VIRUS EVOLUTION

*Virus Evolution*, 2023, **9(1)**, 1–17

DOI: https://doi.org/10.1093/ve/veac123

Advance access publication date: 23 December 2022

**Research Article**

# Reekeekee- and roodoodooviruses, two different *Microviridae* clades constituted by the smallest DNA phages

Eric Olo Ndela,[1,†] Simon Roux,[2,‡] Christian Henke,[3,4] Alexander Sczyrba,[3,4] Télesphore Sime Ngando,[1] Arvind Varsani,[5,6,§] and François Enault[1,*]

[1]Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Genome et Environnement, Clermont-Ferrand F-63000, France, [2]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [3]Computational Metagenomics, Bielefeld University, Universitätsstraße 27, Bielefeld 30501, Germany, [4]Center for Biotechnology, Bielefeld University, Universitätsstraße 27, Bielefeld 33615, Germany, [5]The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA and [6]Structural Biology Research Unit, Department of Integrative Biomedical Sciences, University of Cape Town, Observatory, Cape Town 7701, South Africa

[†]https://orcid.org/0000-0002-4241-1161

[‡]https://orcid.org/0000-0002-5831-5895

[§]https://orcid.org/0000-0003-4111-2415

*Corresponding author: E-mail: francois.enault@uca.fr

## Abstract

Small circular single-stranded DNA viruses of the *Microviridae* family are both prevalent and diverse in all ecosystems. They usually harbor a genome between 4.3 and 6.3 kb, with a microvirus recently isolated from a marine Alphaproteobacteria being the smallest known genome of a DNA phage (4.248 kb). A subfamily, *Amoyvirinae*, has been proposed to classify this virus and other related small Alphaproteobacteria-infecting phages. Here, we report the discovery, in meta-omics data sets from various aquatic ecosystems, of sixteen complete microvirus genomes significantly smaller (2.991–3.692 kb) than known ones. Phylogenetic analysis reveals that these sixteen genomes represent two related, yet distinct and diverse, novel groups of microviruses—*amoyviruses* being their closest known relatives. We propose that these small microviruses are members of two tentatively named subfamilies *Reekeekeevirinae* and *Roodoodoovirinae*. As known microvirus genomes encode many overlapping and overprinted genes that are not identified by gene prediction software, we developed a new methodology to identify all genes based on protein conservation, amino acid composition, and selection pressure estimations. Surprisingly, only four to five genes could be identified per genome, with the number of overprinted genes lower than that in phiX174. These small genomes thus tend to have both a lower number of genes and a shorter length for each gene, leaving no place for variable gene regions that could harbor overprinted genes. Even more surprisingly, these two *Microviridae* groups had specific and different gene content, and major differences in their conserved protein sequences, highlighting that these two related groups of small genome microviruses use very different strategies to fulfill their lifecycle with such a small number of genes. The discovery of these genomes and the detailed prediction and annotation of their genome content expand our understanding of ssDNA phages in nature and are further evidence that these viruses have explored a wide range of possibilities during their long evolution.

**Key words:** virus; smallest DNA phages; microviridae; gene overprinting; gene calling.

## 1. Introduction

Phages in the family *Microviridae* are important entities in viral communities with broad prevalence and diversity. They are frequently identified in metagenomes derived from all types of ecosystems and are known to infect very different bacterial hosts. These phages have small icosahedral capsids (Caps; ∼30 nm diameter) and a circular single-stranded DNA genome ranging in size from 4.3 to 6.3 kb (Doore and Fane 2016). Although most cultured microviruses were strictly lytic like their first discovered representative phiX174, proviruses have been recently discovered in bacterial genomes and successfully cultivated (Krupovic and Forterre 2011; Kirchberger and Ochman 2020), testifying that these viruses can also have a temperate lifestyle. The family *Microviridae* is classified in the *Monodnaviria* domain, and it is the sole family of ssDNA viruses in the *Sangervirae* kingdom (Koonin et al. 2020) whose members all share a major Cap protein harboring a single jelly roll (SJR) structure and a rolling-circle replication (Rep) endonuclease that initiates genome (Rep). The family *Microviridae* includes two subfamilies, *Bullavirinae* and *Gokushovirinae*, which differ in their genome composition and Cap structure. Bullaviruses are environmentally rare but have been intensely studied, as these Gammaproteobacteria-infecting viruses are represented by phiX174, a major molecular biology model organism (Sanger et al. 1977). On the other hand, gokushoviruses

are often abundant in the environment and infect various hosts including *Spiroplasma, Chlamydia, Bdellovibrio* (Chipman et al. 1998; Brentlinger et al. 2002; Everson et al. 2002), or *Enterobacteriaceae* (Kirchberger and Ochman 2020). Thousands of complete genomes of microviruses have been assembled in metagenomic studies from diverse natural habitats (Roux et al. 2012; Labonté and Suttle 2013; Székely and Breitbart 2016) or from animal feces and tissues (Roux et al. 2012). These environmental microviruses both shed light on the huge diversity of the members of the *Gokushovirinae* subfamily and led to the proposal of several putative subfamilies including *Alpavirinae* found abundantly in the human gut (Krupovic and Forterre 2011), *Pichovirinae* in various ecosystems (Roux et al. 2012), *Aravirinae* and *Stokavirinae* in peatland systems (Quaiser et al. 2015), *Sukshmavirinae* from the termite gut (Tikhe and Husseneder 2018), Group D associated to dragonflies (Rosario et al. 2012), as well as groups such as pequenoviruses identified in methane seep sediments (Bryson et al. 2015). Additionally, three subfamilies were very recently discovered: *Occultatumvirinae, Tainavirinae*, and *Libervirinae* (Zhang et al. 2021a; Zucker et al. 2022).

Recently, a new subfamily, named *Amoyvirinae*, was proposed for a new group made of both proviruses and isolates infecting Alphaproteobacteria (Zheng et al. 2018). Among them, vB_RpoMi-Mini, isolated from a marine Rhodobacteraceae, was reported as the smallest ssDNA phage with a 4,248-nucleotide (nt)-long genome (Zhan and Chen 2019). Other members of *Amoyvirinae*—for example vB_Cib_ssDNA_P1 that infects a marine Sphingomonadaceae (Zheng et al. 2018) and RHph_N39 that infects a soil *Rhizobiaceae* (Van Cauwenberghe et al. 2021)—both have small genomes (respectively 4,360 and 4,781 nt). Viruses in the proposed subfamily *Amoyvirinae* appear to be quite distant from the other members of the family *Microviridae* based on a recent phylogenomic analysis (Kirchberger, Martinez, and Ochman 2022) that proposed the existence of three major groups—*Amoyvirinae* being in a group containing only 63 out of more than 13,000 genomes. In terms of gene content, four genes were predicted on the same strand for vb_RpoMi-Mini: the major Cap protein, the Rep endonuclease, an M15 family peptidase-related protein, and a large gene of unknown function. Three of these predicted genes are conserved among all amoyviruses, while the peptidase was horizontally acquired on several occasions, and thus this peptidase is not ubiquitous but found in other microviral clades (Roux et al. 2012). In addition to these four genes, vB_Cib_ssDNA_P1 and RHph_N39 are predicted to encode two additional non-conserved proteins of unknown function. The four amoyviral core genes are non-overlapping, which might reflect an erroneous gene prediction. Indeed, overlapping and even overprinted genes are often not predicted by protein prediction software. For example, only seven of the eleven experimentally determined phiX174 proteins are predicted using Prodigal (Hyatt et al. 2010).

In this study, we assembled 2,946 viral metagenomes and searched for the presence of undescribed small genomes with similarities to members of the family *Microviridae*. Sixteen genomes ranging from 2.9 to 3.6 kb were identified in aquatic and sediment samples, as well as associated with fish microbiome. A phylogenetic analysis revealed two groups related to *Amoyvirinae*. For these sixteen newly found microvirus genomes, we conducted an in-depth analysis of all open reading frames (ORFs), taking into account their sequence conservation, amino acid composition, and single nucleotide polymorphisms (SNPs), in order to find all *bona fide* genes and try to better understand how these organisms function with such small genomes.

## 2. Materials and methods

### 2.1 Public virome data set

A data set of 2,946 publicly available viromes was downloaded for this analysis, two of them being only available as assembled contigs. For the other 2,944 viromes, reads were first cleaned by removing potential adapters using Cutadapt v2.8 (Martin 2011) and trimmed using Trimmomatic v0.36 (Bolger, Lohse, and Usadel 2014). Depending on the sequencing technology used to generate reads, the data sets were individually assembled into contigs using Newbler (Silva et al. 2013), IDBA_UD (Peng et al. 2012), or MEGAHIT v1.1.3 (Li et al. 2015) with default parameters. Details of these viromes (database source, ecosystem type, sampling location, associated publication, etc.) are provided in Supplementary Table S1. Circular contigs were detected based on identical 5′ and 3′ ends of at least ten nucleotides.

### 2.2 Identifying small circular contigs similar to *Microviridae*

To identify new microvirus genomes, a data set of thirty-three reference viruses was generated that represent all known members of the *Microviridae* subfamilies described so far. The major Cap protein and Rep initiation protein of the viruses in this data set were aligned using MAFFT v7.470 (*–genafpair* with 1,000 iterations; Katoh et al. 2002), and two hidden Markov model (HMM) profiles were generated. Three different sources were used to search for small potential microviruses. First, sequences described as complete genomes in GenBank (January 2020; Clark et al. 2016) were retrieved using BLASTp (Altschul et al. 1990) with the Cap proteins of phiX174 (GenBank accession #AF299307) and Chlamydia phage Chp1 (GenBank accession #NC001741) as bait (bit-score >40). Second, contigs from publicly available assembled metagenomes from IMG/JGI (Markowitz et al. 2012) were screened with the two reference HMM profiles using HMMER v3.1b2 (*hmmsearch* e-value > 0.001; Eddy 1998). Finally, complete microvirus genomes were searched among the circular contigs from the 2,946 in-house assembled viromes. To this end, contig proteins were predicted using Prodigal v2.6.2 (Hyatt et al. 2010) and compared to Cap and Rep HMM profiles of reference microviruses using HMMER.

### 2.3 Genomic and phylogenetic analysis of the small microviruses

First, intergenomic similarities were determined between the sixteen identified small microviruses and the thirty-three reference genomes using VIRIDIC (Moraru, Varsani, and Kropinski 2020). Genera and species were defined using classical thresholds of 70 and 95 per cent of intergenomic identity, respectively. Following this, genes were predicted using Prodigal (Hyatt et al. 2010) and Cap and Rep proteins were identified in the sixteen genomes. The three motifs conserved in rolling-circle Rep endonucleases were identified in Rep proteins (Chandler et al. 2013): (1) the HUH motif made of two histidine residues (H) separated by a bulky hydrophobic residue (U), (2) the UUTU motif, U being hydrophobic residues, and (3) the YxxKY motif, x being any amino acid. The structure of the major Cap proteins was predicted using ColabFold (Mirdita et al. 2022), allowing to identify the position on the protein sequence of the eight beta strands assembled in two antiparallel sheets constituting the SJR fold, characteristic of such Cap proteins (Liljas 1991; Koonin et al. 2020). Together with the proteins from the thirty-three references, a multiple alignment was computed independently for these two core proteins. Amino acid

identity was calculated using these two alignments and plotted as a heatmap using the ggplot R package (Wickham 2009). Three phylogenies were then computed, one for each of the two multiple alignments and one based on their concatenation, using RAxML v8.2.12 (Stamatakis 2014) with 100 bootstrap replicates and automatic model selection (PROTGAMMAUTO) and visualized using iTOL (Letunic and Bork 2021). Midpoint rooting was applied to all phylogenies.

## 2.4 Potential host prediction and detection of dif-motifs

To try and identify prophages closely related to the sixteen genomes that could help identifying their host, sequences similar to one of the major Cap proteins of all newly assembled microvirus genomes were searched in RefSeq Bacteria (tBLASTn). In addition, CRISPR spacer sequences similar to our sixteen microviruses, traces of a previous viral infection in bacterial genomes, were searched using BLASTn against the IMG spacer database (Markowitz et al. 2012) and using SpacePHARER (Zhang et al. 2021b).

To assess the possibility of a lysogenic lifestyle, sequences recognized by the XerCD bacterial recombination machinery (dif-motifs) were searched in all newly found small complete microvirus genomes using nhmmer from HMMER v3.1b2 (Wheeler and Eddy 2013). For this search of dif-motifs, two HMMs were independently built and used: one with the bacterial dif-motifs (Kono, Arakawa, and Tomita 2011: 578 motifs) and one with the first fifteen nucleotides of dif-motifs detected in enterogokushoviruses (Kirchberger and Ochman 2020: 81 motifs).

