

RESEARCH ARTICLE

# Epigenome analysis of an algae-infecting giant virus reveals a unique methylation motif catalogue

Alexander R. Truchon<sup>1</sup>, Erik R. Zinser, Steven W. Wilhelm<sup>1</sup>✉\*

Department of Microbiology, University of Tennessee, Knoxville, Tennessee, United States of America

\* [wilhelm@utk.edu](mailto:wilhelm@utk.edu)



## OPEN ACCESS

**Citation:** Truchon AR, Zinser ER, Wilhelm SW (2025) Epigenome analysis of an algae-infecting giant virus reveals a unique methylation motif catalogue. PLoS One 20(12): e0330887. <https://doi.org/10.1371/journal.pone.0330887>

**Editor:** Sebastian D. Fugmann, Chang Gung University, TAIWAN

**Received:** August 7, 2025

**Accepted:** November 21, 2025

**Published:** December 12, 2025

**Copyright:** © 2025 Truchon et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data availability statement:** All sequencing data generated in this study have been submitted to NCBI under the BioProject accession PRJNA1292256 as sequence read archives with the accession numbers SRR34564230-SRR34564239. All python scripts used in

## Abstract

DNA methylation can epigenetically alter gene expression and serve as a mechanism for genomic stabilization. Advancements in long-read sequencing technology have allowed for increased exploration into the methylation profiles of various organisms, including viruses. Studies into the *Nucleocytoviricota* phylum of giant dsDNA viruses have revealed unique strategies for genomic methylation. However, given the diversity across this phylum, further inquiries into specific lineages are necessary.

*Kratosvirus quantuckense* (formerly known as *Aureococcus anophagefferens* Virus, AaV) is predicted to encode six distinct methyltransferases, which bear homology to other methyltransferases across the many clades of *Nucleocytoviricota*. We found that the virus' DNA is methylated with high consistency, including nine different motifs targeted for DNA adenine methylation. Methylation levels varied depending on the associated motif. Likewise, distinct motifs were enriched within unique genomic regions. Collectively our data suggest that each methyltransferase targets unique DNA regions, suggesting they have varying functionality. This work reveals an array of methyltransferase activity in *Kratosvirus quantuckense* and implicates the importance of DNA methylation to the *Nucleocytoviricota* infection cycle.

## Introduction

The role of DNA methylation has been characterized largely through robust genetic analysis in a handful of model organisms [1–4]. DNA methyltransferases (MTases) catalyze the addition of a methyl group typically onto either an adenine, yielding N<sup>6</sup>-methyladenine (6mA), or a cytosine, yielding often 5-methylcytosine (5mC) in either a CpG or non-CpG context, and less often N<sup>4</sup>-methylcytosine (4mC) or 5-hydroxymethylcytosine (5hmC) [4–7]. These modifications can have significant effects on the functional potential and structure of an organism's genome. While 5mC methylation has been classically attributed to repressing gene promoter activity, particularly in animals [8], this type of decoration has been noted to silence viral genes

this study are available at <https://github.com/Wilhelmlab/Truchon2025>.

**Funding:** This work was supported by grants from the Simons Foundation (735007) as well as the National Science Foundation (IOS1922958).

**Competing interests:** The authors have declared that no competing interests exist.

and retrotransposons in other eukaryotes [9,10]. In contrast 6mA methylation has been shown to exhibit a myriad of functions in prokaryotes [11]. In addition to regulating gene expression [12], DNA adenine methylation (Dam) has been associated with restriction modification (RM) systems that in tandem methylate an organism's DNA while targeting unknown DNA for degradation *via* restriction endonuclease activity [5,11,13,14]. While RM systems have not been identified in eukaryotes, evidence of 6mA methylation has been identified in single-celled eukaryotes [15,16], though certain examples may indicate residual methylation signatures driven by commensal bacteria [17]. Despite this, evidence of ancestral 6mA methyltransferases exists in multiple lineages of single-celled eukaryotes [18], including within the green algae *Chlamydomonas reinhardtii* [16].

Few analyses of the *Nucleocytoviricota* have delved into the methylation of viral genomes, despite a high abundance of DNA MTases encoded within their genomes [19,20]. Of particular significance is *Paramecium bursaria Chlorella Virus 1* (PBCV-1), which encodes five predicted DNA MTases, two of which have strongly defined functionality [19,21,22]. These two adenine-specific MTases, M. CviAI and M. CviAII, are flanked by restriction endonucleases which target GATC and CATG cut sites, respectively [21]. Methylation of the virus's own genome provides protection from these self-encoded restriction enzymes while allowing for degradation of host DNA, similarly to the Dam RM systems typical of bacteria which target invading phage DNA. For this reason, a high proportion of GATC or CATG sites are fully methylated on each strand, with most of the remaining sites being hemi-methylated [19]. Given the importance of evading restriction endonucleases for proper replication of the viral genome, universal genomic methylation is an expected functionality of these types of MTases.

The three other DNA MTases encoded by PBCV-1 have not yet been shown to be functional, though if they are, they likely target cytosine sites rather than adenine [21]. They are also not genetically colocalized with known restriction enzymes. This is the case for many DNA MTases encoded by the *Nucleocytoviricota* [20], thus bringing their function into question. Several DNA MTases were characterized in the *Pandoraviruses*, a lineage of *Nucleocytoviricota* with particularly large genomes [23], which methylate a variety of cytosine motifs and bear high phylogenetic identity to *Acanthamoeba* spp. MTases, suggesting these are genes are host derived [20]. However, *Mollivirus sibericum*, *Cedratvirus kamchatka*, and several *Marseilleviruses* encode for adenine-specific MTases and do methylate their own genomes [20]. In *Marseilleviruses* in particular, this appears to again be associated with an RM system [20,24]. Outside of this information the role of DNA methylation in viral infection is still largely unclear.

This is particularly significant regarding *Kratosvirus quantuckense*, a virus of the *Nucleocytoviricota* which infects the eukaryotic brown alga *Aureococcus anophagefferens* [25,26]. *K. quantuckense* encodes for several DNA-specific MTases [25]. However, these MTases are not homologous to the functionally defined MTases of PBCV-1, nor do they co-occur with any identifiable restriction endonucleases. The high density of MTases encoded by *K. quantuckense* merits further analysis of their

function during infection, their activity on genomic viral DNA, and the variability of sites that are targeted for methylation. Moreover, the diversity among these MTases may imply divergent functionality, possibly involving genomic stability within the virocell and regulation of DNA packaging.

Here we used Nanopore long-read sequencing to define the methylation landscape of *Kratosvirus quantuckense* strain AaV. Repeated whole genome sequencing of four biological replicates of viral DNA revealed consistent methylation patterns of both adenines and cytosines. Analyses of these methylation patterns across the genome revealed the presence of several motifs which are likely targeted by different MTases encoded by AaV. We present this information in the context of virus-host interactions, the diversity of methylation strategies across the *Nucleocytoviricota*, and how this generally overlooked aspect of the viral genome shapes the potential success of this pathogen.

## Methods

### Culture conditions

Three 750 mL batch cultures of non-axenic *Aureococcus anophagefferens* CCMP1984 were grown in ASP<sub>12</sub>A media [27] on a 12:12 light:dark cycle at 19° C and an irradiance of ~90 μmol photons m<sup>-2</sup> s<sup>-1</sup>. After one week of logarithmic growth, cultures were diluted 10:1 in ASP<sub>12</sub>A to 1 L to prevent nutrient starvation. Two days following dilution, cultures were infected with 10 mL of fresh *K. quantuckense* strain AaV in ASP<sub>12</sub>A [25]. Fresh AaV had been prepared through infection of a 25 mL culture of *A. anophagefferens* CCMP1984 with AaV, after which lysate was pushed through a Durapore 0.45-μM nominal pore-size PVDF-membrane syringe filter (MilliporeSigma; Burlington, MA) before inoculation. Once infected cultures were completely lysed (~3 d), lysate was stored at 4° C.

### Preparation of Viral Particles for DNA extraction

Lysate was prefiltered through an Isopore™ 0.4-μM nominal pore-size 25-mm diameter polycarbonate membrane syringe filters (MilliporeSigma; Burlington, MA) to remove lysed cellular material and surplus heterotrophic bacteria. Each liter of filtered lysate was then concentrated by tangential flow filtration (TFF) as previously described [28]. Briefly, lysate was sequentially concentrated with a Pellicon XL 30 kDa Cassette (MilliporeSigma; Burlington, MA) using a LabScale TFF System (MilliporeSigma; Burlington, MA). Lysate was concentrated from 1 L to approximately 25 mL, yielding a theoretical 40-fold increase in viral particle concentration. Concentrated lysate was once again filtered through a Durapore 0.45 μm pore-size PVDF membrane syringe filter (MilliporeSigma; Burlington, MA). Lysate was enumerated on a CytoFLEX flow cytometer (C07821) by gating on the Violet Side Scatter channel using a 405 nm violet laser [29] to ensure production of viral particles. To further clean AaV particles, lysate was centrifuged at 2,000 xG for 5 min to pellet remaining bacterial cells. The supernatant was moved to a clean tube and 10% Triton X-100 was added at a final concentration of 1%. To concentrate particles, the supernatant was then centrifuged at 60,000 xG for 75 min to pellet viral particles, which were subsequently resuspended in 400 μL ASP<sub>12</sub>A.

### DNA extraction and sequencing

Lysozyme (120 μL of 20 mg/mL) and 1 μL RNase (10 mg/mL) were added to the concentrated viral particles and incubated at 37° C for 30 min. Viral particles were then treated with 10 μL of proteinase K solution (20 mg/mL, 3 mM CaCl<sub>2</sub>, 200 mM Tris buffer) along with 40 μL lysis solution (0.5% SDS, 10 mM EDTA, 20 mM sodium acetate) and incubated at 55° C for 2 hours while gently shaking. DNA was extracted using a standard phenol-chloroform method. DNA concentration and purity was determined on a NanoDrop spectrophotometer and size of extracted DNA was visualized using gel electrophoresis.

DNA extracts from each replicate were individually sequenced using a MinION Mk1B sequencing device fitted with a Flongle adapter for Flongle Flow Cells (R10.4.1) (Oxford Nanopore Technologies; Oxford, UK). Sequencing libraries were prepared according to manufacturer's instructions using a V14 Ligation Sequencing Kit (SQK-LSK114) and the Flongle Sequencing Expansion (EXP-FSE002) (Oxford Nanopore Technologies; Oxford, UK).

To serve as a negative control, whole genome amplification (WGA) was performed on AaV DNA, a process which, through random hexamer amplification, generates thousands of undecorated copies of a genomic region, thus dampening the methylation signal. DNA was extracted from a fourth biological replicate of concentrated viral lysate. This DNA was diluted to approximately 3 ng/μL before performing WGA with the NEB phi29-XT WGA Kit (New England Biolabs, Ipswich, MA). This WGA DNA and an aliquot of the original, undiluted DNA from the fourth replicate were then sequenced on tandem Flongle Flow Cells (R10.4.1). All sequencing runs performed for this analysis are summarized in Table S1 in [S2 File](#).

