Received: 18 September 2023 | Revised: 21 December 2023 | Accepted: 15 January 2024

DOI: 10.1111/mec.17287

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Dinucleotide biases in the genomes of prokaryotic and eukaryotic dsDNA viruses and their hosts

Diego Forni | Uberto Pozzoli | Rachele Cagliani | Manuela Sironi

Scientific Institute IRCCS E. MEDEA, Bioinformatics, Bosisio Parini, Italy

Correspondence
Diego Forni, Scientific Institute IRCCS E. MEDEA, Bioinformatics, Bosisio Parini 23842, Italy.
Email: diego.forni@lanostrafamiglia.it

Funding information
Ministero della Salute

Handling Editor: Sebastien Calvignac-Spencer

Abstract

The genomes of cellular organisms display CpG and TpA dinucleotide composition biases. Such biases have been poorly investigated in dsDNA viruses. Here, we show that in dsDNA virus, bacterial, and eukaryotic genomes, the representation of TpA and CpG dinucleotides is strongly dependent on genomic G+C content. Thus, the classical observed/expected ratios do not fully capture dinucleotide biases across genomes. Because a larger portion of the variance in TpA frequency was explained by G+C content, we explored which additional factors drive the distribution of CpG dinucleotides. Using the residuals of the linear regressions as a measure of dinucleotide abundance and ancestral state reconstruction across eukaryotic and prokaryotic virus trees, we identified an important role for phylogeny in driving CpG representation. Nonetheless, phylogenetic ANOVA analyses showed that few host associations also account for significant variations. Among eukaryotic viruses, most significant differences were observed between arthropod-infecting viruses and viruses that infect vertebrates or unicellular organisms. However, an effect of viral DNA methylation status (either driven by the host or by viral-encoded methyltransferases) is also likely. Among prokaryotic viruses, cyanobacteria-infecting phages resulted to be significantly CpG-depleted, whereas phages that infect bacteria in the genera *Burkolderia* and *Staphylococcus* were CpG-rich. Comparison with bacterial genomes indicated that this effect is largely driven by the general tendency for phages to resemble the host's genomic CpG content. Notably, such tendency is stronger for temperate than for lytic phages. Our data shed light into the processes that shape virus genome composition and inform manipulation strategies for biotechnological applications.

KEYWORDS
CpG, dinucleotide composition, DNA virus, GC content, UpA

1 | INTRODUCTION

It has been known for many years that the genomes of cellular organisms display important dinucleotide composition biases. Vertebrate, but not invertebrate, genomes are characterized by a depletion of CpG (a cytosine followed by a guanine in the 5′ to 3′ direction) dinucleotides (Bird & Taggart, 1980; Burge et al., 1992; Gentles & Karlin, 2001; Gonçalves-Carneiro et al., 2021; Karlin & Burge, 1995; Provataris

et al., 2018; Simmonds et al., 2013). This is thought to be at least partially due to methylation and mutational loss of cytosines at CpG sites (in vertebrates) or lack thereof (in invertebrates). In microbial and plant genomes, the representation of CpG dinucleotides is highly variable ranging from under- to over-representation (Karlin & Burge, 1995). Conversely, TpA dinucleotides (UpA in RNA) are under-represented across the tree of life (Beutler et al., 1989; Burge et al., 1992; Karlin & Burge, 1995; Simmen, 2008; Simmonds et al., 2013).

As obligate intracellular parasites, viruses, especially those infecting animals, were often reported to have evolved genomes that resemble the genomic dinucleotide composition of their hosts.

(Karlin et al., 1994; Lobo et al., 2009; Rima & McFerran, 1997; Sexton & Ebel, 2019; Simmonds et al., 2013). This is thought to partially result from the ability of cellular defence mechanisms to sense nucleic acids with a molecular signature different from the hosts' – that is, with a different dinucleotide composition. In vertebrates, the identification of toll-like receptor 9 (TLR9) as a sensor of DNA containing unmethylated CpG dinucleotides, and of the zinc-finger antiviral protein (ZAP) as a restriction factor of viruses with elevated CpG content, supported this view (Bowie & Unterholzner, 2008; Luo et al., 2020; Takata et al., 2017). However, the abundance and role of CpG (and TpA) dinucleotides in the genomes of viruses that infect invertebrates and unicellular organisms remains under-investigated or poorly explained. Also, a survey of animal RNA viruses indicated that the dinucleotide composition of their genomes more closely reflects viral family than host associations (Giallonardo et al., 2017).

Compared to RNA viruses, the dinucleotide composition of DNA viruses has been investigated in shallower details. Nonetheless, DNA viruses cover an impressive diversity in terms of host spectrum, genome size, and biological characteristics. Also, double-strand DNA (dsDNA) viruses that infect animals pose an exceptional health burden to human and animal populations, whereas those infecting prokaryotes play important roles as modulators of microbial ecosystems (Brown et al., 2022; Koskella et al., 2022; Puxty & Millard, 2023). Moreover, viruses that infect bacteria (bacteriophages, phages) are regarded as a promising strategy to control infections, especially in the light of the rising problem of antibiotic resistance (Hatfull et al., 2022). Herein, we use base composition- and phylogeny-aware methods to explore dinucleotide biases in the full set of representative dsDNA viruses.

2 | MATERIALS AND METHODS

2.1 | Genome datasets and proteomic trees

A list of exemplar viruses was retrieved from the ICTV Virus Metadata Resource (https://ictv.global/vmr, VMR_MSL38_v1 of 04/25/2023). This list includes information regarding virus name, host source, isolate designation, taxonomical classification, and GenBank accession entries. We only retained dsDNA viruses, including virophages, longer than 500 nucleotides (Table S1). Due to their limited representation, reverse-transcribing viruses were excluded.

The list of bacterial genomes was derived from the BV-BRC site (https://www.bv-brc.org/) by selecting entries corresponding to representative or reference strains (Table S2).

A list of eukaryotic genomes was retrieved from the NCBI Genome Table for eukaryotes (https://www.ncbi.nlm.nih.gov/genome/browse#!/eukaryotes/). We used the filter utility to select organism groups (green algae, birds, fishes, insects, mammals, protists, reptiles/amphibia) and, for each group, one representative per genus was randomly included. Out of the remaining genomes, 50 in each group were randomly selected (with the exception of Algae, for which only 48 genomes were available after filtering). In the case of arthropods, we selected insects as a group, as the majority of arthropod-infecting viruses in our dataset were sequenced from these animals (Table S3).

Viral genome sequences were retrieved by using the GenBank accession entries from the ICTV Virus Metadata tables and applying the read.GenBank function from the APE R tool package (Paradis & Schliep, 2019). Eukaryotic and prokaryotic host genome sequences were retrieved using the ncbi-genome-download tool (Blin, 2023).

The proteomic reference trees of prokaryotic and eukaryotic viruses were downloaded from the GenomeNet/ViPTree server (https://www.genome.jp/viptree/). Pioneered by Rohwer & Edwards (2002), proteomic trees are distance trees based on genome-wide sequence similarities computed by tBLASTx. Specifically, all-against-all distances are calculated on the basis of the length-normalized bit score of tBLASTx. Pairwise genome distances are then used to generate trees using BIONJ (Bhunchoth et al., 2016; Gascuel, 1997; Nishimura, Watai, et al., 2017). Reference proteomic trees were pruned to retain only tips corresponding to genomes present in our database using the APE R tool package and the keep.tip.phylo function (Table S1).

2.2 | Virus metadata

Kingdom/clade host associations were retrieved from the ICTV Virus Metadata Resource. For eukaryotic viruses, a finer level of taxonomic definition was obtained through manual inspection of individual entries. For prokaryotic viruses, host genera were retrieved, when available, from the Virus–Host DB (Mihara et al., 2016; https://www.genome.jp/virushostdb/). For the remaining viruses, information was retrieved from GenBank accessions.

Information about lifestyle (bioinformatic prediction) and GCF mode was retrieved from a previous study (Mavrich & Hatfull, 2017; Table S4).

2.3 | Dinucleotide observed/expected ratio

To investigate dinucleotide biases, we calculated the observed/expected ratio for all dinucleotides. Specifically, the frequency of each dinucleotide in each genome (i.e. the observed frequency) was divided by the product of the frequencies of the contributing

nucleotides (i.e. the expected frequency). For instance, for CpG, we calculated the number of CpG along the genome divided by the number of all possible dinucleotides; this frequency was then divided by the product of C and G frequencies. Dinucleotides were also counted in the reverse complement of the sequence.

Dinucleotide composition was calculated using the compseq tool (https://www.bioinformatics.nl/cgi-bin/emboss/), by setting the size of word equal to 2 and using the 'calcfreq' parameter, so that the dinucleotide expected frequencies are calculated from the observed frequency of single bases.

For eukaryotic genome analysis, each sequence was divided into chunks of 50 kb, and 100 regions per genome were randomly selected. For each selected chunk, G+C content and dinucleotide composition were calculated. Sequences were processed using the APE R tool package and the functions therein, and dinucleotide composition was calculated using compseq, as explained above.

2.4 | Linear regression, ancestral state reconstruction, phylogenetic ANOVA, and Euclidean distances

We modelled the relationship between CpG (TpA) ratios and G+C content with linear regressions using the lm function in the stats R package (Wilkinson & Rogers, 1973). The residuals were obtained from the models using the residual function from stats.

Ancestral state reconstruction was performed by maximum likelihood using the FastAnc function in the phytools R package (Revell, 2012, 2023), and pyhlogenetic trees were plotted using the contMap and setMap functions.

The phylogenetic ANOVA and post-hoc tests were performed using the phylANOVA function in phytools (Garland et al., 1993; Harmon et al., 2008; Revell, 2012, 2023) and *p*-values were calculated with 10,000 simulations.

Euclidean distances were calculated between the resCpG of each two points representing a phage and a bacterial genome. Specifically, for each bacteriophage sequence, we calculated the euclidean distance to all bacteria hosts, and we then mediated distances among bacteria genera it infects, or not. Statistical analysis was performed through paired Wilcoxon-rank sum tests followed by Bonferroni correction for multiple tests (R package stats).

2.5 | Dinucleotide composition in viral ORFs

Viral coding sequences were downloaded using the datasets command line tool (https://www.ncbi.nlm.nih.gov/datasets/docs/v2/reference-docs/command-line/datasets/) and processed using the APE R tool package and the functions therein. The G+C content for each ORF was calculated using the GC.content function and dinucleotide composition was calculated using compseq, this time without considering the reverse complement strand. For each virus, the dinucleotide observed values of all ORFs were added together and then divided by the sum of the viral ORF expected values; in this way we took into account the different composition both in G+C content and in length of each ORF. Correlation between resCpG and resTpG or resApG was calculated using Spearman's test implemented in the cor.test function of the R package stats.