## 2.5 Clustering all ORFs

To identify all protein coding genes in these small genomes, including overlapping and overprinted ones, all ORFs of at least eighteen residues were identified in these forty-nine microvirus genomes using getorf from the EMBOSS suite (Rice, Longden, and Bleasby 2000). The 4,719 identified ORFs were organized into orthologous groups (OGs) in a two-step procedure as in Olo Ndela, Enault, and Toussaint (2021): (1) all proteins were compared to each other using MMseqs2 (Steinegger and Söding 2017) (bit-score $\geq$50 and reciprocal coverage >70%) and clustered using MCL v14-137 (Enright, Van Dongen, and Ouzounis 2002) (inflation 2.0); and (2) for each protein cluster, a multiple alignment was built using Clustal Omega (Sievers et al. 2011) and transformed into an HMM profile and all profiles were compared to each other using HHsearch of the HH-suite toolkit (Steinegger et al. 2019) (version 2.0.16, cut-offs: probability $\geq$90% and coverage $\geq$50%). Protein singletons were also individually compared to all these HMM profiles. Using significant similarities, singletons and protein clusters were then again clustered using MCL (inflation 2.0). This second step involving HMM comparisons was then performed twice, with no threshold on the coverage for the last iteration, to group clusters that were similar but not grouped yet. Finally, HMM profiles of the OGs were annotated by comparison to the Pfam database (Finn et al. 2014).

## 2.6 Defining coding ORFs

In order to define which ORF corresponds to a protein coding gene, additional indices were computed to complement the ORF conservation. First, the percentage of high-degeneracy amino acids (leucine-L, arginine-R, and serine-S, namely LRS%) and the number of tyrosines were calculated for each ORF of the sixteen small microvirus genomes. Then, to assess the selective constraints on each ORF, single nucleotide variations were identified in the sixteen genomes. Reads of the virome from which each genome was assembled were mapped to their corresponding genome using *bwa mem* v0.7.17 (Li and Durbin 2009) with default parameters. SNPs were called using the bcftools v1.14 commands *mpileup* and *call* successively (Li et al. 2009). Only SNPs present in a minimum of two reads and representing more than 1 per cent of the total depth were considered for downstream analysis. An SNP in a protein coding gene can be synonymous or non-synonymous, *i.e.* leading to an amino acid substitution. In the BLOSUM62 substitution matrix, a score is attributed to each amino acid substitution and reflects the likelihood that this substitution was produced by an evolutionary trajectory over time rather than by chance (Cline and Karchin 2011). The idea here is that synonymous SNPs or SNPs leading to amino acid substitutions with positive scores are more consistent with evolutionary trends and are more likely to happen in coding regions than in non-coding ORFs, whereas SNPs leading to substitutions with negative scores will be more likely to happen in non-coding regions. The number and proportion of synonymous (no amino acid change), positive (substitution score >0 in the BLOSUM62 matrix), negative (score between –2 and 0), and strongly negative (score <–2) amino acid substitutions were computed. In order to find additional SNPs, 6,329 contigs (Supplementary Table S2) similar to the sixteen complete small microviruses were retrieved from 2,946 assembled viromes and metagenomes/metatranscriptomes from the IMG database (tBLASTn of the contigs against all ORFs of the sixteen newly identified microviruses, e-value < 0.001). To further evaluate selective pressure on ORFs, the pN/pS ratio was computed for each of them, dividing the number of non-synonymous mutations (pN) by the number of synonymous ones (pS). The identified contigs, mostly partial microvirus genomes, were then mapped to the sixteen genomes of interest, and SNPs were assessed using the methodology described previously. All measured parameters, namely LRS%, SNPs, pN/pS, and conservation, are summarized for all ORFs in Supplementary Table S5.

To test whether the gene conservation, composition, and SNP analysis could help finding *bona fide* genes, we performed these analyses on the bullavirus phiX174 whose gene content is experimentally validated. For the gene conservation analysis, fifteen microvirus genomes ranging from 0 to 100 per cent intergenomic identity with phiX174 were randomly chosen from GenBank sequences and our in-house virome contigs. For these fifteen genomes and phiX174, ORFs were predicted and our two-step clustering strategy was applied on them. For the SNP analysis, twenty-five phiX-like genomes were selected from GenBank (intergenomic identity > 95%; Supplementary Table S4) and mapped to the phiX174 genome to detect SNPs. These SNPs were used as references to calculate the pN/pS ratio on the experimentally verified genes and the predicted ORFs of phiX174. Finally, the percentage of high-degeneracy amino acids (LRS%) was computed for each ORF and gene of phiX174.

For each of the sixteen genomes and phiX174, a genomic map was drawn and colored according to ORF conservation, LRS%, and pN/pS values using the Python library *DNA Features Viewer* (Zulkower and Rosser 2020). SNPs were plotted with ggplot (Wickham 2009) as the sum of ratios of SNP depths over total depths across a sliding window of 15 nt, divided by 15 to normalize the window size. Additionally, the number of tBLASTn hits was evaluated as a heatmap using a 15-nt window, also plotted with *DNA Features Viewer*, in order to emphasize the conserved zones

of each ORF. All these results were used to manually predict *bona fide* genes and refine their 5′ ends. Once all genes were defined, secondary structures of intergenic regions were predicted using RNAfold v2.5.0 (Lorenz, Hofacker, and Stadler 2016), with default DNA parameters, and visualized with RiboSketch (Lu et al. 2018). Quality assessment was done via the entropy values associated with each nucleotide as a stability metric for folded structures via hydrogen bonds, alongside sequence conservation in related microviruses. Finally, for three representative genomes, the secondary structures of all their proteins were tentatively predicted using the ColabFold web server (Mirdita et al. 2022) with default parameters.

## 3. Results

### 3.1 The sixteen newly detected genomes are significantly shorter than any known phages

Using the HMM profiles of the Rep and Cap proteins from thirty-three reference microviruses, a total of 12,495 circular contigs and genomes similar to microvirus ones were identified in sequence data from GenBank, in IMG microbial metagenomes and metatranscriptomes (Nowinski et al. 2019), and in 2,946 assembled viral metagenomes. Although many contigs (1,314) were shorter than the smallest genome of a cultivated microvirus (4.248 kb), only a handful of contigs stood out as extremely small (Supplementary Fig. S1). Eleven contigs smaller than 3.5 kb were thus recovered, and their Cap proteins were compared to the other 12,484 sequences in order to identify related ones (BLASTp > 50 on bit-score). Five additional contigs, themselves small (<3.692 kb), had a significant similarity with one of the former eleven smallest ones, leading to a total of sixteen sequences. These sixteen circular contigs (Supplementary Table S3) had a region similar to at least one of the two HMMs, all being similar to the Rep profile (probability > 99%) and eleven similar to the Cap profile (probability > 99%). Two genomes were found in the GenBank database (assembled from viromes), three in microbiomes, one in a metatranscriptome, and ten in the set of public viromes we assembled. These sixteen genomes originated from fifteen different data sets derived from marine water ($n = 7$), sediments ($n = 1$), freshwater river ($n = 1$), wastewater systems ($n = 4$), and samples associated with cellular aquatic hosts, *i.e.* marine sponge ($n = 1$) and freshwater fishes (minnow, $n = 2$). These contigs, from 2.991 to 3.692 kb in length, were significantly smaller than vB_RpoMi-Mini, so far considered as the shortest known DNA phage (4.248 kb, GenBank accession #MF101922) (Zhan and Chen 2019). These phages are also smaller than the shortest known RNA phage, the 3.405-kb-long *Enterobacteria* phage M that belongs to the *Leviviricetes* class.

### 3.2 The two core proteins of the *Microviridae* family were (not so easily) identified in these sixteen genomes

As described previously, a protein similar to the Rep protein of known microviruses was detected in all sixteen new genomes, all three known motifs of Rep proteins being found in the sixteen proteins (Supplementary Fig. S2). Yet, a Cap protein was only found in eleven genomes, when all ORFs were individually compared to the Cap protein HMM profile. As other viral families contain similar (yet distant) Rep initiation proteins, the identification of the typical Cap protein is necessary to ensure the remaining five viral genomes are microviruses. Thus, we wanted to determine whether the Cap protein was really absent or just too distant to be identified in these genomes. Only one ORF was considered as long enough (>200 residues) to be a putative Cap-encoding in each of the five genomes. These five proteins were compared and were indeed similar to each other, although sometimes distant with a minimum BLASTp bit-score of 124. For the search of Cap proteins to be more sensitive, an HMM profile was built out of the multiple alignment of these five proteins and was compared to the HMM profile built from the alignment of the thirty-three reference Cap proteins and that of the eleven Cap proteins detected earlier. Although only distantly related, these two HMM profiles were significantly similar (e-value of $10^{-5}$). In addition, protein structure prediction confirmed the presence of an SJR domain in these candidate Cap proteins. Thus, the two core proteins of microviruses were detected in all sixteen newly detected genomes. As a final verification, these sixteen Cap and Rep proteins were compared to all viral proteins from the KEGG (Kyoto Encyclopedia of Genes and Genomes; Kanehisa and Goto, 2000) database (BLASTp, e-value $<10^{-5}$) and all proteins identified as similar were proteins from microviruses. All these elements allow us to state that these genomes undoubtedly belong to the *Microviridae* family.

### 3.3 The sixteen genomes are separated into two clades, distantly related to *Amoyvirinae*

To place the sixteen new genomes in the known microvirus diversity, phylogenies derived from multiple sequence alignments (MSAs) on the Rep and Cap proteins of the thirty-three reference microviruses and sixteen new genomes were inferred. All groups of closely related virus sequences (amino acid identity of Rep or Cap protein >50%) in one phylogeny formed monophyletic groups with high bootstrap values in the other phylogeny (Supplementary Fig. S3). In contrast, relationships between distant groups, corresponding to long branches on the trees, were more variable and had low bootstrap values in both phylogenies. This suggests that each protein alone allows us to identify the same groups of closely related viruses but does not carry enough information to decipher distant relationships on its own. To maximize the information, the two alignments were thus concatenated to generate a more robust phylogeny (Fig. 1).

Considering this phylogeny, the five genomes with a distant Cap protein formed a clade related to the *Amoyvirinae* subfamily, and the other eleven clustered in a clade related to the first one. Branch lengths inside these two new clades were comparable to the ones of the known subfamilies, suggesting that they represent two related yet distinct and diverse groups of microviruses. We thus tentatively propose two new subfamilies to classify these viruses, *Reekeekeevirinae* (meaning 'tiny' in French) and *Roodoodooviri nae* (meaning 'little sweet' in French). To ensure that our two new groups are different from the recently discovered Occultatumvirinae, Tainavirinae, and Libervirinae subfamilies, we generated an additional Cap protein phylogeny (Supplementary Fig. S4). Reekeekee- and roodoodooviruses were clearly separated from these three clades on this phylogeny.

