

Metagenomic sequencing reveals a lack of virus exchange between native and invasive freshwater fish across the Murray–Darling Basin, Australia

Vincenzo A. Costa,¹ Jonathon C.O. Mifsud,^{1,†} Dean Gilligan,²
Jane E. Williamson,¹ Edward C. Holmes,^{3,‡} and Jemma L. Geoghegan^{1,4,5,*§}

¹Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia, ²NSW Department of Primary Industries, Batemans Bay Fisheries Office, Batemans Bay 2536, Australia, ³Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Life and Environmental Sciences and Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia, ⁴Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand and ⁵Institute of Environmental Science and Research, Wellington, Porirua 5022, New Zealand

*Corresponding author: E-mail: jemma.geoghegan@otago.ac.nz

†<https://orcid.org/0000-0001-8934-9193>

‡<https://orcid.org/0000-0001-9596-3552>

§<https://orcid.org/0000-0003-0970-0153>

Abstract

Biological invasions are among the biggest threats to freshwater biodiversity. This is increasingly relevant in the Murray–Darling Basin, Australia, particularly since the introduction of the common carp (*Cyprinus carpio*). This invasive species now occupies up to ninety per cent of fish biomass, with hugely detrimental impacts on native fauna and flora. To address the ongoing impacts of carp, cyprinid herpesvirus 3 (CyHV-3) has been proposed as a potentially effective biological control agent. Crucially, however, it is unknown whether CyHV-3 and other cyprinid herpesviruses already exist in the Murray–Darling. Further, little is known about those viruses that naturally occur in wild freshwater fauna, and the frequency with which these viruses jump species boundaries. To document the evolution and diversity of freshwater fish viromes and better understand the ecological context to the proposed introduction of CyHV-3, we performed a meta-transcriptomic viral survey of invasive and native fish across the Murray–Darling Basin, covering over 2,200 km of the river system. Across a total of thirty-six RNA libraries representing ten species, we failed to detect CyHV-3 nor any closely related viruses. Rather, meta-transcriptomic analysis identified eighteen vertebrate-associated viruses that could be assigned to the *Arenaviridae*, *Astroviridae*, *Bornaviridae*, *Caliciviridae*, *Coronaviridae*, *Chuviridae*, *Flaviviridae*, *Hantaviridae*, *Hepeviridae*, *Paramyxoviridae*, *Picornaviridae*, *Poxviridae*, *Reoviridae* and *Rhabdoviridae* families, and a further twenty-seven that were deemed to be associated with non-vertebrate hosts. Notably, we revealed a marked lack of viruses that are shared among invasive and native fish sampled here, suggesting that there is little virus transmission from common carp to native fish species, despite co-existing for over

fifty years. Overall, this study provides the first data on the viruses naturally circulating in a major river system and supports the notion that fish harbour a large diversity of viruses with often deep evolutionary histories.

Key words: virome; meta-transcriptomics; fish virus; freshwater fish; virus evolution; RNA sequencing.

1. Introduction

Anthropogenic stressors such as pollution, climate change and the introduction of exotic species continue to pose a significant threat to freshwater habitats, with almost one-third of all fish species threatened by extinction (World Wildlife Fund 2021). The Murray–Darling Basin, the largest freshwater river system in Australia, harbours at least twelve exotic freshwater fish species (Barrett, Bamford, and Jackson 2014). Key among these are eastern mosquitofish (*Gambusia holbrooki*), redbfin perch (*Perca fluviatilis*) and, most notably, common carp (*Cyprinus carpio*). Common carp (also known as European carp) were initially introduced into Australia during the mid-1800s for aquaculture operations and again on several occasions throughout the 1900s (Koehn 2004; Forsyth et al. 2013). During extensive flooding events during the 1970s, carp spread across much of the basin and now represent up to ninety per cent of total fish biomass (Forsyth et al. 2013).

The invasion of carp has been hugely detrimental to Australian freshwater ecosystems (Koehn 2004; Forsyth et al. 2013). Impacts include increased water turbidity, decreased light penetration, erosion of riverbanks, changes in the abundance and diversity of native invertebrate communities and outcompeting native fish species for habitat and resources (Koehn 2004; Forsyth et al. 2013; Vilizzi et al. 2014). Several control methods have been proposed to control invasive carp; nevertheless, their resilience and high fecundity create significant challenges (Hayes et al. 2014). This has stimulated research into biological control methods, such as deployment of the virus cyprinid herpesvirus 3 (CyHV-3) (McColl et al. 2017; McColl, Cooke, and Sunarto 2014).

CyHV-3 is a double-stranded DNA virus (family *Alloherpesviridae*, order *Herpesvirales*) first isolated from farmed carp in the late 1990s (Matsui et al. 2008). Since its discovery, it has been responsible for large disease outbreaks worldwide with a mortality rate of up to 80 per cent in domestic carp (Michel et al. 2010). CyHV-3 is transmitted horizontally through direct contact with skin lesions or secretion of viral particles in freshwater where it can survive for up to 3 days (Shimizu et al. 2006). The host range of CyHV-3 is currently limited to koi and common carp (Michel et al. 2010). While CyHV-3 DNA has been identified in goldfish (*Carrasius auratus*) (Bergmann et al. 2010), it is still relatively unclear whether infection occurs in these species (Ilouze et al. 2011; Tolo et al. 2021).

Initial laboratory trials suggest that CyHV-3 is safe for non-target species (McColl et al. 2017). However, little is known about the viruses that naturally circulate in Australian native freshwater fauna, including any prior evidence for the existence of CyHV-3 (Kopf et al. 2019), nor on the time-scales and frequency with which viruses jump between fish hosts. To completely assess the safety and efficacy of any virus biocontrol agent, including CyHV-3, a comprehensive assessment of the viruses that naturally infect both native and invasive species is required.

Following the advent of meta-transcriptomic sequencing, it is now possible to characterize the entire set of viruses—the virome—within a given host (Shi, Zhang, and Holmes 2018;

Zhang, Shi, and Holmes 2018). Fish, in particular, harbour a high abundance and diversity of viruses often with deep evolutionary histories (Shi et al. 2018; Zhang et al. 2018). However, despite the antiquity and diversity of fish viruses, there are few studies of virus diversity and evolution in wild freshwater fish populations, particularly in the context of biological invasions.

Determining the viromes of invasive freshwater fish like the common carp will enhance our understanding of the broad-scale factors that influence virus emergence and evolution. As the date and site of their introduction is well-documented in Australian waters, these species can potentially provide important information on the both rate of cross-species transmission and how frequently viruses might move between invasive and native species. In addition, despite representing a small fraction of the earth's surface water, freshwater environments serve as a habitat for forty to fifty per cent of total fish species, harbouring the greatest biodiversity per land area (Dudgeon et al. 2006). Such habitats are subject to rapid environmental change, which may significantly impact species connectivity (Johnson and Paull 2011). Since contact and exposure between hosts are vital for cross-species transmission of viruses (Parrish et al. 2008), these species may also inform us on the ecological factors that impact virome composition within a given host.

We performed a meta-transcriptomic viral survey of invasive and native freshwater fish species across the Murray–Darling Basin in Australia to document the diversity and evolution of freshwater fish viromes and, from this, better understand the ecological drivers of virus evolution and emergence. To the best of our knowledge, this is the largest survey of freshwater fish viruses undertaken to date. In particular, we aimed to determine whether CyHV-3 is already present in common carp in Australia (Kopf et al. 2019), and whether there is evidence for transmission of existing viruses between exotic and native species. As such, we provide important information on the ecological and evolutionary context for the potential release of future virus biocontrols.

2. Methods

2.1 Ethics

Fish sampling was conducted with animal ethics approval (ref: 2019/035) from the Animal Ethics Committee (AEC) at Macquarie University, Sydney, NSW. Biosafety was approved by Macquarie University (ref: 5201700856).

2.2 Sample Collection

We compared the viromes of native and invasive fish species occupying different areas across the Murray–Darling Basin, Australia (Fig. 1). Sampling occurred between January and March 2020. A total of seven native fish species were collected: bony herring (*Nematalosa erebi*) (n = 20), spangled perch (*Leiopotherapon unicolor*) (n = 1), Australian smelt (*Retropinna semoni*) (n = 12), Murray–Darling rainbowfish (*Melanotaenia fluviatilis*) (n = 17), flat-headed gudgeon (*Philypnodon grandiceps*) (n = 9), western carp-gudgeon (*Hypseleotris* spp.) (n = 20) and unspotted

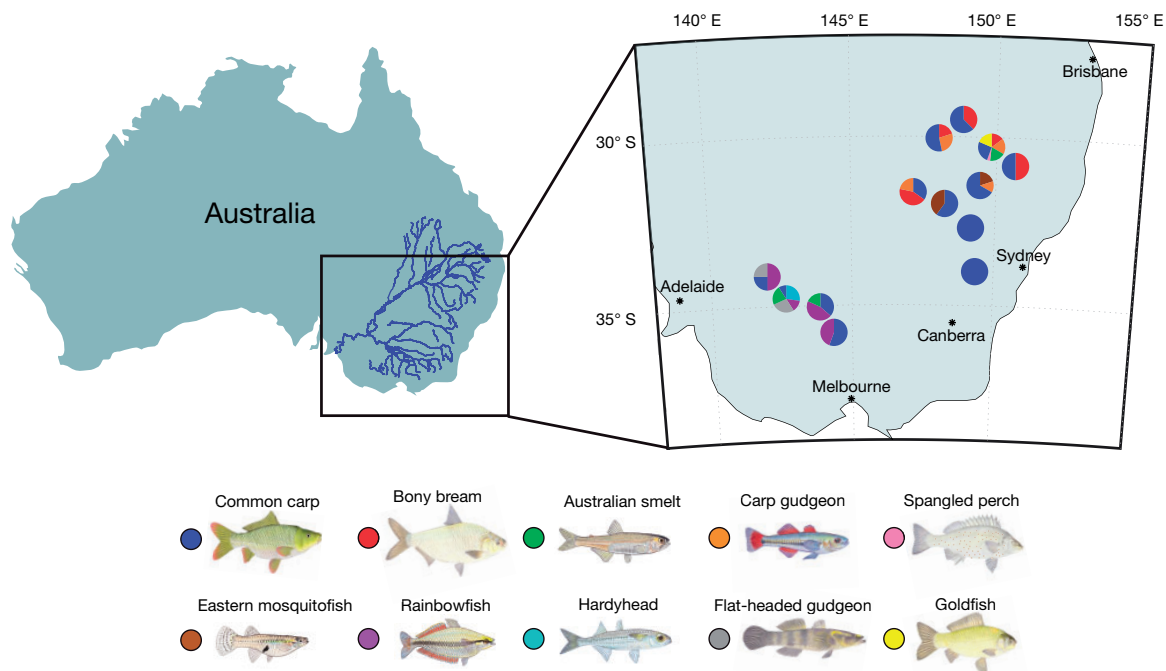


Figure 1. Fish sampling locations. Map illustrating freshwater sampling locations across the Murray–Darling Basin, Australia. Pie charts show the abundance and diversity of fish species captured at each site. Colours correspond to the fish species sampled as illustrated below the map.

hardyhead (*Craterocephalus fulvus*) ($n=6$). Three species of invasive fish were also collected: common carp (*C. carpio*) ($n=74$), goldfish (*C. auratus*) ($n=5$) and eastern mosquitofish (*G. holbrooki*) ($n=15$). All fish caught were apparently healthy, with no signs of disease. Fish were caught using boat electrofishing, euthanized and dissected immediately upon capture. Tissue specimens (liver and gills) were placed in RNALater and stored in a portable -80°C freezer, then later in an -80°C freezer in the laboratory at Macquarie University, Sydney. Tissue selection was based on previous studies (Shi et al. 2018; Geoghegan et al. 2018; Geoghegan et al. 2021), which show that liver and gill tissue serve as a rich source of viruses. To facilitate virus discovery, multiple individuals (one to ten) were pooled according to species and the location in which they were captured (Supplementary Table S2 and Fig. S1).

2.3 Total RNA Extraction and Transcriptome Sequencing

Frozen samples of liver and gill tissue were processed together as a single extraction for each sample. The combined tissues were placed in 600 μl of lysis buffer containing 0.5 per cent foaming reagent (Reagent DX, Qiagen) and 1 per cent of β -mercaptoethanol (Sigma-Aldrich). Submerged tissue samples were homogenized with TissueRuptor (Qiagen) for one minute at 5,000 rpm. To further homogenize tissue samples and remove tissue residues, the homogenate was centrifuged at full speed for three minutes. The homogenate was carefully removed and RNA from the clear supernatant was extracted using the RNeasy Plus Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. Extracted RNA was quantified using NanoDrop (ThermoFisher) and RNA from each species was

pooled corresponding to the site in which they were captured, resulting in thirty-six sample libraries (Supplementary Table S2). RNA libraries were constructed using the Truseq Total RNA Library Preparation Protocol (Illumina). To enhance viral discovery and reduce the presence of non-viral reads, host ribosomal RNA (rRNA) was depleted using the Ribo-Zero-Gold Kit (Illumina) and paired-end sequencing (150bp) was performed on the NovaSeq 500 platform (Illumina). Sample library construction, rRNA depletion and RNA sequencing were performed at the Australian Genome Research Facility.