3 | RESULTS

3.1 | Dinucleotide biases in dsDNA virus genomes

We assembled a dataset of 4885 dsDNA viruses from the ICTV (International Committee on Taxonomy of Viruses) virus metadata resource (Table S1). To investigate composition biases, we calculated the observed/expected (O/E) ratio for all dinucleotides. Specifically, the expected dinucleotide frequency in a sequence is simply the product of the frequencies of the contributing nucleotides. Thus, ratios higher or lower than 1 indicate that a dinucleotide is over- or under-represented, respectively. In particular, ratios lower than 0.78 and higher than 1.23 are generally considered to define significant depletion and enrichment (Karlin et al., 1994; Karlin & Mrázek, 1997). Across our virus genome dataset, TpA dinucleotides were the most depleted, whereas CpG dinucleotides, although not generally under-represented, showed the widest dispersion (Figure 1a).

Previous analyses of vertebrate genomes and RNA viruses detected a positive correlation between the O/E ratio of CpG (O/E CpG) and G+C content. A negative correlation with G+C content was instead detected for O/E TpA (Duret & Arndt, 2008; Odon et al., 2022; Simmen, 2008; Simmonds et al., 2013). By fitting linear models to the virus dataset, we also detected strong and significant relationships (Figure 1b). This basically implies that CpG and TpA dinucleotide abundance cannot be interpreted without taking G+C content into account. Thus, to further investigate how these biases relate to other virus properties, we used the fitted regression lines to calculate the residuals of O/E CpG and O/E TpA over G+C content (hereafter referred to as resCpG and resTpA). Thus, resCpG and resTpA are measures of dinucleotide representation which account for genomic G+C content. Comparison of such measures indicated that resCpG have much wider distribution than resTpA (Figure 1c). In fact, G+C content was associated to 54% of the variability in TpA representation and 29% in CpG (Figure 1b). As a comparison, we also calculated resGpC and resGpG/CpC and analysed their distributions, which resulted narrower than that of resCpG (Figure S1). Overall, these data indicate that the representation of TpA dinucleotides is more influenced by G+C content than that of CpG, suggesting that additional factors drive the abundance of CpG dinucleotides in viral genomes.

3.2 | A role for the host or for taxonomy?

To explore additional forces that may explain CpG representation in the viral genomes, we grouped viruses by host or by phylum

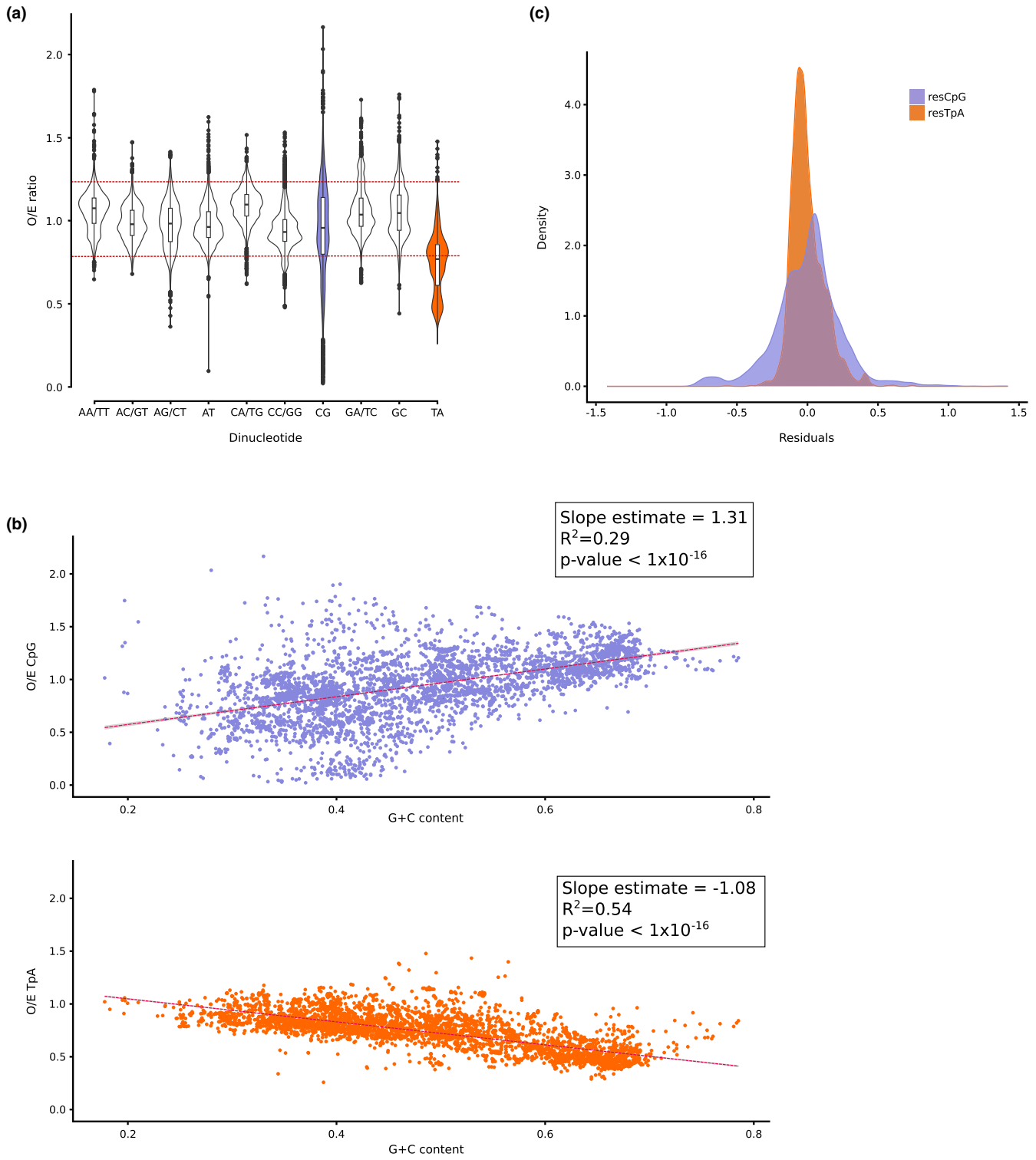

FIGURE 1 Dinucleotide representation in dsDNA viruses. (a) Violin plots with boxplots of the observed/expected ratio (O/E) for all dinucleotides. The horizontal hatched lines correspond to ratios of 0.78 and 1.23, which generally define significant depletion and enrichment (Karlin et al., 1994; Karlin & Mrázek, 1997). (b) Linear models of O/E CpG (purple) and O/E TpA (orange) on G+C content. Regression lines are shown with confidence intervals and the results of the models are included in each plot. (c) Distribution of residuals for the O/E CpG model (resCpG) and the O/E TpA model (resTpA).

(Figure 2a,b). In both cases, important differences were evident. For instance, vertebrate-infecting viruses were the most depleted, although with a wide variability. Conversely, viruses infecting invertebrates showed no suppression of CpG dinucleotides (Figure 2a). Similarly, *Cossaviricota* viruses showed remarkable CpG depletion, whereas others did not (Figure 2b). Clearly, though, host associations and taxonomy

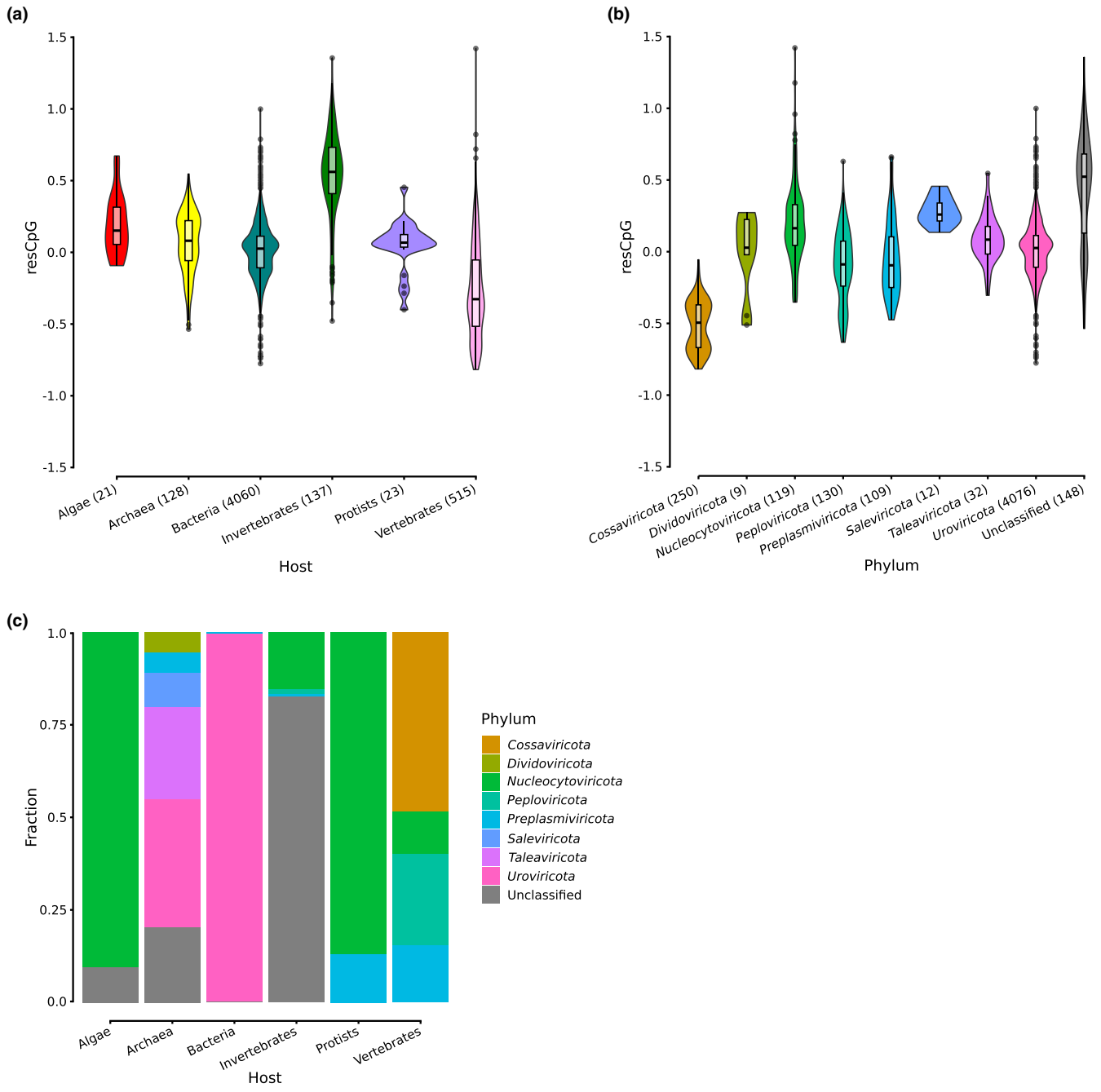

FIGURE 2 CpG representation as a function of host associations and viral taxonomy. Violin plots with boxplots of resCpG for viruses that infect different hosts (a) or classified in different phyla (b). In panel (a), African swine fever virus was not included in the analysis as it infects both invertebrate and vertebrate hosts. (c) Taxonomic classification of viruses that infect different hosts.

are not independent and it is thus imperative to disentangle the two effects. In fact, *Cossaviricota* viruses are found in vertebrate hosts only (Figure 2c).