Genome length within each group was coherent, with reekeekeeviruses being constituted by the five smallest genomes. Based on VIRIDIC analysis, only three and two genomes were clustered into a single species (and genus) in reekeekee- and roodoodooviruses, respectively; in both cases the genomes have been sampled from very similar ecosystems, *i.e.* wastewater (Brinkman et al. 2018) and marine water (Kim, Aw, and Rose 2016), respectively (Fig. 1 and Supplementary Fig. S5). Considering more distant pairs of genomes, the fourth genome from wastewater also exhibited some nucleotidic similarity with the three genomes forming

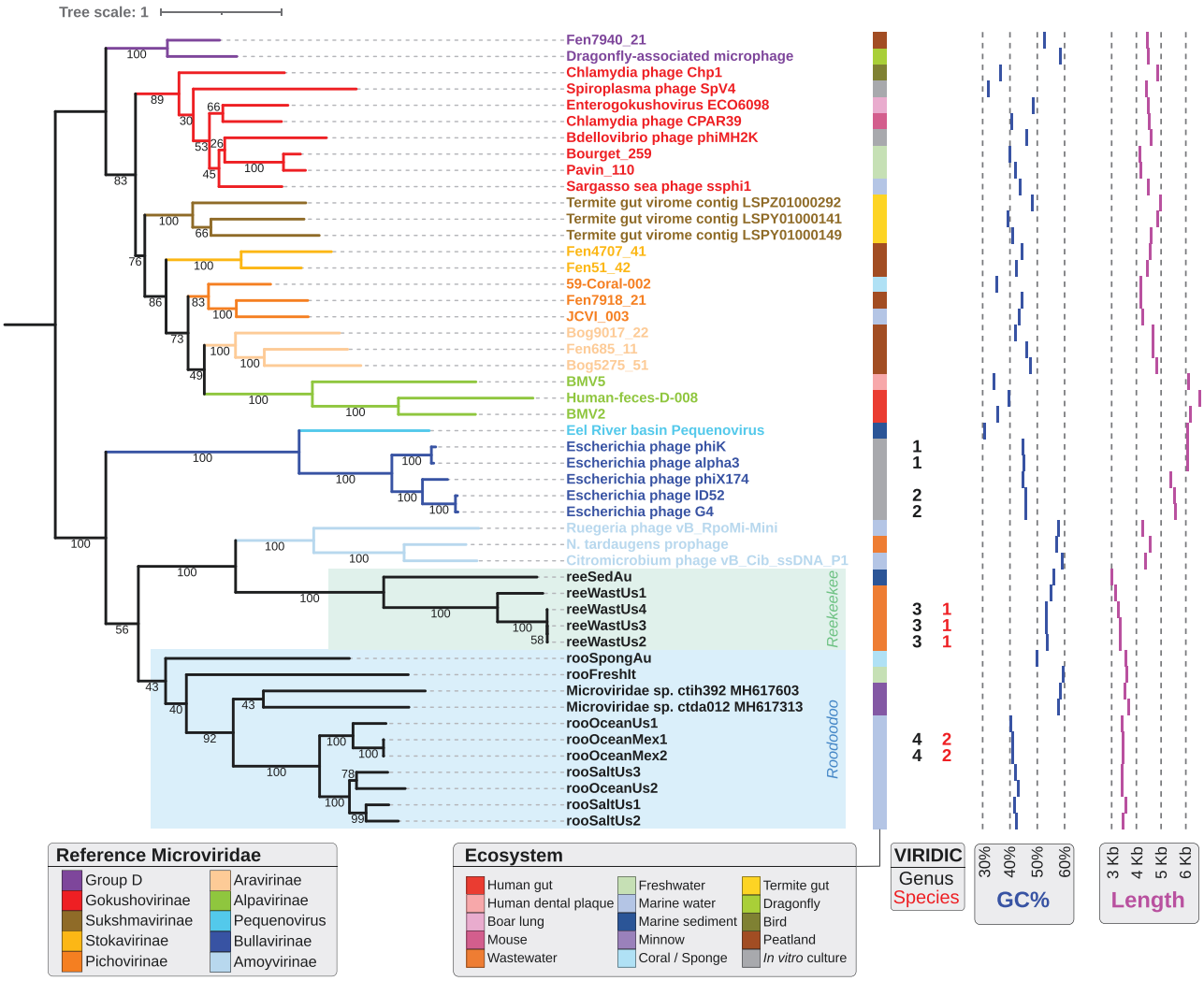


**Figure 1.** Phylogeny of the concatenated major Cap and Rep proteins. Branches were colored according to the reference subfamily, while the two novel groups (reekeekeeviruses and roodoodooviruses) at the bottom of the phylogeny have a colored background. Four tracks were appended next to the phylogeny: (1) sampled ecosystems, (2) VIRIDIC genera and species, (3) GC%, and (4) genome length. Bootstrap values are written in the middle of each internal branch.

a species in the reekeekee clade (Supplementary Fig. S5). Indeed, the VIRIDIC intergenomic similarity was of 33.1 per cent on average between reeWastUs1 and its three closest relatives; these genomes have regions with ~80 per cent nucleotidic identity, and these regions cover ~40 per cent of their total lengths. In the same way, all seven marine-derived genomes of roodoodooviruses were related with a minimum of 14.2 per cent intergenomic similarity and formed two groups of three and four genomes in the phylogeny with a similarity of at least 41.5 per cent between genome pairs inside each group. The guanine-cytosine (GC) content of these genomes was also consistent with the phylogeny (Fig. 1). The five reekeekeeviruses all have a GC content between 53 and 56 per cent. Roodoodooviruses appear to be more diverse, with (1) a genome distant from others that has a GC content of 50 per cent, (2) a subgroup of three genomes with a GC content between 57 and 59 per cent, and (3) the subgroup of the seven marine genomes with 40–43 per cent GC. No dif-motif suggesting a lysogenic lifestyle was found in these genomes, nor hits in bacterial genomes or CRISPR spacers that could have pointed out to their potential host.

The two groups were also different in terms of their core protein length. Indeed, Cap proteins from reekeekeeviruses were between 456 and 499 amino acids (aa) in length (Supplementary Fig. S6), thus almost as large as Cap proteins from gokushoviruses or pichoviruses, whereas roodoodooviruses have smaller Cap proteins (384 to 418 aa), even smaller than those of bullaviruses. Similarly, reekeekeeviruses encode a Rep protein that is comparable in length to that of gokushoviruses (average 288 and 308 aa in length, respectively) and smaller than that of bullaviruses (average 522 aa in length), and roodoodooviruses have even smaller Rep proteins (average 247 aa in length). For the sixteen small genomes, only five protein coding genes were predicted at most using Prodigal, including Cap and Rep (Supplementary Fig. S7A). Even though genomes with as few as two genes are known for eukaryotic viruses (Finsterbusch and Mankertz 2009), microviruses with longer genomes such as phiX174 are known to encode many overprinted genes. As gene prediction software do not predict overprinted and overlapping genes (Vanderhaeghen et al. 2018), a methodology was developed to find all *bona fide* genes in the sixteen genomes.

## 3.4 At least four genes were probably encoded in each of these sixteen genomes but possibly not many more

### 3.4.1 *Characteristics of non-coding and coding ORFs in phiX174*

In order to identify all genes in the sixteen new genomes, including overprinted ones and ones with long overlaps, three different strategies were used. These strategies were first tested on phiX174, as all genes of this microvirus have been determined experimentally.

First, a strong argument for an ORF to be a *bona fide* gene is its conservation among related viruses. Using fifteen genomes that were ranging from nearly identical (>95% intergenomic similarity) to having no nucleotide similarity with phiX174 (0% intergenomic similarity), the ten genes, excluding A* as it is in frame with the Rep gene, of phiX174 were found to be conserved and detected in all microvirus genomes with an intergenomic similarity higher than 25 per cent, five of them being the only ORFs conserved in all genomes with at least 15 per cent similarity (Fig. 2). Only two additional non-coding ORFs were conserved in all >25 per cent similar genomes: one small ORF in a region inside two already overprinted genes (but in the third frame) and one ORF overlapping one of the three conserved domains of the Rep gene.

Second, codons of highly degenerate residues (namely leucine-L, arginine-R and serine-S, each encoded by six different codons) represent 18 out of the 61 possible codons, thus accounting for 29.5 per cent of codons if chosen randomly in genomes with a 50-percent GC content. In contrast, these three residues (LRS) represent only 15 per cent of the existing twenty amino acids, even though they are abundant in proteins with a total of 20.07 per cent in known viruses, leucine being the most used residue (8.28%). This apparent clear signal for discriminating coding (∼20% of LRS) from non-coding ORFs (∼30% of LRS) may not fully apply to overprinted genes that are enriched in highly degenerated residues, as these allow more synonymous mutations and flexibility for overprinted regions (Pavesi, Magiorkinis, and Karlin 2013). Yet, the usage of LRS residues in overprinted genes is still lower than that in non-coding ORFs as it was described to be 23 per cent on average in known overprinted viral genes (Pavesi, Magiorkinis, and Karlin 2013). Thus, the proportion of highly degenerate residues (L, R, and S) was computed for each ORF of phiX174, as well as the expected proportion of their codons, depending on the GC content of the genome. The eight ORFs that exhibited an LRS proportion lower than expected corresponded to seven *bona fide* genes and a non-coding ORF, here again overlapping a Rep conserved motif. The remaining three true genes (K, C, and E) of phiX174 had much higher LRS proportion (35.9 on average) and corresponded to overprinted genes in three different regions.

In addition to the sequence conservation and composition in LRS, the variability (*i.e.* SNPs) observed for each ORF was also used to further help discriminate between coding and non-coding ORFs. Different software exists to perform similar tasks such as PhyloCSF (Lin, Jungreis, and Kellis 2011) or RNAcode (Washietl et al. 2011), and these methods are based on the analysis of a multiple alignment of at least ten similar genomes. Unfortunately, the sixteen genomes analyzed here do not have similar genomes at the nucleotidic level in any database, and the VIRIDIC analysis showed that these sixteen genomes are not even similar to each other at the nucleotidic level. Yet, high levels of microdiversity were shown for some viral species in a given environment and even in a given metagenome (Martinez-Hernandez et al. 2017; Gregory et al. 2019). Thus, metagenome reads and contigs, even those representing only partial genomes, carry some microdiversity that could be used for detecting SNPs and then distinguish coding from non-coding ORFs. As phiX174's genome was not assembled from metagenomic data, its SNPs were identified using twenty-five genomes from GenBank that are very similar to the phiX174 genome (at least 95% genome identity; Supplementary Table S4). All non-coding ORFs had a number of negative SNPs higher than the number of synonymous ones. This result is expected, as negative SNPs are more frequent randomly than synonymous ones (Supplementary Fig. S8A). Indeed, 168 negative and 138 synonymous SNPs are found when considering all codons and their nine possible SNPs. Conversely, the ten validated genes had more synonymous than negative SNPs. This result reflects a purifying selection on these coding regions at the codon level, with the selective removal of amino acid substitutions that are deleterious. Considering positive SNPs that are the most frequent randomly ($n = 174$), most genes here again favored synonymous mutations over them. Yet, three phiX174 genes had more positive mutations than synonymous ones, the overprinted and/or overlapping genes (B, E, and K) ($n = 14.3$ *vs* $n = 7.6$) (Supplementary Fig. S8A–C). For these ORFs, the high rate of non-synonymous changes, similarly to overlapping genes of RNA viruses (Hughes et al. 2001; Narechania, Terai, and Burk 2005), is likely due to concurrent dominance of synonymous substitutions in the alternative frame. Yet, these non-synonymous changes were mostly positive in overprinted and overlapping genes of phiX174. The pN/pS ratio was then calculated for each of the protein coding and non-coding ORFs of phiX174. As expected, all coding and non-overlapping ORFs had a pN/pS value under 1, all non-coding ORFs above 1, and overprinted genes B, E, and K had values over 2.5 due to their enrichment in positive mutations. Since positive mutations are frequent in non-coding ORFs as well as in overlapping genes, these mutations were excluded when calculating the proportion of non-synonymous mutations (pN). Considering these modified pN/pS ratios, genes without overlapping had values lower than 1, overprinted genes lower than 1.37, and all non-coding ORFs greater than 1.42 (Supplementary Fig. S8D, E).

Last, as each ORF being the longest stretch between a start codon and a stop codon, the true beginning of the corresponding gene can often be another start codon inside the ORF. Thus, mapping results were used to refine the identification of the correct start position of each gene, falsely predicted regions being enriched by negative substitutions over synonymous ones. In addition, a careful inspection of the conservation of the ORF beginnings within MSAs and prediction of ribosomal binding sites using Prodigal (Hyatt et al. 2010) were done to identify these start codons. Accordingly, for the six phiX174 ORFs that represent genes but with an incorrect start position, this information allowed us to find back the right start codon further down on the genome, corresponding to the experimentally defined coordinates as reported in GenBank.