### Methylation calling and initial analysis

POD5 files generated from long-read sequencing were aligned to the most recent version of the AaV genome [30] and called for 5mC and 6mA methylation using Bonito basecaller with the dna\_r10.4.1\_e8.2\_400bps\_hac@v5.0.0 model with a minimum q-score of 9 [31]. Supplemental basecalling was performed to verify methylated sites using Nanopore dorado v1.2 (ONT, Oxford, UK). Called and aligned reads were sorted and indexed using Samtools [32]. BedMethyl tables, which display total methylation at all genomic sites, were created using the Modkit pileup command (Oxford Nanopore Technologies). To avoid considering methylation at sites with only a small number of reads mapped (i.e., single nucleotide polymorphisms) python scripts were used to filter the bedMethyl file down to only sites that contain the respective nucleotide (either adenine for 6mA methylation or cytosine for 5mC methylation). Methylation frequency (i.e., proportion of nucleotides at a given site methylated) for each individual site was calculated using the equation:

$$\text{methylation frequency} = \frac{\# \text{ methylated nucleotides at site}}{\text{total reads mapped to site}}$$

Likewise, genomic methylation fraction was calculated by determining the proportion of all sites in the genome that meet the threshold to be considered methylated, which, for the purpose of this study, is 70% methylation frequency. For methylated site enrichment scores of specific genomic regions, every 1000 bp frequently methylated sites were counted within a sliding window of 5000 bp before being normalized by the total number of the respective site within the region. Z-scores from enrichment scores were calculated and used for Circos heatmaps.

To determine WGA-corrected methylation frequency for each site, methylation frequency scores from the WGA-amplified library BedMethyl files were subtracted from the methylation frequency scores from the respective unamplified library. Python scripts were used to determine the frequency of nucleotides surrounding highly methylated sites (>80% for 6mA and >50% for 5mC). Methylation maps were generated using python scripts and imaged using Circos.

### Methylation targeted motif identification and characterization

To identify DNA sequences likely to be targeted for methylation in the AaV genome, corrected BedMethyl tables were run through the motif detection software Nanomotif using the motif discovery function [33]. This process was performed for all three libraries and any motifs identified were retained for future analysis. Overall, nine putative targeted motifs were discovered, five of which were palindromic. For palindromic motifs, python scripts were used to pair scores belonging to the same palindrome on opposite strands to determine hemi-methylation of these sites.

To identify genomic regions in which specific motifs are enriched, the web server DistAMo was used to visualize motif distribution. Coding regions that were overrepresented with a specific motif with a Z-score  $\geq 2$  were identified for each motif. Genes that were found to be associated with a specific motif were clustered using Cytoscape where every line connects a gene to any motifs that are overrepresented in the region.

### Phylogenetic analysis of viral methyltransferases

To better characterize the DNA MTases, genes were called from sequenced *Nucleocytoviricota* genomes (including AaV) using the Nanomotif tool MTase-linker, which compares coding regions of each respective genome to the entire REBASE

database and flags likely DNA MTases. The number of DNA MTases identified through this process was normalized to both the length of the respective virus's genome as well as the number of identified coding regions for comparison across all viruses.

To generate a phylogenetic framework of *Nucleocytoviricota* DNA MTases, protein sequences for each gene were aligned using MAFFT v7.520 using a maximum number of iterative refinements of 1000[34]. Sequence alignments were then trimmed using trimAl v1.4.rev15 with a gap threshold of 0.9 and a 25% conservation[35]. A maximum-likelihood protein tree was constructed using IQ-TREE version 2.2.0.3[36] with 1000 bootstraps and visualized in IToL.

### Restriction enzyme digestion

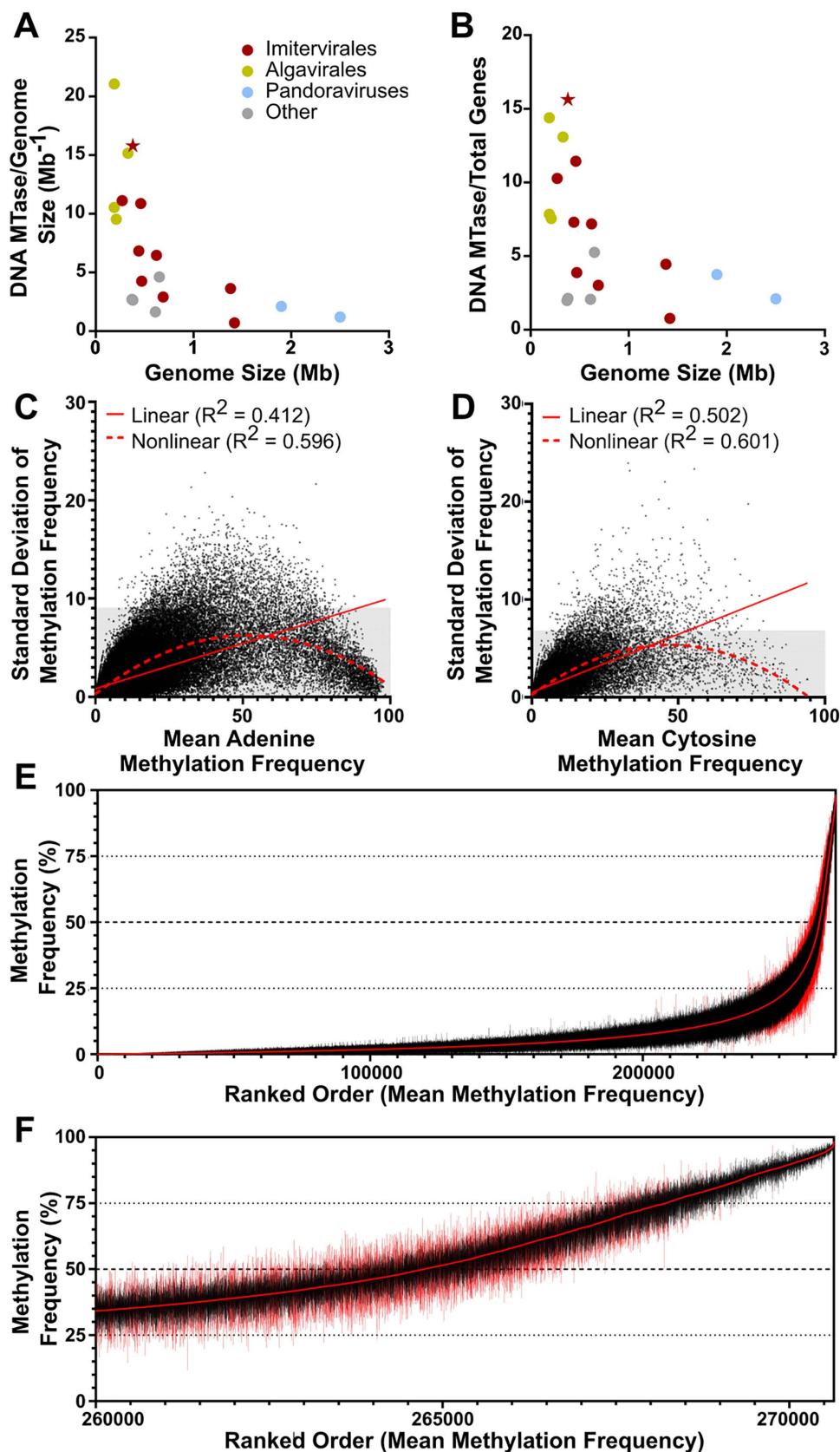
To verify results noted from motif analysis, three restriction enzymes were used to verify the presence of methylation on AaV adenines. These included Hpy166II, which targets the GTNNAC motif, Xhol, which targets the CTCGAG motif (i.e., a proxy for CTNNAG), and XbaI, which targets the TCTAGA motif (i.e., a proxy for CTAG). DpnI and DpnII were used as controls, which are expected to digest methylated and unmethylated motifs respectively. All digestions were performed with either AaV, whole genome amplified AaV, or *A. anophagefferens* DNA in rCutSmart Buffer (New England Biolabs, Ipswich, MA) for 15 min at 37° C. Reactions were inactivated at 65° C for 20 min and then visualized via gel electrophoresis.

## Results

AaV encode seven nucleotide-associated MTases, one of which is likely to act as an RNA methylase based on homologous sequences [25,30]. Among the six remaining MTases, five appear in the Restriction Enzyme Database (REBASE), a catalog identifying restriction endonucleases and MTases [37]. All are considered type II MTases, meaning they exist as distinct ORFs with no associated endonuclease domains. All have the characteristic domains of an MTase [i.e., fgg, dppy, and the DNA target recognition domain (TRD)] (Table S2 in [S2 File](#)) [38,39]. Four of these MTases fit into the  $\gamma$  subclass, signifying a motif order of fgg-dppy-TRD, while the final MTase belongs to the  $\beta$  subclass, signifying a motif order of dppy-TRD-fgg [40]. When characterizing the function of these MTases using the Nanomotif MTase-caller tool, the five genes described above are characterized as adenine-specific MTases, while an additional MTase was identified as potentially cytosine-specific. This final gene, which was not described in REBASE, contains three of the six characteristic domains of a cytosine-specific MTase (fgg, pc, and env), though it appears truncated and lacks the three terminal domains that are consistent with cytosine specific MTases (qrr, ix, and x) [41]. This gene bears a high identity to two MTases encoded by the *A. anophagefferens* mitochondrion (59.48% amino acid identity), one of which is also truncated while the other contains all six motifs. As it is unclear if the protein product for this gene is functional, it will not be further analyzed in this study. Furthermore, none of the MTases appear to be packaged in the viral capsid [42], meaning they are likely only active after transcription initiation.

When compared to other giant viruses, AaV encodes more MTases in the context of the entire genome. Despite being one of the smaller *Nucleocytoviricota* with a genome size of just ~380 kb, AaV encodes six MTases out of 384 ORFs, giving it a high ratio of MTases per genome ([Fig 1A](#)) and MTases per total encoded genes ([Fig 1B](#)) as compared to other similar viruses (Table S3 in [S2 File](#)). Of the viruses examined, these represent some of the highest rates of encoded DNA MTases, in a similar fashion to *Ostreococcus lucimarinus* Virus (OIV; 4 DNA MTases in a 190 kb genome), PBCV-1 (5 DNA MTases in a 330 kb genome), and *Phaeocystis globosa* Virus (5 DNA MTases in a 460 kb genome). While the larger *Nucleocytoviricota* like the pandoraviruses or *Bodo saltans* Virus do encode DNA MTases, none display the density seen in AaV (Table S3 in [S2 File](#)). Moreover, many amoebal *Mimiviridae*, like *Acanthamoeba castellanii* Mimivirus (APMV), do not appear to encode any traditional DNA MTases according to REBASE standards (Table S3 in [S2 File](#)).

Methylation frequency of both adenines (6mA) and cytosines in the AaV genome was consistent between each successive sequencing run (Figure S1 in [S1 File](#)). The standard deviation of the methylation frequency is lower than 10% for 99% of AaV adenines ([Fig 1C](#)) and lower than 7% for 99% of cytosines ([Fig 1D](#)). The standard deviation of the methylation frequency also follows a nonlinear quadratic regression in relation to the mean methylation frequency for



**Fig 1. Base methyltransferase abundance and methylation occurrence in AaV.** Encoded DNA methyltransferases of various *Nucleocytoviricota* normalized to genome size (A) and total coding potential (B). AaV is indicated with a star. Distribution of methylation frequency average scores and standard deviations ( $n=3$ ) among every AaV adenine (C) and cytosine (D) within the AaV genome. Gray regions of the graph represent the lower 99% of standard deviations. R-squared statistics of linear and non-linear quadratic regressions are denoted. Adenines in AaV sorted by ranked order of lowest to highest mean methylation frequency (E-F). Mean methylation scores are denoted as the red line in the center, while lines extending from the center represent the range for each respective site. Sites within the 99<sup>th</sup> percentile of standard deviation have red ranges. [Fig 1F](#) displays a magnified view of the most highly methylated sites in 1E.