Taxonomic classifications largely reflect phylogenetic relatedness. We thus aimed to explore the variation in resCpG resulting from host associations after accounting for phylogenetic relationships. However, viral genomes do not contain universally conserved genes that can be used for phylogenetic reconstruction (Rohwer & Edwards, 2002). We thus exploited viral proteomic trees, which do not depend on gene alignments but rather rely on the whole set of proteins encoded by viral genomes (Nishimura, Yoshida, et al., 2017). Proteomic trees are expected to largely reflect evolutionary histories, although they are affected by some limitations. These include the challenge of reliably detecting remote homologies among proteins and the noise introduced by horizontal gene transfer (Adriaenssens et al., 2015; Chibani et al., 2019; Medvedeva et al., 2023; Rohwer & Edwards, 2002; Zhang et al., 2023). We downloaded the reference prokaryotic and eukaryotic virus trees from the GenomeNet/ViPTree server. Trees were then pruned to retain only genomes present in our dataset.

For prokaryotes, we only retained bacteriophages, as the number of archaea viruses was limited. We eventually obtained two trees with 3428 (bacteriophages) and 576 (eukaryotic viruses) tips (Figure 3, Table S1).

Using these trees, we reconstructed the maximum likelihood ancestral state of resCpG. In both trees, clear associations between CpG content and phylogenetic relationships were evident (Figure 3). For instance, among eukaryotic viruses, papillomaviruses and polyomaviruses were strongly CpG depleted. The opposite was true for baculoviruses, whereas other viral families (e.g. *Adenoviridae* and *Orthoherpesviridae*) were more heterogeneous (Figure 3a). Among bacteriophages, there were a large number of viruses with no family association. However, families with generally high (*Drexlerviridae*, *Rountreeviridae*) or low (e.g. *Kyanoviridae*, *Schitoviridae*) resCpG were also evident (Figure 3b).

We next set out to explore host associations by taking phylogeny into account. To this purpose, we defined hosts at a finer level of taxonomic definition (e.g. birds, fish, reptiles, amphibians, and mammals rather than vertebrates; for bacteria we used genera as the taxonomic level) and we exploited the trees to run *phylogenetic ANOVA analyses* (Garland et al., 1993).

For the eukaryotic viral dataset, we limited analysis to hosts with at least five infecting viruses and the phylogenetic ANOVA detected significant differences among groups ($F = 117.78$, $p = .001$). After post-hoc tests, it was evident that most significant pairwise comparisons involved viruses infecting arthropods (Figure 4). In these viruses, CpG dinucleotides were significantly more represented than in viruses infecting not only vertebrates, but also protists and algae. Conversely, no significant difference was observed among viruses that infect different vertebrate hosts or between the latter and protist-infecting viruses (Figure 4).

For the prokaryotic viruses, the situation is complicated by the large number of viruses, as well as of bacterial genera that such viruses infect. We thus limited analysis to host genera with at least 10 associated viruses. A significant phylogenetic ANOVA was obtained for bacteriophages, as well ($F = 21.65$, $p = .011$). Pairwise comparisons indicated that four host genera accounted for most differences (Figure 4). Thus, viruses infecting bacteria in the *Burkholderia* and, to a lesser extent, in the *Staphylococcus* genera tended to be enriched in CpG dinucleotides. The opposite was true for viruses that infect cyanobacteria (*Prochlorococcus* and *Synechococcus* genera).

3.3 | Host and methylation status may drive CpG abundance in eukaryotic virus genomes

Several previous studies showed that the genomes of vertebrates are more CpG depleted than invertebrate genomes (Bird & Taggart, 1980; Burge et al., 1992; Gentles & Karlin, 2001; Gonçalves-Carneiro et al., 2021; Provataris et al., 2018; Simmonds et al., 2013). However, most analyses focused on the genomes of a few model organisms. Also, differences among vertebrates were poorly explored and data on other eukaryotes (e.g. protists, algae) are scant. To relate the results obtained in the phylogenetic ANOVA for eukaryotic viruses to host genome features, we calculated O/E CpG and G+C content for 50 genomes from each host groups analysed above. Specifically, for each genome, we randomly extracted 100 regions 50 Kb in length (Gentles & Karlin, 2001; see Section 2,2.1: Materials and Methods).

Results indicated a similar dependency of O/E CpG on G+C content in all host groups (i.e. similar slopes of the regression lines), with some uncertainty in the case of insects, as the majority of genomes have a G+C content in the 0.3–0.4 range (Figure 5). However, the degree of CpG depletion was very different across groups, with most hosts showing some level of under-representation, strongest in mammals and birds, and insects having CpG contents similar to or higher than the expectation based on nucleotide frequencies (i.e. O/E CpG equal or higher than 1) (Figure 5). Thus, the results of the phylogenetic ANOVA, which detected significant differences between arthropod-infecting viruses and viruses infecting other eukaryotes, have parallels in the host genome composition.

Among eukaryotic viruses, strong CpG depletion was observed for papillomaviruses and polyomaviruses, which infect different vertebrates (mammals, birds, fishes, and reptiles) and are known to undergo CpG methylation during latency (Hoelzer et al., 2008; Figure 3a). Likewise, gammaherpesviruses (all of them infecting mammals), which are also methylated in their latent phases, were the most CpG depleted among orthoherpesviruses (Hoelzer et al., 2008; Lieberman, 2013; Figure 3a). On the other hand, poxviruses, which infect a broad range of hosts including vertebrates and invertebrates, showed little or no CpG depletion (Figure 3a). These viruses replicate in the cytoplasm and are not known to be methylated (Hoelzer et al., 2008).

Whereas papillomaviruses, polyomaviruses, gammaherpesviruses, and adenoviruses are methylated by host enzymes, some viruses in the families *Iridoviridae*, *Alloherpesviridae*, *Phycodnaviridae*, and *Nudiviridae* are known to encode their own 5-cytosine methyltransferases (Hoelzer et al., 2008; Wagner et al., 1985; Xu, 1998). We thus inspected viral genomes for the presence of methyltransferases. These were detected in nine iridoviruses, two ranid herpesviruses, seven phycodnaviruses, and three nudiviruses. These genomes displayed variable levels of CpG depletion (Figure 3a). It is thus possible that methylation status interacts with host association and other factors in driving CpG content. Unfortunately, lack of information on the methylation status of individual viruses hampers formal statistical testing of this hypothesis (see also Section 4: Discussion).

3.4 | CpG dinucleotide abundance in eukaryotic virus coding sequences

Previous analyses of vertebrate genomes detected stronger and additional selection pressures imposed on mRNA sequences compared to non-cytoplasmically expressed DNA sequences (Simmonds et al., 2013). Viral genomes are mostly occupied by coding sequences and, as expected, we found similar levels of CpG abundance