2.4 Virus discovery

Raw Illumina sequence reads (forward and reverse) were initially quality trimmed with Trimmomatic v.0.39 (Bolger, Lohse, and Usadel 2014) then assembled into contigs *de novo* using Trinity RNA-seq v.2.8.5 (Haas et al. 2013), with the default parameter settings. Assembled contigs were annotated and compared against the NCBI nucleotide (nt) and non-redundant protein (nr) databases with an e-value threshold of 1×10^{-5} using BLASTn and Diamond (BLASTX) (Buchfink, Xie, and Huson 2014). To initially distinguish between invertebrate and vertebrate-associated viruses, contigs that matched viral sequences were inspected using Geneious v.11.1.5 (Kearse et al. 2012) and translated into amino acid sequences. Amino acid sequences were then used as a single query in additional sequence comparisons against the NCBI nt and nr databases using BLAST algorithms. This method was also used to remove false-positives (e.g. host genes and endogenous viral elements) from our analyses. To help exclude instances of index hopping, viral sequences that were identified in multiple libraries were also

inspected using Geneious Prime (www.geneious.com) and amino acid pairwise alignments between viral sequences were performed with Multiple Alignment using Fast Fourier Transform (MAFFT) v.7.450 (Kato and Standley 2013), using the E-INS-i algorithm. Abundances of identical viral transcripts were then calculated (see below) and sequences that were present at frequency of less than one per cent of that of the number of reads present in the dominant library were excluded. To determine whether a virus was novel, we followed the broad criteria specified by The International Committee on Taxonomy of Viruses (<http://www.ictvonline.org/>).

2.5 Inferring the evolutionary history of novel viral sequences

To determine the evolutionary history of the viruses identified in this study and further distinguish between vertebrate and invertebrate-associated viruses (which are usually phylogenetically distinct), we estimated phylogenetic trees using amino acid sequences of stable genomic regions such as RNA-dependent RNA polymerase (RdRp) or DNA polymerase for DNA viruses. To this end, we combined our sequences with background sequences for each respective virus family taken from NCBI/GenBank. Amino acid sequences were aligned with MAFFT v.7.450 using the E-INS-i algorithm. To remove ambiguous regions in the sequence alignment, amino acid sequences were trimmed using trimAl v.1.2 (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009). To estimate phylogenetic trees, selection of the best-fit model of amino acid substitution was determined using the Akaike information criterion (AIC), corrected AIC, and the Bayesian information criterion with the ModelFinder function (-m MFP) in IQ-TREE (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017). Sequence data were analysed using a maximum likelihood (ML) approach in IQ-TREE, with 1,000 bootstrap replicates. Phylogenetic trees were annotated with FigTree v.1.4.2. and further edited using Adobe Illustrator (<https://www.adobe.com>).

2.6 Virome composition

To quantify the relative abundance of viral transcripts within the host transcriptome, the RNA-Seq by Expectation (RSEM) value was estimated using Trinity (Haas et al. 2013), and raw counts from each transcript were standardized against the total number of reads within the given sequencing library. We also used this approach to estimate the relative abundance of a host reference gene, ribosomal protein S13 (RPS13), which is stably expressed in fish. To assess any differences in virome composition between hosts and sites, we calculated alpha diversity (virome richness and Shannon diversity) using Rhea packages (Lagkouravos et al. 2017). Generalized linear models (GLM) were used to identify the impact of host taxonomy (i.e. species), host geography (i.e. site), water temperature, water pH, water turbidity and species origin (i.e. invasive or native) on both vertebrate-associated virus composition (abundance, richness and diversity) and those viruses likely associated with non-fish hosts: the latter should not be affected by aspects of fish biology and hence effectively constitute a negative control. All GLM models were tested using a likelihood-ratio test (χ^2) and a Tukey's post hoc analysis (glht) was performed using the *multcomp* package (Hothorn, Bretz, and Westfall 2008). To assess viral diversity between samples, we calculated beta diversity using a Bray Curtis dissimilarity with the *phyloseq* package (McMurdle and Holmes 2013). Differences in virome

composition between native and invasive species were calculated using permanova (Adonis test), with the *vegan* package (Dixon 2003). All statistical analyses were carried out on RStudio V1.2.1335 and plotted using the *ggplot2* package (Valero-Mora 2015).

3. Results

We characterized the viromes of ten freshwater ray-finned fish species across seven taxonomic orders (two invasive and five native) at thirteen locations across the Murray–Darling Basin in Australia. Total RNA-sequencing was performed on thirty-six libraries, resulting in a median of 76,528,534 (range 66,015,138–95,168,951) reads per library. *De novo* assembly of the sequencing reads resulted in a median of 617,588 contigs (range 198,446–1,989,596) per library, with a total of 23,976,218 contigs generated. Analysis of the host reference gene, RPS13, revealed abundances of 0.000001–0.0002 per cent, suggesting an inconsistent sequence coverage across all RNA libraries, which may have impacted virus discovery (Fig. 2).

3.1 Abundance and diversity of viruses

We identified eighteen viral sequences that were associated with vertebrate hosts and a further twenty-seven that were likely associated with algae, invertebrates and protists in the freshwater environment (Supplementary Figs. S1 and S2). Because such non-vertebrate viruses were likely derived from diet or contamination of gill tissue, we primarily focused on vertebrate-associated viruses.

Among the likely vertebrate-associated viruses, we identified viral sequences from fourteen viral families. With the exception of a novel poxvirus (family *Poxviridae*), a double-stranded DNA virus, all the viruses identified possessed RNA genomes. The most abundant vertebrate-associated viral transcripts were those assigned to the *Arenaviridae* (49% of all vertebrate-associated viruses), *Hepeviridae* (20%), *Chuviridae* (21%), *Astroviridae* (3%) and *Flaviviridae* (2%) families. Other likely vertebrate viral transcripts detected were assigned to the *Coronaviridae* (<1%), *Caliciviridae* (<1%), *Picornaviridae* (<1%), *Paramyxoviridae* (<1%), *Hantaviridae* (<1%), *Bornaviridae* (<1%), *Poxviridae* (<1%), *Reoviridae* (<1%) and *Rhabdoviridae* (<1%) families. The most common vertebrate-associated viruses identified were astroviruses, detected in three host species (eastern mosquitofish, Murray–Darling rainbowfish, spangled perch). In addition, arenaviruses were detected in two host species (western carp-gudgeon, eastern mosquitofish) along with hepeviruses (common carp, eastern mosquitofish). All other viruses were identified in one host species (Fig. 3).

Among the viruses likely associated with non-vertebrate hosts (i.e. those infecting arthropods, fungi, plants and protozoans), a large proportion (70%) were unclassified, comprising picorna-like viruses, rhabdo-like viruses, tombus-like viruses and narna-like viruses (Shi et al. 2016) (Supplementary Fig. S1). We also detected viral transcripts that could be assigned to the *Nodaviridae* (27.1%), *Permutotetraviridae* (1.3%), *Dicistroviridae* (1.2%) and *Phenuiviridae* (<1%) families. Although viruses within the *Nodaviridae* have been shown to infect fish (Hameed et al. 2019), all of the nodavirus sequences identified here clustered with viruses from invertebrate hosts (Supplementary Fig. S2), strongly suggesting they were similarly associated with fish diet or contamination of gill tissue.

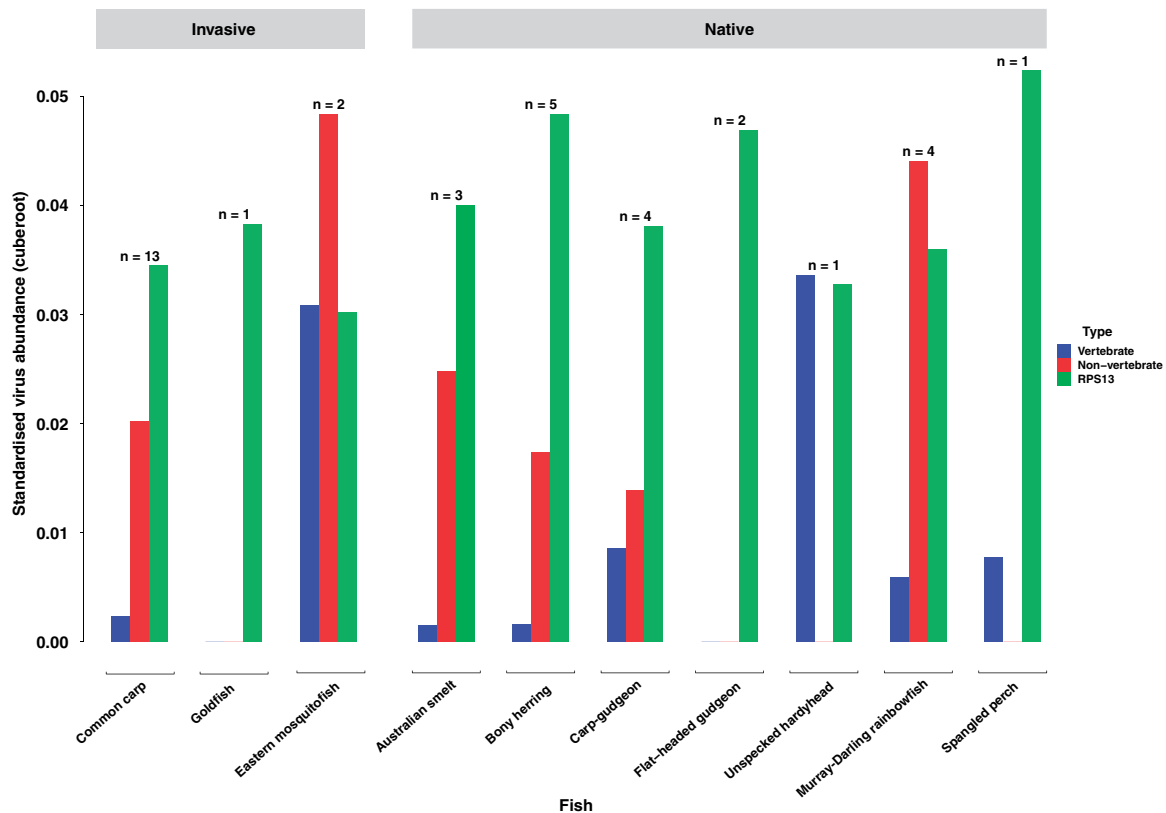


Figure 2. Mean standardized viral abundance across all libraries. Clustered bar chart reveals differences in viral abundance between invasive and native fish species. Blue bars represent vertebrate-associated viruses; red bars represent non-vertebrate-associated viruses; and green bars represent host reference gene RPS13. Number of sequencing libraries for each fish species is displayed above bars.

3.2 Phylogenetic relationships of vertebrate-associated viruses

To infer the phylogenetic relationships and hence the evolutionary history of the viruses newly identified here, we focused on stable genomic regions such as the RdRp in RNA viruses and DNA polymerase in the case of the novel poxvirus. Using these genomic regions, we identified seven negative-sense single-stranded RNA (-ssRNA) viruses (families *Arenaviridae*, *Bornaviridae*, *Chuviridae*, *Hantaviridae*, *Paramyxoviridae*, *Rhabdoviridae*), nine positive-sense single-stranded RNA (+ssRNA) viruses (families *Astroviridae*, *Caliciviridae*, *Coronaviridae*, *Flaviviridae*, *Hepeviridae*, *Picornaviridae*), one double-stranded RNA (dsRNA) virus (family *Reoviridae*) and one double-stranded DNA (dsDNA) virus (family *Poxviridae*). We now describe each of these groups in turn.

3.2.1 -ssRNA viruses

We identified -ssRNA viruses that occupied phylogenetic positions that were broadly indicative of long-term virus–host co-divergence, with many fish viruses falling basal to reptile, avian and mammalian viruses (Fig. 4). Notably, we identified two novel arenaviruses that clustered with members of the newly formed *Antennavirus* genus that includes fish hosts (Shi et al. 2018; Abudurexiti et al. 2019). *Western carp-gudgeon arenavirus* found at Narrabri creek shared 36.9 per cent amino acid sequence similarity with its closest relative, *Wenling frogfish arenavirus 1* (Shi et al. 2018) and grouped with recently discovered

arenaviruses from pygmy gobies (*Eviota zebrina*) sampled from an Australian coral reef (Geoghegan et al. 2021). *Eastern mosquitofish arenavirus* found in the Macquarie River shared 84.5 per cent amino acid sequence similarity with its closest available relative, *Wenling frogfish arenavirus 1*.

The divergent hantavirus detected in the Murray–Darling rainbowfish falls basal to mammalian hantaviruses (orthohantaviruses) and clustered with members of the *Actantavirinae* and *Agantavirinae* subfamilies that include ray-finned and jawless fish hosts (Shi et al. 2018; Abudurexiti et al. 2019) (Fig. 4). This virus had only 27.3 per cent amino acid similarity with its closest relative, *Bern perch virus* (NCBI/GenBank: QGM12349.1). Broad patterns of virus–host co-divergence can similarly be seen in the cultivirus identified in carp from Lake Burrendong. BLAST analysis identified *sharpbelly cultivirus* (Shi et al. 2018) as the closest relative of all genomic regions, including the L gene (93% amino acid similarity), glycoprotein (86.3%) and nucleoprotein (92.9%).