<https://doi.org/10.1371/journal.pone.0330887.g001>

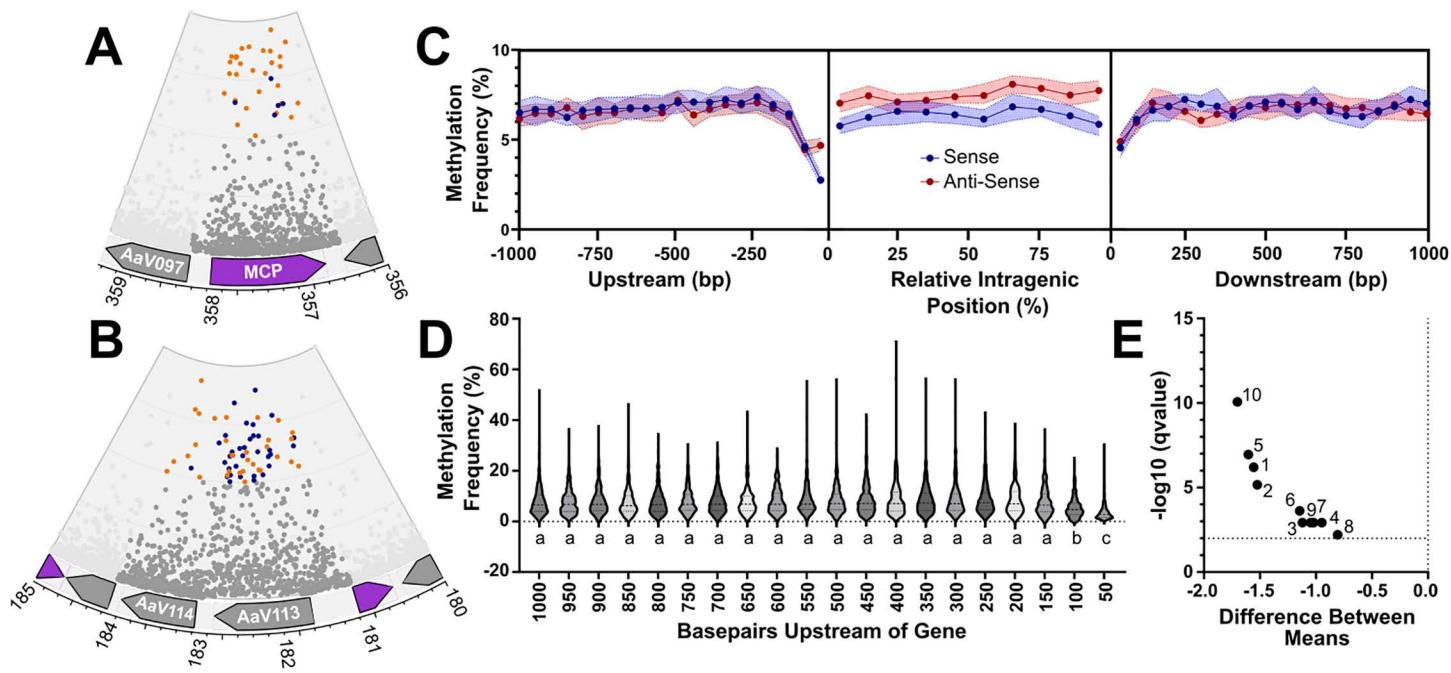
both adenines and cytosines, showing that both low methylation and high methylation sites have reduced variability ([Fig 1C-D](#)). A vast majority of adenines have low methylation scores, i.e., below 25%, while the range of methylation scores for a given site increases as the mean increases ([Fig 1E](#)). However, while it is expected for there to be a higher standard deviation around larger scores, the absolute highest methylation frequencies have very consistent scores across all libraries ([Fig 1F](#)).

Genomic maps of the methylation of AaV are detailed in [Fig 2](#). 6mA methylated adenines (Figure S2A in [S1 File](#)) occur consistently throughout much of the genome at relatively even rates. Several specific genes are highly methylated on adenine residues within the intragenic region, including the major capsid protein (MCP) in which 5.7% of adenines are methylated at a frequency over 75% (compared to 0.90% genome-wide) ([Fig 2A](#)). Cytosines returning high methylation signals are frequently found in repeat containing proteins, one of which 9.1% of cytosines are supposedly methylated at a frequency over 50% (compared to 0.63% genome-wide) ([Fig 2B](#)). Given the propensity for 5mC signals to appear in these repetitive, error-prone regions, and the inability to determine whether the AaV cytosine methyltransferase is functional, these basecalls may not represent true 5mC modifications and may instead be sequencing artifacts. Regardless, enrichment of 6mA methylation occurs in different locations depending on the strand (Figure S3 in [S1 File](#)). Adenines in the AaV genome are methylated at a rate of approximately 6.55% with 1.917% of adenines being methylated at a rate higher than 50%.

Methylation across intragenic regions, as well as the 1 kb upstream region and 1 kb downstream region for each gene were characterized collectively across the entire AaV genome. Intragenic regions for each gene were normalized into ten distinct regions, while each 50 bp window within the upstream and downstream regions were considered for methylation frequency, as per previous studies [\[43\]](#). Average methylation frequency for these sites was determined based on the total adenines in a given region, and the averaged methylation scores across said region. Notably, methylation frequency across the gene body was consistently lower on the sense strand as compared to the anti-sense strand, which averaged 6–7% and 7–9% respectively ([Fig 2C,E](#)). Despite this, methylation did not vary largely within different intragenic regions ([Fig 2C](#)).

In regions both 100 bp upstream and 50 bp downstream of coding regions, a decrease in average methylation frequency was visible ([Fig 2C](#)). This decrease was most notably exacerbated within the 50-bp immediately upstream of the gene body on the sense strand, where most AaV genes exhibit very infrequent methylation. This drop in methylation was significantly lower than the 19 other sites upstream of the gene body ([Fig 2D](#)). A similar pattern is notable on the antisense strand upstream of the gene body as well as on both strands downstream of the gene body, if not to a lesser extent.

The drop in methylation frequency immediately upstream of coding regions may be attributable to a promoter motif that has been associated with a subset of AaV genes expressed early during infection: “[AT][AT][AT][TA]AAAAATGAT[ATG] [AG][AC]AAA[AT]” [\[44\]](#). This motif lacks any of the methylation motifs defined in this study, which may explain the general decrease in methylation frequency in the promoter region. However, we found no connection between methylation in the promoter region and temporal gene expression (i.e., the time point at which a gene was first detected as transcribed) as defined by Moniruzzaman, et al., 2018 (Figure S4 in [S1 File](#)). In fact, several genes that are expressed early in the infection cycle, including MCP ([Fig 2A](#)), have multiple highly methylated sites in the promoter region.



**Fig 2. Trends in methylation frequency relative to coding regions.** The major capsid protein shows an abundance of highly methylated adenines (A), while other regions are densely populated with methylated cytosines (B). The inner ring represents 100% methylation, and the outer ring represents 0%. Sites below 50% methylation frequency are shaded gray, cytosines above 50% are colored blue and adenines above 50% are colored orange. In 2A and 2B sites within coding regions outside those of interest are not shown. Genes on the sense strand are shown in purple while genes on the antisense strand are gray. Median methylation frequency is associated with intragenic regions along with the 1000 bp upstream and downstream regions for all coding regions of AaV (C), with 95% confidence intervals around the site shown in the respective color of the strand. Distribution of methylation frequencies in the upstream sense region for each gene is displayed in violin plots (D), with different letters above each plot signifying significant statistical difference ( $p < 0.0001$ ). The difference between averages of each normalized intragenic region is displayed in a volcano plot (E), with numbers representing the position along the gene (i.e., the terminal 5' end represented by 1 and the terminal 3' end represented by 10).

<https://doi.org/10.1371/journal.pone.0330887.g002>

### Nucleotide bias surrounding adenines influences methylation

6mA methylation was found to be highly dependent on the nucleotides surrounding the adenine, indicating motif dependent DNA binding typical of MTases (Figure S5 in [S1 File](#)). Sequence specific analysis revealed that a disproportionate number of methylated adenines occur in the CTAG motif, a 4-mer which boasts an average methylation frequency ranging from 34.48 to 57.49%, a stark difference from the average methylation rate of adenines throughout the genome. While methylation frequency at these sites is relatively high, this motif is not common in the viral genome, with the “CTAGG” motif occurring approximately once every 4500 bases, while the analogous “TCAGG” motif occurs once every 360 bases. Likewise, polyadenine regions were largely unmethylated, a pattern that is particularly exacerbated by the presence of one or more adenines downstream of the adenine in question. Classical Dam motifs GATC and CATG, which have been identified in the Chlorovirus PBCV-1 and are typically associated with RM systems, bear slight methylation signals at frequencies of 8.3–18.5% and 12.3–14.7% respectively.

To test for co-occurrence of adenine methylation, the methylation frequencies of adenines within ten nucleotides of highly methylated adenines (methylation frequency  $> 80\%$ ) were collectively examined. Sites that are located 2–3 nucleotides upstream and 2 nucleotides downstream on the same strand of the highly methylated adenine are frequently hypomethylated (Figure S6 in [S1 File](#)). Interestingly, sites that are three and five nucleotides upstream of the highly methylated adenine display elevated methylation frequencies on the opposite strand (Figure S6 in [S1 File](#)). Biases that exist in favor of nucleotide occurrence surrounding a methylated adenine can also be identified using this approach.

When considering only sites surrounding a methylated adenine, the proportion of adenines at 5 and 4 bases upstream and 1 base downstream, and the proportion of guanines 1 base upstream, steadily decreased as a function of minimum methylation frequency (Figure S7 in [S1 File](#)). Likewise, the proportion of thymine at 5 bases upstream, guanine at 4 bases upstream, and cytosine 1 base upstream, increase (Figure S7 in [S1 File](#)).

### Several complex motifs identified display variable methylation levels

To search for more complicated motifs, Nanomotif was used to identify motifs in the genome that were enriched for methylation, using a stringent cut-off of 70% methylation frequency to define methylated sites. Nine motifs were identified between the three sequencing libraries ([Table 1](#)), including the CTAG motif previously identified. Other derivative motifs to CTAG were also identified, either containing additional nucleotides within (CTNNAG) or flanking the motif (CTAGY). These variations of the CTAG motif represented seven out of nine of those identified, not including GTNNAC and TGNNCA. However, these two motifs, along with CTNNAG, were identified as targeted for methylation in each library. These motifs were highly consistent in methylation frequency regardless of modification-detecting basecaller (Figure S8 in [S1 File](#)).