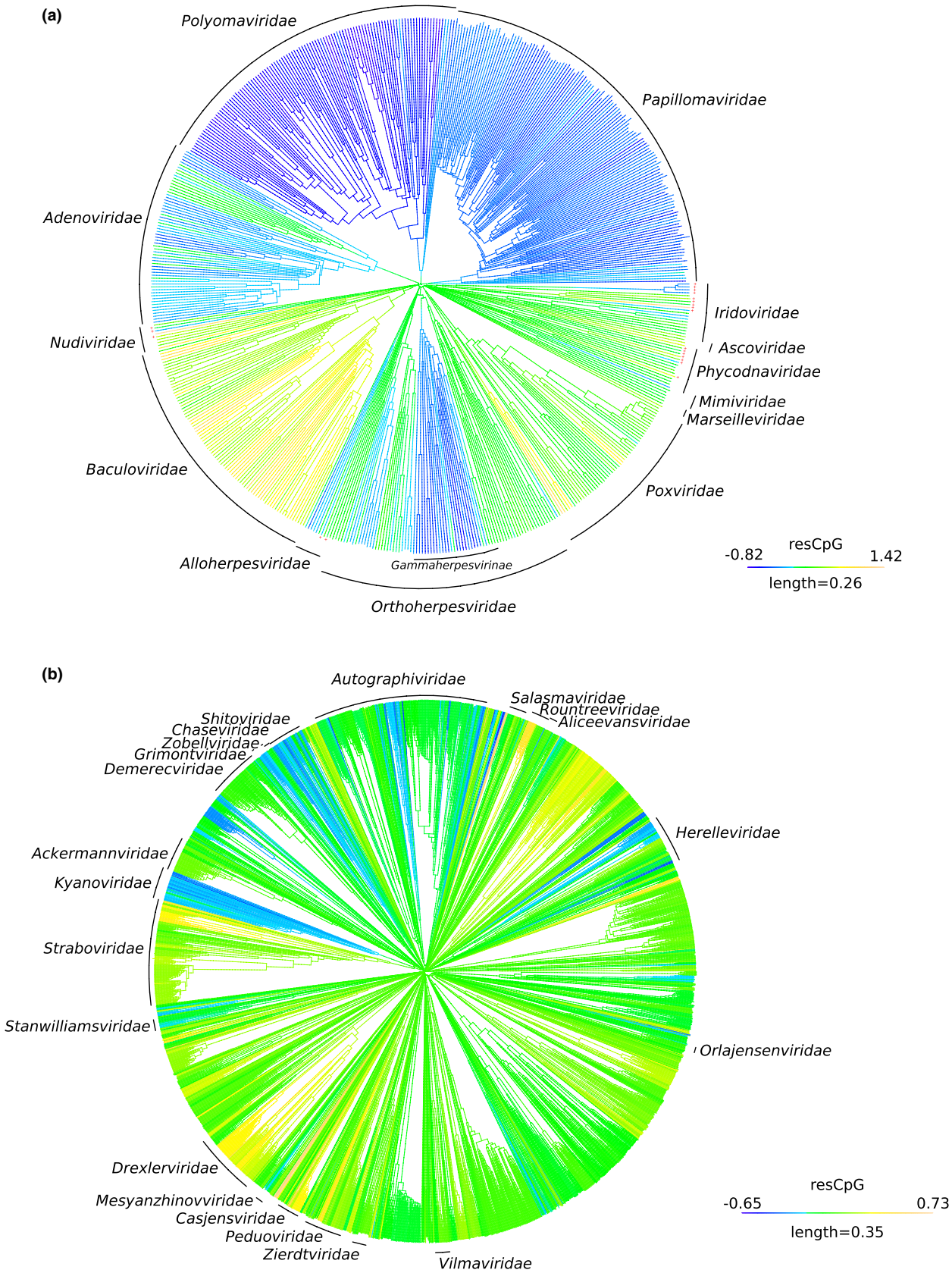

FIGURE 3 Ancestral state of resCpG for eukaryotic and prokaryotic viruses. Phylogenetic trees of eukaryotic (a) and prokaryotic (b) viruses with branches coloured by ancestral state reconstruction of resCpG calculated by maximum likelihood. When available, the classification into viral families (the most numerous only) is reported on both trees. Red crosses indicate eukaryotic viruses that encode their own methyltransferases.

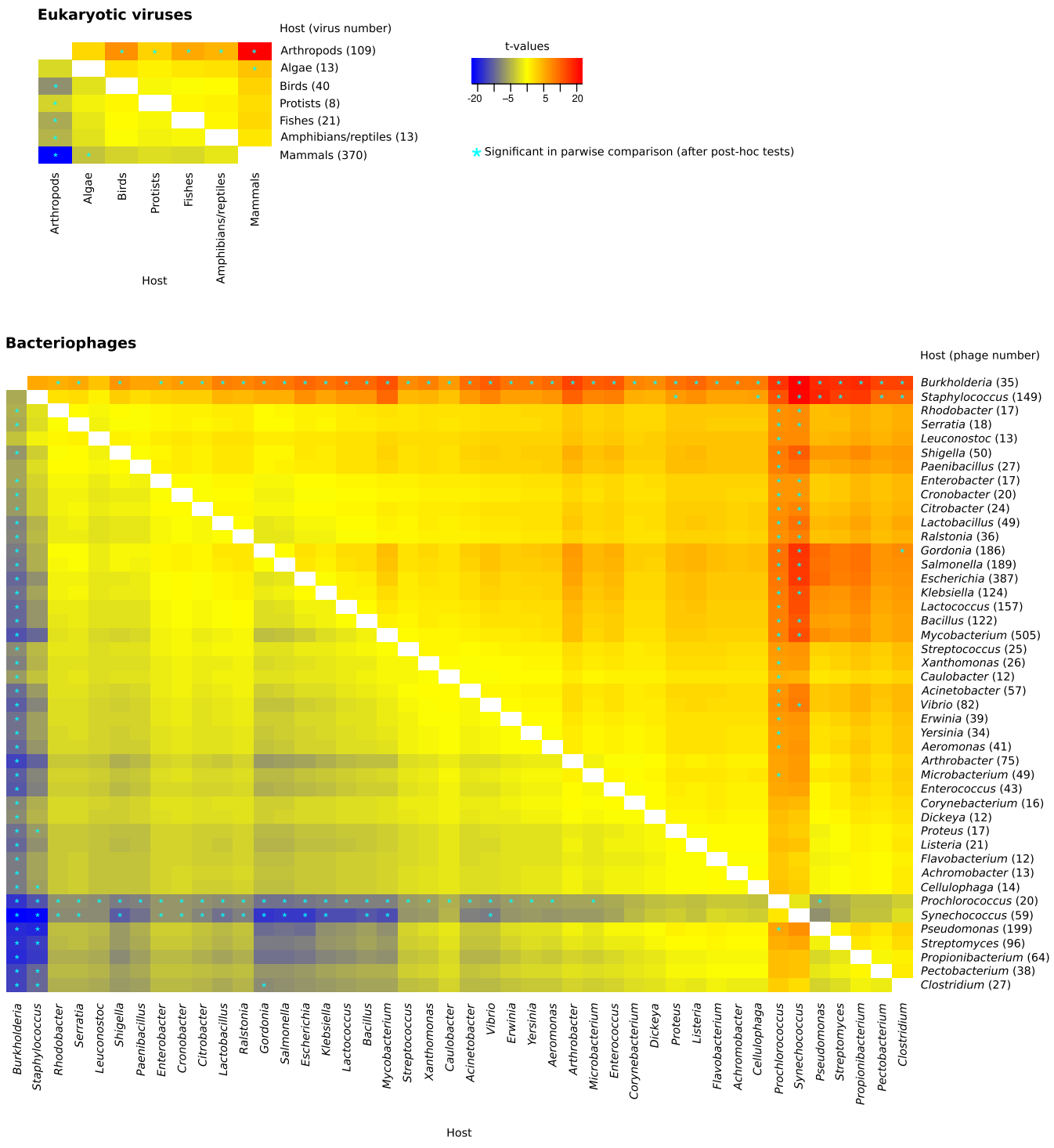

FIGURE 4 Results of the phylogenetic ANOVA. The results of the phylogenetic ANOVA of resCpG against host group are shown for eukaryotic viruses and bacteriophages. Specifically, the pairwise comparisons of *t*-values are shown as matrices and colour-coded as in the legend. Comparisons that remained significant after pos-hoc tests are marked with an asterisk.

in whole genomes and in the coding portion only (Figure S2). We nonetheless wished to gain further insight into the mutational biases responsible for CpG depletion. In vertebrates, loss of CpG dinucleotides caused by methylation-induced mutation occurs through CpG→TpG changes. However, mRNA sequences from organisms with methylated genomes were shown to display additional selection against CpG through CpG→ApG mutations (Simmonds et al., 2013). Because O/E TpG and O/E ApG also depend on G+C content, we calculated resTpG and resApG for the coding sequences of viruses infecting vertebrates or invertebrates. For both hosts, negative and significant correlations between resCpG and resTpG or resApG were detected (Figure 6). However, the correlations were stronger for vertebrate-infecting viruses, suggesting that the sequences of their mRNAs are subject to similar selective pressures as host transcripts.

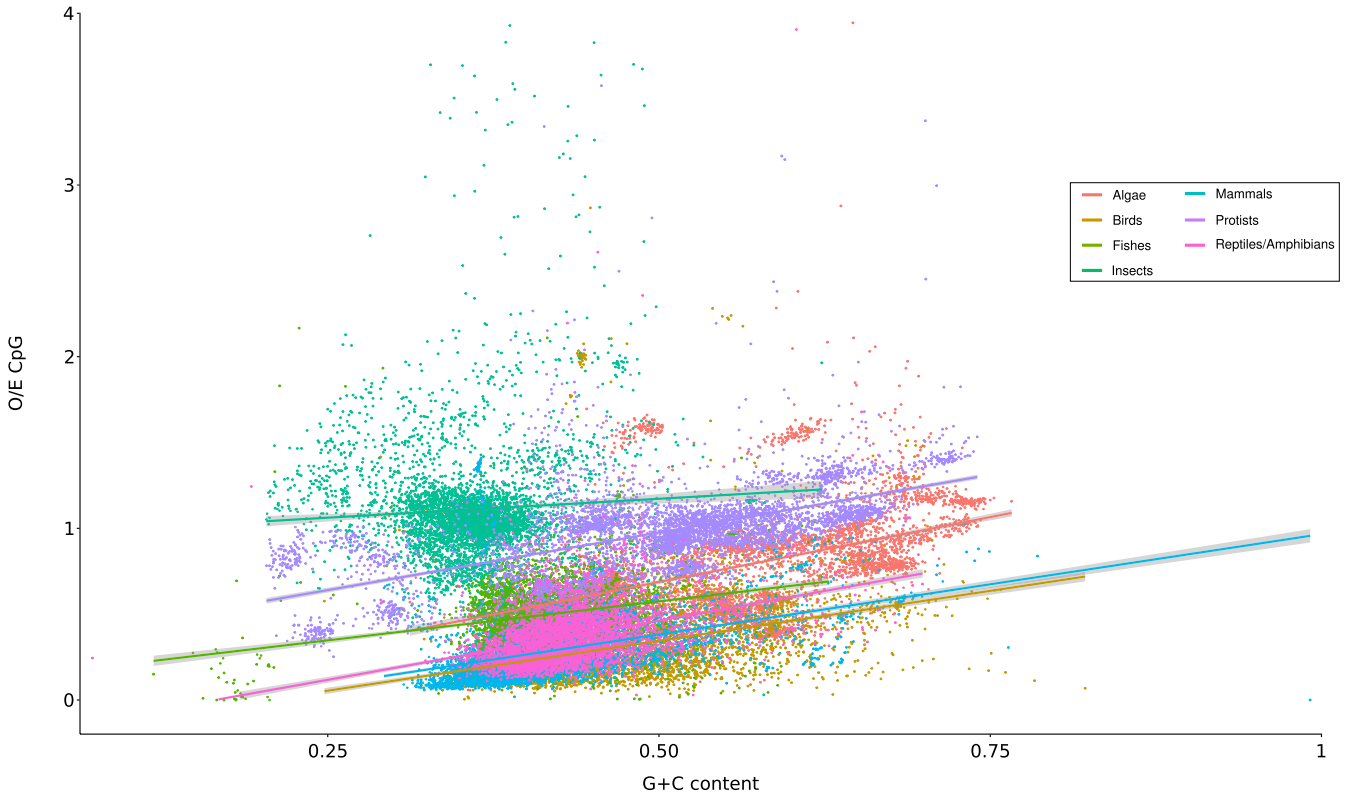

FIGURE 5 CpG distribution in eukaryote genomes. Linear models of O/E CpG on G+C content for a set of randomly selected genomic regions from different eukaryotes. Regression lines are shown with confidence intervals.