### 3.4.2 *Using conservation, pN/pS, and LRS%, sixty-nine ORFs were identified as coding*

The 478 putative ORFs determined in the 16 genomes clustered into 45 OGs and 307 singletons. The two largest OGs, conserved in all sixteen genomes, corresponded to Cap and Rep proteins. Only two additional OGs contained more than four ORFs and included ORFs of similar length in ten of the eleven roodoodooviruses (OG11 and OG6 in Fig. 4). These four OGs conserved in distant microviruses were considered as coding ORFs. Furthermore, the

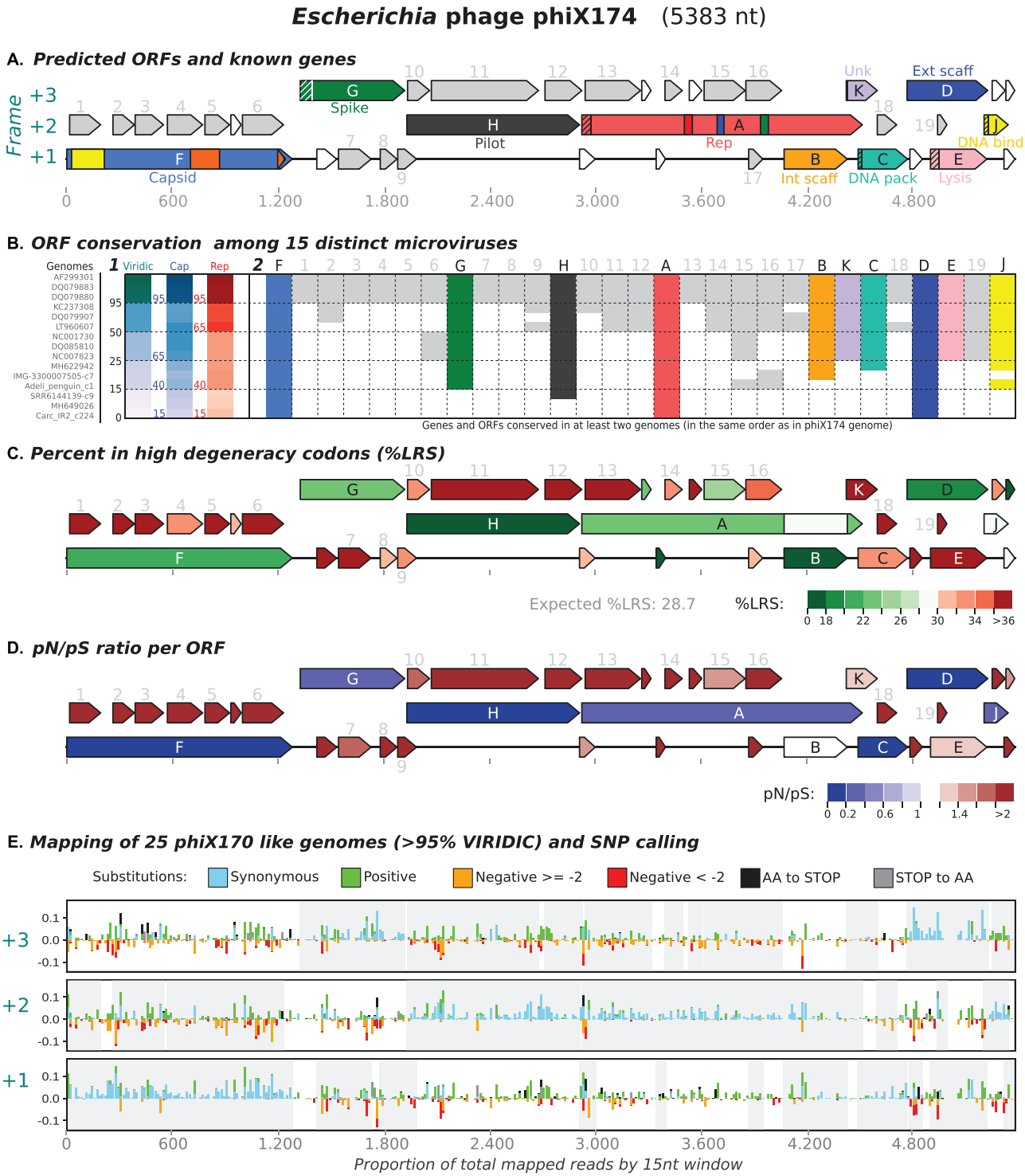


**Figure 2.** Genomic map of phiX174 (5383 nt) and characteristic of its genes and non-coding ORFs. (A) Genomic map with known genes colored (the three conserved motifs in both Cap and Rep are indicated), and non-coding ORFs in grey if present in other microviruses (otherwise in white). 5′-ends of ORFs were cross-hatched when a false START codon was predicted. (B) Conservation of phiX174 ORFs/genes in fifteen microvirus genomes: (1) on the left, the distance between phiX174 and each genome (one line per genome) using VIRIDIC intergenomic distance and amino acid identity for Cap and Rep, and (2) on the right, phiX174 genes/ORFs displayed as columns in the order of the genome and the fifteen genomes as lines, cells being colored if the genes/ORFs are present in the corresponding genomes. (C) For each ORF/gene, the proportion of the L, R, and S amino acids is computed. The expected random proportion of LRS considering the phiX174 GC content is indicated and colored in white, LRS% values lower than this value are in green, and those greater in red. (D) Considering the SNPs detected (see below), ORFs are colored according to pN/pS values, from blue (<1) to red (>2). (E) For all ORFS, SNPs were detected among GenBank genomes very similar to phiX174 and indicated in blue if they lead to synonymous mutations, and green, orange, and red if they lead to positive, negative, and strongly negative mutations, respectively, according to the BLOSUM62 matrix. A bar chart displaying the abundance of each type of SNPs is plotted, with grey backgrounds delimiting the predicted ORFs.

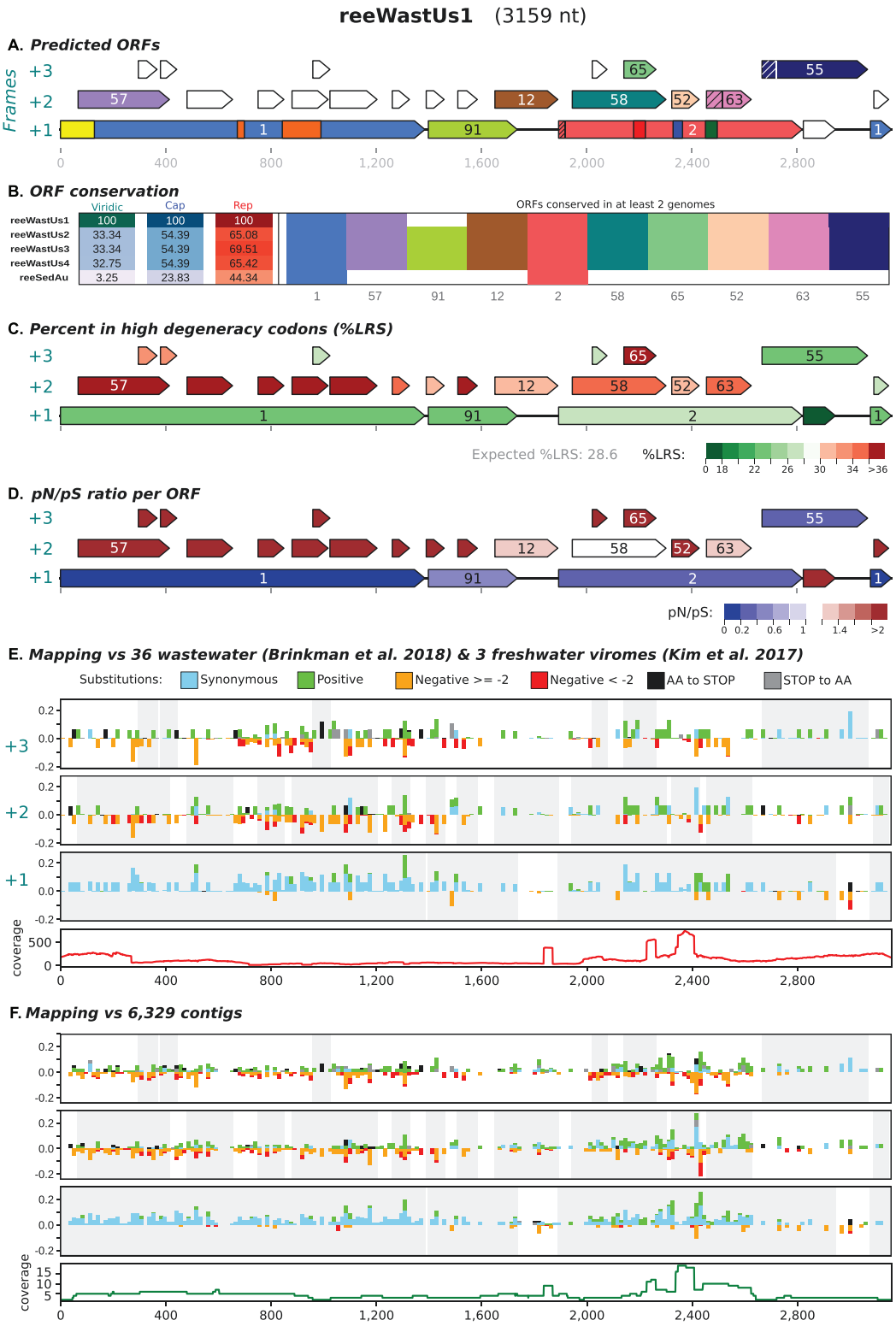


**Figure 3.** Genomic maps of rooOceanUs1 (3408 nt) genome of each clade. (A) Genomic map with all predicted ORFs, in color if conserved in other microviruses. 5′-ends of ORFs were cross-hatched when a false START codon was predicted. (B) Conservation of ORFs: (1) on the left, the distance between the studied genome and all other similar genome (one line per genome) using VIRIDIC intergenomic distance and amino acid identity percentage for Cap and Rep; and (2) on the right, ORFs are displayed as columns in the order of the genome. (C) For each gene, the LRS% is computed and displayed. (D) Considering the SNPs detected (see below), ORFs are colored according to pN/pS values. (E) For all ORFs, SNPs were detected using virome reads and (F) using a set of similar contigs. A bar chart displaying the abundance of each type of SNPs is plotted for virome reads and contigs, with grey backgrounds delimiting the predicted ORFs. See Fig. 2 legend for additional details.

overall pN/pS values of these OGs were all below 0.25, each having more than 300 SNPs detected in metagenome reads or contigs, indicating here again that they correspond to genes under purifying selection. Although no ORF from the rooSpongAu genome was initially clustered in OG11, a small motif of OG11 was retrieved in one of its ORFs (residues 330 to 431; 32 on

*hhsearch* bit-score), and this ORF is of similar length and position as the other ones from OG11. Additionally, even if no significant similarities were detected against OG6 in the rooFreshIt genome, one ORF of concordant length and position was considered as coding as it presented a similar amino acid composition enriched in basic residues at its N-terminus and was supported by a low pN/pS value (0.24). Among the fifty-four ORFs in these four OGs, only four had a pN/pS value greater than 1: two having only one and ten SNPs respectively, and the other two being from the only genome (rooSpongAu) for which the SNPs and pN/pS were visibly not coherent and thus not further considered.

For 11 of the 16 genomes that have at least 1 closely related genome (>25% intergenomic similarity), 208 of the 330 ORFs were not conserved and were not considered further, as all genuine phiX174 genes are conserved in such closely related genomes. Among these 208 discarded ORFs (average pN/pS of 27.2), only 22 had a pN/pS value lower than 1.4, from which only one had more than 21 SNPs detected.

Sequence conservation results allowed us to retain 54 ORFs and exclude 208. For the remaining 216 ORFs for which conservation information was not sufficient, the pN/pS values were then considered. Out of these, 118 ORFs had enough detected SNPs, especially 80 ORFs that had more than 10 SNPs or were in OGs with more than 10 detected SNPs. Thirteen ORFs belonging to four OGs had pN/pS values lower than 1.5 and represent likely coding sequences (OG55, OG63, OG91, and OG103, see Fig. 3 and Fig. 4). Yet, OG103 was made of three quite different ORFs; the end of one ORF was only similar on fifteen residues to the beginning of the other two ORFs and was then considered as not conserved and discarded. OG91 was additionally found in reeWastUs1 at a concordant position on the genome by BLASTp (30 on bit-score *vs* reeWastUs4 OG91 protein).

Finally, the remaining ninety-eight ORFs were manually scrutinized considering their conservation, position on the genome, overlap with a conserved region of the SJR structure of Cap or of the three motifs of Rep, and their LRS%, and none were considered as coding ORFs.

The ORFs identified as coding (seven OGs and sixty-six ORFs) in these sixteen genomes were used to identify three additional coding ORFs in the reeSedAu genome that had no closely related genome nor usable pN/pS results. Even if no gene of reeSedAu was similar to the three OGs shared by the four other reekeekeeviruses, coding sequences were most likely present at similar loci in this genome, *i.e.* OG91 located between Cap and Rep, OG55 at the end of the genome, and OG63 overprinted at the end of Rep.

Ultimately, using SNP results, residue conservation at the beginning of MSAs, and ribosome-binding site prediction, the start codon of each coding ORF was then manually checked and was modified for sixteen ORFs out of the sixty-nine ORFs freshly defined as coding. For reeWastUs1, three ORFs were refined; for example, the retained start of the ORF in OG63 was defined twenty residues further on the genome because this region was not conserved, had a high pN/pS value in reeWastUs1 (1.43 vs 0.875 with the final residues), and overlapped one of the conserved motifs of Rep protein.