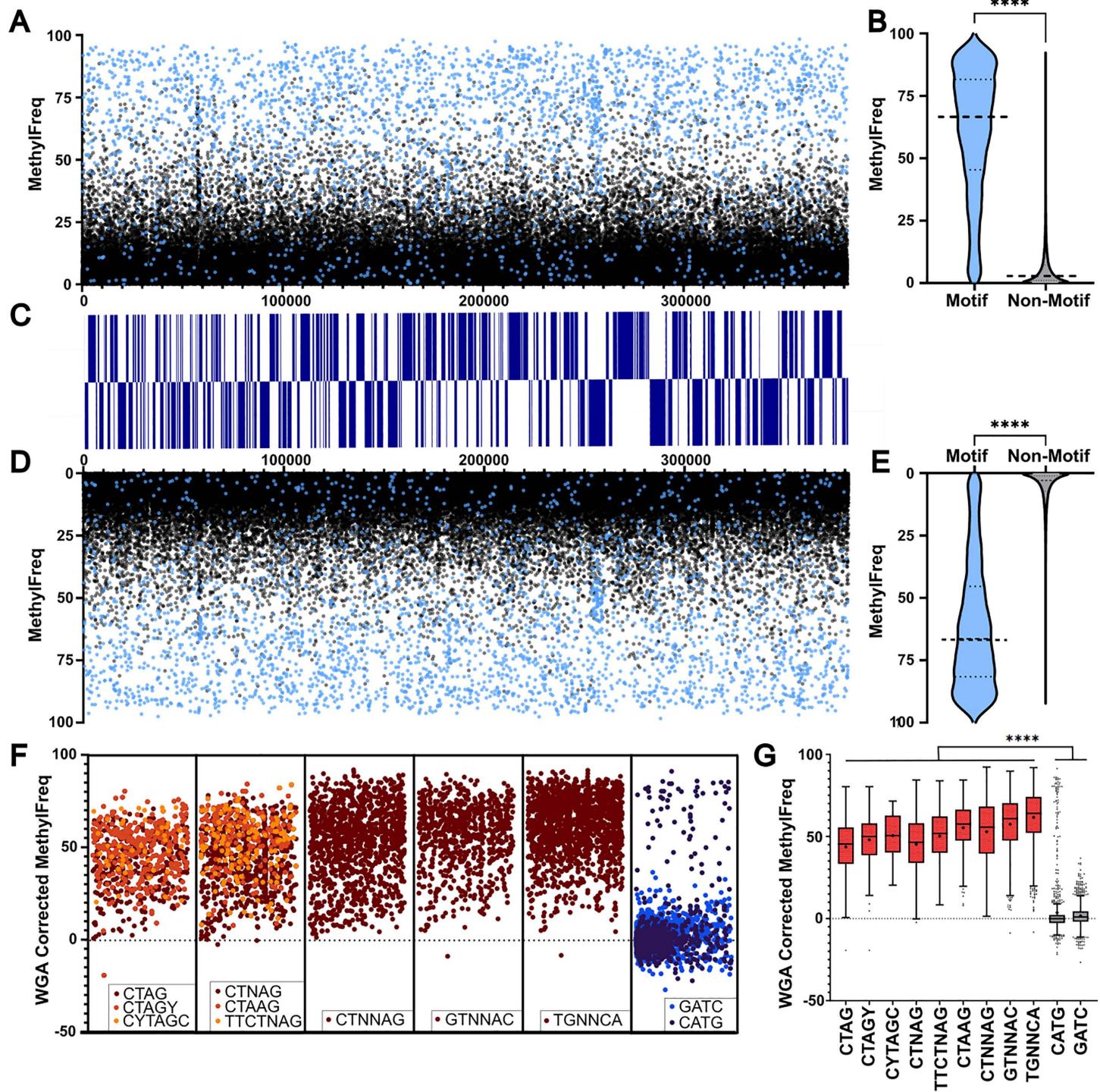
Across the nine motifs, GTNNAC and TGNNCA were also the most highly methylated, with an average methylation frequency of 74.52% and 74.08% respectively. Meanwhile, CTAG and CTNAG exhibited the lowest average methylation frequency of identified motifs, at 58.78% and 58.06% respectively. High genomic methylation fraction (*i.e.*, the proportion of sites meeting a minimum methylation threshold) generally cooccurs with methylation frequency ([Table 1](#)). Collectively, the identified motifs represent a vast majority of the methylated sites, including approximately 74% of sites with a methylation frequency > 50% and approximately 92% of sites with a methylation frequency greater than 70% ([Fig 3A–E](#)). Motifs distribution is also largely consistent with highly methylated sites found globally throughout the genome. To rule out false signals that might be attributed to sequencing errors, a WGA library was used to correct methylation frequency scores. Most sites attributed to motifs remained significantly methylated, as compared to classical methylation motifs CATG and GATC ([Fig 3F–G](#)). Interestingly, while the motifs enjoy a more cosmopolitan distribution, the CATG motif is clearly more abundant at one end of the linearized genome ([Fig 3F](#)).

Most sites displayed similar distributions regarding variance, with most standard deviations below 10% and barely any above 20% ([Fig 4](#)). Distribution of methylation frequency means for each individual site varied more depending on the motif, however ([Fig 4](#)). In certain cases, motifs targeted for methylation contain ambiguous nucleotides (*e.g.*, the ambiguous CTNAG) in one library while being defined in other libraries (*e.g.*, the specified CTAAG). Comparing the shifts in mean methylation frequency and standard deviation between ambiguous motifs and their respective specified motifs reveals that the specified motifs are both more consistently and frequently methylated ([Fig 4A–B](#)). Still, a proportion of the highly

**Table 1.** Motifs detected in at least one of three sequencing libraries of AaV. Detection in a library is represented with an X. Methylation frequency and genomic methylation fraction are averaged across all three libraries.

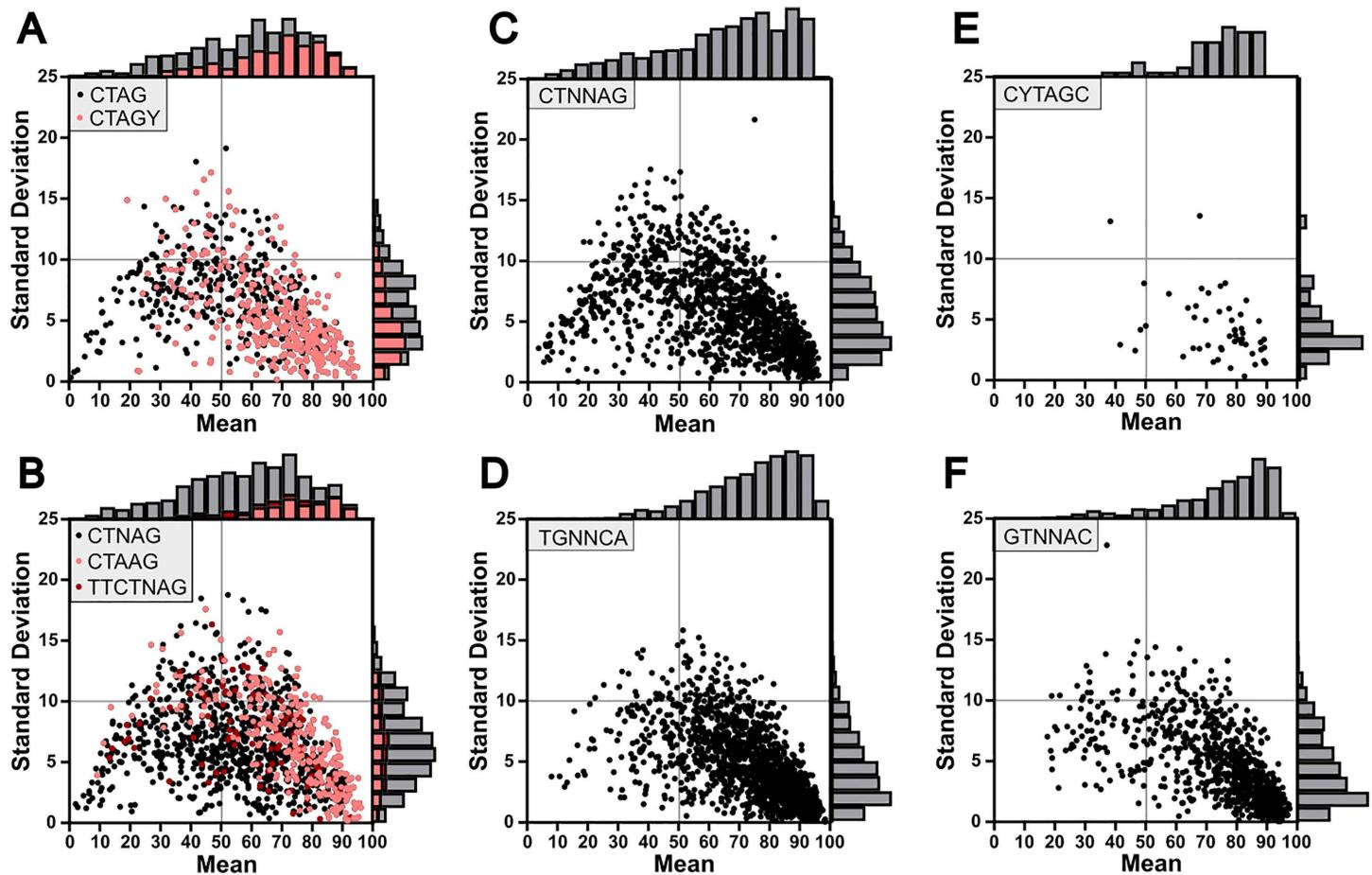
Motif	Library 1	Library 2	Library 3	Methylation Frequency (%)	Genomic Methylation Fraction (70%)
CTAG		X	X	58.78	34.92
CTAGY	X			67.39	51.44
CTNAG	X	X		58.06	33.06
TTCTNAG			X	66.94	52.00
CTAAG			X	72.60	65.04
CTNNAG	X	X	X	63.73	47.19
CYTAGC		X	X	73.56	68.82
GTNNAC	X	X	X	74.52	69.34
TGNNA	X	X	X	74.08	66.03

<https://doi.org/10.1371/journal.pone.0330887.t001>



**Fig 3. Methylation frequency of characterized 6mA motifs on a genomic scale.** Genomic maps show mean methylation frequency for each site on the positive strand (A) and negative strand (D) colored based on whether the site is associated with one of the 9 identified motifs (blue) or not (black). Respective coding regions for each strand are shown in blue (C). Violin plots for each strand detail the differences between motif-associated and non-motif sites on positive (B) and negative (E) strands. Scores for the identified motifs as well as the two classical recognition motifs GATC and CATG were corrected with WGA library scores (F). Each box in F represents a linearized genome meaning each site corresponds to its respective position. Tukey's box plots were generated to show the distribution of WGA-corrected methylation frequencies, with means denoted as plus symbols (G). \*\*\*\*:  $p < 0.0001$ .

<https://doi.org/10.1371/journal.pone.0330887.g003>



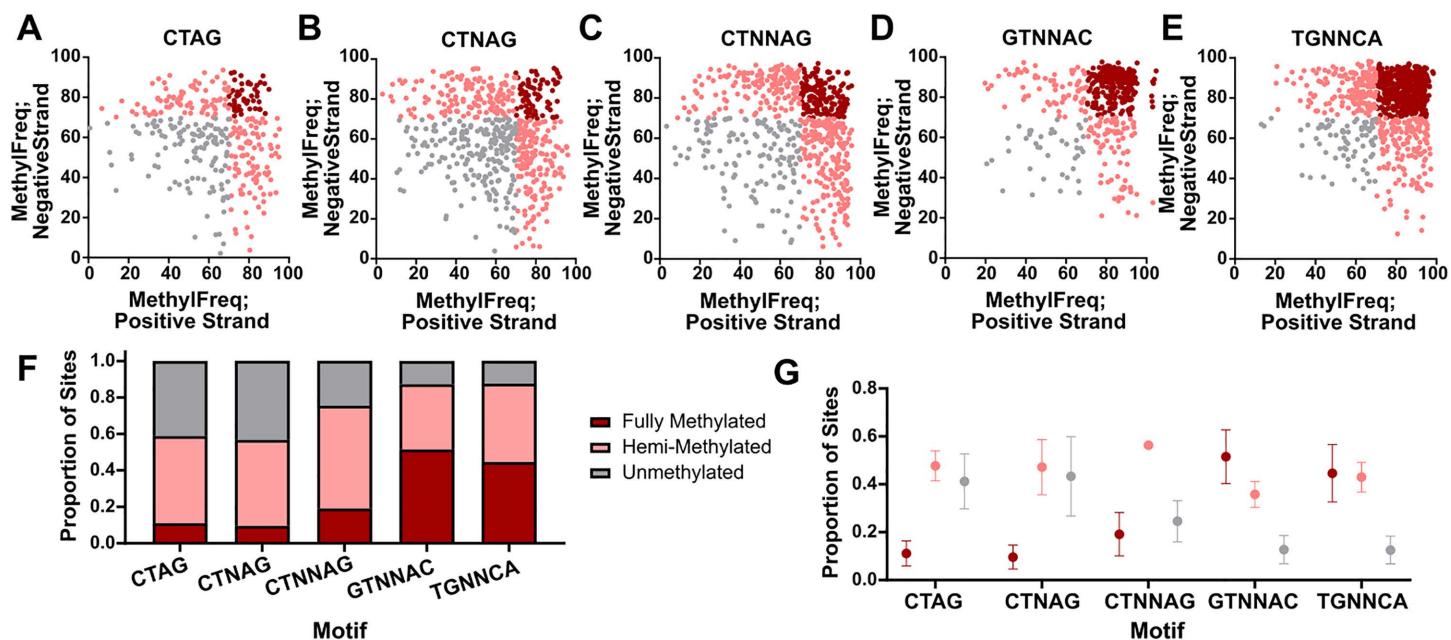
**Fig 4. Distribution of methylation frequency for AaV methyltransferase targeted motifs.** Distribution of methylation score averages versus standard deviations for each site belonging to a respective motif, with derivative motifs clustered into single plots (A-B). Histograms detail total counts of sites within respective regions.