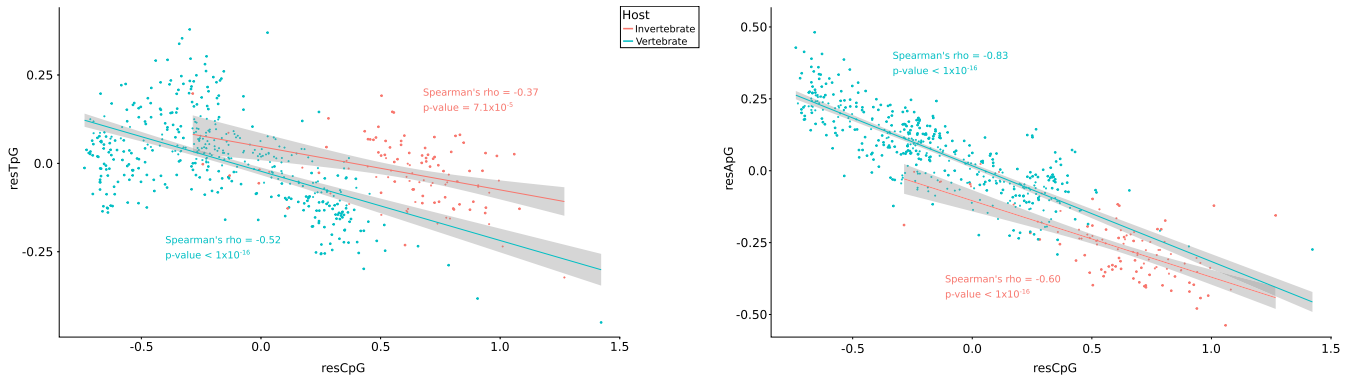

FIGURE 6 CpG depletion through TpG and ApG changes. Correlation plots between resCpG and resTpG (left panel) or resApG (right panel) in coding sequences of eukaryotic viruses. Viruses are divided based on the hosts they infect (see legend). Spearman's rho coefficients, along with their *p*-values, are also reported.

One possible source of selective pressure on mRNAs is binding by ZAP. The preferred virus RNA targets of ZAP are CpGs located between 12 and 32 nucleotides apart (Gonçalves-Carneiro et al., 2022). We thus aimed to investigate whether viral coding sequences were targeted by pressures to space CpG dinucleotides at distances other than those preferred by ZAP. We divided viruses based on the host, because mammals, birds, and reptiles/amphibia possess ZAP, whereas fishes and invertebrates do not (Gonçalves-Carneiro et al., 2021). We next calculated the distances between all consecutive CpG dinucleotides in each viral ORF and, as a comparison, the distances between GpC dinucleotides. No differences were observed in any host between the distribution of CpG and GpC distances (Figure S3), suggesting that such distances largely depend on the length of individual ORFs, with no or little effect from ZAP.

3.5 | Host and lifestyle modulate CpG dinucleotide abundance in phage genomes

To investigate whether differences in CpG abundance in phage genomes result from adaptation to the host genome composition, we calculated the G+C content and O/E CpG for all reference genomes

($n=2890$) of bacteria in the genera infected by the viruses we analysed. Several bacterial genomes were found to be CpG-depleted and a strong correlation between G+C content and O/E CpG was observed (linear regression, estimate=0.98, p-value $<2\times10^{16}$; Figure 7a). As expected, given that we stratified by host taxonomy, clustering by genus was evident for bacterial genomes and much less for their phages (Figure 7a). However, comparison of the G+C content and O/E CpG for phages and their hosts revealed some level of correspondence. For instance, the genomes of bacteria in the genus *Burkholderia* and of their phages were in the upper range of the distribution of G+C content and O/E CpG. Likewise, phages that infect *Synechococcus* cells, as well as the single reference bacterial genome in the genus *Prochlorococcus*, were in the low range of G+C content and O/E CpG (Figure 7a).

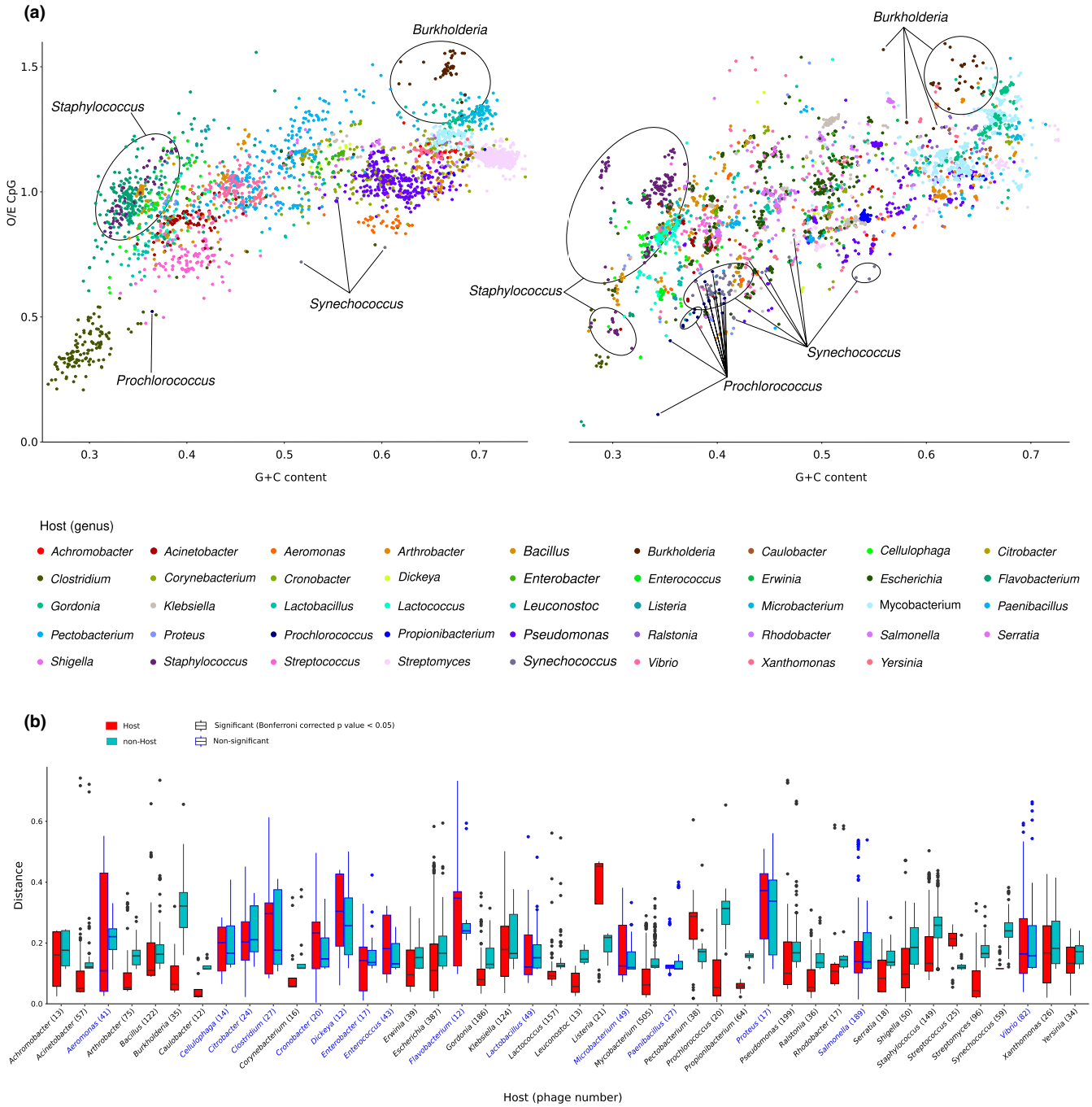

FIGURE 7 Similarity in genome composition between phages and their hosts. (a) Scatter plot of G+C content and O/E CpG for reference bacterial genomes (left) or phage genomes (right). Genomes are colour-coded based on the bacterial genus or on the infected bacterial host (genus). Host genera significantly different in the phylogenetic ANOVA analysis are highlighted. (b) Boxplot of Euclidean distances between phages and bacteria. For each phage, the average of the distances from all genomes of bacteria in the genus it infects (red), as well as from bacterial genomes in other genera (green) are shown. Statistical analysis was performed through paired Wilcoxon-rank sum tests followed by Bonferroni correction for multiple tests. A p-value <.05 was considered significant.

To formally determine whether phage genomes are more similar to the genomes of their hosts in terms of CpG abundance, irrespective of G+C content, we used the same approach as above and fitted a linear regression to the bacterial genome data. We next used the residuals of this regression to calculate Euclidean distances to the resCpG of phage genomes. Specifically, for each phage genome, we calculated the average distances from all genomes of bacteria in the genus it infects, as well from bacterial genomes in other genera. We found that, in many instances, phages have a CpG representation significantly more similar to their hosts than to other bacteria (Figure 7b). This was the case of phages that infect bacteria in the genera *Burkholderia*, *Staphylococcus*, *Synchlorococcus* and *Prochlorococcus*, but also of several others (Figure 7b). These results indicate a widespread tendency of phages to resemble the dinucleotide composition of their hosts and suggest that the results of the phylogenetic ANOVA are explained by this effect (as bacteria in the genera *Burkholderia* and *Staphylococcus* are CpG-rich, whereas those in the genera *Synchlorococcus* and *Prochlorococcus* are CpG-poor). However, the opposite pattern (i.e. phage genomes significantly less similar in CpG content to their host) was also observed for phages that infect bacteria in the genera *Listeria*, *Klebsiella*, *Pectobacterium*, and *Streptococcus*.

Previous works have shown that bacteriophages differ in terms of lifestyle (i.e. lytic and temperate) and in the extent of horizontal gene transfer (i.e. high- and low-gene content flux, HGCF and LGCF; Mavrich & Hatfull, 2017). These two features are not independent and distribute unequally among phages that infect bacteria in distinct genera (Mavrich & Hatfull, 2017). We thus obtained information about lifestyle and GCF for 928 phages in our dataset from a previous work (Mavrich & Hatfull, 2017). As expected, the two features were unevenly distributed among the phages we analysed, as well (Figure 8a,b).