## 3.5 Genomes of the two new clades have conserved gene content

Five coding ORFs were inferred in all genomes of reekeekeeviruses and four in roodoodooviruses (Fig. 5), and except for Cap and Rep protein genes, no gene is shared between the two clades (Fig. 5). Yet, within the two clades, both gene content and order are conserved. For reekeekeeviruses, besides Cap and Rep proteins, three other OGs (55, 63, and 91; Fig. 3) were conserved in four of the five genomes and had low pN/pS (0.43, 1.02, and 0.72 on average, respectively) and low LRS% values (21.3% and 25% on average for OG55 and OG91, respectively), except for OG63 that had a much higher LRS% value likely because of its overprinted nature (36.8%). These three OGs likely exist in the more distant reeSedAu genome as coding sequences were identified at similar positions in this genome. For roodoodooviruses, strong sequence conservation, low pN/pS, and LRS% allowed to undoubtedly infer OGs 6 and 11 as coding in addition to Cap and Rep proteins, with OG6 having a long overlap with OG11. Their genome architecture is different from the reekeekeevirus ones, as (1) Cap and Rep genes are only separated by a short region (between 32 and 117 bp), in which no ORFs were predicted as coding, (2) Rep genes are closely followed by a large gene (OG11, Fig. 5), and (3) one last gene (OG6) is conserved in all eleven roodoodoovirus genomes, overlapping or not with OG11. Thus, as GC contents, genome lengths, and Cap sizes, genome architectures of these sixteen genomes here again advocate for a separation into two homogeneous yet diverse groups. It has to be noted here that even though the reekeekeevirus clade is more closely related to the amoyviruses in the phylogeny, the only OG besides Rep and Cap from the sixteen genomes conserved in another microvirus group is OG11 that is found in roodoodooviruses and in amoyviruses (Supplementary Fig. S9). The major differences between roodoodooviruses and the amoyvirus vB_RpoMi-Mini genomes are the presence of a peptidase encoding gene between Cap and Rep and a larger Rep gene (Fig. 5).

## 3.6 Our strategy helped predict overprinted genes and overlapping regions

The gene prediction tool Prodigal (Hyatt et al. 2010), frequently used for phage genomes, missed five of the eleven genes of OG6 (colored in orange) in roodoodooviruses and three of the four genes of OG55 (dark blue) in reekeekeeviruses (Supplementary Fig. S7A). This result is not surprising as the beginnings of these genes are overlapping with the preceding gene on the genome. When using PHANOTATE (McNair et al. 2019; Supplementary Fig. S7B), a tool specifically designed for phages and better able to handle genomes with high coding density, OG6 genes were found in nine of the eleven roodoodooviruses yet restricted to their non-overlapping part. Instead of the overlapping OG55 genes, four genes in another reading frame were preferred by PHANOTATE, all these having high pN/pS values pointing to a false prediction. Rep was missing in one genome and was replaced by four genes in the reverse strand. Finally, the OG63 overprinted in Rep in reekeekeeviruses was not predicted by any *in silico* automated software. Overall, PHANOTATE predicted small genes in all intergenic regions longer than 100 nt, leading to a total of eighty genes instead of the sixty-nine genes. The five overprinted genes were not detected by automated software and were only recovered using our custom approach.

## 3.7 Tentative annotation of each conserved gene

The custom gene prediction strategy resulted in the identification of seven proteins conserved in all or part of the sixteen genomes and three singletons found in the smallest genome reeSedAu. Among these proteins, only Rep and Cap were similar to any Pfam domains and to proteins of known function encoded by members of the *Microviridae* family. For all these OGs, especially the ones for which the function remained unknown, their secondary structure was tentatively predicted using ColabFold (Mirdita et al. 2022) and

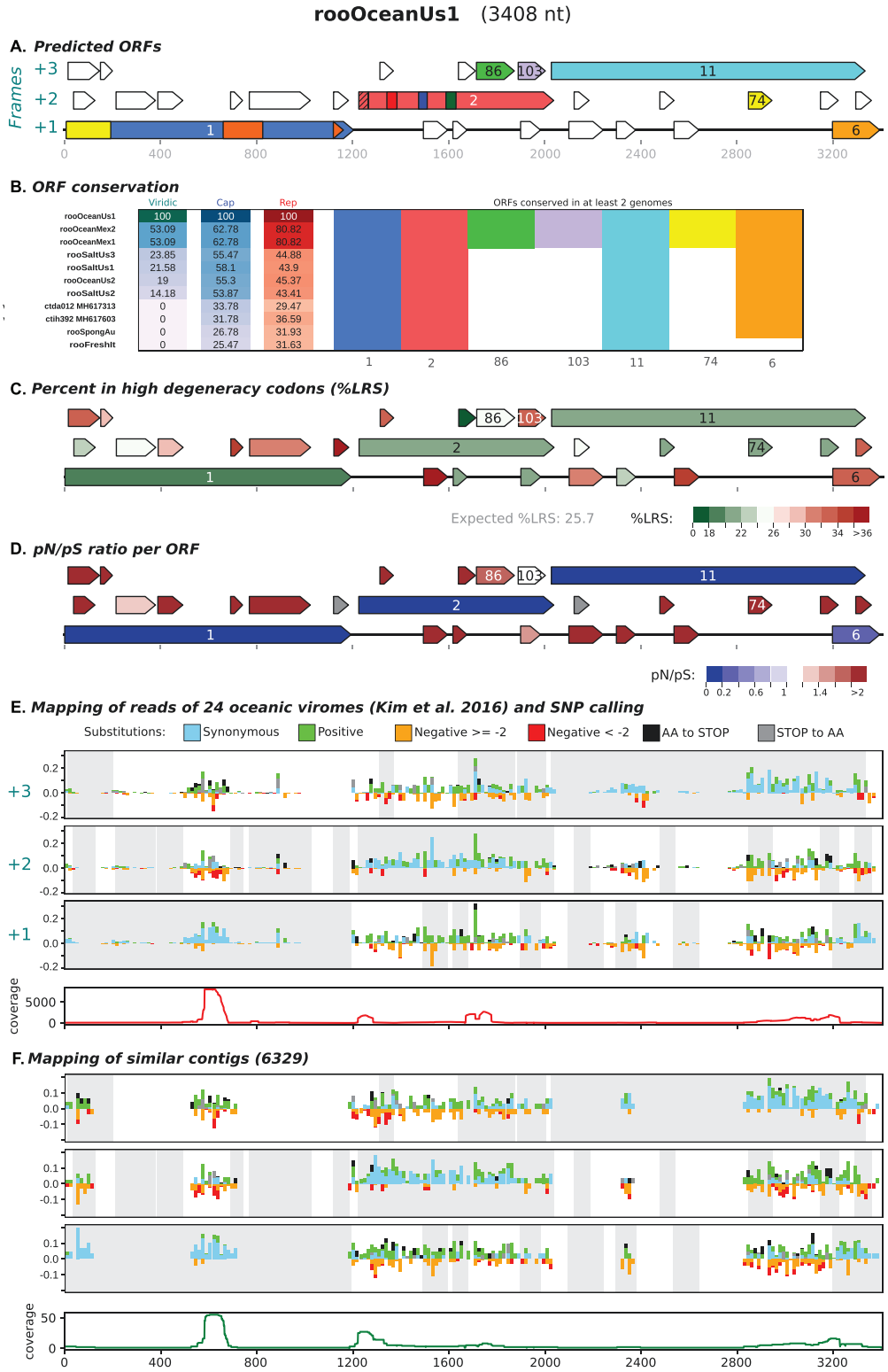


**Figure 4.** Genomic maps of reeWastUs1 (3159 nt) genome of each clade. (A) Genomic map with all predicted ORFs, in color if conserved in other microviruses. 5′-ends of ORFs were cross-hatched when a false START codon was predicted. (B) Conservation of ORFs: (1) on the left, the distance between the studied genome and all other similar genome (one line per genome) using VIRIDIC intergenomic distance and amino acid identity percentage for Cap and Rep; and (2) on the right, ORFs are displayed as columns in the order of the genome. (C) For each gene, the LRS% is computed and displayed. (D) Considering the SNPs detected (see below), ORFs are colored according to pN/pS values. (E) For all ORFs, SNPs were detected using virome reads and (F) using a set of similar contigs. A bar chart displaying the abundance of each type of SNPs is plotted for virome reads and contigs, with grey backgrounds delimiting the predicted ORFs. See Fig. 2 legend for additional details.

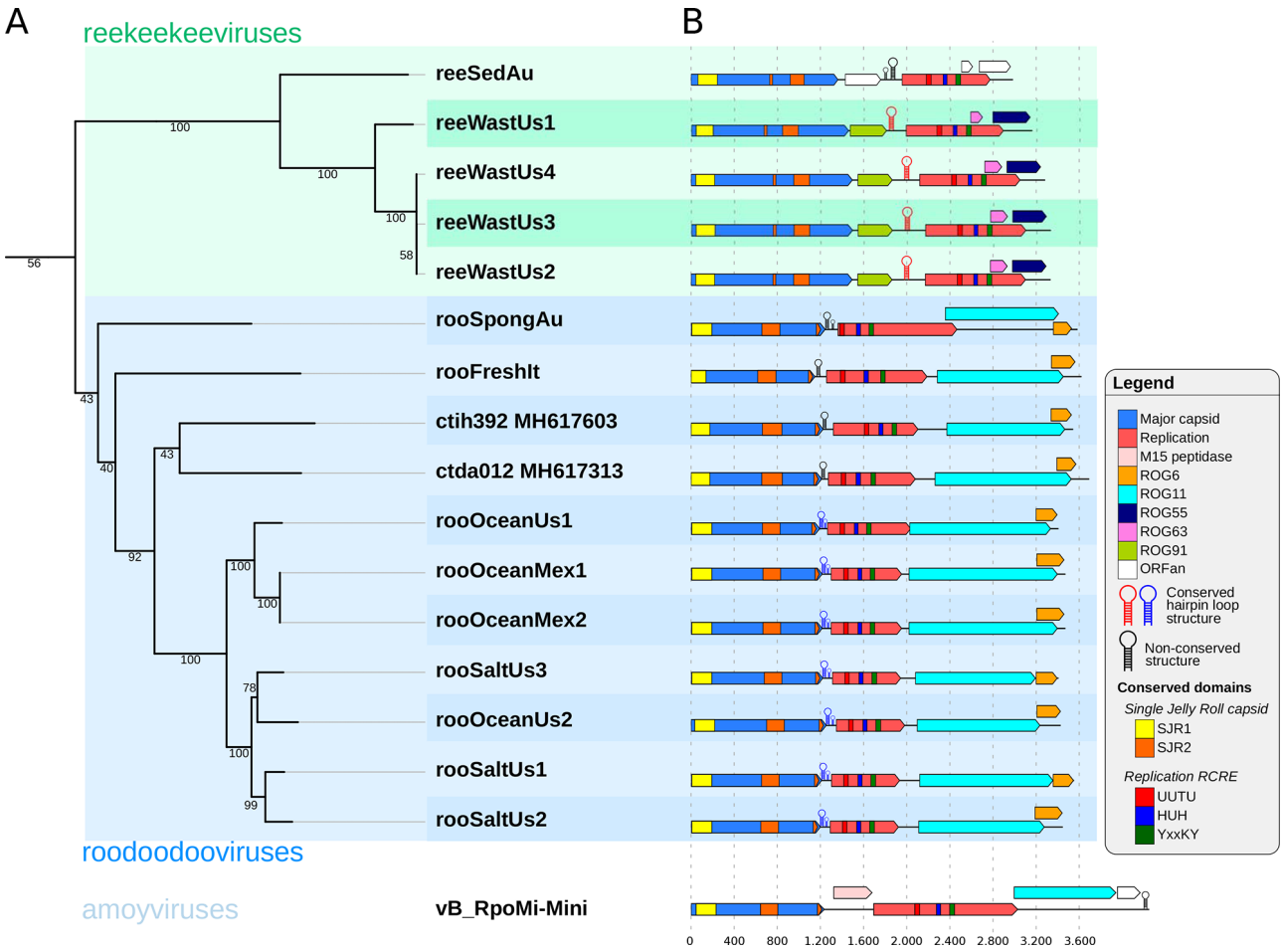


**Figure 5.** Final genomic maps of reekeekee- and roodoodooviruses. (A) Phylogeny on the concatenated major Cap and Rep proteins with 100 bootstrap replicates. (B) Curated genomic maps of the sixteen small microviruses discovered here and Prodigal prediction of amoyvirus vB_RpoMi-Mini, the smallest DNA phage. When predicted, hairpin loops are colored by sequence conservation; blue and red hairpins are conserved among marine roodoodooviruses and wastewater reekeekeeviruses, respectively, while black hairpins are predictions with no sequence similarity between viruses.

then compared to the known structures of phiX174 proteins (Protein Data Bank ids: 2BPA, 1CD3, 4JPP; McKenna et al. 1992; Dokland et al. 1999; Berman et al. 2000; Sun et al. 2014).