Our virological survey revealed the complete genome of a novel chuvirus in the unspecked hardyhead in the Edward River. This *Hardyhead chuvirus* displayed three open reading frames, representing the L protein (RdRp), glycoprotein and nucleoprotein. Our analysis identified *Guangdong red-banded snake chuvirus* (Shi et al. 2018) as the closest relative of the L protein (44% amino acid similarity), *Wenling fish chu-like virus* (Shi et al. 2018) as the closest relative of the glycoprotein (41%) and *Herr Frank virus 1* (Argenta et al. 2020) as the closest relative of the

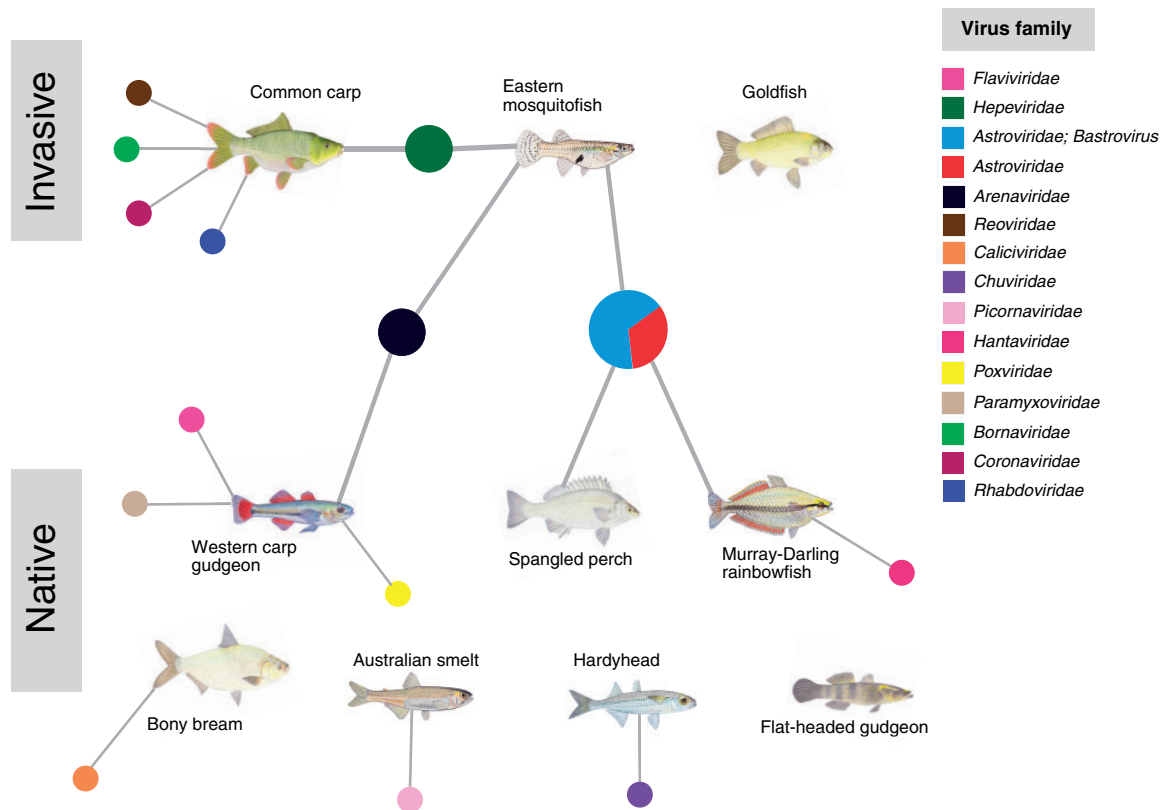


Figure 3. Network diagram displaying vertebrate-associated viruses identified in native and invasive freshwater fish. Colours of each node represent a virus family. Both goldfish and flat-headed gudgeon contained non-vertebrate-associated viruses.

nucleoprotein (34%). *Hardyhead* chuvirus formed a distinct phylogenetic clade with all other vertebrate-associated chuviruses (Fig. 5).

We also detected a novel paramyxovirus in western carp-gudgeon in the Bogan River. This divergent viral sequence shared 35.2 per cent L gene amino acid similarity with its closest relative, *Wenling tonguesole paramyxovirus* (genus *Cynoglossusvirus*, family *Paramyxoviridae*) (Shi et al. 2018). These viruses grouped with *Wenzhou pacific spadenose shark paramyxovirus* (genus *Scoliodonvirus*), together falling basal to other members of the *Paramyxoviridae* family. In addition, a novel rhabdovirus in common carp similarly formed a distinct clade, basal to other fish-infecting rhabdoviruses. This virus shared 35.7 per cent amino acid L gene sequence similarity with *Beihai dimarhabdovirus* that was also identified in fish (Shi et al. 2018) and clustered with other dimarhabdoviruses, including those found in the spotted paddle-tail newt from China (Shi et al. 2018) and the big brown bat (*Eptesicus fuscus*) from the USA (NCBI/Genbank: QPO14166.1). Across all genera within the *Rhabdoviridae*, lyssaviruses were the closest relatives to this clade (Fig. 4), with *Murray-Darling carp rhabdovirus* sharing 31.6 per cent amino acid L gene similarity with *rabies lyssavirus* (Tordo et al. 1988).

3.2.2 +ssRNA viruses

We identified a viral sequence in common carp that shared 50.7 per cent RdRp sequence similarity with *Pacific salmon nidovirus* (family *Coronaviridae*) (Mordecai et al. 2019). *Murray-Darling carp*

letovirus also exhibited sequence similarity (46.2%) with gamma-coronaviruses, including *bottlenose dolphin coronavirus* and *beluga whale coronavirus* (Mihindukulasuriya et al. 2008; Woo et al. 2014). This virus grouped with both *Pacific salmon nidovirus* and *Microhylla letovirus* (Bukhari et al. 2018), which together form an outgroup to all other coronaviruses (Fig. 6).

We also identified a novel flavivirus (genus *Flavivirus*, family *Flaviviridae*) in western carp-gudgeon in the Bogan River. This viral sequence exhibited thirty-three to thirty-six per cent NS5 amino acid sequence similarity with its closest relatives, *Cyclopterus lumpus virus* (Skoge et al. 2018), *Tamana bat virus* (de Lamballerie et al. 2002), *salmon flavivirus* (Soto et al. 2020) and *Wenzhou shark flavivirus* (Shi et al. 2018). All these viruses fall basal to vector-borne viruses within the genus *Flavivirus* (Fig. 7).

Among other positive-sense RNA viruses identified, a novel astrovirus, calicivirus and picornavirus all grouped with other fish hosts and expanded the phylogenetic diversity of these virus families (Fig. 7). The novel astrovirus identified in *Murray-Darling rainbowfish* shared forty per cent RdRp amino acid similarity with *Wuhan astro-like virus* (Shi et al. 2018). This virus clustered with other astro-like viruses discovered in fish, including *Beihai mudskipper astro-like virus* (Shi et al. 2018) and *Guangdong catfish astro-like virus* (Shi et al. 2018). This was similarly observed in *Australian smelt picornavirus*, which clustered with picornaviruses found in other freshwater fish, including those from eels (*Anguilla anguilla*) (Fichtner et al. 2013) and carp sampled from China (Shi et al. 2018). The *Caliciviridae* includes two genera that infect fish: *saloviruses* associated with

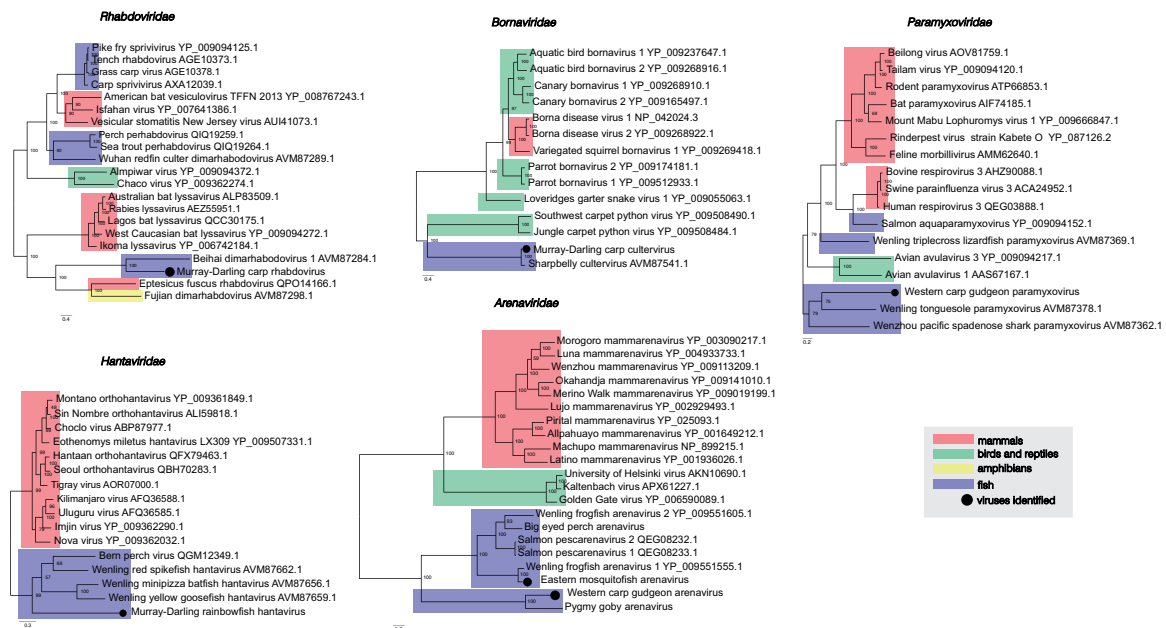


Figure 4. Phylogenetic relationships of negative-sense single-stranded vertebrate-associated viruses identified in this study. Viruses identified here are shown as a black circle. ML trees were estimated using amino acid sequences of the RdRp gene and mid-point rooted for clarity only. Bootstrap values are represented as a percentage with branches scaled to amino acid substitutions per site. Tree branches are highlighted to represent host class: red, mammals; green, birds and reptiles; yellow, amphibians; and blue, fish. Tip labels represent virus name and NCBI/GenBank accession numbers.

salmonid hosts and minoviruses associated with cyprinid hosts (Vinjé et al. 2019). Recently, several caliciviruses have been discovered in ray-finned and jawless fish (Mikalsen et al. 2014; Shi et al. 2018). The novel calicivirus identified in bony herring expands the diversity of fish viruses as it shared eighty per cent amino acid similarity with Atlantic salmon calicivirus (Mikalsen et al. 2014) and clustered with other freshwater fish caliciviruses, including Wuhan carp calicivirus and Guangdong pseudohemiculter dispar calicivirus (Shi et al. 2018).

3.2.3 dsRNA viruses

We identified a novel dsRNA virus in carp in the Castlereagh River that could be assigned to the *Reoviridae*. This divergent virus shared forty per cent RdRp amino acid similarity with its closest relative, *Wenling scaldfish reovirus* (Shi et al. 2018), together forming a clade basal to the genus *Aquareovirus* that are known to cause considerable disease in some fish species (Chen et al. 2018) (Fig. 8).

3.2.4 dsDNA viruses

A key observation of our study was the absence of Cyprinid herpesviruses, including CyHV-3, as well as other *Alloherpesviridae*, in any of the thirty-six RNA libraries. Similarly, although members of the *Hepadnaviridae* are commonly detected in fish (Dill et al. 2016; Lauber et al. 2017; Geoghegan et al. 2018; Geoghegan et al. 2021) they were notably absent in our samples. The only DNA virus detected in this study was a novel poxvirus (*Poxviridae*) identified in western carp-gudgeon. This virus shared DNA polymerase amino acid sequence similarity with salmon gill poxvirus (SGPV) (61%) (Gjessing et al. 2015). We also detected other genomic regions such as DNA-dependent RNA polymerase subunit rpo22 (49.5%), DNA-dependent RNA polymerase subunit rpo19 (40%), DNA binding virion core protein I1L

(28.1%), A16L (32.9%) and SGPV079 (40%) (Supplementary Table S1). Both SGPV and western carp-gudgeon poxvirus form a highly divergent clade within the subfamily *Chordopoxvirinae* that is strongly indicative of virus-host co-divergence (Fig. 9).

3.3 Virome composition, ecological and environmental factors

We next examined whether and how vertebrate virome composition in a range of native and introduced Murray–Darling Basin fish was associated host ecological factors, namely host species, geography (i.e. sampling site), water quality (temperature, pH and turbidity). GLMs revealed host species ($\chi^2 = 7.5^{-6}$, $df = 9$, $P = 0.001$) as the best predictor of viral abundance (i.e. the standardized number of viral sequencing reads) (Fig. 2). In particular, the eastern mosquitofish had significantly higher viral abundance compared to Australian smelt (Tukey: $z = 3.976$, $P = 0.002$), bony herring (Tukey: $z = 4.334$, $P = 0.001$), western carp-gudgeon (Tukey: $z = 4.019$, $P = 0.002$), common carp (Tukey: $z = 4.665$, $P = 0.001$), flat-headed gudgeon (Tukey: $z = 3.632$, $P = 0.001$), goldfish (Tukey: $z = 3.632$, $P = 0.001$) and rainbowfish (Tukey: $z = 4.110$, $P = 0.001$). However, the high viral abundance in the eastern mosquitofish was driven by one sample containing an extremely high abundance of arenaviruses, accounting for seventy-six per cent of its total vertebrate virome and forty-nine per cent of all vertebrate-associated viral reads. We found no evidence for an association between viral abundance and host geography ($P = 0.111$), water turbidity ($P = 0.804$), water temperature ($P = 0.709$) nor water pH ($P = 0.141$).