We first asked whether lifestyle or GCF mode affect the degree of similarity of phage and host genomes in terms of CpG content (measured as Euclidean distances, as above). Results indicated that temperate phage genomes are significantly more similar to their host genomes than lytic phages (Wilcoxon Rank Sum test, $p=4.0\times10^{-06}$). No difference was instead observed for GCF mode (Kruskal-Wallis Rank Sum test, $p=.73$; Figure 8c). We next analysed the effect of lifestyle for phages that infect bacteria in different genera. To this aim, we retained only genera for which at least three temperate and three lytic phages were available. Results indicated that the tendency of temperate phages to have CpG content more similar to their host's than lytic phages is observed for many genera. However, significant results (after Bonferroni correction for multiple tests) were only obtained for phages that infect bacteria in the genera *Listeria* and *Staphylococcus* (Figure 8d). In many instances, the failure to reach statistical significance is likely due to lack of power due to small sample sizes. This observation helps explain why phages that infect bacteria in the genera *Listeria*, *Klebsiella*, and *Pectobacterium* have genomes significantly less similar in CpG content to their hosts. In the case of *Listeria*, the effect is likely mediated by the large number of lytic phages, which are the only type infecting bacteria in the genera *Klebsiella* and *Pectobacterium* (Figure 8b).

4 | DISCUSSION

Gaining insight into the processes that shape virus genome composition is central not only to understand virus evolution and emergence but also to inform effective manipulation strategies for the development of vaccine strains and biotechnological applications (e.g. phage therapy). Herein we used the full set of reference genomes of dsDNA viruses to test different hypotheses regarding their dinucleotide composition. Our data indicate that, in dsDNA virus genomes, as well as in bacterial and eukaryotic genomes, the representation of TpA and CpG dinucleotides is strongly dependent on G+C content. This feature, also observed for RNA viruses (Ibrahim et al., 2019; Odon et al., 2022), thus seems a general characteristic of biological sequences. Whereas the underlying reasons are presently unclear, the existence of such relationships implies that dinucleotide biases cannot be interpreted without taking G+C content into account. Thus, the classical dinucleotide O/E ratio does not fully capture the variation of dinucleotide biases across genomes. Hence our use of the residuals of the linear regression models as a measure of dinucleotide representation that accounts for variation in G+C content. A caveat that should be taken into account in this respect is that such residuals are not absolute measures of dinucleotide abundance (e.g. resCpG=0 does not indicate that the representation of CpG dinucleotides corresponds to the expected based on nucleotide frequency). Indeed, resCpG and resTpA depend on the G+C content and on the distributions that are fitted by the regression.

In our viral dataset, a larger portion of the variance in TpA abundance was explained by G+C content compared to CpG. We thus explored which additional factors drive the distribution of CpG dinucleotides. Our working hypotheses were: (1) CpG representation largely reflects viral taxonomy; (2) adaptation to the dinucleotide representation of host genomes is the major driver of CpG dinucleotide content. To address these non-mutually exclusive hypotheses, we resorted to viral proteomic trees (Nishimura, Yoshida, et al., 2017). The general advantage is that highly diverse genomes, which do not share a universally common gene, can be included in the same tree. For the purpose of the analyses herein, another favourable aspect of proteomic trees is that they avoid possible circularity issues (i.e. the analysis of dinucleotide composition using a tree derived from nucleotide/protein alignments). Unfortunately, though, such trees are unrooted and thus prevent formal quantification of the phylogenetic signal (e.g. by use of Pagel's lambda (Pagel, 1999) or similar tests). Nonetheless, inspection of both the eukaryotic virus and prokaryotic virus trees indicated an important role of phylogeny in driving the distribution of resCpG values. Thus, in order to account for the phylogenetic inertia, we studied the role of host associations using phylogenetic ANOVA analyses (Garland et al., 1993), which revealed that relatively few hosts account for significant variations in CpG abundance.

Among eukaryotic viruses, most significant differences were observed between arthropod-infecting viruses and viruses that infect vertebrates, protists, and algae. The latter two hosts were represented by few viruses, limiting the power of the ANOVA.

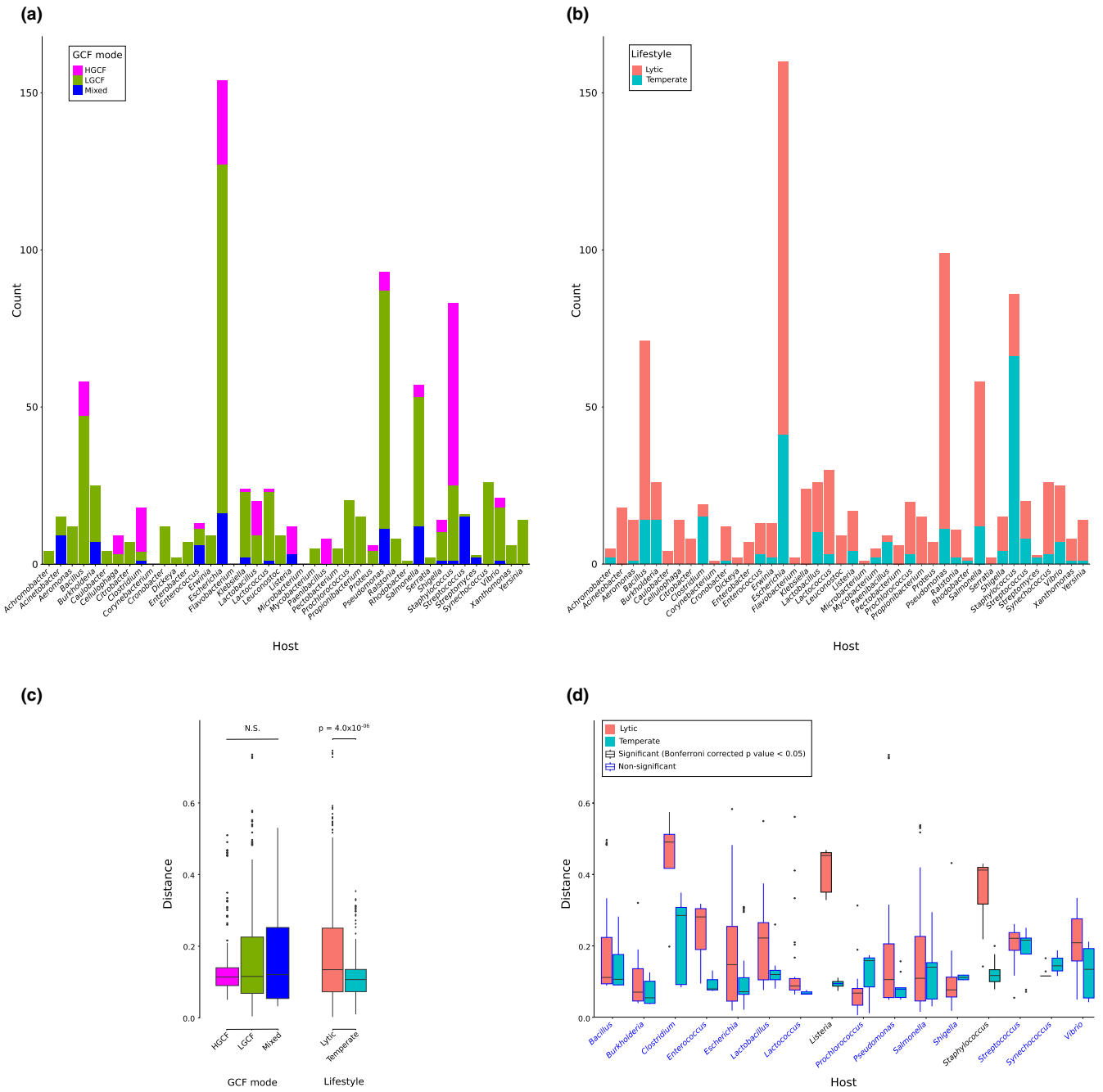

FIGURE 8 The effect of lifestyle and gene content flow on bacteriophage CpG content. (a) Barplots of the mode of horizontal gene transfer among bacteriophages (high- and low-gene content flux, HGCF and LGCF). (b) Barplots of the different bacteriophage lifestyles. (c) Boxplot of Euclidean distances between phages and their hosts grouped by GCF mode or lifestyle. For the lifestyle grouping the Wilcoxon Rank Sum test, *p*-value is also reported. (d) Boxplot of Euclidean distances between phages and their hosts grouped by lifestyle and host genus. Only genera for which at least three temperate and three lytic phages were available are reported. Statistical significance was assessed by Wilcoxon Rank Sum tests and by Bonferroni correction for multiple testing.

However, the differences among viruses somehow reflected the divide in terms of CpG representation among hosts, as insects had the lowest level of genomic CpG depletion. Whereas vertebrate genomes are known to be highly methylated, insect genomes display no or low level CpG methylation (Bewick et al., 2016; Zemach et al., 2010). Methylation status has been poorly investigated in protists and algae. Although several protists were reported to display little or no genome methylation (Drewell et al., 2023; Gissot et al., 2008; Pays et al., 1984; Ponts et al., 2013; Rojas & Galanti, 1991), *Entamoeba histolytica* (Fisher, 2004) and a genus of marine dinoflagellates (*Symbiodinium*; De Mendoza et al., 2018) do methylate cytosines at CpG sites. The heterogeneity of genome methylation in protists might thus explain the intermediate level of CpG depletion we observed in their genomes. Interestingly, marine

heterothrophic protists and amoeboid protists are thought to represent the natural hosts of giant viruses in the families *Mimiviridae* and *Marseilliviridae*, which showed some degree of CpG depletion, as well. Thus, as the CpG suppression of gammaherpesviruses, polyomaviruses, papillomaviruses, and to a lesser extent adenoviruses also suggests, the susceptibility to host methyltransferases might represent the driver of CpG suppression, either as a result of methyl-cytosine mutation or to avoid gene silencing (Hoelzer et al., 2008; White et al., 2009). This conclusion may seem in contrast with the variable level of suppression observed in viruses that encode their own 5-cytosine methyltransferases (Hoelzer et al., 2008; Wagner et al., 1985; Xu, 1998). However, the very reason that methylation is driven by viral enzymes suggests that these viruses experience no pressure to avoid CpG dinucleotide usage. Conversely, in these viruses, some loss of CpG might result from mutation of methyl-cytosines. In this respect, it is worth noting that some giant viruses that infect protists (e.g. viruses in the family *Marseilliviridae*) also encode their own methyltransferase, but these do not target CpG sites (Jeudy et al., 2020).