As previously described, the Cap proteins of the five reekeekeeviruses were so distantly related to the one of the known microviruses that they were difficult to identify even using remote homology tools. The structure prediction of these five Cap proteins allowed us to better understand how these Cap proteins differed from known Cap proteins. Indeed, even though the presence of the SJR domain was clearly identified (61.6 for the local distance difference test (LDDT) on average), no structural prediction could be found for any other part of the sequence (34.1 LDDT). This structural prediction allowed us to identify the regions of the Cap proteins encoding this SJR in reekeekeeviruses and other microviruses, highlighting that the last beta strand of the SJR is encoded by a different part of the protein (Supplementary Fig. S10).

For the rest of the five reekeekeevirus proteins, the Rep and OGs 91 and 55 (depicted in green and dark blue) of the four wastewater genomes had structural predictions similar to Rep proteins and to the two singleton proteins encoded by genes at similar positions in reeSedAu genome (short alpha helices with unknown overall conformation). The last gene of this clade, OG63, overprinted in Rep, had two different structural predictions in wastewater and reeSedAu genomes, as three beta strands were confidently predicted for the four reeWastUs genomes (78.3 LDDT) and only a short alpha helix was predicted for the gene at a similar position in reeSedAu (84.6 LDDT) (Fig. 6).

For roodoodooviruses, the Cap proteins are the smallest among microviruses with a confident structure prediction similar to an SJR structure spanning almost the entire length of the protein (Fig. 6). Important structural components of the Rep protein were also identified for these viruses. For the OG6 proteins conserved in all roodoodooviruses, only a small alpha helix was predicted. Because the amino acid composition and the mostly unfolded nature of this protein are coherent with that of phiX174's DNA-binding protein, with an N-terminus enriched in basic amino acids and the C-terminus presenting mostly hydrophobic and aromatic residues, this OG6 protein was tentatively annotated as a DNA-binding protein (Supplementary Fig. S11). For the OG11 protein conserved in all roodoodooviruses, both alpha helices and short beta strands were predicted.

## 3.8 DNA secondary structures

In phiX174, the primosome, *i.e.* the protein complex containing a primase and helicase from the bacterial host, initiator of the complementary negative-strand synthesis, recognizes a hairpin region ahead of the spike coding gene G (Shlomai and Kornberg 1980).

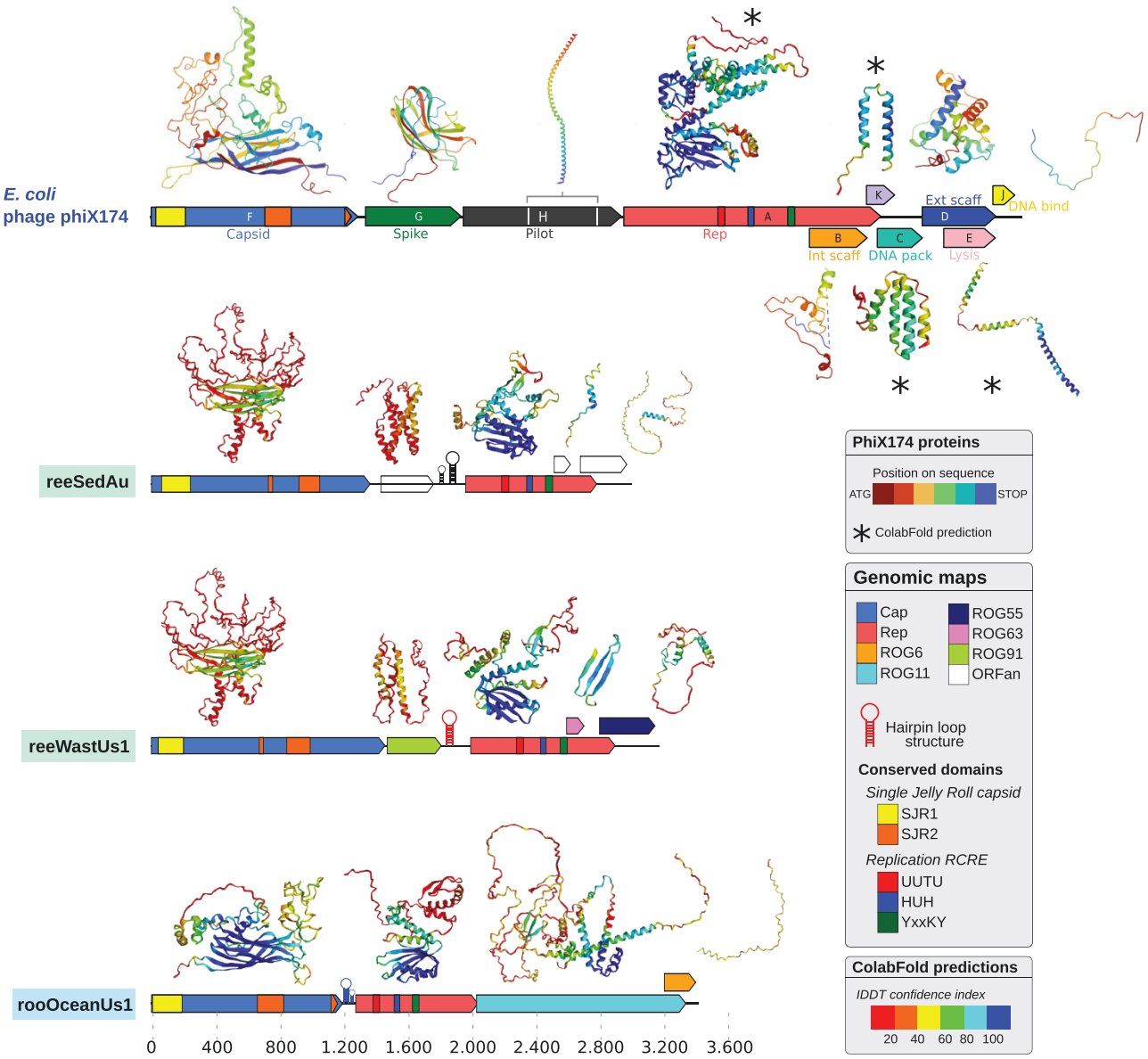


**Figure 6.** Protein structure prediction for the final inferred genes. PhiX174 is on top with structures retrieved from the Protein Data Bank (PDB): Cap, spike, and DNA-binding proteins were retrieved from the PDB entry 2BPA, while scaffolding proteins are solved in 1CD3, and the pilot protein is described in entry 4JPP. When not available, a prediction was run with ColabFold and marked with a star on the figure. Then, one representative genome was chosen for reekeekee- and roodoodooviruses, alongside the smallest assembled genome, to predict the structure of each of the proteins from the final curated genomes. ColabFold predictions are colored based on the LDDT confidence index, while PDB structures are colored according to the N- and C-termini of the sequences. Hairpin loop structures are displayed using the same colors as in Fig. 5.

This structure is probably necessary for any microvirus, as complementary strand synthesis is the beginning of their Rep cycle (Cherwa and Fane 2011). We consequently attempted to find hairpin-like structures in all intergenic regions and unveiled two different patterns among the two new clades. RNAfold predicted two significant hairpin structures between the Cap and the Rep proteins for roodoodooviruses (20–22 and 13–15 nt, average entropy 0.16) and a long hairpin in reekeekeeviruses between OG91 (the green gene) and Rep (64 nt, entropy 0.05) (Supplementary Fig. S12). These predictions were supported by sequence conservation at the nucleotide level among marine-derived roodoodooviruses and wastewater-derived reekeekeeviruses and corresponded to unusual amino acid sequence conservation in these locations, with the random presence of stop codons. Hairpin-like structures are more variable in marine-derived roodoodoovirus, both in the stem and the loop. Indeed, on the one hand, loop and stem size present a 1-nt standard deviation, accepting insertions, deletions, and substitutions (Supplementary Fig. S12A). On the other hand, although reeWastUs1 is only 33 per cent similar by VIRIDIC to the other wastewater representatives, the loop sequence is identical in all four genomes, while stem sequences are 72 per cent identical (Supplementary Fig. S12B). Notably, in both clades, compensatory substitutions were identified across both arms, further advocating a conservation of this structure. As stem-loop structures with different functions exist in phiX174, the involvement of the detected hairpin-like structures in the initiation of negative-strand synthesis cannot be certified.

### 3.9 Quantification of these new clades in the environment

The abundance of these two new clades was evaluated in the fifteen data sets in which the sixteen genomes were assembled, as well as in thirty-five additional viromes described in the same publications. For each metagenome, reads were compared using BLASTx to an in-house database of 13,390 *Microviridae* genomes that included the sixteen genomes. Unsurprisingly, the proportion of microviruses in viromes (Supplementary Fig. S13 right) was found to be highly variable (between 0.01 and 49.91%). Among them, *Gokushovirinae* and *Alpavirinae* were almost always the most abundant groups (Supplementary Fig. S13, left), followed by *Pichovirinae* and Group D. Even among the viromes from which they were assembled and despite the use of multiple displacement amplification in seven out of fifteen data sets, the reekeekee- and roodoodooviruses represent a small fraction of all microviruses (only up to $2 \times 10^{-4}$ and 13%, respectively). Although only a handful of reads (between 100 and 2,000) allowed the assembly of these short genomes, as many as 100 K reads that had roodoodooviruses as closest neighbors were identified in other viromes in which no roodoodoovirus genome was collected.

## 4. Discussion

### 4.1 These circular contigs likely represent existing complete genomes

The results of the analyses we conducted all led to the conclusion that these sixteen contigs represent complete genomes. Indeed, these circular contigs were found in data sets from diverse ecosystems, obtained through different experimental protocols, sequencing technologies, and assembly software. Despite these diverse origins, their genome length, gene content, and phylogeny all led to coherent conclusions. As these small contigs are not similar to larger ones, which could have indicated an assembly artifact, these results strongly suggest that these circular contigs represent complete microvirus genomes. Furthermore, roodoodoovirus genomes exhibit properties similar to the smallest genome of a cultivated microvirus, the amoyvirus vB_RpoMi-Mini. These genomes are significantly smaller than the shortest cultivated DNA phage (vB_RpoMi-Mini, 4.248 kb) and RNA phages (phage M, 3.405 kb) that belong to the Leviviricetes class. It seems likely that shorter RNA phage genomes exist. Yet, as these genomes are linear, the presence of direct terminal repeats (DTRs) is necessary to prove the completeness of genomes assembled from environmental samples. Considering the 44,779 RNA phage contigs recently assembled from oceanic metatranscriptomes (Zayed et al. 2022), DTRs were detected in only eleven contigs, and only one of these contigs was smaller than 3.3 kb, with a size of 2.434 kb.

### 4.2 Microdiversity used for the first time to help identify all genes

Although viruses that only encode a Rep and a Cap gene in their genome are known for eukaryotic viruses (Finsterbusch and Mankertz 2009), microviruses with longer genomes such as phiX174 are known to encode several overprinted genes or genes with large overlapping regions. As these last genes are not predicted by gene prediction software, we developed a methodology based on protein conservation, composition, and mutation. Protein conservation alone did not prove discriminative enough to separate all genes from non-coding ORFs. Indeed, non-coding ORFs are sometimes as conserved as genes; for example, non-coding ORFs found in DNA regions conserved because of the presence of secondary DNA structures, or ones overlapping highly conserved gene regions such as Rep motifs. Concerning the protein composition, even though the LRS% is a strong measure to discriminate genes with no overlap from non-coding ORFs, some genes with long overlaps proved to have LRS%, sometimes greater than non-coding ORFs. Finally, the most discriminative feature between coding and non-coding regions was the effective selective constraints. Although such methods exist, such as RNAcode and PhyloCSF, they are based on the identification of homologous regions in other species and the computation of MSAs. Here, our method based on SNP detection was more suited for metagenomic contigs, as (1) very few or no closely related genomes exist, thus impeding the construction of a MSA, and (2) viral population often exhibit a non-negligible level of microdiversity—this microdiversity being present in virome reads. Using virome reads allowed us to get access to enough SNPs to evaluate selective constraints on each ORF, even when no related genome is available (see rooFreshIt on Supplementary Fig. S14). In addition, tools like Prodigal and PHANOTATE understandably strongly penalize gene overlaps, and considering ORFs allowed us to better identify the beginning of overlapping genes. Considering the huge amount of sequence data already available and the ever growing size of metagenomes, this new methodology could be useful to predict genes more accurately, a crucial step on which genome analysis depends.

