molecular Cell* 85.2 (Jan. 16, 2025), pp. 225–237. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2024.10.034.

[320] C. Alabert et al. “Two distinct modes for propagation of histone PTMs across the cell cycle”. In: *Genes Dev* 29.6 (Mar. 15, 2015), pp. 585–590. ISSN: 0890-9369. DOI: 10.1101/gad.256354.114.

[321] R. E. Sobel et al. “Conservation of deposition-related acetylation sites in newly synthesized histones H3 and H4”. In: *Proc Natl Acad Sci U S A* 92.4 (Feb. 14, 1995), pp. 1237–1241. ISSN: 0027-8424. DOI: 10.1073/pnas.92.4.1237.

[322] I. Müller and K. Helin. “Keep quiet: the HUSH complex in transcriptional silencing and disease”. In: *Nat Struct Mol Biol* 31.1 (Jan. 2024), pp. 11–22. ISSN: 1545-9985. DOI: 10.1038/s41594-023-01173-7.

[323] T. G. Müller et al. “HIV-1 uncoating by release of viral cDNA from capsid-like structures in the nucleus of infected cells”. In: *eLife* 10 (Apr. 27, 2021). Ed. by W. I. Sundquist, S. L. Sawyer, and E. M. Campbell, e64776. ISSN: 2050-084X. DOI: 10.7554/eLife.64776.

[324] A. Pupo et al. “AAV vectors: The Rubik’s cube of human gene therapy”. In: *Mol Ther* 30.12 (Dec. 7, 2022), pp. 3515–3541. ISSN: 1525-0016. DOI: 10.1016/j.ymthe.2022.09.015.

[325] J. Becker, J. Fakhiri, and D. Grimm. “Fantastic AAV Gene Therapy Vectors and How to Find Them—Random Diversification, Rational Design and Machine Learning”. In: *Pathogens* 11.7 (July 3, 2022), p. 756. ISSN: 2076-0817. DOI: 10.3390/pathogens11070756.

[326] X. Yue et al. “Simultaneous profiling of histone modifications and DNA methylation via nanopore sequencing”. In: *Nat Commun* 13.1 (Dec. 24, 2022), p. 7939. ISSN: 2041-1723. DOI: 10.1038/s41467-022-35650-2.

[327] D. Duan, Z. Yan, Y. Yue, and J. F. Engelhardt. “Structural analysis of adeno-associated virus transduction circular intermediates”. In: *Virology* 261.1 (Aug. 15, 1999), pp. 8–14. ISSN: 0042-6822. DOI: 10.1006/viro.1999.9821.

[328] L. V. Crawford and M. J. Waring. “The supercoiling of papilloma virus DNA”. In: *J Gen Virol* 1.3 (July 1967), pp. 387–390. ISSN: 0022-1317. DOI: 10.1099/0022-1317-1-3-387.

[329] V. B. Zhurkin. “Periodicity in DNA primary structure is defined by secondary structure of the coded protein.” In: *Nucleic Acids Res* 9.8 (Apr. 24, 1981), pp. 1963–1971. ISSN: 0305-1048.

[330] *Gaius-Augustus/BRAKER*. Apr. 17, 2025. URL: https://github.com/Gaius-Augustus/BRAKER (visited on 04/22/2025).

[331] C. Baumgartl et al. *Adenovirus maturation establishes the transcription competent packaging of its genome*. Apr. 8, 2025. DOI: 10.1101/2025.04.08.647783.

[332] A. Tarakhovsky and R. K. Prinjha. “Drawing on disorder: How viruses use histone mimicry to their advantage”. In: *The Journal of Experimental Medicine* 215.7 (July 2, 2018), p. 1777. DOI: 10.1084/jem.20180099.

[333] C. Beren et al. “Genome organization and interaction with capsid protein in a multipartite RNA virus”. In: *Proc Natl Acad Sci U S A* 117.20 (May 19, 2020), pp. 10673–10680. ISSN: 1091-6490. DOI: 10.1073/pnas.1915078117.

[334] S. Sarker et al. “Structural insights into the assembly and regulation of distinct viral capsid complexes”. In: *Nat Commun* 7.1 (Oct. 4, 2016), p. 13014. ISSN: 2041-1723. DOI: 10.1038/ncomms13014.

[335] M. S. Chapman and M. G. Rossmann. “Single-stranded DNA–protein interactions in canine parvovirus”. In: *Structure* 3.2 (Feb. 1, 1995), pp. 151–162. ISSN: 0969-2126. DOI: 10.1016/S0969-2126(01)00146-0.

[336] R. A. López-Astacio et al. “The Structures and Functions of Parvovirus Capsids and Missing Pieces: the Viral DNA and Its Packaging, Asymmetrical Features, Nonprotein Components, and Receptor or Antibody Binding and Interactions”. In: *J Virol* 97.7 (June 2023), e00161–23. ISSN: 0022-538X. DOI: 10.1128/jvi.00161-23.

[337] J. J. Penzes, M. Holm, S. A. Yost, and J. T. Kaelber. "Cryo-EM-based discovery of a pathogenic parvovirus causing epidemic mortality by black wasting disease in farmed beetles". In: *Cell* 187.20 (Oct. 3, 2024), 5604–5619.e14. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2024.07.053.

[338] S. Kronenberg et al. "A Conformational Change in the Adeno-Associated Virus Type 2 Capsid Leads to the Exposure of Hidden VP1 N Termini". In: *J Virol* 79.9 (May 2005), pp. 5296–5303. ISSN: 0022-538X. DOI: 10.1128/JVI.79.9.5296-5303.2005.