We calculated alpha diversity to assess any differences in virome composition (abundance and diversity) between hosts and sites. This included the observed virus species richness (the number of viruses found in each sequencing library) and Shannon diversity (both the number of viral families and

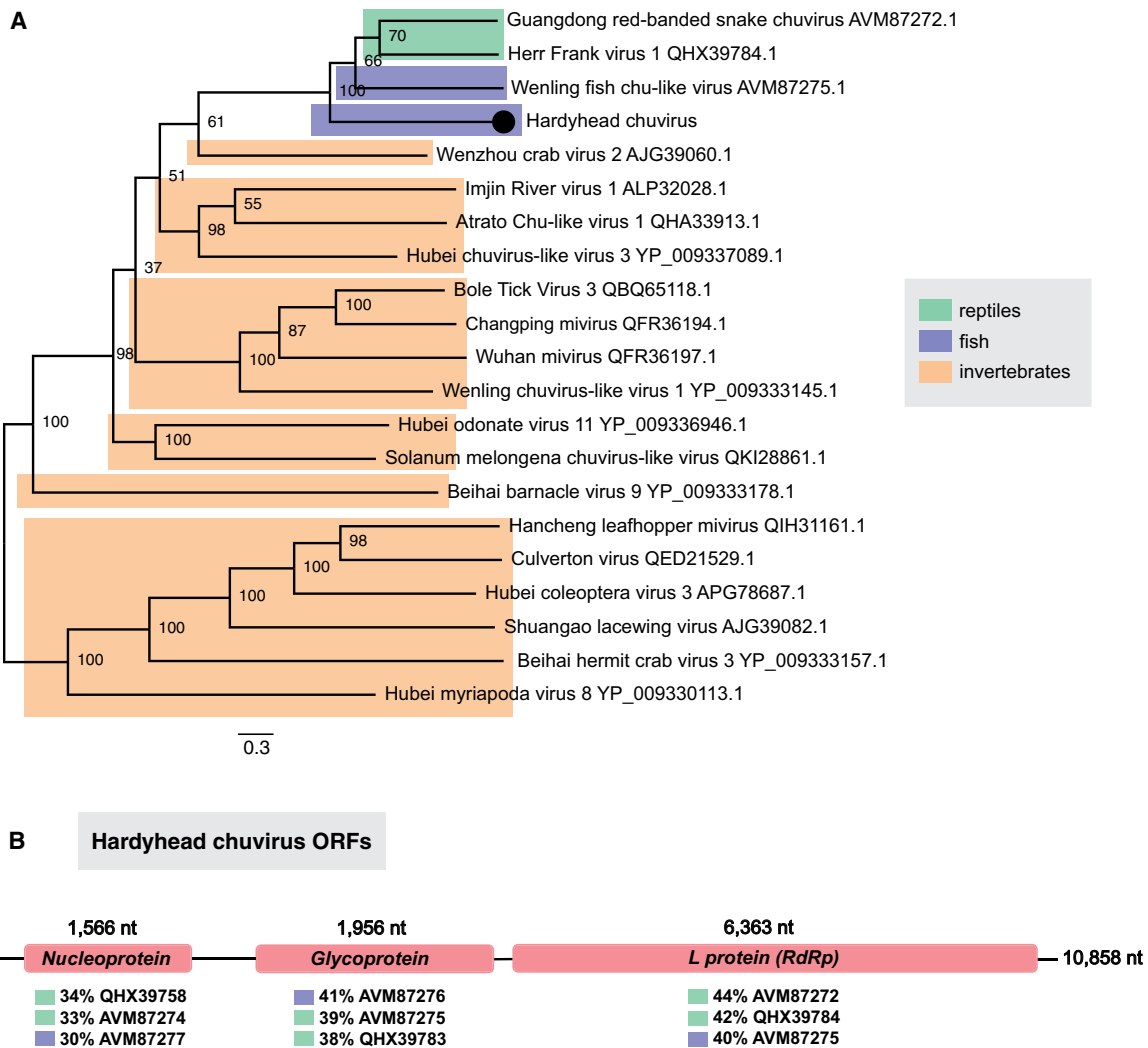


Figure 5. Phylogenetic relationships and genome organization of *hardyhead chuvirus*. (A) Phylogenetic relationships between viruses within the *Chuviridae*. Novel chuvirus identified in the unspecked hardyhead is represented as a black circle. The phylogenetic tree was estimated using amino acid sequences of the RdRp gene and mid-point rooted for clarity only. The scale bar represents amino acid substitutions per site. Bootstrap values are shown as a percentage. Tree branches are highlighted to distinguish between vertebrate and invertebrate-associated viruses: green, reptiles; blue, fish; and orange, invertebrates. Tip labels represent virus name and NCBI/GenBank accession numbers. (B) Genome organization of *hardyhead chuvirus*. Diagram illustrates the structure and length of each genomic segment. Percentages below each segment reveal the closest relatives (NCBI/GenBank accession number), with the L protein more related to reptile chuviruses; glycoprotein more related to fish chuviruses; and nucleoprotein more related to reptile chuviruses.

abundance of viral reads in a given host). We found no association between observed viral species richness and host species ($P = 0.286$), host geography ($P = 0.748$), water turbidity ($P = 0.826$), water temperature ($P = 0.625$) and water pH ($P = 0.115$). Similarly, there was also no observed association between Shannon diversity and host taxonomy ($P = 0.117$), host geography ($P = 0.551$), water turbidity ($P = 0.546$), water temperature ($P = 0.206$) and water pH ($P = 0.039$).

3.4 Virome composition of native versus invasive fish species

While carp and native fish species were sampled together at ten out of thirteen sites (Fig. 1), they shared no vertebrate-

associated viruses. Similarly, no vertebrate-associated viruses were shared among any of the native fish species. Although we identified hepeviruses in common carp and eastern mosquitofish, these were highly divergent and exhibited only twenty per cent amino acid similarity such that they reflect ancient divergence events. These viruses were also distinct in that *Murray-Darling carp hepevirus* clustered with ray-finned fish hosts (Shi et al. 2018), while *eastern mosquitofish hepevirus* formed a distinct basal clade with amphibian (Reuter et al. 2018), jawless fish and cartilaginous fish hosts (Shi et al. 2018) (Fig. 7).

While the native and invasive fish species largely had distinct viromes, two vertebrate-associated virus families were present in both: arenaviruses (western carp-gudgeon and

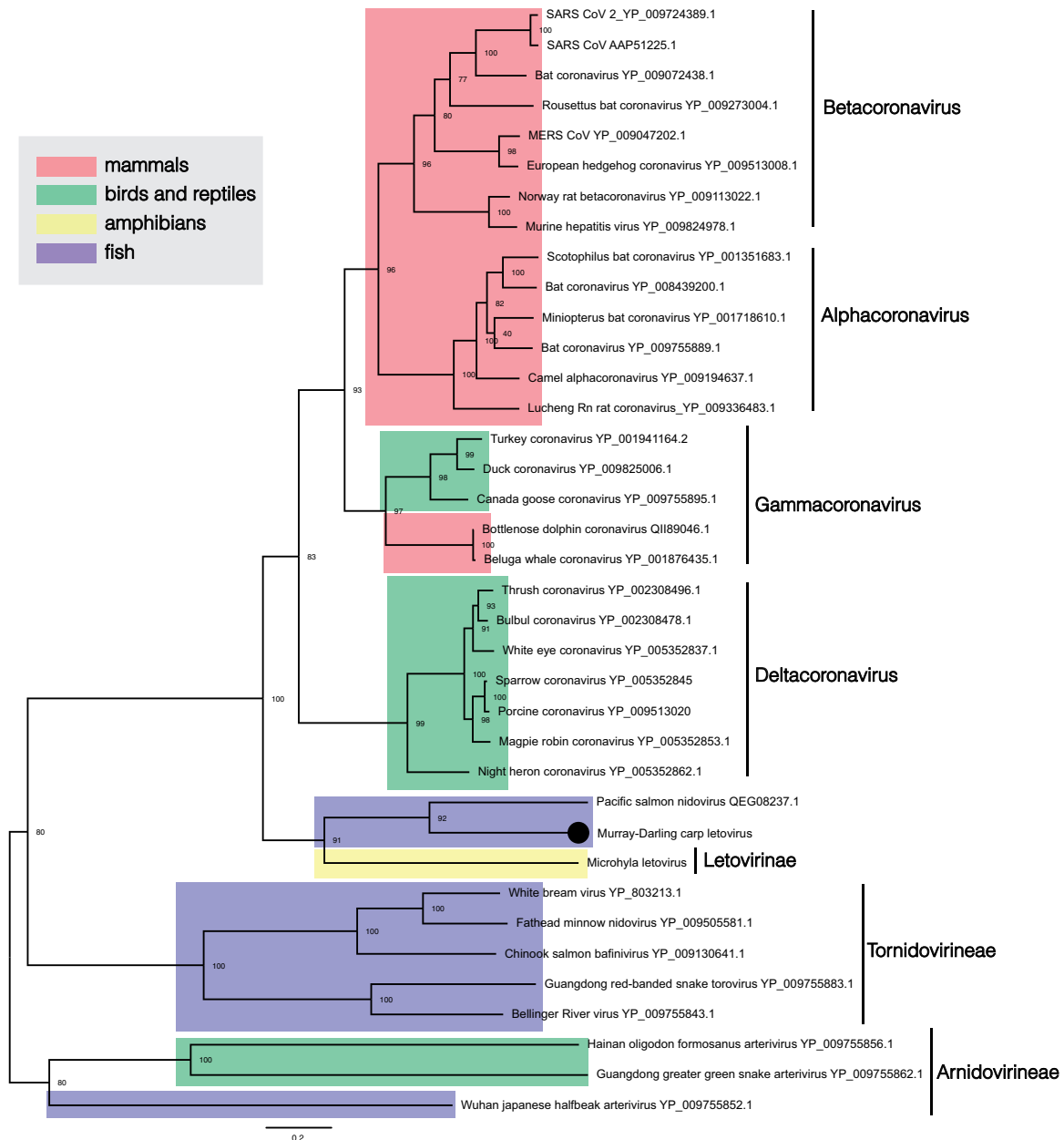


Figure 6. Phylogenetic relationships within the virus order Nidovirales. The novel letovirus detected in common carp is labelled with a black circle. The phylogenetic tree was estimated using the *orf1ab* polypeptide (polymerase associated) and midpoint rooted for clarity alone. Scale bar represents amino acid substitutions per site. Bootstrap values are shown as a percentage. Taxon labels are highlighted: blue, fish; red, mammals; green, birds and reptiles; yellow, amphibians. Virus taxonomic names are labelled to the right. Tip labels represent virus name and NCBI/GenBank IDs.

eastern mosquitofish) and bastroviruses (spangled perch and eastern mosquitofish) (Fig. 3). However, the arenaviruses identified in the eastern mosquitofish and western carp-gudgeon were highly divergent, exhibiting only 27.9 per cent amino acid similarity, with *western carp-gudgeon arenavirus* falling basal to *eastern mosquitofish arenavirus* (Fig. 4). The bastroviruses detected in eastern mosquitofish and spangled perch shared 57.1 per cent RdRp amino acid similarity and formed a distinct clade with other bastrovirus sequences identified in *Culex*

mosquitos (Sadeghi et al. 2018), bats (NCBI/GenBank: NC_035471.1) and sewage samples in Brazil (Dos Anjos, Nagata, and Melo 2017) (Supplementary Fig. S3). Because bastroviruses have genomic features that resemble hepeviruses (Reuter et al. 2018), both these contigs had matches to invertebrate and vertebrate-associated hepeviruses such that their true hosts could not be easily determined.

We also examined whether there were any differences in alpha and beta diversity between native and invasive

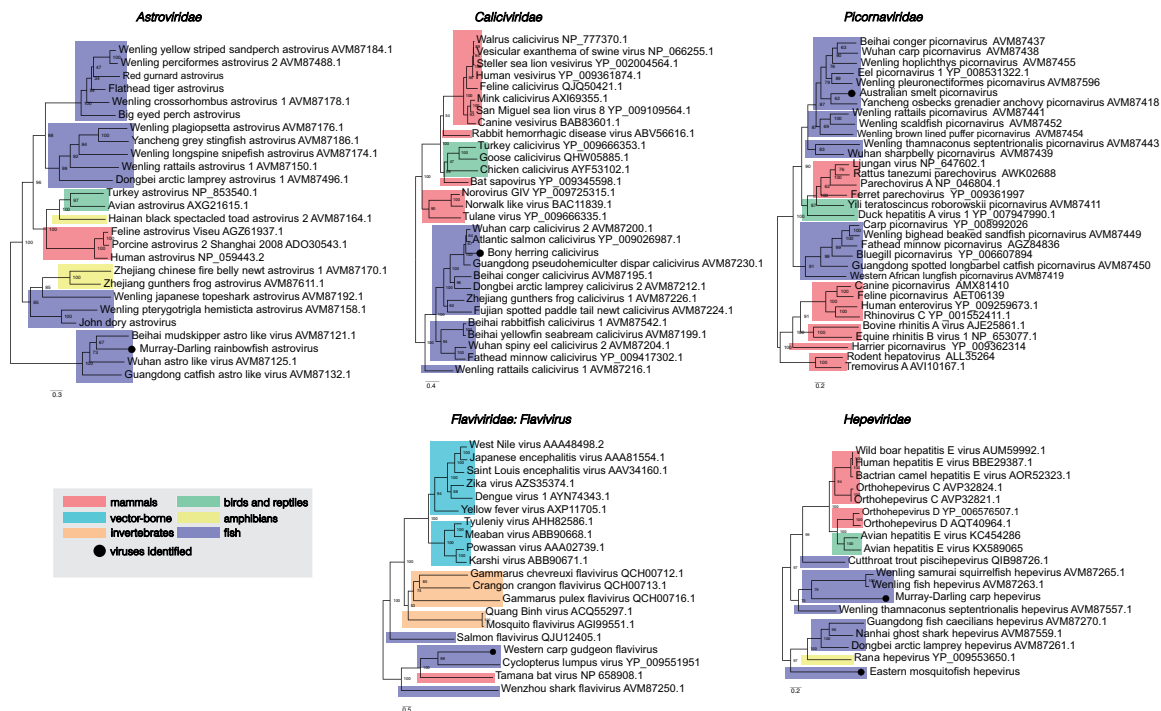


Figure 7. Phylogenetic relationships of positive-sense single-stranded vertebrate-associated viruses identified in this study. Viruses identified here are shown as a black circle. ML trees were estimated using amino acid sequences of the RdRp gene and NS5 gene for the novel flavivirus. Trees were mid-point rooted for clarity only and bootstrap values are represented as a percentage with branches scaled to amino acid substitutions per site. Tree branches are highlighted to represent host class: red, mammals; cyan, vector-borne viruses; green, birds and reptiles; yellow, amphibians; blue, fish; and orange, invertebrates. Tip labels represent virus name and NCBI/GenBank accession IDs.

freshwater fish. Accordingly, we found no association between host origin (i.e. invasive or native) and virome abundance ($P = 0.390$). When assessing alpha diversity, we similarly found no association between host origin and virome richness ($P = 0.626$) nor Shannon diversity ($P = 0.425$). Likewise, this result was also observed when examining beta diversity ($P = 0.602$).