The possible role of methylation on CpG abundance is also supported by the analysis of coding sequences in viruses that infect vertebrates or invertebrates. In fact, a stronger negative correlation in the former compared to the latter was observed between CpG and TpG, suggesting a larger loss of CpG through methylation-induced mutation. However, as previously shown for mammalian transcripts, we also detected a negative correlation between CpG and ApG, again stronger for vertebrate-infecting viruses (Simmonds et al., 2013). This suggests the existence of mechanisms other than methylation in the suppression of CpG dinucleotides in coding sequences. Clearly, a possibility is that CpG avoidance in vertebrate-infecting viruses represents a strategy to avoid host immune sensing. Vertebrates use ZAP and TLR9 to sense CpG-containing non-self nucleic acids (i.e. CpG-rich RNAs and unmethylated CpG dinucleotides in DNA; Bowie & Unterholzner, 2008; Luo et al., 2020; Takata et al., 2017). These molecules are not encoded by invertebrate genomes, in line with their lack of CpG suppression. However, albeit little is known of defence responses in algae and protists, specific mechanisms that target CpG-containing DNA have not been reported in these organisms either. Thus, the presence of dedicated immune-response pathways does not represent an overarching explanation for the representation of CpG dinucleotides in eukaryotic viruses. In line with this view, we also failed to detect an effect on the spacing of CpG dinucleotides in viruses that infect hosts possessing ZAP (Gonçalves-Carneiro et al., 2022).

In prokaryotes, defence mechanisms that target viral nucleic acids are highly widespread and include restriction-modification (R-M) and CRISPR-Cas systems (Bernheim & Sorek, 2020). R-M systems are based on the cleavage of phage DNA through recognition of specific sequence motifs that are modified (usually by methylation) in the host DNA. Other systems (e.g. BREX and DISARM) based on DNA methylation were also identified but are still poorly understood (Goldfarb et al., 2015; Ofir et al., 2017). Although widespread R-M-mediated methylation at CpG sites was described in *Mycoplasma penetrans*, analysis of 230 diverse bacteria revealed no specific prevalence of CpG methylation by R-M systems (Blow et al., 2016; Wojciechowski et al., 2013). Likewise, the BREX and DISARM pathways are not known to preferentially target CpG dinucleotides (Goldfarb et al., 2015; Ofir et al., 2017). Moreover, the association of specific R-M systems with their hosts is often transient (Bernheim & Sorek, 2020). Thus, there is little evidence that bacterial immune responses underlie the variable representation of CpG nucleotides we observed in phage (and bacterial) genomes.

The phylogenetic ANOVA analysis identified cyanobacteria-infecting phages as CpG-depleted, whereas phages that infect bacteria in the genera *Burkolderia* and *Staphylococcus* as CpG-rich.

Comparison with bacterial genomes suggested that this effect is largely driven by the tendency for these phages to resemble the host's genomic CpG content. The reason(s) why these latter differ in terms of CpG content remain presently unknown. The tendency of phages to resemble their host genome composition was common to several other genera, although CpG content was not unexceptionally high or low (and thus did not contribute significant differences in the phylogenetic ANOVA). In this respect, it is worth mentioning that we used the genus as the taxonomic level of classification to assign host associations. However, growing evidence suggests that several phages have relatively broad host ranges, which can span more than one bacterial genus, whereas others may be restricted to individual bacterial species (de Jonge et al., 2019). Also, host range can evolve quickly (Meyer et al., 2016). It is thus possible that we detected significant similarities for phages that have stronger and long-lasting host associations (at the species of genus level). Indeed, this most likely is the case of temperate phages, which engage in long-term, mutually beneficial interactions with their hosts (Argov et al., 2017). Although the host range of temperate and lytic phages has not been systematically compared, several temperate phages integrate into bacterial chromosomes using integrases. The latter often rely on distinct target sequences and, as a consequence, temperate phages might have an overall narrower host range than lytic phages (de Jonge et al., 2019).

In line with our results, temperate bacteriophages were reported to adapt to their hosts in terms of gene orientation and DNA motifs involved in chromosome segregation and organization (Bobay et al., 2013). Previous analyses also indicated that the G+C content of bacteria genomes and their phages' are significantly correlated (Almpanis et al., 2018). Whereas we measured CpG representation after accounting for G+C content, it is possible that both represent specific adaptations of phages to their hosts, possibly to optimize gene expression levels or regulation. Although these observations partially shift the investigation of the determinants of CpG content from phages to bacteria, they shed light into a distinct tier of selection at the dinucleotide level and might have important practical implications. Indeed, phages are increasingly regarded as important tools to modulate bacterial communities in different settings. In particular, with the rise of antibiotic resistance among pathogens worldwide, phage therapy is increasingly considered an attractive alternative for application both in medicine and in the food industry (Chen et al., 2019;

Hatfull et al., 2022; Mahler et al., 2023; Meile et al., 2022; Mitsunaka et al., 2022). Clearly, the development of effective and safe approaches hinges upon accurate understanding of phage-host interactions and their possible evolution. Our results suggest that phage engineering approaches might benefit from taking G+C content and CpG abundance into account to better tailor host specificity and efficacy.

AUTHOR CONTRIBUTIONS

Conceptualization, M.S. and D.F.; Methodology, M.S., U.P., and D.F; Investigation, M.S., U.P., D.F., R.C.; Writing – Original draft, M.S., D.F.; Review & editing, M.S., U.P., and R.C.; Funding acquisition, M.S.

FUNDING INFORMATION

This work was supported by the Italian Ministry of Health ('Ricerca Corrente' to MS). APC funded by Bibliosan.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The lists of viral strains and host genomes analysed in this study with their Accession IDs are provided in Tables S1–S3.

ORCID

Diego Forni https://orcid.org/0000-0001-9291-5352

REFERENCES

Adriaenssens, E. M., Edwards, R., Nash, J. H. E., Mahadevan, P., Seto, D., Ackermann, H.-W., Lavigne, R., & Kropinski, A. M. (2015). Integration of genomic and proteomic analyses in the classification of the Siphoviridae family. *Virology*, *477*, 144–154.

Almpanis, A., Swain, M., Gatherer, D., & McEwan, N. (2018). Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microbial Genomics*, *4*(4), e000168.

Argov, T., Azulay, G., Pasechnek, A., Stadnyuk, O., Ran-Sapir, S., Borovok, I., Sigal, N., & Herskovits, A. A. (2017). Temperate bacteriophages as regulators of host behavior. *Current Opinion in Microbiology*, *38*, 81–87.

Bernheim, A., & Sorek, R. (2020). The pan-immune system of bacteria: Antiviral defence as a community resource. *Nature Reviews Microbiology*, *18*(2), 113–119.

Beutler, E., Gelbart, T., Han, J. H., Koziol, J. A., & Beutler, B. (1989). Evolution of the genome and the genetic code: Selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(1), 192–196.

Bewick, A. J., Vogel, K. J., Moore, A. J., & Schmitz, R. J. (2016). Evolution of DNA methylation across insects. *Molecular Biology and Evolution*, *34*, 654–665.

Bhunchoth, A., Blanc-Mathieu, R., Mihara, T., Nishimura, Y., Askora, A., Phironrit, N., Leksomboon, C., Chatchawankanphanich, O., Kawasaki, T., Nakano, M., Fujie, M., Ogata, H., & Yamada, T. (2016). Two asian jumbo phages, *ϕ*RSL2 and *ϕ*RSF1, infect Ralstonia solanacearum and show common features of *ϕ*KZ-related phages. *Virology*, *494*, 56–66.

Bird, A. P., & Taggart, M. H. (1980). Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acids Research*, *8*(7), 1485–1497.

Blin, K. (2023). *Ncbi-genome-download* (0.3.3) [computer software]. Zenodo. https://doi.org/10.5281/ZENODO.8192432

Blow, M. J., Clark, T. A., Daum, C. G., Deutschbauer, A. M., Fomenkov, A., Fries, R., Froula, J., Kang, D. D., Malmstrom, R. R., Morgan, R. D., Posfai, J., Singh, K., Visel, A., Wetmore, K., Zhao, Z., Rubin, E. M., Korlach, J., Pennacchio, L. A., & Roberts, R. J. (2016). The Epigenomic landscape of prokaryotes. *PLoS Genetics*, *12*(2), e1005854.

Bobay, L.-M., Rocha, E. P. C., & Touchon, M. (2013). The adaptation of temperate bacteriophages to their host genomes. *Molecular Biology and Evolution*, *30*(4), 737–751.

Bowie, A. G., & Unterholzner, L. (2008). Viral evasion and subversion of pattern-recognition receptor signalling. *Nature Reviews. Immunology*, *8*(12), 911–922.

Brown, T. L., Charity, O. J., & Adriaenssens, E. M. (2022). Ecological and functional roles of bacteriophages in contrasting environments: Marine, terrestrial and human gut. *Current Opinion in Microbiology*, *70*, 102229.

Burge, C., Campbell, A. M., & Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(4), 1358–1362.

Chen, Y., Batra, H., Dong, J., Chen, C., Rao, V. B., & Tao, P. (2019). Genetic engineering of bacteriophages against infectious diseases. *Frontiers in Microbiology*, *10*, 954.

Chibani, C. M., Farr, A., Klama, S., Dietrich, S., & Liesegang, H. (2019). Classifying the unclassified: A phage classification method. *Viruses*, *11*(2), 195.

de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J., & Dutilh, B. E. (2019). Molecular and evolutionary determinants of bacteriophage host range. *Trends in Microbiology*, *27*(1), 51–63.

de Mendoza, A., Bonnet, A., Vargas-Landin, D. B., Ji, N., Li, H., Yang, F., Li, L., Hori, K., Pflueger, J., Buckberry, S., Ohta, H., Rosic, N., Lesage, P., Lin, S., & Lister, R. (2018). Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. *Nature Communications*, *9*(1), 1341.

Drewell, R. A., Cormier, T. C., Steenwyk, J. L., Denis, J. S., Tabima, J. F., Dresch, J. M., & Larochelle, D. A. (2023). The *Dictyostelium discoideum* genome lacks significant DNA methylation and uncovers palindromic sequences as a source of false positives in bisulfite sequencing. *NAR Genomics and Bioinformatics*, *5*(2), lqad035.

Duret, L., & Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics*, *4*(5), e1000071.

Fisher, O. (2004). Characterization of cytosine methylated regions and 5-cytosine DNA methyltransferase (Ehmeth) in the protozoan parasite Entamoeba histolytica. *Nucleic Acids Research*, *32*(1), 287–297.

Garland, T., Dickerman, A. W., Janis, C. M., & Jones, J. A. (1993). Phylogenetic analysis of covariance by computer simulation. *Systematic Biology*, *42*(3), 265–292.