### 4.3 Strategies to accommodate a small genome

Our initial hypothesis was that a smaller genome length was coupled with, or even made possible, by an increase in overprinting. The results obtained here point toward a different scenario. Indeed, although genes with significant overlaps and even overprinted genes were detected, their number is lower than that in phiX174. Furthermore, the core gene lengths of these new genomes are small compared to the ones of known microviruses. Indeed, Cap proteins from roodoodooviruses are the smallest ones among all microviruses and reekeekeevirus ones are also smaller than the ones of most microvirus subfamilies. Rep proteins from these two groups (288 and 247 residues on average) are also among the smallest within microviruses, with a length comparable to the ones of gokushoviruses (308 residues). These Rep proteins are much smaller than bullavirus ones (522 residues)—both ends surrounding the conserved region that contains the three essential motifs being shorter. For these small Rep proteins, only reekeekeevirus ones presented an overprinted gene within Rep C-terminal end, coherently with the known overprinted gene encoded in this less conserved region of Rep in phiX174 (Pavesi, Magiorkinis, and Karlin 2013).

Regarding the number of genes, only five protein coding genes were predicted at most for all sixteen genomes described here, including Cap and Rep, representing a twofold reduction in comparison with phiX174. Even though our gene prediction might have missed some genes, it is likely that these phages with small genomes have both a smaller number of genes and smaller genes compared to other microviruses, rather than a greater number of overprinted genes. Accordingly, the small size of Rep and Cap proteins in these microviral genomes is perhaps due to the fact that only the essential (and less flexible) regions of these proteins are present, limiting gene overlap and overprinting. These microviruses are thus limited to Rep and Cap proteins, and only two and three additional genes remained unannotated in roodoodoo- and reekeekeeviruses, respectively, without any sequence similarity to a known protein family. One of the two unannotated proteins (OG6) in roodoodoovirus had characteristics

similar to DNA-binding protein, and the other unannotated proteins are likely DNA-binding, DNA-packaging, lysis, spike, pilot, or scaffolding proteins. Concerning these last proteins, maybe these viruses manage to assemble without any scaffolding proteins as in most viruses with a $T = 1$ icosahedral Cap architecture, even though bullaviruses have one internal and one external scaffolding protein and gokushoviruses have an identified scaffolding protein. Although crucial for host cell penetration in isolated microviruses, the DNA pilot protein was not found here—either its sequence is too divergent to be detectable, or a potentially different DNA delivery mechanism is at play. Additional experimental work have to be performed in order to decipher the function of these remaining genes.

## 4.4 Additional spike protein or capsid-encoded protrusion?

Interestingly, major Cap proteins of reekeekeeviruses were composed of between 456 and 499 aa (Supplementary Fig. S6) and are thus significantly larger than bullavirus Cap proteins and almost as large as those of gokushoviruses and pichoviruses. Cap proteins from these last two subfamilies contain an insertion loop that forms 'mushroom-like' surface protrusion at each icosahedral threefold axis (Chipman et al. 1998; Roux et al. 2012), while bullaviruses such as phiX174 are decorated by an additional spike protein. Accordingly, the MSA of Cap proteins reveal that the region encoding this protrusion is only conserved in gokushoviruses and their closely related subfamilies (green-colored zone, Supplementary Fig. S10). Thus, the Cap proteins of reekeekee- or roodoodooviruses are not decorated by protrusions homologous to the one of the gokushoviruses. Roodoodooviruses are seemingly not decorated by a different protrusion in their Cap protein as (1) they encode the smaller Cap proteins among microviruses; (2) the residues of their Cap proteins are well aligned with phiX174 ones—for example 358 residues of the 393 of rooOceanUs1 Cap are aligned to one of the 427 residues of phiX174 Cap; and (3) no region longer than ten residues in their Cap proteins is absent of phiX174 Cap. Despite their small protrusion-less Cap proteins, roodoodooviruses have genomes larger than reekeekeeviruses, suggesting additional coding regions and possibly the presence of a spike protein.

Concerning reekeekeeviruses, their longer Cap proteins unveil a new type of revamped microviral Cap. Indeed, the SJR beta strands are encoded by the regions of the Cap proteins different from the other microviruses and their Cap proteins include several large insertions compared to other microviruses (Supplementary Fig. S10). Thus, the decoration of reekeekeevirus Caps remains elusive. Although it was impossible to determine their Cap protein structure, the major changes in their Cap sequence suggest profound changes in their Cap shape and internal size.

## 4.5 *Reekeekeevirinae* and *Roodoodoovirinae*: two related yet diverse and different clades

The analyses conducted on these genomes, whether phylogeny or gene content, show that the sixteen genomes form two different groups. Although these two groups only shared the Rep and Cap genes, they are related in the microvirus tree and form a monophyletic group with *amoyviruses*. Even though reekeekeeviruses are more closely related to the *amoyviruses* in phylogenies, roodoodooviruses have a gene content and gene size more similar to the ones of *amoyviruses*. Indeed, they both share three long genes, namely Rep, Cap, and OG11, which represent a large part of their respective genomes. Furthermore, roodoodooviruses and *amoyviruses* have Cap proteins of similar size that align well (Supplementary Fig. S10). Thus, despite their monophyly and clear separation from known microviruses, roodoodooviruses are related to *amoyviruses*, a group to which the smallest cultured genome belongs. Despite their monophyly with amoyviruses, reekeekeeviruses are distantly related to these last viruses based on the branch lengths on the Cap and Rep protein phylogenies. Reekeekeeviruses stand out as a group distant from the other *Microviridae* groups, as they only share Rep and Cap genes with the other microviruses, they have major differences in their Cap proteins, and their genomes are significantly smaller than all other ones.

Evolutionary reconstruction is challenging for these two groups. Indeed, they could originate from reduction and simplification of ancestral microviruses with large genomes or they could be contemporary versions of ancestral microviruses with small genomes. Although we cannot conclude about their evolutionary history, it is interesting to note that these two separate and different groups have similar trends in the genome architecture. As discussed previously, both clades were found to encode a small number of genes, small versions of essential genes, and reduced number of overlapping genes. This suggests that these characteristics, likely acquired independently, are important for microviruses to have a small genome.

## 4.6 Potential hosts of these viruses

Since no CRISPR spacers similar to these sixteen phages were found, nor highly similar prophages, potential hosts of these viruses remain unknown. Microvirus clades were, however, shown to be host specific as sequences from different microviruses infecting the same host tend to cluster together (Székely and Breitbart 2016). These new genomes forming a monophyletic group with the amoyviruses that infect only Alphaproteobacteria suggests that these new viruses also infect Alphaproteobacteria. Indeed, members of the *Amoyvirinae* were discovered on marine Sphingomonadaceae cultures, and prophages found within the same bacterial family expanded this tentative subfamily (Zheng et al. 2018). Microviruses infecting Rhodobacterales and Rhizobiales were also assigned to amoyviruses, confirming their specificity to Alphaproteobacteria (Zucker et al. 2022). Moreover, it was proposed recently to elevate microviruses to the level of an order that consists of three suborders: one containing bullaviruses and pequenoviruses (60 genomes), one containing the amoyviruses (63 genomes), and one containing all the other existing subfamilies (12,980 genomes) (Kirchberger, Martinez, and Ochman 2022). Two of the sixteen genomes described here (ctih392 and ctda012) were part of this study and were classified in the suborder containing amoyviruses. Moreover, viruses belonging to *Libervirinae*, a recently described subfamily related to amoyviruses, infect the *Liberibacter* genus of Rhizobiaceae, an Alphaproteobacteria family (Zhang et al. 2021a).

Thus, the two clades described here are only linked to microviruses infecting *Alphaproteobacteria*, and although their relationships with these viruses are distant, the most plausible hypothesis is that they also infect Alphaproteobacteria. As microvirus genomes have a GC content similar to their host's (Roux et al. 2015; Zhan and Chen 2019), the variable GC content (between 40 and 59%) suggest that they infect different Alphaproteobacteria hosts: (1) reekeekeeviruses and some roodoodooviruses (55.3% on average) could hypothetically infect *Caulobacteridae*, an Alphaproteobacteria subclass with GC-rich genomes; and (2) the seven marine water monophyletic roodoodooviruses (40.5% on average) possibly infecting *Ricketsiidae*, an Alphaproteobacteria subclass with AT-rich genomes, potentially

the free-living and abundant *Pelagibacterales* (Muñoz-Gómez et al. 2019).

The Alphaproteobacteria have an ancient origin, estimated to ~1,900 million years (Wang and Luo 2021) and have extensively diversified since its origin with a rapid divergence of their major clades. In addition to their diversity in terms of colonized habitats and lifestyle, they are numerically dominant in many marine ecosystems, representing 40–50 per cent of bacterioplankton cells in sunlit oceans (Sunagawa et al. 2015), and are also ubiquitous freshwater systems although less numerous (Newton et al. 2011). It is tempting to imagine that one (or more) microviruses infected the common ancestor of Alphaproteobacteria more than 2 billion years ago, and that these ancestral viruses co-evolved with these organisms and experienced their eventful evolutionary history from the inside.

## 5. Conclusion

The family *Microviridae* is composed of abundant and diverse small circular DNA phages, to which belong the isolated DNA phage with the smallest genome. Despite the assembly and description of tens of thousands of complete circular microvirus genomes, the lower bound of their genome size has not been investigated. This study identified sixteen genomes in diverse aquatic environmental data sets, with sizes between 2.991 and 3.692 kb, that separate in two related yet distinct groups. Since cultured microvirus genomes contain many overlapping or even overprinting genes, it was particularly important to identify all genes in these genomes to better understand the functioning of these microviruses with small genomes. As such embedded genes were not predicted by automated gene identification tools, we thus performed a very thorough search of all coding regions, calculating for each ORF, their level of conservation, variations (SNPs), composition, and location in the genome. This allowed us to determine only four to five genes, with only one overprinted gene in five genomes and one gene with a significant overlap with another gene in each genome. Therefore it seems as though these viruses encode a minimal set of genes. For the two core genes Cap and Rep, which are the minimal gene set in some eukaryotic viruses, the length of each gene was the smallest among microviruses, only the essential regions of these essential genes being encoded. The small size of these genes, particularly for Rep whose variable region in phiX174 is known to be a nest for overprinted genes, likely prevents the presence of such regions that allow the appearance of very flexible coding regions. Even more surprising, the two *Microviridae* groups found here had specific and different gene content, and major differences in their conserved protein sequences highlight that these two related groups found different solutions to fulfill their lifecycle with such small numbers of genes.

## Data availability

Nucleotide sequence data reported are available in the Third Party Annotation Section of the DDBJ/ENA/GenBank databases under the accession numbers TPA: BK061417 and BK061430.

## Supplementary data

Supplementary data are available at *Virus Evolution* online.

## Acknowledgements

The authors are very grateful to François Hoh for his expertise on protein structure prediction and helpful discussions.

## Funding

This work was supported by the EU Horizon 2020 Framework Programme for Research and Innovation ('Virus-X', Project No. 685778). The work conducted by the US Department of Energy Joint Genome Institute (S.R.), a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy operated under Contract No. DE-AC02-05CH11231.

**Conflict of interest:** None declared.

## References

Altschul, S. F. et al. (1990) 'Basic Local Alignment Search Tool', *Journal of Molecular Biology*, 215: 403–10.

Berman, H. M. et al. (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28: 235–42.

Bolger, A. M., Lohse, M., and Usadel, B. (2014) 'Trimmomatic: A Flexible Trimmer for Illumina Sequence Data', *Bioinformatics*, 30: 2114–20.

Brentlinger, K. L. et al. (2002) 'Microviridae, a Family Divided: Isolation, Characterization, and Genome Sequence of φMH2K, a Bacteriophage of the Obligate Intracellular Parasitic Bacterium *Bdellovibrio bacteriovorus*', *Journal of Bacteriology*, 184: 1089–94.

Brinkman, N. E. et al. (2018) 'Reducing Inherent Biases Introduced during DNA Viral Metagenome Analyses of Municipal Wastewater', *PLoS One* 13: e0195350.

Bryson, S. J. et al. (2015) 'A Novel Sister Clade to the Enterobacteria Microviruses (Family Microviridae) Identified in Methane Seep Sediments', *Environmental Microbiology*, 17: 3708–21.

Chandler, M. et al. (2013) 'Breaking and Joining Single-stranded DNA: The HUH Endonuclease Superfamily', *Nature Reviews Microbiology*, 11: 525–38.

Cherwa, J. E., and Fane, B. A. (2011) 'Microviridae: Microviruses and Gokushoviruses', *eLS*. Chichester: John Wiley & Sons, Ltd.

Chipman, P. R. et al. (1998) 'Structural Analysis of the Spiroplasma Virus, SpV4: Implications for Evolutionary Variation to Obtain Host Diversity among the Microviridae', *Structure*, 6: 135–45.