3.5 Associations between host ecology and non-vertebrate viruses

To assess any associations between host ecology and non-vertebrate viruses, we similarly performed GLMs using the aforementioned ecological factors as a negative control. As expected, this revealed no association between non-vertebrate viral abundance and host taxonomy ($P = 0.200$), host geography ($P = 0.101$), host origin ($P = 0.998$), water turbidity ($P = 0.421$), water temperature ($P = 0.282$) and water pH ($P = 0.343$). We similarly found no association between non-vertebrate virome richness and host taxonomy ($P = 0.204$), host geography ($P = 0.090$), host origin ($P = 0.675$), water turbidity ($P = 0.398$), water temperature ($P = 0.072$) and water pH ($P = 0.461$). We also found no evidence for an association between Shannon diversity and host taxonomy ($P = 0.691$), host geography ($P = 0.173$), host origin ($P = 0.876$), water turbidity ($P = 0.571$), water temperature ($P = 0.334$) and water pH ($P = 0.578$). This was also observed when assessing statistical associations between beta diversity and host species ($P = 0.684$), host origin ($P = 0.239$) and host geography ($P = 0.501$).

4. Discussion

Our meta-transcriptomic viral survey of native and invasive fish across the Murray–Darling Basin, Australia, revealed a high diversity and abundance of viruses, including the identification of 45 novel virus species that infected seemingly healthy fish or non-vertebrate hosts in the freshwater environment. Crucially, however, we observed no clear examples of recent cross-species transmission among any fish hosts, including between invasive and native species, nor any evidence for the presence of CyHV-3 from a total of thirty-six RNA sequencing libraries. Hence, these data provide further evidence of the absence of CyHV-3 in Australia (McColl and Crane 2013; McColl, Cooke, and Sunarto 2014; McColl et al. 2017). Similarly, our analysis failed to detect other cypriniviruses (i.e. CyHV-1, CyHV-2), despite previous reports of the presence of CyHV-2 in the Murray–Darling Basin (Stephens, Roldal, and Jones 2004; Becker et al. 2014).

While we observed no cypriniviruses in our analysis, such viruses typically induce latent infections in fish hosts and may become transcriptionally inactive (Michel et al. 2010). Thus, cypriniviruses in a latent stage may be difficult to detect using the transcriptome-based methods described here. Despite this, a viral survey of CyHV-1, CyHV-2 and CyHV-3 was previously conducted on carp across eight sites in the Murray–Darling Basin (McColl and Crane 2013). Using PCR amplicon sequencing, this study similarly failed to identify all three cypriniviruses in 849 carp DNA samples. This suggests that our findings may reflect a true absence of CyHV-3 in Australian waterways, although additional surveillance studies will be required to further address this issue.



Figure 8. Phylogenetic relationships within the *Reoviridae*. Novel reovirus identified here shown as a black circle. ML tree was estimated using amino acid sequences of the RdRp gene. Tree was mid-point rooted for clarity only and bootstrap values are represented as a percentage with branches scaled to amino acid substitutions per site. Tree branches are highlighted to represent host class: red, mammals; cyan, vector-borne viruses; green, birds and reptiles; blue, fish; and orange, invertebrates. Tip labels represent virus name and NCBI/GenBank accession IDs.

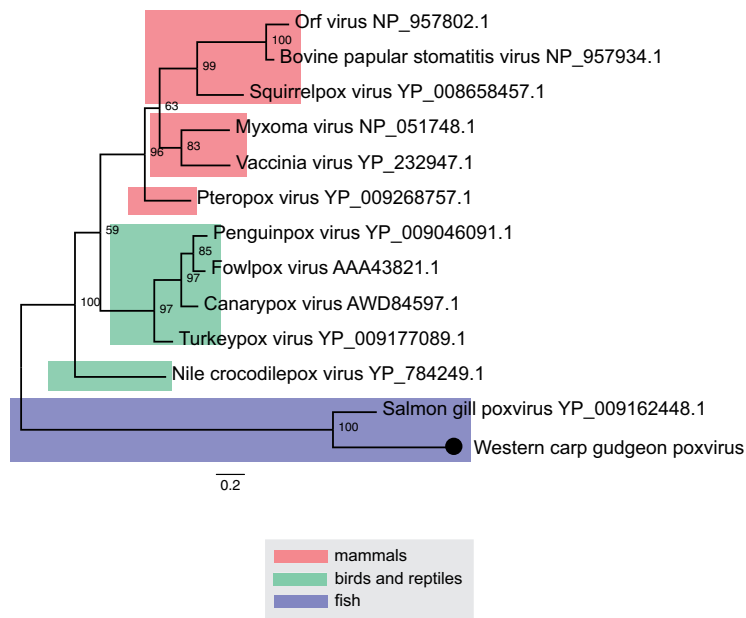


Figure 9. Phylogenetic relationships within the subfamily *Chordopoxvirinae* (family *Poxviridae*). Phylogenetic tree reveals virus-host co-divergence, with fish viruses falling basal to reptilian, avian and mammalian poxviruses. Novel poxvirus identified here shown as a black circle. ML tree was estimated using amino acid sequences of the DNA polymerase gene. Tree was mid-point rooted for clarity only and bootstrap values are represented as a percentage with branches scaled to amino acid substitutions per site. Tree branches are highlighted to represent host class: red, mammals; green, birds and reptiles; blue, fish. Tip labels represent virus name and NCBI/GenBank accession IDs.

Although carp and native fish have co-existed in the Murray–Darling Basin for over 50 years (Forsyth et al. 2013), they shared no vertebrate-associated viruses. The only instance of co-occurrence of viruses from the same family in both invasive and native species was the presence of arenaviruses in native carp-gudgeon and invasive mosquitofish. However, these viruses were so divergent that they likely represent ancient common ancestry rather than recent cross-species transmission (Fig. 4). Eastern mosquitofish, introduced into Australia during the early 1920s to control mosquito populations (Ayres, Pettigrove, and Hoffman 2010), are now widespread across the Murray–Darling Basin and have become a successful invasive species (Macdonald et al. 2012). Their abundance primarily impacts smaller native fish (such as carp-gudgeon, rainbowfish and hardyheads) since they typically outcompete these species and disrupt food webs (Macdonald et al. 2012). As well as being highly divergent, western carp-gudgeon arenavirus formed a basal clade with a recently discovered arenavirus in the pygmy goby sampled from an Australian coral reef (Geoghegan et al. 2021), a member of the same fish order (Gobiiformes). These data further suggest that arenaviruses may have been circulating in Gobiiform fishes (gobies and gudgeons) in Australia prior to the introduction of eastern mosquitofish.

We also identified a novel coronavirus—Murray–Darling carp letovirus—that shared 50.7 per cent amino acid sequence similarity with Pacific salmon nidovirus and 46.2 per cent amino acid similarity with gammacoronaviruses. The Coronaviridae (order Nidovirales) can be split into two subfamilies: the Orthocoronavirinae, associated with birds and mammals, and the Letovirinae associated with amphibians (Bukhari et al. 2018). Both Murray–Darling carp letovirus and Pacific salmon nidovirus formed a sister clade to only member of the Letovirinae subfamily, *Microhyla letovirus* (genus *Alphaletovirus*) identified in the ornamental pygmy frog (*Microhyla fissipes*) (Fig. 6), suggests that fish may be common and ancient hosts for the Letovirinae. It is also notable that Murray–Darling carp letovirus and Pacific salmon nidovirus are highly divergent from the other Nidovirales that are known to infect fish (e.g. Chinook salmon bafinivirus) (Cano et al. 2020).

The phylogenetic range of the Chuviridae largely incorporates invertebrate hosts with diverse genomes (segmented, unsegmented and circular) (Jun-Hua et al. 2015). Recently, chuviruses have been discovered in vertebrates, all possessing three segments (Shi et al. 2018; Argenta et al. 2020). The novel chuvirus detected here in the unspecked hardyhead displayed these genomic features with the L gene (RdRp), S gene (glycoprotein) and N gene (nucleoprotein) all related to fish and reptile viruses (Fig. 5). The phylogenetic position of this vertebrate clade suggests the ancestors of the viruses may be of invertebrate origin, particularly those that inhabit aquatic ecosystems. For instance, the closest related invertebrate viruses were Wenzhou crab virus (Jun-Hua et al. 2015), Imjin River virus (mosquitos) (Hang et al. 2016) and Atrato chu-like virus (mosquitos). Similarly, chuvirus endogenous viral elements have been detected in several freshwater fish species (Shi et al. 2018).

We identified a novel flavivirus in western carp-gudgeon from the Bogan River. This virus falls basal to mammalian vector-borne viruses in phylogenetic trees, grouping with viruses from other vertebrate hosts including *Cyclopterus lumpus virus*, *Tamania bat virus* and *Wenzhou shark flavivirus*. Although western carp-gudgeon flavivirus was detected in apparently healthy fish, in vivo flavivirus replication was recently demonstrated in Chinook salmon (*Oncorhynchus tshawytscha*) that were associated with abnormal mortalities in the Eel River, California (Soto

et al. 2020). While there is still no clear link between flavivirus infection, transmission and disease in aquatic hosts, these data suggest that flaviviruses may be common in fish species. Moreover, the basal phylogenetic positions of aquatic flaviviruses also suggest that these viruses may be the ancestors of notable vector-borne viruses (Fig. 7). Nevertheless, gaps still remain in the evolutionary history of the genus *Flavivirus* and will likely be bridged with additional metagenomic studies.

In broad terms, the evolutionary histories of many vertebrate viral families appear to generally follow patterns of long-term virus–host co-divergence, albeit with regular cross-species transmission (Shi et al. 2018; Zhang et al. 2018). This evolutionary pattern can be observed in the phylogenies of the cultervirus, poxvirus and arenaviruses identified here. The *Bornaviridae* contain three genera with eleven currently classified viral species that infect mammals, birds and reptiles (Amarasinghe et al. 2019). The only fish virus identified to date falls within the genus *Cultervirus*, comprising *Sharpbelly cultervirus* from China (Shi et al. 2018). We identified this virus (i.e. transcripts with 93% L gene amino acid similarity) in common carp in Australia. Intriguingly, both fish hosts are members of the *Cyprinidae* that date as far back as the Cretaceous to Jurassic periods (Hughes et al. 2018; Betancur et al. 2017). Molecular clock dating using endogenous viral elements also showed that culterviruses likely emerged early on during the course of vertebrate evolution, more than fifty million years ago (Shi et al. 2018).

Patterns of long-term virus–host co-divergence can also be seen in the evolutionary history of the *Chordopoxvirinae*. Western carp-gudgeon poxvirus expands the host range of the *Chordopoxvirinae* subfamily within the *Poxviridae*, forming a highly divergent clade with the only other fish-infecting chordopoxvirus discovered to date, SGPV (Fig. 9). Since its classification in 2015, several cases of SGPV have been identified in farmed salmon with complex gill disease, although the reservoir host is unknown (Gjessing et al. 2017). The phylogeny of the *Chordopoxvirinae* mirrors that of vertebrate hosts, strongly suggesting long-term virus–host co-divergence (Fig. 9). Similarly, the phylogeny of the *Arenaviridae* displays a basal fish clade that is characterized by long branches with a large degree of divergence (Fig. 4).

On this evolutionary backbone of ancient virus–host co-divergence, we also detected cases of cross-species virus transmission during evolutionary history, although the time-scales of these events are uncertain. For example, we discovered a novel reovirus that formed a basal divergent clade to other fish viruses within the genus *Aquareovirus* (Fig. 8). Murray–Darling carp reovirus was more closely related to viruses that infect scaldfish (Shi et al. 2018) rather than other cyprinid hosts, which are highly susceptible to reovirus infection (e.g. 80% mortality in grass carp) (Chen et al. 2018). These patterns were also observed in the phylogeny of the *Rhabdoviridae*, with Murray–Darling carp rhabdovirus forming a distinct phylogenetic clade with other recently discovered rhabdoviruses in fish, the big brown bat (NCBI/Genbank: QP014166.1) and the spotted paddle-tail newt (Shi et al. 2018). Rhabdoviruses exhibit a very broad host range including invertebrates, plants, mammals, fish, amphibians, birds and reptiles (Bourhy et al. 2005). Notable among the *Rhabdoviridae* are the lyssaviruses that can cause high mortality in human populations (e.g. *rabies lyssavirus*). Intriguingly, Murray–Darling carp rhabdovirus and its closest relatives form a sister clade to the genus *Lyssavirus*, suggesting these viruses may have a fish-infecting ancestor (Fig. 4).

Although carp are widespread and abundant across the Murray–Darling Basin, they displayed lower viral abundance

than some of the other hosts sampled (Fig. 2). This could be partly explained by the large phylogenetic distance between carp and other fish in the Murray–Darling Basin. For instance, aside from bony herring, all the fish studied here are members of the *Acanthopterygii* (Percomorpha). This could also explain why invasive mosquitofish harboured similar viruses to native gudgeon species (e.g. arenaviruses). Although not always the case (Parrish et al. 2008), cross-species virus transmission often occurs between phylogenetically related hosts, particularly those that display conserved cell receptors (Longdon et al. 2014; Dennehy 2017). In addition, it has been widely suggested that introduced populations are associated with a lower pathogen prevalence and diversity than native species (Phillips et al. 2010; Mark et al. 2003). For example, because invasive species are often established from a small founder population, they likely carry and acquire only a small proportion of pathogens in the novel environment. Once a species rapidly becomes invasive, the diversity of pathogens in this population should remain small, such that the lack of disease likely facilitates the success of invasive species (Mark et al. 2003; Phillips et al. 2010).

It is important to note, however, that there were necessary variations within our sampling. For instance, carp and native fish species were sampled together at ten out of thirteen sites, with carp sampled from all thirteen sites (Fig. 1). In addition, all other fish species were sampled from one to five sites. While an artefact of the distribution of the fish, such gaps obviously limit the power of our statistical analyses and perhaps prevent the detection of ecological associations on virome composition within host species, including between invasive and native fish. In addition, due to animal ethics constraints, we were limited to only a subset of native fish species. Nevertheless, the native species examined in this study are generally those present in the highest densities.