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–695.

Gentles, A. J., & Karlin, S. (2001). Genome-scale compositional comparisons in eukaryotes. *Genome Research*, *11*(4), 540–546.

Giallonardo, F. D., Schlub, T. E., Shi, M., & Holmes, E. C. (2017). Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species. *Journal of Virology*, *91*(8), e02381-16. Print 2017 Apr 15.

Gissot, M., Choi, S.-W., Thompson, R. F., Greally, J. M., & Kim, K. (2008). *Toxoplasma gondii* and *Cryptosporidium parvum* lack detectable DNA cytosine methylation. *Eukaryotic Cell*, *7*(3), 537–540.

Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., Afik, S., Ofir, G., & Sorek, R. (2015). BREX is a novel phage resistance system widespread in microbial genomes. *The EMBO Journal*, *34*(2), 169–183.

Gonçalves-Carneiro, D., Mastrocola, E., Lei, X., DaSilva, J., Chan, Y. F., & Bieniasz, P. D. (2022). Rational attenuation of RNA viruses with zinc finger antiviral protein. *Nature Microbiology*, *7*(10), 1558–1567.

Gonçalves-Carneiro, D., Takata, M. A., Ong, H., Shilton, A., & Bieniasz, P. D. (2021). Origin and evolution of the zinc finger antiviral protein. *PLoS Pathogens*, *17*(4), e1009545.

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., & Challenger, W. (2008). GEIGER: Investigating evolutionary radiations. *Bioinformatics*, *24*(1), 129–131.

Hatfull, G. F., Dedrick, R. M., & Schooley, R. T. (2022). Phage therapy for antibiotic-resistant bacterial infections. *Annual Review of Medicine*, *73*(1), 197–211.

Hoelzer, K., Shackelton, L. A., & Parrish, C. R. (2008). Presence and role of cytosine methylation in DNA viruses of animals. *Nucleic Acids Research*, *36*(9), 2825–2837.

Ibrahim, A., Fros, J., Bertran, A., Sechan, F., Odon, V., Torrance, L., Kormelink, R., & Simmonds, P. (2019). A functional investigation of the suppression of CpG and UpA dinucleotide frequencies in plant RNA virus genomes. *Scientific Reports*, *9*(1), 18359.

Jeudy, S., Rigou, S., Alempic, J.-M., Claverie, J.-M., Abergel, C., & Legendre, M. (2020). The DNA methylation landscape of giant viruses. *Nature Communications*, *11*(1), 2657.

Karlin, S., & Burge, C. (1995). Dinucleotide relative abundance extremes: A genomic signature. *Trends in Genetics: TIG*, *11*(7), 283–290.

Karlin, S., Doerfler, W., & Cardon, L. R. (1994). Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *Journal of Virology*, *68*(5), 2889–2897.

Karlin, S., & Mrázek, J. (1997). Compositional differences within and between eukaryotic genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(19), 10227–10232.

Koskella, B., Hernandez, C. A., & Wheatley, R. M. (2022). Understanding the impacts of bacteriophage viruses: From laboratory evolution to natural ecosystems. *Annual Review of Virology*, *9*(1), 57–78.

Lieberman, P. M. (2013). Keeping it quiet: Chromatin control of gammaherpesvirus latency. *Nature Reviews Microbiology*, *11*(12), 863–875.

Lobo, F. P., Mota, B. E. F., Pena, S. D. J., Azevedo, V., Macedo, A. M., Tauch, A., Machado, C. R., & Franco, G. R. (2009). Virus-host coevolution: Common patterns of nucleotide motif usage in Flaviviridae and their hosts. *PLoS One*, *4*(7), e6282.

Luo, X., Wang, X., Gao, Y., Zhu, J., Liu, S., Gao, G., & Gao, P. (2020). Molecular mechanism of RNA recognition by zinc-finger antiviral protein. *Cell Reports*, *30*(1), 46–52.

Mahler, M., Costa, A. R., Van Beljouw, S. P. B., Fineran, P. C., & Brouns, S. J. J. (2023). Approaches for bacteriophage genome engineering. *Trends in Biotechnology*, *41*(5), 669–685.

Mavrich, T. N., & Hatfull, G. F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology*, *2*(9), 17112.

Medvedeva, S., Borrel, G., Krupovic, M., & Gribaldo, S. (2023). A compendium of viruses from methanogenic archaea reveals their diversity and adaptations to the gut environment. *Nature Microbiology*, *8*(11), 2170–2182.

Meile, S., Du, J., Dunne, M., Kilcher, S., & Loessner, M. J. (2022). Engineering therapeutic phages for enhanced antibacterial efficacy. *Current Opinion in Virology*, *52*, 182–191.

Meyer, J. R., Dobias, D. T., Medina, S. J., Servilio, L., Gupta, A., & Lenski, R. E. (2016). Ecological speciation of bacteriophage lambda in allopatry and sympatry. *Science*, *354*(6317), 1301–1304.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., & Ogata, H. (2016). Linking virus genomes with host taxonomy. *Viruses*, *8*(3), 66.

Mitsunaka, S., Yamazaki, K., Pramono, A. K., Ikeuchi, M., Kitao, T., Ohara, N., Kubori, T., Nagai, H., & Ando, H. (2022). Synthetic engineering and biological containment of bacteriophages. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(48), e2206739119.

Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., Blanc-Mathieu, R., Yamamoto, K., Hingamp, P., Sako, Y., Sullivan, M. B., Goto, S., Ogata, H., & Yoshida, T. (2017). Environmental viral genomes shed new light on virus-host interactions in the ocean. *mSphere*, *2*(2), e00359-16.

Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., & Goto, S. (2017). ViPTree: The viral proteomic tree server. *Bioinformatics*, *33*(15), 2379–2380.

Odon, V., Fiddaman, S. R., Smith, A. L., & Simmonds, P. (2022). Comparison of CpG- and UpA-mediated restriction of RNA virus replication in mammalian and avian cells and investigation of potential ZAP-mediated shaping of host transcriptome compositions. *RNA (New York, N.Y.)*, *28*(8), 1089–1109.

Ofir, G., Melamed, S., Sberro, H., Mukamel, Z., Silverman, S., Yaakov, G., Doron, S., & Sorek, R. (2017). DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nature Microbiology*, *3*(1), 90–98.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*(6756), 877–884.

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, *35*(3), 526–528.

Pays, E., Delauw, M. F., Laurent, M., & Steinert, M. (1984). Possible DNA modification in GC dinucleotides of *Trypanosoma brucei* telomeric sequences; relationship with antigen gene transcriptiond. *Nucleic Acids Research*, *12*(13), 5235–5247.

Ponts, N., Fu, L., Harris, E. Y., Zhang, J., Chung, D.-W. D., Cervantes, M. C., Prudhomme, J., Atanasova-Penichon, V., Zehraoui, E., Bunnik, E. M., Rodrigues, E. M., Lonardi, S., Hicks, G. R., Wang, Y., & Le Roch, K. G. (2013). Genome-wide mapping of DNA methylation in the human malaria parasite *Plasmodium falciparum*. *Cell Host & Microbe*, *14*(6), 696–706.

Provataris, P., Meusemann, K., Niehuis, O., Grath, S., & Misof, B. (2018). Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. *Genome Biology and Evolution*, *10*(4), 1185–1197.

Puxty, R. J., & Millard, A. D. (2023). Functional ecology of bacteriophages in the environment. *Current Opinion in Microbiology*, *71*, 102245.

Revell, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things): *Phytools: R package*. *Methods in Ecology and Evolution*, *3*(2), 217–223.

Revell, L. J. (2024). *Phytools 2.0: An updated R ecosystem for phylogenetic comparative methods (and other things)*. *PeerJ*, *12*, e16505.

Rima, B. K., & McFerran, N. V. (1997). Dinucleotide and stop codon frequencies in single-stranded RNA viruses. *The Journal of General Virology*, *78*(Pt 11), 2859–2870.

Rohwer, F., & Edwards, R. (2002). The phage proteomic tree: A genome-based taxonomy for phage. *Journal of Bacteriology*, *184*(16), 4529–4535.

Rojas, M. V., & Galanti, N. (1991). Relationship between DNA methylation and cell proliferation in *Trypanosoma cruzi*. *FEBS Letters*, *295*(1–3), 31–34.

Sexton, N. R., & Ebel, G. D. (2019). Effects of arbovirus multi-host life cycles on dinucleotide and codon usage patterns. *Viruses*, *11*(7), 643.

Simmen, M. W. (2008). Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. *Genomics*, *92*(1), 33–40.

Simmonds, P., Xia, W., Baillie, J. K., & McKinnon, K. (2013). Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla–selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. *BMC Genomics*, *14*, 610.

Takata, M. A., Gonçalves-Carneiro, D., Zang, T. M., Soll, S. J., York, A., Blanco-Melo, D., & Bieniasz, P. D. (2017). CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature*, *550*(7674), 124–127.

Wagner, H., Simon, D., Werner, E., Gelderblom, H., Darai, C., & Flügel, R. M. (1985). Methylation pattern of fish lymphocystis disease virus DNA. *Journal of Virology*, *53*(3), 1005–1007.

White, M. K., Safak, M., & Khalili, K. (2009). Regulation of gene expression in primate polyomaviruses. *Journal of Virology*, *83*(21), 10846–10856.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, *22*(3), 392.

Wojciechowski, M., Czapinska, H., & Bochtler, M. (2013). CpG underrepresentation and the bacterial CpG-specific DNA methyltransferase M.MpeI. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 105–110.

Xu, M. (1998). Cloning, characterization and expression of the gene coding for a cytosine-5-DNA methyltransferase recognizing GpC. *Nucleic Acids Research*, *26*(17), 3961–3966.

Zemach, A., McDaniel, I. E., Silva, P., & Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, *328*(5980), 916–919.

Zhang, W., Wang, R., Zou, X., Gu, C., Yang, Q., He, M., Xiao, W., He, L., Zhao, M., & Yu, Z. (2023). Comparative genomic analysis of alloherpesviruses: Exploring an available genus/species demarcation proposal and method. *Virus Research*, *334*, 199163.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Forni, D., Pozzoli, U., Cagliani, R., & Sironi, M. (2024). Dinucleotide biases in the genomes of prokaryotic and eukaryotic dsDNA viruses and their hosts. *Molecular Ecology*, *33*, e17287. https://doi.org/10.1111/mec.17287