<https://doi.org/10.1371/journal.pone.0330887.g004>

methylated sites belong to the ambiguous motif only, implying that the specified motif does not completely account for all the associated methylation.

Considering that five motifs are palindromic (*i.e.*, the reverse complement sequence is identical), we sought to determine rates of full and hemi-methylation. To be considered a fully methylated motif, both respective adenines on the positive and negative strand had to reach a minimum methylation frequency of 70%, whereas only one strand being methylated at this level was considered hemi-methylation. The palindromic motifs that are derivatives of the CTAG motif, thus CTAG itself, CTNAG, and CTNNAG, all had low levels of full methylation (<20%) and were hemi-methylated over 50% of the time (Figure 5A-C). The GTNNAC and TGNNCA motifs, while still displaying high rates of hemi-methylation, were much more likely to be fully methylated (40–60% of sites) while also being much less likely to be completely unmethylated than the CTAG derivatives (Fig 5D-E). In these respective motifs, a significant increase in proportional full methylation was detected as compared to the CTAG derivatives (Fig 5F-G). Variation in reciprocal methylation patterns tended to appear higher across sequencing libraries as compared to other measures of variation throughout the study (Fig 5G), though this is likely driven by the hard cutoff of 75% used to define complete methylation.

To further verify the authenticity of the defined methylation motifs, a series of restriction enzyme digests were performed using the enzymes Hpy166II (targeting GTNNAC), Xhol (targeting CTCGAG), and XbaI (targeting TCTAGA). DpnI



**Fig 5. Patterns in reciprocal methylation among palindromic motifs.** Reciprocal methylation plots for all palindromic motifs with fully methylated sites displayed in dark red, hemi-methylated sites displayed in pink, and unmethylated sites displayed in gray. Proportion of all sites for each motif (F) and average and the standard deviation of each proportion (G) are denoted.

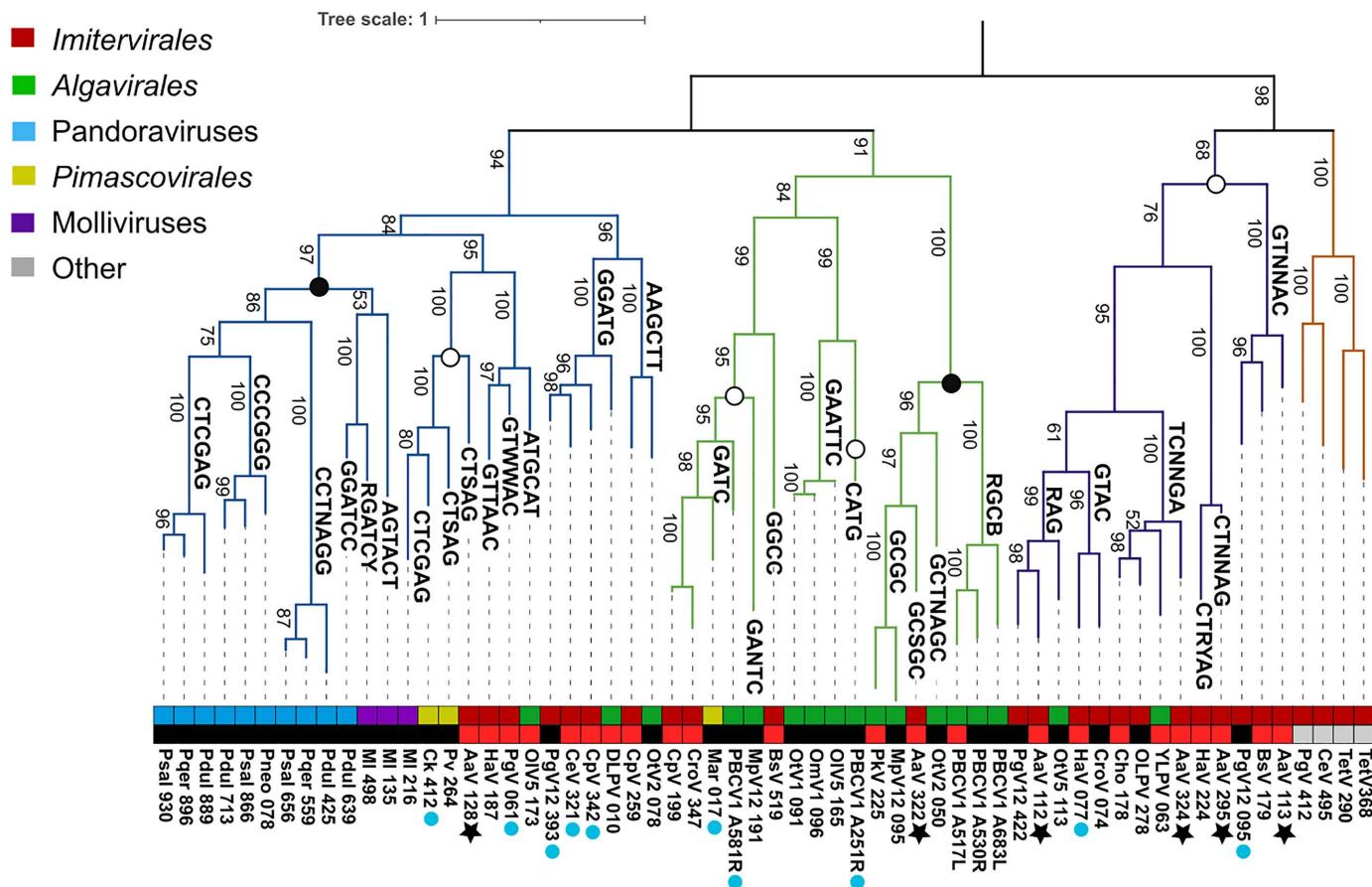
<https://doi.org/10.1371/journal.pone.0330887.g005>

and DpnII only cleave GATC in which the adenine is either methylated or unmethylated, respectively. Considering GATC sites are apparently unmethylated in AaV, these enzymes were used as negative and positive controls for the restriction digests. None of these enzymes were capable of degrading raw viral DNA, though degradation by DpnII showed the viral genome is definitively not methylated at GATC sites (Figure S9 in [S1 File](#)). WGA viral DNA was clearly digested by Hpy166II, with potential digestion by Xhol and XbaI as well, compared to the negative control (Figure S9 in [S1 File](#)). What's more, *A. anophagefferens* DNA was heavily digested by both Hpy166II and Xhol, implicating divergent methylation patterns between host and virus. As neither Xhol nor XbaI showed strong digestion on AaV or amplified DNA, likely due to the motifs' infrequency in the viral genome, both were used concurrently. While the two enzymes together showed digestion of amplified DNA, unamplified DNA was unaffected (Figure S10 in [S1 File](#)).

### Phylogenetics of AaV methyltransferases may signify unique origins of methylation targets

The six DNA MTases encoded by AaV were placed into a phylogenetic tree of *Nucleocytoviricota* MTases which have either been shown to methylate a specific site [20] or are predicted to do so based on homology to defined sequences in REBASE ([Fig 6](#)). The resulting phylogeny can be broken into four distinct clades, of which three contain AaV MTases.

Clades I and II contain one AaV MTase each (Figure S11 in [S1 File](#)). Clade I contains one AaV MTase (AaV\_128) which is predicted to target the CTSAG motif based on REBASE homology. This is also the only type II  $\beta$  MTase encoded by AaV, as all others are of the  $\gamma$  subclass. Despite other *Imitervirales* and *Algavirales* methyltransferases in this clade, the AaV MTase bears relatively increased similarity (approximately 35%–40% amino acid similarity) to those of Cedratvirus, Pithovirus and *Mollivirus sibericum*, which again target either a CTNNAG motif derivative or CTNAG motif derivative. Notably, while REBASE initially predicted that the Cedratvirus MTase would target CTSAG as well, PacBio sequencing revealed that it instead targets the CTCGAG motif [20]. This may imply that the related AaV MTase either targets the CTNAG motif or the CTNAG motif.



**Fig 6. Protein tree of *Nucleocytoviricota* type II DNA methyltransferases.** Branches which have a defined or predicted motif are labeled as such. The upper colored bar defines the respective family for each virus while the lower colored bar defines the quality of the predicted target sequence, black meaning the MTase is predicted to target the corresponding sequence by REBASE and red meaning the MTase is predicted to target the corresponding sequence based on homology to other MTases alone. Grey boxes had no homology to any MTases in REBASE. Blue dots below sequence IDs represent the presence of a neighboring restriction endonuclease. Black stars indicate AaV MTases. Dots on nodes signify the targeted base of all known MTases within the clade, either cytosine (black) or adenine (white). Bootstraps are displayed on branches. Psal: Pandoravirus salinus; Pdul: Pandoravirus dulcis; Pqer: Pandoravirus quercus; Pneo: Pandoravirus neocaledonia; Ml: Mollivirus; Ck: Cedratvirus kamtchatka; Pv: Pithovirus sibericum; AaV: *Aureococcus anophagefferens* Virus; PgV: *Phaeocystis globosa* Virus; CeV: *Chrysochromulina ericina* Virus; CpV: *Chrysochromulina parva* Virus; DLPV: *Dishui Lake Phycodnavirus*; EhV: *Emiliania huxleyi* Virus; MpV: *Micromonas pusilla* Virus; Mar: *Marseilleivirus*; CroV: *Cafeteria roenbergensis* Virus; OIV: *Ostreococcus lucimarinus* Virus; OtV: *Ostreococcus tauri* Virus; OmV: *Ostreococcus mediterraneus* Virus; HaV: *Heterosigma akashiwo* Vius; TetV: *Tetraselmis* Virus; BsV: *Bodo saltans* Virus.

<https://doi.org/10.1371/journal.pone.0330887.g006>

Interestingly, Clade I contains a cluster of cytosine specific MTases (subclass IA; Figure S11 in [S1 File](#)), which belong to either the Pandoraviruses or Molliviruses. A monophyletic sub-clade of these genes belongs exclusively to the Pandoraviruses with zero homologs outside of this specific viral family. While these enzymes perform 5mC methylation [20], two of the described motifs mirror the CTNAG and CTNNAG motifs identified in the AaV methylome. In fact, the CTCGAG motif appears in both several Pandoraviruses as well as *Cedratvirus kamtchatka*, though cytosine and adenine are targeted respectively between the two viral MTases.

Clade II is primarily comprised of MTases of the *Algavirales*, among which is the PBCV-1 gene M.CviAll, which targets the adenine in the CATG motif and M.CviAll which targets the adenine in the GATC motif [45]. Despite this, this clade varies in both the predicted motif for methylation as well as type of methylation (6mA or 5mC). Still, distinct lineages including

the Marseilleviruses and the *Imitervirales* are present throughout this clade. The truncated cytosine-specific MTase AaV\_322 falls within subclade IIC (Fig 6, Figure S11 in [S1 File](#)), which has 100% bootstrap support as distinct from the adenine-specific enzymes in this clade.