Finally, our analysis detected no viruses that are listed as reportable notifiable aquatic diseases in the Murray–Darling Basin (Australian Health Committee 2020). Such notifiable aquatic diseases include epizootic haematopoietic necrosis virus (EHNV—*Iridoviridae*) and spring viraemia of carp virus (SVCV—*Rhabdoviridae*). EHNV is known to cause high-impact infections in redfin perch and is capable of infecting other freshwater fish in the Murray–Darling Basin, including silver perch (*Bidyanus bidyanus*), Macquarie perch (*Macquaria australasica*), Murray–Darling rainbowfish, freshwater catfish (*Tandanus tandanus*) and invasive mosquitofish (Becker et al. 2013). Although thought to be endemic to the Murray–Darling Basin (upper Murrumbidgee River), EHNV was last reported in 2012 (Becker et al. 2019). Similarly, we did not detect the emerging dwarf gourami iridovirus (*Iridoviridae*) that causes infectious spleen and kidney necrosis in several species of native Australian fish (Go et al. 2006; Rimmer et al. 2017).

In sum, our metagenomic surveillance revealed a marked lack of virus exchange between native and invasive fish species in the Murray–Darling Basin, including those viruses found in invasive common carp. At face value, these data suggest that there is minimal virus transmission from common carp to native fish species, although more extensive sampling is needed to fully address this issue. By investigating the viromes of native and invasive fish, we provide the first data on viruses that naturally circulate in a 2,200 km river system, enhancing our understanding of the evolutionary history of vertebrate viruses.

Supplementary data

Supplementary data are available at Virus Evolution online.

Acknowledgements

We thank Justin Stanger, Patrick Martin, Karen Scott, Michael Rodgers, Duncan McLay and Nick O'Brien from the Department of Primary Industries for their help with fish sampling. We also thank efishalbum.com for providing us with the fish images used in Figs. 1 and 3, which were used with permission.

Funding

This work was funded by a Recreational Fishing Trust grant (SF024) awarded to J.L.G., J.E.W. and E.C.H.; an ARC Discovery grant (DP200102351) awarded to E.C.H. and J.L.G.; a Macquarie University grant awarded to J.L.G.; and the Department of Biological Sciences at Macquarie University. E.C.H. is funded by ARC Australian Laureate Fellowship (FL170100022). J.L.G. is funded by a New Zealand Royal Society Rutherford Discovery Fellowship (RDF-20-UOO-007).

Data availability

All sequence reads generated in this project are available under the NCBI Short Read Archive (SRA) under BioProject PRJNA701716 and all consensus virus genetic sequences have been deposited in GenBank under accessions MW645018–MW645046.

Conflict of interest: None declared.

References

- Abudurexiti, A. et al. (2019) 'Taxonomy of the Order *Bunyavirales*: Update 2019', *Official Journal of the Virology Division of the International Union of Microbiological Societies*, 164: 1949–65.
- Amarasinghe, G. et al. (2019) 'Taxonomy of the Order *Mononegavirales*: Update 2019', *Official Journal of the Virology Division of the International Union of Microbiological Societies*, 164: 1967–80.
- Argenta, F. F. et al. (2020) 'Identification of Reptarenaviruses, Hartmaniviruses, and a Novel Chuvirus in Captive Native Brazilian Boa Constrictors with Boid Inclusion Body Disease', *Journal of Virology*, 94: e00001–20.
- Australian Health Committee. (2020) Australia's National List of Reportable Diseases of Aquatic Animals, <<https://www.agriculture.gov.au/animal/aquatic/reporting/reportable-diseases#fnfish>> accessed 23 Feb 2021.
- Ayres, R., Pettigrove, V., and Hoffmann, A. (2010) 'Low Diversity and High Levels of Population Genetic Structuring in Introduced Eastern Mosquitofish (*Gambusia holbrooki*) in the Greater Melbourne Area, Australia', *Biological Invasions*, 12: 3727–44.
- Barrett, J., Bamford, H., and Jackson, P. (2014) 'Management of Alien Fishes in the Murray–Darling Basin', *Ecological Management & Restoration*, 15: 51–6.
- Becker, J. A. et al. (2013) 'Experimental Infection of Australian Freshwater Fish with Epizootic Haematopoietic Necrosis Virus (EHNV)', *Journal of Aquatic Animal Health*, 25: 66–76.
- et al. (2014) 'Incursions of Cyprinid Herpesvirus 2 in Goldfish Populations in Australia despite Quarantine Practices', *Aquaculture*, 432: 53–9.
- et al. (2019) 'Geographic Distribution of (EHNV) in Freshwater Fish in South Eastern Australia: Lost Opportunity

- for a Notifiable Pathogen to Expand Its Geographic Range', *Viruses*, 11: 315.
- Bergmann, S. M. et al. (2010) 'Susceptibility of Koi × Crucian Carp and Koi × Goldfish Hybrids to Koi Herpesvirus (KHV) and the Development of KHV Disease (KHVD)', *Journal of Fish Diseases*, 33: 267–72.
- Betancur, R. R. et al. (2017) 'Phylogenetic Classification of Bony Fishes', *BMC Evolutionary Biology*, 17:
- Bolger, A. M., Lohse, M., and Usadel, B. (2014) 'Trimmomatic: A Flexible Trimmer for Illumina Sequence Data', *Bioinformatics*, 30: 2114–20.
- Bourhy, H. et al. (2005) 'Phylogenetic Relationships among Rhabdoviruses Inferred Using the L Polymerase Gene', *The Journal of General Virology*, 86(Pt 10): 2849–58.
- Buchfink, B., Xie, C., and Huson, D. H. (2014) 'Fast and Sensitive Protein Alignment Using DIAMOND', *Nature Methods*, 12: 59–60.
- Bukhari, K. et al. (2018) 'Description and Initial Characterization of Metatranscriptomic Nidovirus-like Genomes from the Proposed New Family Abyssoviridae, and from a Sister Group to the Coronavirinae, the Proposed Genus Alphaletovirus', *Virology*, 524: 160–71.
- Cano, I. et al. (2020) 'Isolation of a Chinook Salmon Bafinivirus (CSBV) in Imported Goldfish L. in the United Kingdom and Evaluation of Its Virulence in Resident Fish Species', *Viruses*, 12: 578.
- Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009) 'trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses', *Bioinformatics*, 25: 1972–3.
- Chen, G. et al. (2018) 'Transcriptomics Sequencing Provides Insights into Understanding the Mechanism of Grass Carp Reovirus Infection', *International Journal of Molecular Sciences*, 19: 488.
- de Lamballerie, X. et al. (2002) 'Genome Sequence Analysis of Tamana Bat Virus and Its Relationship with the Genus Flavivirus', *The Journal of General Virology*, 83(Pt 10): 2443–54.
- Dennehy, J. J. (2017) 'Evolutionary Ecology of Virus Emergence', *Annals of the New York Academy of Sciences*, 1389: 124–46.
- Dill, J. A. et al. (2016) 'Distinct Viral Lineages from Fish and Amphibians Reveal the Complex Evolutionary History of Hepadnaviruses', *Journal of Virology*, 90: 7920–33.
- Dixon, P. (2003) 'VEGAN, a Package of R Functions for Community Ecology', *Journal of Vegetation Science*, 14: 927–30.
- Dos Anjos, K., Nagata, T., and Melo, F. L. d. (2017) 'Complete Genome Sequence of a Novel Bastrovirus Isolated from Raw Sewage', *Genome Announcements*, 5:
- Dudgeon, D. et al. (2006) 'Freshwater Biodiversity: Importance, Threats, Status and Conservation Challenges', *Biological Reviews*, 81: 163–82.
- Fichtner, D. et al. (2013) 'Characterization of a Novel Picornavirus Isolate from a Diseased European Eel (*Anguilla anguilla*)', *Journal of Virology*, 87: 10895–9.
- Forsyth, D. et al. (2013) 'Population Dynamics of Invading Freshwater Fish: Common Carp (*Cyprinus Carpio*) in the Murray-Darling Basin, Australia', *Biological Invasions*, 15: 341–54.
- Geoghegan, J. L. et al. (2018) 'Hidden Diversity and Evolution of Viruses in Market Fish', *Virus Evolution*, 4: vey031.
- et al. (2021) 'Virome Composition in Marine Fish Revealed by Meta-Transcriptomics', *Virus Evolution*, 7: veab005.
- Gjessing, M. C. et al. (2015) 'Salmon Gill Poxvirus, the Deepest Representative of the Chordopoxvirinae', *Journal of Virology*, 89: 9348–67.
- et al. (2017) 'Salmon Gill Poxvirus, a Recently Characterized Infectious Agent of Multifactorial Gill Disease in Freshwater- and Seawater-Reared Atlantic Salmon', *Journal of Fish Disease*, 40: 1253–65.
- Go, J. et al. (2006) 'The Molecular Epidemiology of Iridovirus in Murray Cod (*Maccullochella Peilii Peilii*) and Dwarf Gourami (*Colisa Lalia*) from Distant Biogeographical Regions Suggests a Link between Trade in Ornamental Fish and Emerging Iridoviral Diseases', *Molecular and Cellular Probes*, 20: 212–22.
- Haas, B. J. et al. (2013) 'De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis', *Nature Protocols*, 8: 1494.
- Hang, J. et al. (2016) 'Genome Sequences of Five Arboviruses in Field-Captured Mosquitoes in a Unique Rural Environment of South Korea', *Genome Announcements*, 1: 4.
- Hayes, K., R. et al. (2014) 'Meeting the Challenge of Quantitative Risk Assessment for Genetic Control Techniques: A Framework and Some Methods Applied to the Common Carp (*Cyprinus Carpio*) in Australia', *Biological Invasions*, 16: 1273–88.
- Hothorn, T., Bretz, F., and Westfall, P. (2008) 'Simultaneous Inference in General Parametric Models', *Biometrical Journal*, 50: 346–63.
- Hughes, L., C. et al. (2018) 'Comprehensive Phylogeny of Ray-Finned Fishes (Actinopterygii) Based on Transcriptomic and Genomic Data', *Proceedings of the National Academy of Sciences*, 115: 6249–54.
- Ilouze, M. et al. (2011) 'The Outbreak of Carp Disease Caused by CyHV-3 as a Model for New Emerging Viral Diseases in Aquaculture: A Review', *Ecological Research*, 26: 885–92.
- Johnson, P. T. J., and Paull, S. H. (2011) 'Ecology and Emergence of Diseases in Fresh Waters', *Freshwater Biology*, 56: 638–57.
- Jun-Hua, T. et al. (2015) 'Unprecedented Genomic Diversity of RNA Viruses in Arthropods Reveals the Ancestry of Negative-Sense RNA Viruses', *eLife*, 4: e05378.
- Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods*, 14: 587–9.
- Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.
- Kearse, M. et al. (2012) 'Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data', *Bioinformatics*, 28: 1647–9.
- Koehn, J. D. (2004) 'Carp (*Cyprinus carpio*) as a Powerful Invader in Australian Waterways', *Freshwater Biology*, 49: 882–94.
- Kopf, R. K. et al. (2019) 'Biocontrol in Australia: Can a Carp Herpesvirus (CyHV-3) Deliver Safe and Effective Ecological Restoration?', *Biological Invasions*, 21: 1857–70.
- Lagkovardos, I. et al. (2017) 'Rhea: A Transparent and Modular R Pipeline for Microbial Profiling Based on 16S rRNA Gene Amplicons', *PeerJ*, 5: e2836.
- Lauber, C. et al. (2017) 'Deciphering the Origin and evolution of Hepatitis B Viruses by Means of a Family of Non-Enveloped Fish Viruses', *Cell Host & Microbe*, 22: 387–99.
- Longdon, B. et al. (2014) 'The Evolution and Genetics of Virus Host Shifts', *PLoS Pathogens*, 10: e1004395.
- Macdonald, J. I. et al. (2012) 'Do Invasive Eastern Gambusia Shape Wetland Fish Assemblage Structure in South-Eastern Australia?', *Marine and Freshwater Research*, 63: 659–71.
- Mark, E. T. et al. (2003) 'Introduced Species and Their Missing Parasites', *Nature*, 421: 628–30.
- Matsui, K. et al. (2008) 'Detection and Significance of Koi Herpesvirus (KHV) in Freshwater Environments', *Freshwater Biology*, 53: 1262–72.
- McColl, K. A., Cooke, B. D., and Sunarto, A. (2014) 'Viral Biocontrol of Invasive Vertebrates: Lessons from the past