[339] L. Govindasamy et al. "Structurally Mapping the Diverse Phenotype of Adeno-Associated Virus Serotype 4". In: *J Virol* 80.23 (Dec. 2006), pp. 11556–11570. ISSN: 0022-538X. DOI: 10.1128/JVI.01536-06.

[340] H.-J. Nam et al. "Structural Studies of Adeno-Associated Virus Serotype 8 Capsid Transitions Associated with Endosomal Trafficking □". In: *J Virol* 85.22 (Nov. 2011), pp. 11791–11799. ISSN: 0022-538X. DOI: 10.1128/JVI.05305-11.

[341] K. Mikals et al. "The structure of AAVrh32.33, a Novel Gene Delivery Vector". In: *J Struct Biol* 186.2 (May 2014), pp. 308–317. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2014.03.020.

[342] S. Halder et al. "Structure of Neurotropic Adeno-Associated Virus AAVrh.8". In: *J Struct Biol* 192.1 (Oct. 2015), pp. 21–36. ISSN: 1047-8477. DOI: 10.1016/j.jsb.2015.08.017.

[343] J. R. Brister and N. Muzyczka. "Rep-mediated nicking of the adeno-associated virus origin requires two biochemical activities, DNA helicase activity and transesterification". In: *J Virol* 73.11 (Nov. 1999), pp. 9325–9336. ISSN: 0022-538X. DOI: 10.1128/JVI.73.11.9325-9336.1999.

[344] Y. G. Kuznetsov et al. "Atomic Force Microscopy Analysis of Icosahedral Virus RNA". In: *Journal of Molecular Biology* 347.1 (Mar. 18, 2005), pp. 41–52. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2005.01.006.

[345] E. Segal and J. Widom. "Poly(dA:dT) Tracts: Major Determinants of Nucleosome Organization". In: *Curr Opin Struct Biol* 19.1 (Feb. 2009), pp. 65–71. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2009.01.004.

[346] A. Marin-Gonzalez, J. G. Vilhena, R. Perez, and F. Moreno-Herrero. "A molecular view of DNA flexibility". In: *Quarterly Reviews of Biophysics* 54 (Jan. 2021), e8. ISSN: 0033-5835, 1469-8994. DOI: 10.1017/S0033583521000068.

[347] C. H. Wong et al. "The estimated annual financial impact of gene therapy in the United States". In: *Gene Ther* 30.10 (Nov. 2023), pp. 761–773. ISSN: 1476-5462. DOI: 10.1038/s41434-023-00419-9.

Acknowledgements

During the past 3.5 years, I was surrounded and influenced by great people without whom this thesis would probably still be far from finished. First and foremost, I want to thank my fiancée for supporting me not just in the past 3.5 years but the past 10 years, even committing to the move to Heidelberg in the foreign country of Baden-Württemberg with me. I also want to thank my parents for sparking and continuously fostering my interest in the natural sciences, without which I would hardly be the person I am today.

I want to express my sincerest gratitude to Dirk Grimm for accepting me as his PhD student. You were never dismissive towards strange ideas and always made room for discussions about even stranger data. Thank you for all the valuable advice and encouragement you have given me throughout these years. I also want to thank Olena Maiakovska, whose idea was the initial impulse for the research undertaken in this thesis. Especially in the first few months in the Lab, your advice was vital for my (metaphorical) survival. I am also grateful for your extensive corrections of my thesis.

I want to thank Felix Bubeck, who was my first point of contact with all things wet lab. The discussions about our data were always stimulating, and I hope to continue with those for some time. Thank you to Jonas Becker, who was always ready to share his hard-earned protocols and samples with me. Thank you to Sabrina Weis for the in-depth discussions about science and miscellaneous topics that often were the highlight of my day. I also owe you for your eagle-eyed proofreading of this thesis. Thank you to Ellen Wiedtke, Chiara Krämer, Emma Gerstmann, and Lena Naber for your never-ending patience in enduring my repeated questions. Thank you to Nico Fischer and Quique Hildebrandt for your camaraderie and the banter I had to endure whenever I wandered into the 'wrong' Lab. Thank you, Jixin Liu, for the refreshing trips we took to escape the Lab. Thank you, Man Xu for the fruitful collaboration and the sport climbing excursions! Thank you to Teng Wei Koay and Yumi Sano for the intense discussions. Thank you to Margara Zayas for your Pep talks, Sergio for the sleepless nights in Vienna, and Kleopatra Rapti for your honesty. A sincere thank you also to all other members of the Lab that I failed to mention by name. You all made my time in this group truly memorable.

I want to express my sincerest gratitude to Julie Garcia Gonzalez-Calero for providing me with a spectacularly clean control sample of empty AAV2 capsids.

I also want to express my thanks to my TAC members, Benedikt Brors and Christoph Plass, for giving genuinely helpful advice and recommendations in our yearly meetings.

Thank you also to Karine Lapouge and the EMBL protein expression and purification core facility for their selfless support in the Mass Photometry measurements.

Although I was supposed to be their supervisor, I want to thank the students whose diverse projects allowed me to learn alongside them and who also contributed to some of the findings in this thesis. In order of your start date, thank you, Darius Szablowski, Anastasyia Vladimirova, Zixuan Huo, Sebastian Buddecke, Amelia Irimies, Sebastian Heß, and Emily Locke!