Clade III contains the four remaining AaV MTases, though they cluster into three distinct subclades supported by high bootstrap values (Fig 6, Figure S11 in [S1 File](#)). Notably, all defined MTases in this clade are predicted to target adenine. Two AaV MTases are in subclade IIIA, which are predicted to target an ambiguous RAG motif and a TCNNGA motif. The MTases identified in this clade belong to both *Algavirales* and *Imitervirales* and do not group taxonomically. While the RAG targeting MTase may be unclear, the TCNNGA targeting MTase may be instead represent the methylation of the TGNNA motif. The only associated MTase that is defined by REBASE is that of the Organic Lake phycodnavirus, which is only characterized based on the prototype restriction endonuclease Hpy178III.

Subclade IIIB and IIIC contain the final two AaV MTases which are predicted to target the GTNNAC motif and the CTN-NAG motif. The MTases identified in these subclades are exclusively *Imitervirales* specifically belonging to the Mesomimiviridae, including Heterosigma akashiwo Virus, Phaeocystis globosa Virus, and Bodo saltans Virus. While the defined motifs for these viruses are largely based on homology alone, the further appearance of the experimentally verified motifs from the AaV methylome in these clades is a strong indication of their targeted sequence. With this information in mind, the expression level of AaV MTases during infection was identified from a previous study [44]. Notably, the truncated cytosine-specific MTase was expressed at the highest level within 12 hours of infection (Figure S12 in [S1 File](#)). Of the adenine-specific MTases, the gene predicted to target the GTNNAC region was also highly expressed at the 12 hpi and the highest expressed MTase at 23 hpi (Figure S12 in [S1 File](#)).

### Genes cluster based on enrichment of specific motifs

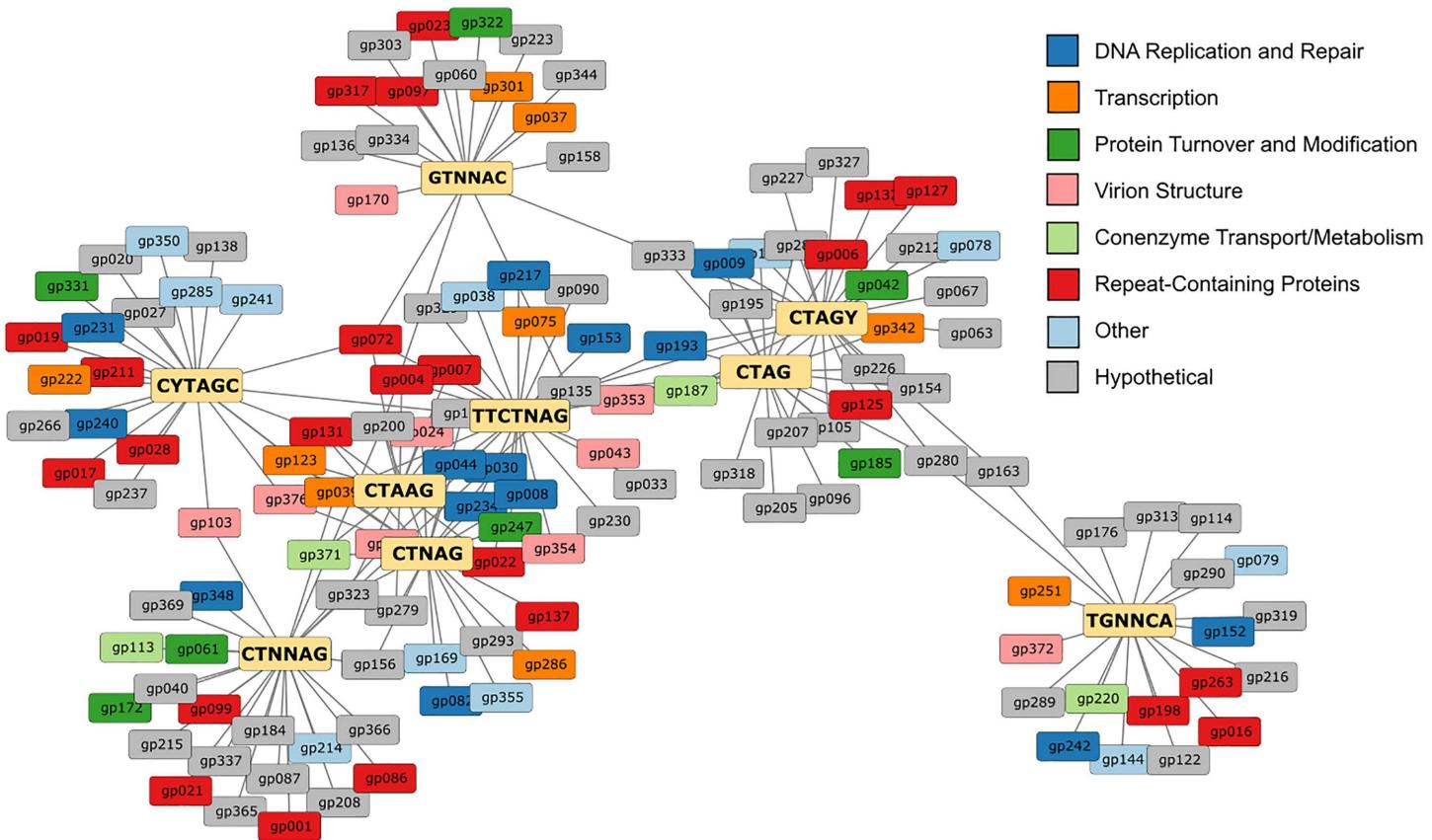
To determine whether the identified motifs were enriched in ORFs encoded by AaV, we used the DistAMo web tool to map the frequency of each motif relative to genomic position. Genes with a z-score greater than two were identified for each motif and annotated functionally from the AaV reference genome (Table S3 in [S2 File](#)). Genes were then clustered in Cytoscape with a single connection between a gene and a motif signifying overrepresentation of the given motif in said gene (Fig 7).

Overall, 135 genes (35% of all predicted coding regions in AaV) were enriched with at least one of the identified motifs. Each individual motif is enriched within 18–25 genes. 55 of the overrepresented genes (41%) could be functionally annotated, which is elevated from the full genome of AaV in which only ~25% of genes can be assigned to functional clusters of orthologous groups (COGs). While there was no obvious relationship in function for genes clustered around a specific motif, several important COG categories were represented across all motifs. Fourteen genes related to DNA replication and repair were identified, including several DNA polymerases, along with nine related to transcription and another nine related to virion structure, including the major capsid protein (Table S3 in [S2 File](#)).

While certain genes are often enriched with multiple motifs, most clustering appears to primarily occur between closely related motifs (i.e., CTAG and CTAGY). Motifs like CTNNAG, CYTAGC, GTNNAC, and TGNNA are largely isolated with few connections to other clusters. Repeat-rich proteins are also detected in every cluster, many of which belong to the domain of unknown function (DUF) 285 family. While these proteins are present in the same family, their genetic makeup seems to differ heavily, with one motif repeating dozens of times and none of the other motifs being detectable.

### Discussion

Laboratory studies as well as metagenomic and metatranscriptomic data from natural systems have yielded remarkable discoveries among the “giant viruses” [46,47]. Yet beyond nucleotide sequence, little is known about their genomes with respect to the relevance of epigenetic modification of *Nucleocytoviricota* DNA. This is of particular importance considering the wide range of *Nucleocytoviricota* species as well as the high density of DNA MTases encoded by certain viruses relative to the total genomic coding potential (see [Fig 1B](#)). It should also be noted that much of the methyltransferase



**Fig 7. Gene clusters formed based on linkages between motif enrichment.** Characterization for motif enrichment within a coding region was determined based on DistAMo results. Only genes with a z-score for motif enrichment greater than 2 were considered. Each motif (yellow) is connected to any genes that are enriched with the given motif. Different COG classifications of genes are denoted by color.

<https://doi.org/10.1371/journal.pone.0330887.g007>

identification in other *Nucleocytoviricota* was performed using the REBASE database which is primarily composed of bacterial restriction modification systems. Thus, there may be additional methyltransferases in these genomes not identified due to database biases.

Methylation of viral DNA has been shown to influence the number of infectious viral particles produced during lysis [48] and allows for resistance to DNA digestion via endonuclease activity [19,20]. Given DNA methylation has also been ascribed to virus-resistance mechanisms in both bacteria [13] and eukaryotes [49], the physical proximity between viral and host genetic material may serve as a means for acquisition and repurposing of host MTases to allow for invading genome survival. Thus, characterization of methylation profiles across viral genomes may improve our ability to predict infection outcomes.

Annotation of a given sequence's methylation state is generally untethered from traditional sequencing approaches: it has historically required lengthy, multi-step analyses like whole genome bisulfite sequencing and restriction enzyme-based analyses. Yet advancements in long-read sequencing have provided the ability to detect methylation through traditional library preparation techniques [50]. Moreover, analysis with the Nanopore MinION Flongle attachment allows for fiscally reasonable high throughput genotyping of small bacterial and viral genomes with both high genomic coverage and read length [51]. This can then allow for the assembly and subsequent read mapping to a complete genome, providing high-confidence methylation scores for adenines and cytosines in a genome [52].

To further our understanding of how giant viruses utilize MTases, we completed a comprehensive genomic sequencing approach of the algal virus *Kratosvirus quantuckense* strain AaV using Nanopore long-read sequencing. By sequencing the AaV genome in biological triplicate, we have provided a perspective on not only the methylation state of the viral genome, but also the consistency of methylation across multiple generations of packaged viral DNA. From this we defined nucleotide biases that influence the targeted sequences for methylation and characterized several unique DNA methylation motifs which can be loosely ascribed to specific AaV MTases. Based on the distribution of methylation across the AaV genome, we suggest genomic methylation may serve as both functional and ancestral traits in *Nucleocytoviricota* genomes.

### Detection of AaV Methylation is Highly Repeatable

Adenine and cytosine methylation of the AaV genome was consistent, with standard deviations for replicate infections of host cultures consistently below 10% for most sites. Considering different environmental conditions can alter genomic methylation [2,53], the homogeneity of the highly methylated sites serves as an important baseline for future studies. The relationship between mean methylation frequency and standard deviation for a given site prefers to follow a quadratic association, with the highest amount of variation detected in the sites that average a moderate methylation frequency, around 50%. This suggests that the sites responsible for the highest methylation frequency scores are universally targeted for methylation, at least under the conditions tested here. From a consistency standpoint, this may indicate a strong selective pressure for methylation of these sites for the propagation of viral progeny. Meanwhile, the sites with moderate levels of methylation may be inconsistently methylated due to a lack of the same pressure, possibly the result of non-specific binding of an MTase to a site other than its preferred recognition domain.

While most nucleotides are definitively unmethylated across the AaV genome, adenine-specific methylation clearly displays higher methylation frequency and genomic methylation fraction compared to all cytosine-specific methylation. This may be attributed to the sheer number of MTases predicted to be adenine-specific (five) compared to only a single cytosine-specific MTase. However, the fact that, within 12 hours of infection, the normalized read abundance of this cytosine-specific MTase is often more than double that of the adenine-specific MTases (Figure S11 in [S1 File](#)), reveals a false equivalence. Higher MTase expression does not correspond to increased methylation levels. Thus, it is likely that the cytosine-specific MTase is either extremely inefficient or provides an additional function for the virus during infection, possibly in nucleotide metabolism or even methylation of the host cytosines as a potential gene silencer [54]. We note that the host *A. anophagefferens* also encodes cytosine-specific MTases, and thus methylated cytosine residues on the viral genome may be a result of host MTase activity.