- Applied to Cyprinid Herpesvirus-3 and Carp (*Cyprinus carpio*) Control in Australia', *Biological Control*, 72: 109–17.
- , and rane, M. S. J. (2013) *Cyprinid Herpesvirus 3 (CyHV-3): Its Potential as a Biological Control Agent for Carp in Australia*, PestSmart Toolkit Publication, Invasive Animals Cooperative Research Centre. https://pestsmart.org.au/wp-content/uploads/sites/3/2020/06/McCollCrane2012_KHV_ResearchReport.pdf
- et al. (2017) 'Cyprinid Herpesvirus 3 as a Potential Biological Control Agent for Carp (*Cyprinus carpio*) in Australia: Susceptibility of Non-Target Species', *Journal of Fish Diseases*, 40: 1141–53.
- McMurdie, P. J., and Holmes, S. (2013) 'Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data', *PLoS ONE*, 8: e61217.
- Michel, B. et al. (2010) 'Cyprinid Herpesvirus 3', *Emerging Infectious Diseases*, 16: 1835–43.
- Mihindukulasuriya, K., A. et al. (2008) 'Identification of a Novel Coronavirus from a Beluga Whale by Using a Panviral Microarray', *The Journal of Virology*, 82: 5084–8.
- Mikalsen, A., B. et al. (2014) 'Characterization of a Novel Calicivirus Causing Systemic Infection in Atlantic Salmon (*Salmo salar* L.): Proposal for a New Genus of Caliciviridae', *PLoS ONE*, 9: e107132.
- Mordecai, G. J. et al. (2019) 'Endangered Wild Salmon Infected by Newly Discovered Viruses', *eLife*, 8: e47615.
- Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.
- Parrish, C., R. et al. (2008) 'Cross-Species Virus Transmission and the Emergence of New Epidemic Diseases', *Microbiology and Molecular Biology Reviews*, 72: 457.
- Phillips, B., L. et al. (2010) 'Parasites and Pathogens Lag behind Their Host during Periods of Host Range Advance', *Ecology*, 91: 872–81.
- Reuter, G. et al. (2018) 'Detection of a Novel RNA Virus with Hepatitis E Virus-like Non-Structural Genome Organization in Amphibian, Agile Frog (*Rana dalmatina*) Tadpoles', *Infection, Genetics and Evolution*, 65: 112–6.
- Rimmer, A., E. et al. (2017) 'Susceptibility of a Number of Australian Freshwater Fishes to Dwarf Gourami Iridovirus (Infectious Spleen and Kidney Necrosis Virus)', *Journal of Fish Diseases*, 40: 293–310.
- Sadeghi, M. et al. (2018) 'Virome of >12 Thousand Culex Mosquitoes from throughout California', *Virology*, 523: 74–88.
- Sahul, H., A. S. et al. (2019) 'ICTV Virus Taxonomy Profile: Nodaviridae', *Journal of General Virology*, 100: 3–4.
- Shi, M. et al. (2018) 'The Evolutionary History of Vertebrate RNA Viruses', *Nature*, 561: 197–202.
- et al. (2016) 'Redefining the Invertebrate RNA Virophere', *Nature*, 540: 539–43.
- , Zhang, Y.-Z., and Holmes, E. C. (2018) 'Meta-Transcriptomics and the Evolutionary Biology of RNA Viruses', *Virus Research*, 243: 83–90.
- Shimizu, T. et al. (2006) 'Survival of Koi Herpesvirus (KHV) in Environmental Water', *Fish Pathology*, 4: 153–7.
- Skoge, R. et al. (2018) 'New Virus of the Family Flaviviridae Detected in Lumpfish (*Cyclopterus lumpus*)', *Official Journal of the Virology Division of the International Union of Microbiological Societies*, 163: 679–85.
- Soto, E. et al. (2020) 'First Isolation of a Novel Aquatic Flavivirus from Chinook Salmon (*Oncorhynchus tshawytscha*) and Its Replication in a Piscine Animal Model', *Journal of Virology*, 94: e00337–20.
- Stephens, F., Raidal, Sr., and Jones, B. (2004) 'Haematopoietic Necrosis in a Goldfish (*Carassius auratus*) Associated with an Agent Morphologically Similar to Herpesvirus', *Australian Veterinary Journal*, 82: 167–9.
- Tolo, I. E. et al. (2021) 'Susceptibility of *Pimephales promelas* and *Carassius auratus* to a Strain of Koi Herpesvirus Isolated from Wild *Cyprinus carpio* in North America', *Scientific Reports*, 11: 1985.
- Tordo, N. et al. (1988) 'Completion of the Rabies Virus Genome Sequence Determination: Highly Conserved Domains among the L (Polymerase) Proteins of Unsegmented Negative-Strand RNA Viruses', *Virology*, 165: 565–76.
- Valero-Mora, P. M. (2010) 'ggplot2: Elegant Graphics for Data Analysis', *Journal of Statistical Software*, 35:
- Vilizzi, L. et al. (2014) 'Ecological Effects of Common Carp (*Cyprinus carpio*) in a Semi-Arid Floodplain Wetland', *Marine and Freshwater Research*, 65: 802–17.
- Vinje, J. et al. (2019) 'ICTV Virus Taxonomy Profile: Caliciviridae', *The Journal of General Virology*, 100: 1469–70.
- Woo, P. C. Y. et al. (2014) 'Discovery of a Novel Bottlenose Dolphin Coronavirus Reveals a Distinct Species of Marine Mammal Coronavirus in Gammacoronavirus', *Journal of Virology*, 88: 1318–31.
- World Wildlife Fund. (2021) The World's Forgotten Fishes <https://www.wwf.panda.org/discover/our_focus/freshwater_practice/> accessed 23 Feb 2021.
- Zhang, Y.-Z., Shi, M., and, and Holmes, E. C. (2018) 'Using Metagenomics to Characterize an Expanding Virophere', *Cell*, 172: 1168–72.
- et al. (2018) 'The Diversity, Evolution and Origins of Vertebrate RNA Viruses', *Current Opinion in Virology*, 31: 9–16.

Slippery when wet: cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the *Coronaviridae*

Allison K. Miller,^{1,†} Jonathon C.O. Mifsud^{2,†} Vincenzo A. Costa,^{2,†} Rebecca M. Grimwood,³ Jane Kitson,⁴ Cindy Baker,⁵ Cara L. Brosnahan,⁶ Anjali Pande,^{5,6} Edward C. Holmes,² Neil J. Gemmell,¹ and Jemma L. Geoghegan^{3,7,*}

¹Department of Anatomy, University of Otago, Dunedin 9016, New Zealand, ²Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, Johns Hopkins Drive, The University of Sydney, Sydney, NSW 2006, Australia, ³Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand, ⁴Kitson Consulting Ltd 9 Black Road, Invercargill 9879, Invercargill/Waihopai, New Zealand, ⁵National Institute of Water and Atmospheric Research, Hamilton 3216, New Zealand, ⁶Animal Health Laboratory and Diagnostic and Surveillance Directorate, Ministry for Primary Industries, Upper Hutt 5018, New Zealand and ⁷Institute of Environmental Science and Research, 34 Kenepuru Drive, Kenepuru, Porirua 5022 Wellington 5018, New Zealand

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: jemma.geoghegan@otago.ac.nz

Abstract

The *Nidovirales* comprise a genetically diverse group of positive-sense single-stranded RNA virus families that infect a range of invertebrate and vertebrate hosts. Recent metagenomic studies have identified nido-like virus sequences, particularly those related to the *Coronaviridae*, in a range of aquatic hosts including fish, amphibians, and reptiles. We sought to identify additional members of the *Coronaviridae* in both bony and jawless fish through a combination of total RNA sequencing (meta-transcriptomics) and data mining of published RNA sequencing data and from this reveal more of the long-term patterns and processes of coronavirus evolution. Accordingly, we identified a number of divergent viruses that fell within the *Letovirinae* subfamily of the *Coronaviridae*, including those in a jawless fish—the pouched lamprey. By mining fish transcriptome data, we identified additional virus transcripts matching these viruses in bony fish from both marine and freshwater environments. These new viruses retained sequence conservation in the RNA-dependant RNA polymerase across the *Coronaviridae* but formed a distinct and diverse phylogenetic group. Although there are broad-scale topological similarities between the phylogenies of the major groups of coronaviruses and their vertebrate hosts, the evolutionary relationship of viruses within the *Letovirinae* does not mirror that of their hosts. For example, the coronavirus found in the pouched lamprey fell within the phylogenetic diversity of bony fish letoviruses, indicative of past host switching events. Hence, despite possessing a phylogenetic history that likely spans the entire history of the vertebrates, coronavirus evolution has been characterised by relatively frequent cross-species transmission, particularly in hosts that reside in aquatic habitats.

Key words: meta-transcriptomics; virus discovery; fish; evolution; phylogeny; *Coronaviridae*; lamprey; coronavirus

1. Introduction

The *Coronaviridae* are a family of enveloped, positive-strand RNA viruses that have a broad host range, although they are predominantly associated with mammals and birds. Several coronaviruses have emerged as highly pathogenic viruses in humans and other animals, the most recent and notable of which is *severe acute respiratory syndrome coronavirus-2* (SARS-CoV-2) (Wu et al. 2020), which, at the time of writing, has infected over 150 million people since its initial identification at the end of 2019. Coronaviruses, like other families within the order *Nidovirales*, possess the longest genomes seen in RNA viruses (27–34 kb in the case of the coronaviruses) and seemingly replicate with higher fidelity than other known positive-sense RNA viruses due to RNA proof-reading associated with an exonuclease domain (Denison et al. 2011).

Most coronaviruses are classified within the sub-family *Orthocoronavirinae* that infect mammals and birds as well as a

single virus identified in a reptilian host (Shi et al. 2018). However, our understanding of the host range of the *Coronaviridae* has recently been enhanced by large-scale metagenomic studies, including the identification of a distinct second sub-family: the *Letovirinae* (*Coronaviridae* Study Group of the International Committee on Taxonomy of Viruses, Gorbalenya et al. 2020). Specifically, an analysis of published transcriptome data identified a novel coronavirus in the ornamented pygmy frog (*Microhyla fissipes*) termed *Microhyla letovirus* (Bukhari et al. 2018). A subsequent virological survey of dead and moribund cultured Chinook salmon (*Oncorhynchus tshawytscha*) identified *Pacific salmon nidovirus* (Mordecai et al. 2019). Phylogenetic analysis revealed that *Microhyla letovirus* and *Pacific salmon nidovirus* formed a sister group to all other known coronaviruses, with the large genetic divergence between them suggesting that this group is likely far larger and under-sampled (Bukhari et al. 2018; Mordecai et al. 2019). It is therefore unsurprising that more recent probes of

public sequencing data have identified additional coronavirus-like sequences in non-mammalian aquatic vertebrate hosts, namely in fish and amphibian transcriptomes (Edgar et al. 2020). Indeed, this extensive data mining study identified six novel viruses within the *Letovirinae*: five in bony fish and one in the axolotl (*Ambystoma mexicanum*) (Edgar et al. 2020). More recently, an additional virus identified in the common carp (*Cyprinus carpio*) in Australia was also found to fall within this group (Costa et al. 2021).

The central aim of the current study was to determine whether coronaviruses might be present in a wider range of fish hosts and from this reveal more of their evolutionary history, particularly whether coronaviruses have co-diverged with their hosts over many millions of years or whether there has been frequent cross-species transmission (i.e. host jumping) as underpins the emergence of SARS-CoV-2 and whether such cross-species transmission occurs in some hosts more frequently than others.

To this end, we screened for coronaviruses using total RNA sequencing (i.e. meta-transcriptomic) data from jawless and bony fish. Jawless fish included the pouched lamprey (*Geotria australis*) sampled from New Zealand—te reo Māori: kanakana (South Island) or piharau (North Island)—that displayed lamprey reddening syndrome (LRS). This syndrome is characterised by red lesions on their gills, fins, and body walls (Brosnahan et al. 2019), but a causative agent has yet to be identified. We combined this investigation with data mining of published RNA sequencing data sampled from fish within the National Centre for Biotechnology Information (NCBI) Sequence Read Archive (SRA). This enabled us to identify new coronavirus sequences that not only significantly expanded the phylogenetic diversity of the coronaviruses but that provided important new insights into the fundamental patterns and processes of viral evolution in this virus family.

2. Results

2.1 Identification of a novel coronavirus in the pouched lamprey

Following a disease investigation into LRS in pouched lamprey from New Zealand by the New Zealand Ministry for Primary Industries (MPI) in 2012, 12 individuals (eight from this investigation and four more recently collected control samples) were subjected to meta-transcriptomic analysis. This led to the identification of a novel and divergent member of the *Coronaviridae* in several individuals. We have named this new virus kanakana letovirus, and as each virus shared >97 per cent sequence similarity at the nucleotide level, they are likely genomic variants of the same virus species.

We detected sequence reads that shared sequence identity to coronaviruses in seven of the eight LRS-affected individuals (confirmed by reverse transcription-polymerase chain reaction (RT-PCR)). No coronaviruses or other viruses were found in the remaining LRS sample nor the four control samples. However, we were only able to generate consensus virus sequences from three of these due to the low and fragmented virus abundance in the other samples. All three samples possessed virus transcripts with sequence similarity to the open reading frames 1a (ORF1a) and 1b (ORF1b) common to the *Coronaviridae*, the latter of which contains the viral RNA-dependent RNA polymerase (RdRp) (Table 1). We also identified transcripts with marked sequence similarity to the *Nidovirales* spike (S) protein. However, the closest sequence match for the S protein transcripts were viruses from *Porcine torovirus* and

Atlantic salmon bafinivirus, both non-coronavirus members of the *Nidovirales* (family *Tobamiviridae*), albeit these were highly divergent sequences with only ~25–27 per cent amino acid sequence similarity (Table 1). Phylogenetic comparisons of the ORF1ab and S genomic regions of kanakana letovirus in the context of its closest relatives illustrated this striking incongruence (Fig. 1). In particular, both *Microhyala letovirus* and kanakana letovirus were more closely related to coronaviruses in the ORF1ab region but not in the S protein where they exhibited greater sequence similarity to the *Tobamiviridae*, suggestive of past recombination events (Fig. 1). In contrast, *Pacific salmon nidovirus* consistently fell as a sister group to the *Coronaviridae* in both gene regions, although with long branch lengths indicative of long periods of evolutionary divergence (Fig. 1).

All three kanakana letovirus sequencing libraries had similar sequencing depths as illustrated by the standardised abundance of the host gene, RSP13, which was detected at a lower abundance than kanakana letovirus (Fig. 2). However, the kanakana letovirus in samples 1 and 3 were the most abundant across the ORF1ab and S proteins. The presence of kanakana letovirus in seven of the individual lampreys was confirmed by RT-PCR (Supplementary Fig. S1). No other viruses were detected in these data. An amino acid sequence comparison with *Microhyala letovirus* across all three genomic regions shows relatively high sequence similarity in ORF1b (Fig. 2).