Clark, K. et al. (2016) 'GenBank', *Nucleic Acids Research* 44: D67–72.

Cline, M. S., and Karchin, R. (2011) 'Using Bioinformatics to Predict the Functional Impact of SNVs', *Bioinformatics*, 27: 441–8.

Dokland, T. et al. (1999) 'The Role of Scaffolding Proteins in the Assembly of the Small, Single-Stranded DNA Virus φX17411 Edited by I. A. Wilson', *Journal of Molecular Biology*, 288: 595–608.

Doore, S. M., and Fane, B. A. (2016) 'The Microviridae: Diversity, Assembly, and Experimental Evolution', *Virology*, 491: 45–55.

Eddy, S. R. (1998) 'Profile Hidden Markov Models', *Bioinformatics*, 14: 755–63.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) 'An Efficient Algorithm for Large-Scale Detection of Protein Families', *Nucleic Acids Research*, 30: 1575–84.

Everson, J. S. et al. (2002) 'Biological Properties and Cell Tropism of Chp2, a Bacteriophage of the Obligate Intracellular Bacterium *Chlamydophila abortus*', *Journal of Bacteriology*, 184: 2748–54.

Finn, R. D. et al. (2014) 'Pfam: The Protein Families Database', *Nucleic Acids Research*, 42: D222–30.

Finsterbusch, T., and Mankertz, A. (2009) 'Porcine circoviruses—Small but Powerful', *Virus Research*, 143: 177–83.

Gregory, A. C. et al. (2019) 'Marine DNA Viral Macro- and Microdiversity from Pole to Pole', *Cell*, 177: 1109–1123.e14.

Hughes, A. L. et al. (2001) 'Simultaneous Positive and Purifying Selection on Overlapping Reading Frames of the Tat and Vpr Genes of Simian Immunodeficiency Virus', *Journal of Virology*, 75: 7966–72.

Hyatt, D. et al. (2010) 'Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification', *BMC Bioinformatics*, 11: 119.

Kanehisa, M., and Goto, S. (2000) 'KEGG: Kyoto Encyclopedia of Genes and Genomes' *Nucleic Acids Research*, 28: 27–30.

Katoh, K. et al. (2002) 'MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform', *Nucleic Acids Research*, 30: 3059–66.

Kim, Y., Aw, T. G., and Rose, J. B. (2016) 'Transporting Ocean Viromes: Invasion of the Aquatic Biosphere', *PLoS One*, 11: e0152671.

Kirchberger, P. C., Martinez, Z. A., and Ochman, H. (2022) 'Organizing the Global Diversity of Microviruses', *mBio* 13: e00588–22.

Kirchberger, P. C., and Ochman, H. (2020) 'Resurrection of a Global, Metagenomically Defined Gokushovirus'. In: M. M. Zambrano, K. Kirkegaard, and M. Breitbart (eds) *eLife*, Vol. IX: p. e51599. eLife Sciences Publications.

Kono, N., Arakawa, K., and Tomita, M. (2011) 'Comprehensive Prediction of Chromosome Dimer Resolution Sites in Bacterial Genomes', *BMC Genomics*, 12: 19.

Koonin, E. V. et al. (2020) 'Global Organization and Proposed Megataxonomy of the Virus World', *Microbiology and Molecular Biology Reviews: MMBR*, 84: e00061–19.

Krupovic, M., and Forterre, P. (2011) 'Microviridae Goes Temperate: Microvirus-Related Proviruses Reside in the Genomes of Bacteroidetes', *PLoS One*, 6: e19893.

Labonté, J. M., and Suttle, C. A. (2013) 'Previously Unknown and Highly Divergent ssDNA Viruses Populate the Oceans', *The ISME Journal*, 7: 2169–77.

Letunic, I., and Bork, P. (2021) 'Interactive Tree of Life (Itol) V5: An Online Tool for Phylogenetic Tree Display and Annotation', *Nucleic Acids Research*, 49: W293–6.

Li, H. et al. (2009) 'The Sequence Alignment/Map Format and SAMtools', *Bioinformatics (Oxford, England)*, 25: 2078–9.

Li, D. et al. (2015) 'MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph', *Bioinformatics*, 31: 1674–6.

Li, H., and Durbin, R. (2009) 'Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform', *Bioinformatics (Oxford, England)*, 25: 1754–60.

Liljas, L. (1991) 'Structure of Spherical Viruses', *International Journal of Biological Macromolecules*, 13: 273–80.

Lin, M. F., Jungreis, I., and Kellis, M. (2011) 'PhyloCSF: A Comparative Genomics Method to Distinguish Protein Coding and Non-coding Regions', *Bioinformatics*, 27: i275–82.

Lorenz, R., Hofacker, I. L., and Stadler, P. F. (2016) 'RNA Folding with Hard and Soft Constraints', *Algorithms for Molecular Biology: AMB*, 11: 8.

Lu, J. S. et al. (2018) 'RiboSketch: Versatile Visualization of Multistranded RNA and DNA Secondary Structure', *Bioinformatics*, 34: 4297–9.

Markowitz, V. M. et al. (2012) 'IMG: The Integrated Microbial Genomes Database and Comparative Analysis System', *Nucleic Acids Research*, 40: D115–22.

Martin, M. (2011) 'Cutadapt Removes Adapter Sequences from High-throughput Sequencing Reads', *EMBnet.journal*, 17: 10–2.

Martinez-Hernandez, F. et al. (2017) 'Single-Virus Genomics Reveals Hidden Cosmopolitan and Abundant Viruses', *Nature Communications*, 8: 15892.

McKenna, R. et al. (1992) 'Atomic Structure of Single-Stranded DNA Bacteriophage ΦX174 and Its Functional Implications', *Nature*, 355: 137–43.

McNair, K. et al. (2019) 'PHANOTATE: A Novel Approach to Gene Identification in Phage Genomes', *Bioinformatics*, 35: 4537–42.

Mirdita, M. et al. (2022) 'ColabFold: Making Protein Folding Accessible to All', *Nature Methods*, 19: 679–82.

Moraru, C., Varsani, A., and Kropinski, A. M. (2020) 'VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses', *Viruses*, 12: 1268.

Muñoz-Gómez, S. A. et al. (2019) 'An Updated Phylogeny of the Alphaproteobacteria Reveals that the Parasitic Rickettsiales and Holosporales Have Independent Origins'. In: A. Rokas, P. J. Wittkopp, and I. Irisarri (eds) *eLife*, viii: p. e42535. eLife Sciences Publications.

Narechania, A., Terai, M., and Burk, R. D. Y. (2005) 'Overlapping Reading Frames in Closely Related Human Papillomaviruses Result in Modular Rates of Selection within E2', *Journal of General Virology*, 86: 1307–13.

Newton, R. J. et al. (2011) 'A Guide to the Natural History of Freshwater Lake Bacteria', *Microbiology and Molecular Biology Reviews: MMBR*, 75: 14–49.

Nowinski, B. et al. (2019) 'Microbial Metagenomes and Metatranscriptomes during a Coastal Phytoplankton Bloom', *Scientific Data*, 6: 129.

Olo Ndela, E., Enault, F., and Toussaint, A. (2021) 'Transposable Prophages in Leptospira: An Ancient, Now Diverse, Group Predominant in Causative Agents of Weil's Disease', *International Journal of Molecular Sciences*, 22: 13434.

Pavesi, A., Magiorkinis, G., and Karlin, D. G. (2013) 'Viral Proteins Originated de novo by Overprinting Can Be Identified by Codon Usage: Application to the "Gene Nursery" of Deltaretroviruses', *PLOS Computational Biology*, 9: e1003162.

Peng, Y. et al. (2012) 'IDBA-UD: A de novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth', *Bioinformatics*, 28: 1420–8.

Quaiser, A. et al. (2015) 'Diversity and Comparative Genomics of Microviridae in Sphagnum- Dominated Peatlands', *Frontiers in Microbiology*, 6: 375.

Rice, P., Longden, I., and Bleasby, A. (2000) 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends in Genetics: TIG*, 16: 276–7.

Rosario, K. et al. (2012) 'Diverse Circular ssDNA Viruses Discovered in Dragonflies (Odonata: Epiprocta)', *The Journal of General Virology*, 93: 2668–81.

Roux, S. et al. (2012) 'Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads'.*PLoS One*, 7: e40418.

Roux, S. et al. (2015) 'Viral Dark Matter and Virus–Host Interactions Resolved from Publicly Available Microbial Genomes', *eLife*, 4: e08490.

Sanger, F. et al. (1977) 'Nucleotide Sequence of Bacteriophage Phi X174 DNA', *Nature*, 265: 687–95.

Shlomai, J., and Kornberg, A. (1980) 'An Escherichia Coli Replication Protein that Recognizes a Unique Sequence within a Hairpin Region in phi X174 DNA', *Proceedings of the National Academy of Sciences*, 77: 799–803.

Sievers, F. et al. (2011) 'Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega', *Molecular Systems Biology*, 7: 539.

Silva, G. G. et al. (2013) 'Combining de novo and Reference-Guided Assembly with Scaffold_builder', *Source Code for Biology and Medicine*, 8: 23.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.

Steinegger, M. et al. (2019) 'HH-suite3 for Fast Remote Homology Detection and Deep Protein Annotation', *BMC Bioinformatics*, 20: 473.

Steinegger, M., and Söding, J. (2017) 'MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets', *Nature Biotechnology*, 35: 1026–8.

Sun, L. et al. (2014) 'Icosahedral Bacteriophage ΦX174 Forms a Tail for DNA Transport during Infection', *Nature*, 505: 432–5.

Sunagawa, S. et al. (2015) 'Structure and Function of the Global Ocean Microbiome', *Science*, 348: 1261359.

Székely, A. J., and Breitbart, M. (2016) 'Single-Stranded DNA Phages: From Early Molecular Biology Tools to Recent Revolutions in Environmental Microbiology', *FEMS Microbiology Letters*, 363: fnw027.

Tikhe, C. V., and Husseneder, C. (2018) 'Metavirome Sequencing of the Termite Gut Reveals the Presence of an Unexplored Bacteriophage Community', *Frontiers in Microbiology*, 8: 2548.

Van Cauwenberghe, J. et al. (2021) 'Spatial Patterns in phage-Rhizobium Co-evolutionary Interactions across Regions of Common Bean Domestication', *The ISME Journal*, 15: 2092–106.

Vanderhaeghen, S. et al. (2018) 'The Novel EHEC Gene *asa* Overlaps the TEGT Transporter Gene in Antisense and Is Regulated by NaCl and Growth Phase', *Scientific Reports*, 8: 17875.

Wang, S., and Luo, H. (2021) 'Dating Alphaproteobacteria Evolution with Eukaryotic Fossils', *Nature Communications*, 12: 3324.

Washietl, S. et al. (2011) 'RNAcode: Robust Discrimination of Coding and Noncoding Regions in Comparative Sequence Data', *RNA*, 17: 578–94.

Wheeler, T. J., and Eddy, S. R. (2013) 'Nhmmer: DNA Homology Search with Profile HMMs', *Bioinformatics*, 29: 2487–9.

Wickham, H. (2009) *Ggplot2*. New York: Springer.

Zayed, A. A. et al. (2022) 'Cryptic and Abundant Marine Viruses at the Evolutionary Origins of Earth's RNA Virome', *Science*, 376: 156–62.

Zhan, Y., and Chen, F. (2019) 'The Smallest ssDNA Phage Infecting a Marine Bacterium', *Environmental Microbiology*, 21: 1916–28.

Zhang, L. et al. (2021a) 'A Novel Microviridae Phage (Clasmv1) from "Candidatus Liberibacter Asiaticus"', *Frontiers in Microbiology*, 12: 754245.

Zhang, R. et al. (2021b) 'SpacePHARER: Sensitive Identification of Phages from CRISPR Spacers in Prokaryotic Hosts', *Bioinformatics*, 37: 3364–6.

Zheng, Q. et al. (2018) 'A Virus Infecting Marine Photoheterotrophic Alphaproteobacteria (*Citromicrobium* spp.) Defines a New Lineage of ssDNA Viruses', *Frontiers in Microbiology*, 9: 1418.

Zucker, F. et al. (2022) 'New Microviridae Isolated from Sulfitobacter Reveals Two Cosmopolitan Subfamilies of ssDNA Phages Infecting Marine and Terrestrial Alphaproteobacteria', *Virus Evolution* 8: veac070.

Zulkower, V., and Rosser, S. (2020) 'DNA Features Viewer: A Sequence Annotation Formatting and Plotting Library for Python', *Bioinformatics*, 36: 4350–2.