Spatial distributions of highly methylated sites imply a functional role in this activity among both cytosines and adenines. Several genome regions are heavily enriched in highly methylated cytosines, with these areas primarily being repeat-dense. This may be representative of altering the steric hindrance around these sites to influence the formation of tertiary structures in the DNA strand during synthesis or affect the binding of transcriptional proteins [55,56]. Likewise, the significant decrease in methylated adenines in the region immediately upstream of coding regions may imply that methylation of viral DNA can act as a deterrent to transcription initiation, at the very least. Methylation shifts in the promoter region are highly common in many eukaryotic systems [1,4]. However, as there are no linkages between promoter region methylation frequency and time of expression during infection, it does not appear to restrict transcription. For example, the major capsid protein, which is expressed within the first five minutes of infection [44], contains an adenine methylated at approximately 70% immediately upstream of the start codon. Thus, transcription can proceed even in the context of high methylation in the promoter.

### Targeted methylation is influenced by divergent nucleotide motifs

While initial attempts at identifying nucleotide biases surrounding methylated sites revealed several preferred nucleotide pentamers, particularly in association with the “CTAGN” motif, very few pentamers had truly high methylation

frequencies. Further motif analysis revealed that this is likely the effect of many of the defined motifs containing multiple ambiguous nucleotides. Among the motifs identified, GTNNAC and TGNNCA were targeted at the highest methylation frequency and genomic methylation fraction, two motifs that have largely not been identified as targets for viral DNA methylation. CATG and GATC, two motifs functionally characterized as methylation targets in PBCV-1 [19], were unmethylated, for the most part, pointing to a unique ancestry of MTases in AaV as well as the possibility of a diversified function.

Seven motifs identified within the AaV methylome ascribe to a CTAG-like structure, with generic and specific forms being identified across the three sequencing libraries. These include the generic CTAG being specified into CTAGY and CYTAGC and the generic CTNAG being specified into CTAAG and TTCTNAG. In all cases, specification of a motif yielded an increase in average methylation frequency and genomic methylation fraction, though specification did not account for all the highly methylated sites of the original generic motif. This may be an indicator of a “leaky methylation” phenotype. As such, there are likely five adenine-specific DNA MTases encoded by the AaV genome, despite the presence of nine identified motifs. This suggests that some MTases are responsible for methylation of multiple motifs. Thus, a MTase may have an optimal DNA recognition sequence which ends up very consistently methylated (*i.e.*, CTAAG), but may also non-specifically bind to other similar derivatives of the same motif, explaining why CTNAG is frequently recognized by motif detection software. Given seven of the nine identified motifs possess at least one ambiguous nucleotide, it is possible that many of the AaV MTases have become leaky with time, allowing for more varied methylation across the genome.

Nucleotide proportion as a function of minimum methylation frequency displays unique distribution depending on the nucleotide and site in question, in some cases appearing largely linear while in other cases displaying polynomial distributions (Figure S7 in [S1 File](#)). This could be reflective of the motif that is causing the distribution. For instance, none of the identified motifs contain adenine immediately following the 6mA site, explaining the rapid decrease in adenine proportion at that site. Meanwhile, a thymine five bases upstream of the 6mA is likely to belong to the TGNNCA motif, which could also be said regarding a cytosine one base upstream. Likewise, a guanine four bases upstream of the 6mA likely represents the GTNNAC motif.

Importantly, methylation of these motifs appears to be highly consistent regardless of the basecaller that is used to identify modified bases (Figure S9 in [S1 File](#)). While Bonito basecaller was used initially, comparing methylation frequencies to what was found using the Dorado basecaller revealed that certain adenines may have had comparatively inflated methylation frequencies. Methylation scores have been noted to vary depending on the basecaller used [57]. However, the fact that the identified motifs retain their high methylation frequency helps to verify our results and imply these sites are in fact targeted within the AaV genome.

### Phylogenetics and genomic distribution of motifs reveals importance of methylation in AaV

Phylogenetic analysis of the AaV adenine-specific DNA MTases has allowed for the association with several possible DNA recognition sequences which correspond to previously identified motifs. The clade I MTase likely targets CTNAG or CTNNAG, the clade IIIC MTase likely targets GTNNAC, and the clade IIIB MTase likely targets CTNNAG. While one clade IIIA MTase is predicted to target TCNNGA, this is instead most likely representative of the TGNNCA motif which was identified in the AaV methylome. The other clade IIIA MTase is predicted to target an RAG sequence, which is unlikely to be an actual motif targeted by any AaV MTases due to its simplicity and abundance in the AaV genome. It is unclear whether this MTase is representative of one of the identified motifs, as many of them contain an adenine followed by a guanine, or whether it instead targets a motif that was unidentified by our analysis.

The presence of a MTase in clade I is particularly intriguing considering there are seemingly no closely related *Mimiviridae* MTases and this gene is more phylogenetically like those of the Pithoviruses and Molliviruses than anything else. The origin of such an MTase is thus perplexing, though these may have been acquired from an ancestral eukaryotic host.

Additionally, confounding is the lack of type II MTases encoded by the classical *Mimiviridae* including Acanthamoeba polyphaga Mimivirus (APMV), Moumouvirus, and Megaviruses [58,59]. While the closely related “extended *Mimiviridae*” or *Mesomimiviridae* including AaV appear to often encode multiple DNA MTases, these amoebal viruses are apparently unmethylated [20]. This seems to suggest that these genes can be lost, at least in the context of some host backgrounds, though most algal viruses appear to encode at least one MTase.

Though adenine-specific motifs are generally distributed evenly throughout the AaV genome, they are overrepresented in certain areas. Identifying genes located in these areas reveals that certain genes may be more likely to be affected by methylation of some motifs as compared to others. Clustering of genes based on their connection to the nine defined motifs showed that while there is an overlap between closely related motifs, divergent motifs largely affect different genes.

Individual clusters of genes were not found to share any connections based on gene ontology or ancestral source (i.e., viral or host-derived). However, the formation of clusters does suggest that in many cases there is not a functional redundancy introduced by the five AaV adenine MTases. In accordance with the phylogenetic analysis, it is unlikely that the MTases target the same motif and thus they end up targeting different regions of the genome as well as affecting different genes. The heterogeneity of MTases in this virus mirrors that of certain bacteria, including *Helicobacter pylori* and *Microcystis aeruginosa*, which may encode between 20 and 50 methylation systems in a single genome [13,60,61]. This creates a paradigm in which the expression of all MTases during infection is important, as absence of one of the MTases may leave certain areas undermethylated. Whether this would result in improper packaging of viral DNA, degradation by host endonucleases, or shifted gene expression is yet unknown, but the prevalence and distribution of methylation across the viral genome still implies functional significance.

### Concluding remarks

Beyond defining the methylome of AaV, our analysis has shown that the annotation and characterization of methylation on a large viral genome can be performed at a relatively low cost and high efficiency. We sequenced the AaV genome a total of four times in this analysis, which yielded three characterized methylomes which displayed high consistency. Our characterization was further justified when WGA-sequencing showed that amplification of the AaV genome (i.e., a negative control) quenched the methylation signal reported in the pipeline, supporting our consensus that the sites described are in fact methylated and are not a byproduct of the sequencing technology used. Collectively this base methylome for the virus provides the opportunity for research into the factors that influence methylation and the downstream effects of differential methylation. Nutrient stress has been associated with changes in methylation in some plant species, and reducing the activity of MTases during infection may reveal more about the nature of methylation in *Nucleocytoviricota* infection. Similar observations like those made in plants as well as other mechanistically focused studies moving forward can now consider how genomic modifications can constrain biological function and/or success beyond the base-code of a genome.

### Supporting information

**S1 File. Epigenome analysis of an algae-infecting giant virus reveals a unique methylation motif catalogue – Supporting Figures.**

(DOCX)

**S2 File. Epigenome analysis of an algae-infecting giant virus reveals a unique methylation motif catalogue – Supporting Data.**

(XLSX)

**S3 File. Original unadjusted gel images included in this study.**

(PDF)

## Acknowledgments

We thank Brittany Zepernick, Robbie Martin, Tim Sparer, Brad Binder, Todd Reynolds, Katelyn Houghton, Laura Smith, and Kennedi Hambrick for discussions regarding this work.

## Author contributions

**Conceptualization:** Alexander R. Truchon, Steven W. Wilhelm.

**Data curation:** Alexander R. Truchon, Erik R. Zinser, Steven W. Wilhelm.

**Formal analysis:** Alexander R. Truchon, Erik R. Zinser.

**Funding acquisition:** Steven W. Wilhelm.

**Investigation:** Alexander R. Truchon, Erik R. Zinser.

**Methodology:** Alexander R. Truchon.

**Project administration:** Steven W. Wilhelm.

**Visualization:** Alexander R. Truchon.

**Writing – original draft:** Alexander R. Truchon.

**Writing – review & editing:** Alexander R. Truchon, Erik R. Zinser, Steven W. Wilhelm.

## References

1. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*. 2006;126(6):1189–201. <https://doi.org/10.1016/j.cell.2006.08.003> PMID: 16949657
2. Mager S, Ludewig U. Massive Loss of DNA Methylation in Nitrogen-, but Not in Phosphorus-Deficient Zea mays Roots Is Poorly Correlated With Gene Expression Differences. *Front Plant Sci*. 2018;9:497. <https://doi.org/10.3389/fpls.2018.00497> PMID: 29725341
3. Hoelzer K, Shackelton LA, Parrish CR. Presence and role of cytosine methylation in DNA viruses of animals. *Nucleic Acids Res*. 2008;36(9):2825–37. <https://doi.org/10.1093/nar/gkn121> PMID: 18367473
4. Schmitz RJ, Lewis ZA, Goll MG. DNA Methylation: Shared and Divergent Features across Eukaryotes. *Trends Genet*. 2019;35(11):818–27. <https://doi.org/10.1016/j.tig.2019.07.007> PMID: 31399242
5. Ratel D, Ravanat J-L, Berger F, Wion D. N6-methyladenine: the other methylated base of DNA. *Bioessays*. 2006;28(3):309–15. <https://doi.org/10.1002/bies.20342> PMID: 16479578
6. He B, Zhang C, Zhang X, Fan Y, Zeng H, Liu J, et al. Tissue-specific 5-hydroxymethylcytosine landscape of the human genome. *Nat Commun*. 2021;12(1):4249. <https://doi.org/10.1038/s41467-021-24425-w> PMID: 34253716
7. Rodriguez F, Yushenova IA, DiCorpo D, Arkhipova IR. Bacterial N4-methylcytosine as an epigenetic mark in eukaryotic DNA. *Nat Commun*. 2022;13(1):1072. <https://doi.org/10.1038/s41467-022-28471-w> PMID: 35228526
8. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*. 2010;328(5980):916–9. <https://doi.org/10.1126/science.1186366> PMID: 20395474
9. Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S, et al. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science*. 2001;292(5524):2077–80. <https://doi.org/10.1126/science.1059745> PMID: 11349138
10. Cabrera A, Edelstein HI, Glykofrydis F, Love KS, Palacios S, Tycko J, et al. The sound of silence: Transgene silencing in mammalian cell engineering. *Cell Syst*. 2022;13(12):950–73. <https://doi.org/10.1016/j.cels.2022.11.005> PMID: 36549273
11. Barras F, Marinus MG. The great GATC: DNA methylation in *E. coli*. *Trends Genet*. 1989;5(5):139–43. [https://doi.org/10.1016/0168-9525\(89\)90054-1](https://doi.org/10.1016/0168-9525(89)90054-1) PMID: 2667217
12. Sitaraman R. Helicobacter pylori DNA methyltransferases and the epigenetic field effect in cancerization. *Frontiers in Microbiology*. 2014;5.
13. Papoulis SE, Wilhelm SW, Talmy D, Zinser ER. Nutrient Loading and Viral Memory Drive Accumulation of Restriction Modification Systems in Bloom-Forming Cyanobacteria. *mBio*. 2021;12(3):e0087321. <https://doi.org/10.1128/mBio.00873-21> PMID: 34060332
14. Gao Q, Lu S, Wang Y, He L, Wang M, Jia R, et al. Bacterial DNA methyltransferase: A key to the epigenetic world with lessons learned from proteobacteria. *Front Microbiol*. 2023;14:1129437. <https://doi.org/10.3389/fmicb.2023.1129437> PMID: 37032876
15. Luo G-Z, Wang F, Weng X, Chen K, Hao Z, Yu M, et al. Characterization of eukaryotic DNA N(6)-methyladenine by a highly sensitive restriction enzyme-assisted sequencing. *Nat Commun*. 2016;7:11301. <https://doi.org/10.1038/ncomms11301> PMID: 27079427

16. Fu Y, Luo G-Z, Chen K, Deng X, Yu M, Han D, et al. N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell*. 2015;161(4):879–92. <https://doi.org/10.1016/j.cell.2015.04.010> PMID: 25936837
17. Kong Y, Cao L, Deikus G, Fan Y, Mead EA, Lai W, et al. Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science*. 2022;375(6580):515–22. <https://doi.org/10.1126/science.abe7489> PMID: 35113693
18. Romero Charria P, Navarrete C, Ovchinnikov V, Sarre LA, Shabardina V, Casacuberta E. Adenine DNA methylation associated to transcription is widespread across eukaryotes. *bioRxiv*. 2024. <https://doi.org/2024.10.28.620566>
19. Coy SR, Gann ER, Papoulis SE, Holder ME, Ajami NJ, Petrosino JF. SMRT Sequencing of Paramecium Bursaria Chlorella Virus-1 Reveals Diverse Methylation Stability in Adenines Targeted by Restriction Modification Systems. *Frontiers in Microbiology*. 2020;11.
20. Jeudy S, Rigou S, Alempic J-M, Claverie J-M, Abergel C, Legendre M. The DNA methylation landscape of giant viruses. *Nature Communications*. 2020;11(1).
21. Agarkova IV, Dunigan DD, Van Etten JL. Virion-associated restriction endonucleases of chloroviruses. *J Virol*. 2006;80(16):8114–23. <https://doi.org/10.1128/JVI.00486-06> PMID: 16873267
22. Chan S, Zhu Z, Dunigan DD, Van Etten JL, Xu S. Cloning of Nt.CviQII nicking endonuclease and its cognate methyltransferase: M.CviQII methylates AG sequences. *Protein Expr Purif*. 2006;49(1):138–50. <https://doi.org/10.1016/j.pep.2006.04.002> PMID: 16737828
23. Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic J-M, et al. Diversity and evolution of the emerging Pandoraviridae family. *Nat Commun*. 2018;9(1):2285. <https://doi.org/10.1038/s41467-018-04698-4> PMID: 29891839
24. Sahmi-Bounsiar D, Rolland C, Aherfi S, Boudjemaa H, Levasseur A, La Scola B. Marseilleviruses: an update in 2021. *Front Microbiol*. 2021;12:648731.
25. Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, et al. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution. *Virology*. 2014;466–467:60–70. <https://doi.org/10.1016/j.virol.2014.06.031> PMID: 25035289
26. Truchon AR, Chase EE, Gann ER, Moniruzzaman M, Creasey BA, Aylward FO. Kratosvirus quantuckense: the history and novelty of an algal bloom disrupting virus and a model for giant virus research. *Frontiers in Microbiology*. 2023;14.
27. Gann ER. ASP12A Recipe for Culturing Aureococcus Anophagefferens. <https://protocols.io>. 2016.
28. Coy SR, Wilhelm SW. Concentrating viruses by tangential flow filtration. <https://protocols.io>. 2020.
29. Chase EE, Truchon AR, Coy SR, Wilhelm SW. Aureococcus anophagefferens virus (AaV)/Kratosvirus quantuckense viral particle count by SYBR green staining and flow cytometry (CytoFLEX S flow cytometer Beckman Coulter) V.2. <https://www.protocols.io>. 2023.
30. Truchon AR, Gann ER, Wilhelm SW. Closed, Circular Genome Sequence of Aureococcus anophagefferens Virus, a Lytic Virus of a Brown Tide-Forming Alga. *Microbiol Resour Announc*. 2022;11(7):e0028222. <https://doi.org/10.1128/mra.00282-22> PMID: 35678577
31. Bonito C. A PyTorch basecaller for Oxford Nanopore reads. 2019.
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352> PMID: 19505943
33. Heidelbach S, Dall SM, Bøjer JS, Nissen J, van der Maas LN, Sereika M, et al. Nanomotif: Leveraging DNA Methylation Motifs for Genome Recovery and Host Association of Plasmids in Metagenomes from Complex Microbial Communities. *Bioinformatics*. 2024;7(21):null. doi: <https://doi.org/10.1101/2024.04.29.591623>
34. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009; 25(15):1972–3.
35. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972–1973. doi: [10.1093/bioinformatics/btp348](https://doi.org/10.1093/bioinformatics/btp348)
36. Wong TKFL-T, N.; Ren, H.; Banos, H.; Roger, A.J.; Susko, E.; Bielow, C.; De Maio, N.; Goldman, N.; Hahn, M.W.; Huttley, G.; Lanfear, R.; Minh, B.Q. IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models. 2025.
37. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE: a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res*. 2023;51(D1):D629–30. <https://doi.org/10.1093/nar/gkac975> PMID: 36318248
38. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE: restriction enzymes and methyltransferases. *Nucleic Acids Res*. 2003;31(1):418–20. <https://doi.org/10.1093/nar/gkg069> PMID: 12520038
39. Klimasauskas S, Timinskas A, Menkevicius S, Butkienė D, Butkus V, Janulaitis A. Sequence motifs characteristic of DNA[cytosine-N4]methyltransferases: similarity to adenine and cytosine-C5 DNA-methylases. *Nucleic Acids Res*. 1989;17(23):9823–32. <https://doi.org/10.1093/nar/17.23.9823> PMID: 2690010
40. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, et al. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res*. 2003;31(7):1805–12. <https://doi.org/10.1093/nar/gkg274> PMID: 12654995
41. Pósfai J, Bhagwat AS, Pósfai G, Roberts RJ. Predictive motifs derived from cytosine methyltransferases. *Nucleic Acids Res*. 1989;17(7):2421–35. <https://doi.org/10.1093/nar/17.7.2421> PMID: 2717398

42. Gann ER, Xian Y, Abraham PE, Hettich RL, Reynolds TB, Xiao C, et al. Structural and Proteomic Studies of the *Aureococcus anophagefferens* Virus Demonstrate a Global Distribution of Virus-Encoded Carbohydrate Processing. *Front Microbiol*. 2020;11:2047. <https://doi.org/10.3389/fmicb.2020.02047> PMID: 33013751

43. Bewick AJ, Hofmeister BT, Powers RA, Mondo SJ, Grigoriev IV, James TY, et al. Diversity of cytosine methylation across the fungal tree of life. *Nat Ecol Evol*. 2019;3(3):479–90. <https://doi.org/10.1038/s41559-019-0810-9> PMID: 30778188

44. Moniruzzaman M, Gann ER, Wilhelm SW. Infection by a Giant Virus (AaV) Induces Widespread Physiological Reprogramming in *Aureococcus anophagefferens* CCMP1984 - A Harmful Bloom Algae. *Front Microbiol*. 2018;9:752. <https://doi.org/10.3389/fmicb.2018.00752> PMID: 29725322

45. Zhang Y, Nelson M, Nietfeldt JW, Burbank DE, Van Etten JL. Characterization of Chlorella virus PBCV-1 CviAll restriction and modification system. *Nucleic Acids Res*. 1992;20(20):5351–6. <https://doi.org/10.1093/nar/20.20.5351> PMID: 1437552

46. VAN Etten JL, Burbank DE, KuczmarSKI D, Meints RH. Virus infection of culturable chlorella-like algae and development of a plaque assay. *Science*. 1983;219(4587):994–6. <https://doi.org/10.1126/science.219.4587.994> PMID: 17817937

47. La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, et al. A giant virus in amoebae. *Science*. 2003;299(5615):2033. <https://doi.org/10.1126/science.1081867> PMID: 12663918

48. Goorha R, Granoff A, Willis DB, Murti KG. The role of DNA methylation in virus replication: inhibition of frog virus 3 replication by 5-azacytidine. *Virology*. 1984;138(1):94–102. [https://doi.org/10.1016/0042-6822\(84\)90150-8](https://doi.org/10.1016/0042-6822(84)90150-8) PMID: 6208681

49. Rowe JM, Dunigan DD, Blanc G, Gurnon JR, Xia Y, Van Etten JL. Evaluation of higher plant virus resistance genes in the green alga, *Chlorella variabilis* NC64A, during the early phase of infection with *Paramecium bursaria chlorella virus-1*. *Virology*. 2013;442(2):101–13. <https://doi.org/10.1016/j.virol.2013.04.018> PMID: 23701839

50. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*. 2017;14(4):407–10. <https://doi.org/10.1038/nmeth.4184> PMID: 28218898

51. Yuen ZW-S, Srivastava A, Daniel R, McNevin D, Jack C, Eyras E. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat Commun*. 2021;12(1):3438. <https://doi.org/10.1038/s41467-021-23778-6> PMID: 34103501

52. Fu Y, Timp W, Sedlazeck FJ. Computational analysis of DNA methylation from long-read sequencing. *Nat Rev Genet*. 2025;26(9):620–34. <https://doi.org/10.1038/s41576-025-00822-5> PMID: 40155770

53. Breckell GL, Silander OK. Growth condition-dependent differences in methylation imply transiently differentiated DNA methylation states in *Escherichia coli*. *G3 Genes|Genomes|Genetics*. 2022;13(2).

54. Brown CM, Bidle KD. Attenuation of virus production at high multiplicities of infection in *Aureococcus anophagefferens*. *Virology*. 2014;466–467:71–81. <https://doi.org/10.1016/j.virol.2014.07.023> PMID: 25104555

55. Charvin M, Halter T, Blanc-Mathieu R, Barraud P, Aumont-Nicaise M, Parcy F, et al. Single-cytosine methylation at W-boxes repels binding of WRKY transcription factors through steric hindrance. *Plant Physiol*. 2023;192(1):77–84. <https://doi.org/10.1093/plphys/kiad069> PMID: 36782389

56. Buitrago D, Labrador M, Arcon JP, Lema R, Flores O, Esteve-Codina A, et al. Impact of DNA methylation on 3D genome structure. *Nat Commun*. 2021;12(1):3243. <https://doi.org/10.1038/s41467-021-23142-8> PMID: 34050148

57. Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol*. 2021;22(1):295. <https://doi.org/10.1186/s13059-021-02510-z> PMID: 34663425

58. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, et al. The 1.2-megabase genome sequence of Mimivirus. *Science*. 2004;306(5700):1344–50. <https://doi.org/10.1126/science.1101485> PMID: 15486256

59. Yoosuf N, Yutin N, Colson P, Shabalina SA, Pagnier I, Robert C, et al. Related giant viruses in distant locations and different habitats: Acanthamoeba polyphaga moumouvirus represents a third lineage of the Mimiviridae that is close to the megavirus lineage. *Genome Biol Evol*. 2012;4(12):1324–30. <https://doi.org/10.1093/gbe/evs109> PMID: 23221609

60. Zhao L, Song Y, Li L, Gan N, Brand JJ, Song L. The highly heterogeneous methylated genomes and diverse restriction-modification systems of bloom-forming *Microcystis*. *Harmful Algae*. 2018;75:87–93. <https://doi.org/10.1016/j.hal.2018.04.005> PMID: 29778228

61. Yano H, Alam MZ, Rimbara E, Shibata TF, Fukuyo M, Furuta Y, et al. Networking and Specificity-Changing DNA Methyltransferases in *Helicobacter pylori*. *Front Microbiol*. 2020;11:1628. <https://doi.org/10.3389/fmicb.2020.01628> PMID: 32765461