2.2 Identification of novel coronaviruses in published SRA data from bony fish

Data mining the SRA for coronavirus-like sequences revealed several potentially novel virus transcripts within sequences obtained from bony fish (Table 2). We detected coronavirus-like reads in 32 sequencing libraries (of the 1279 that met our search criteria) that were further investigated. First, we identified four new viruses that again shared ORF1b sequence similarity to *Pacific salmon nidovirus* (52–56 per cent). These sequences were identified in blenny (*Acanthemblemaria* sp.), bluespotted mud hopper (*Boleophthalmus pectinirostris*), *Schizothorax waltoni*, which were sampled from China, as well as the broadnosed pipefish (*Syngnathus typhle*), sampled from Europe. The new viruses were named blenny letovirus, mud hopper letovirus, waltoni letovirus, and broadnosed pipefish letovirus, respectively. Second, we identified a viral transcript in the trout perch (*Percopsis omiscomaycus*) that shared 39 per cent amino acid sequence similarity to *Shrew coronavirus* in ORF1b and 20 per cent amino acid sequence similarity to *Bat coronavirus* in the S protein (Table 2), although it is important to note that the vast majority of viruses in the *Letovirinae* are not yet included in the non-redundant (nr) protein database. This was termed 'trout perch letovirus'. Finally, two potentially novel virus sequences were identified that, upon further inspection, shared greater sequence similarity to viruses within the *Tobamiviridae*, in a similar manner to the S protein of kanakana letovirus. These two novel viruses, found in giant mudskipper (*Periophthalmodon schlosseri*) and golden Chinese loach (*Botia supercilialis*), had >50 per cent sequence similarity in the RdRp and ~40 per cent sequence similarity in the S protein to other fish viruses in the *Tobamiviridae* (see Table 1). These new viruses were named giant mudskipper bafinivirus and golden Chinese loach bafinivirus, respectively. The remaining 25 libraries that were further investigated did not possess transcripts with sequence similarity to the *Coronaviridae*.

Table 1. Novel virus transcripts identified in this study from pouched lamprey (*Geotria australis*), their corresponding closest genetic match on GenBank, and standardised abundances.

Virus	LRS symptoms	Host	Gene	Sequence length (nt)	Closest match on GenBank (accession); e-value	Sequence identity (%)	Standardised viral abundance	RSP13 abundance
<i>Kanakana letovirus</i> (1)	Light haemorrhaging around gill slits.	<i>Geotria australis</i>	ORF1a	9,270	Pacific salmon nidovirus ORF1a (QEG08236.1); e-value = 1×10^{-155}	26.36	6.86×10^{-5}	5.52×10^{-6}
			ORF1b	6,507	Pacific salmon nidovirus ORF1b (QEG08237.1); e-value = 0	46.53	4.67×10^{-5}	
			Spike	5,022	Porcine torovirus spike protein (QJ02013.1); e-value = 2×10^{-10}	27	4.24×10^{-5}	
<i>Kanakana letovirus</i> (2)	Small area of haemorrhaging near the ventral fin.	<i>Geotria australis</i>	ORF1a	9,255	Pacific salmon nidovirus ORF1a (QEG08236.1); e-value = 5×10^{-49}	27.81	1.48×10^{-5}	6.75×10^{-6}
			ORF1b	6,501	Pacific salmon nidovirus ORF1b (QEG08237.1); e-value = 4×10^{-175}	46.30	1.26×10^{-5}	
			Spike	4,989	Atlantic salmon bafinivirus spike protein (AUE23862.1); e-value = 2×10^{-4}	24.07	6.4×10^{-6}	
<i>Kanakana letovirus</i> (3)	Haemorrhaging on dorsal fins and gill slits. Pale liver with haemorrhaging, swollen kidney, adhesions between organs and muscles.	<i>Geotria australis</i>	ORF1a	9,270	Pacific salmon nidovirus ORF1a (QEG08236.1); e-value = 2×10^{-156}	26.22	6.38×10^{-5}	6.40×10^{-6}
			ORF1b	6,509	Pacific salmon nidovirus ORF1b (QEG08237.1); e-value = 0	43.97	4.20×10^{-5}	
			Spike	5,022	Porcine torovirus spike protein (QJ01983.1); e-value = 1×10^{-9}	27.18	4.48×10^{-5}	

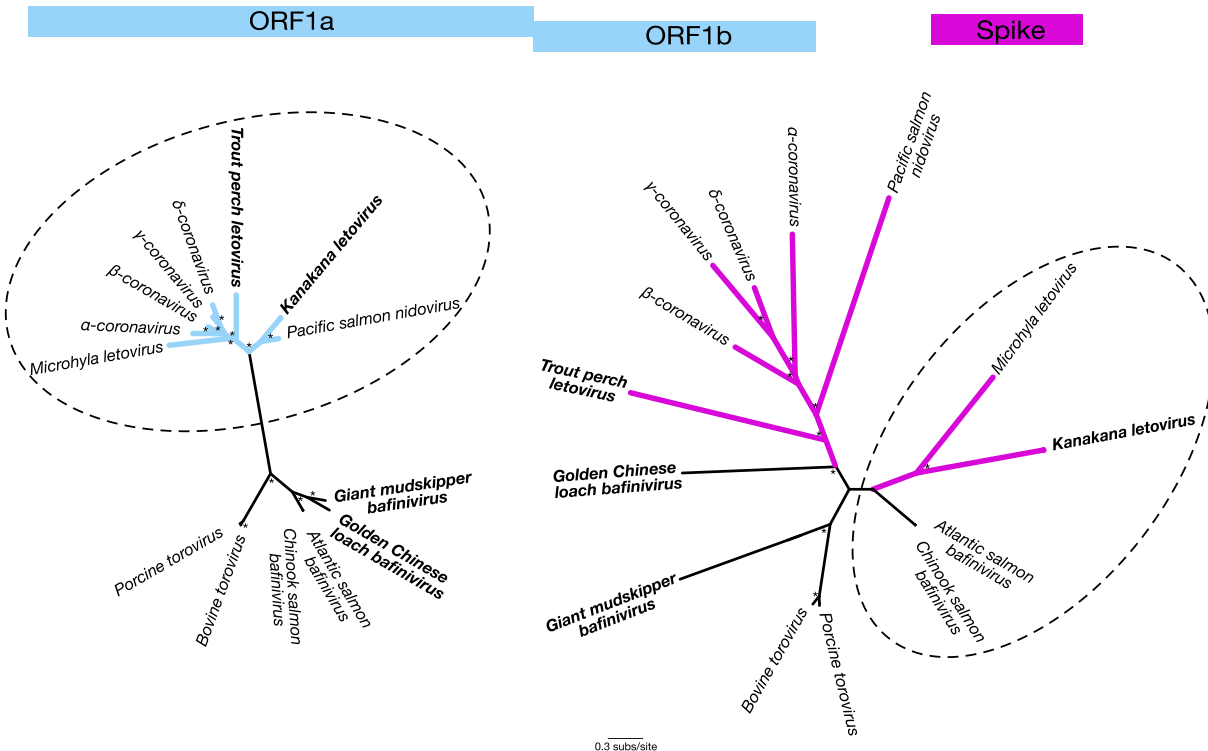


Figure 1. Unrooted phylogenetic trees for the open reading frame 1ab region (blue) and the spike protein (pink). Branches for viruses classified within the *Tobaniviridae* are shown in black while all other branches are coloured. A dashed line circles those taxa that are closest known relatives of kanakana letovirus in each case. Four of the viruses discovered in this study in which spike proteins were identified are in bold. All branches are scaled according to the number of amino acid substitutions per site. Percentage node labels with bootstrap support >70 per cent are indicated with an asterisk (*). Supplementary Table S2 provides the accession numbers for each sequence used in the trees. Amino acid alignments are available at https://github.com/jemmageoghegan/Fish_coronaviridae.

2.3 The evolutionary history of novel coronaviruses in aquatic vertebrates

We next inferred the evolutionary relationships of all the viruses newly identified here and related *Nidovirales* through phylogenetic analysis of amino acid sequences of ORF1b that includes the highly conserved RdRp. As noted above, two viruses identified here fell within the *Tobaniviridae*, rather than the *Coronaviridae*, and were close genetic relatives of other fish bafiniviruses (Cano et al. 2020) (Fig. 3B). This analysis also revealed that fish coronaviruses fell only within the subfamily *Letovirinae* and were distinct from the *Orthocoronavirinae*. Specifically, the hosts of viruses in the *Letovirinae* comprised only aquatic vertebrates: both jawless and bony fish as well as viruses from amphibians (Fig. 3C).

Although there are broad patterns of host clustering across the coronaviruses as a whole with, for example, most of the mammalian coronaviruses falling in the genera *Alphacoronavirus* and *Betacoronavirus*, and the genus *Deltacoronavirus* comprising avian-associated viruses only, it is notable that the phylogenetic position of the *Letovirinae* does not follow that of their hosts (Fig. 4A). Of particular note, kanakana letovirus sampled from New Zealand pouched lamprey falls as a relative ingroup in this subfamily (Fig. 3). Similar patterns can be seen for viruses in amphibians, indicative of major cross-species transmission events. For example, letoviruses identified from amphibian hosts are highly diverse, with the axolotl letovirus more closely related to viruses in fish and the frog letovirus falling in a more divergent

position, forming an outgroup to other viruses in this sub-family (Fig. 3). Indeed, phylogenetic reconciliation analysis found that host-jumping was consistently the most frequent evolutionary event, although there was also clear evidence for host-virus co-divergence that likely reflects deeper branching events in the evolutionary history of these viruses (Fig. 4B).

2.4 Amino acid conservation across the *Coronaviridae*

Compatible with their phylogenetic relationships, the new viruses identified here possessed several motifs that are conserved among the coronaviruses, particularly within the RdRp domain in ORF1b (Fig. 5). These included the amino acid sequences 'GWDYPKCD', 'SGDATTAYANS', and 'ILSDDGV' that are mostly conserved among the *Coronaviridae* but not the *Nidovirales* as a whole (Fig. 5). Indeed, these conserved domains are essential for metal ion chelation and binding of the primer 3'-end template complex (Snijder et al. 2003) and have recently been identified as potential 'barcodes' for coronavirus identification (Babaian and Edgar 2021).

3. Discussion

Virological sampling of jawless fish combined with mining publicly available sequence data of bony fish transcriptomes enabled us to identify eight new members of the *Nidovirales*, six of which fell within the *Coronaviridae* and a further two fell within

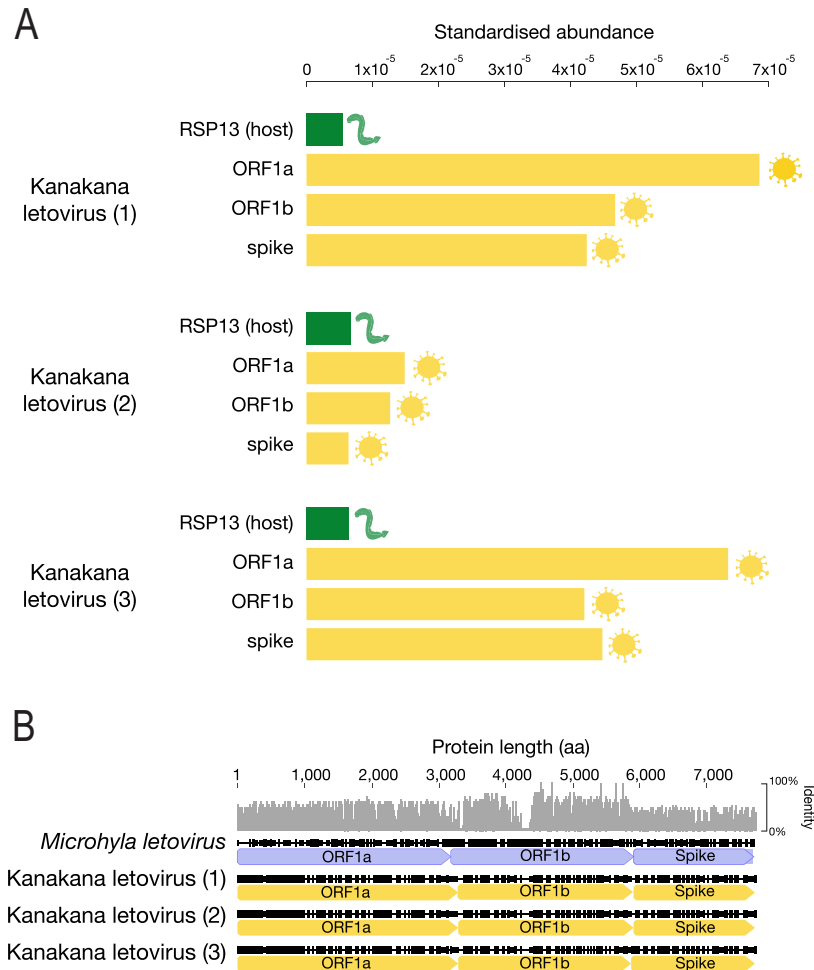


Figure 2. (A) Standardised abundance of kanakana letovirus transcripts (yellow) in comparison to the lamprey host gene, RSP13 (green), across the three lamprey samples that contained the highest abundance of the novel coronavirus. (B) Schematic alignment of all three transcripts against *Microhyla letovirus*, along with the percentage sequence identity.

the related *Tobaniviridae*. The six new coronavirus-like transcripts were close genetic relatives of viruses classified within the *Letovirinae* in the RdRp, including *Pacific salmon nidovirus*, *Microhyla letovirus*, and other viruses recently identified in fishes and amphibians, forming a sister clade to all other known coronaviruses.

The discovery of these new viruses not only expands the phylogenetic diversity of the *Letovirinae*, but it also sheds important new light on the evolution and host range in the *Coronaviridae* as a whole. In particular, our study revealed that coronavirus hosts include both freshwater and marine fish spanning multiple taxonomic orders. This includes an ancient lineage of jawless fish, herein represented by the pouched lamprey from New Zealand, that represents one of the evolutionary oldest vertebrates still to exist and one that has experienced little morphological change for around 360 million years (Gess, Coates, and Rubidge 2006), although its true age has recently been debated (Miyashita et al. 2021). These jawless and bony fish hosts were also sampled across a range of geographic localities: New Zealand, China, Europe and North America. Although divergent in their sequences,

conserved motifs across the entire virus family strongly support the inclusion of these new viruses within the *Coronaviridae* and more specifically the *Letovirinae*. With the expansion of virological surveys of wildlife, it is likely that additional viruses within the *Coronaviridae* will be discovered. Indeed, the long branch lengths separating the divergent families within the *Nidovirales* are indicative of very limited sampling across this order of RNA viruses.

The genomic structure of both *Pacific salmon nidovirus* and *Ambystoma mexicanum nidovirus* includes a short, non-coding region between the ORF1ab and the remaining genes, including the S protein (Edgar et al. 2020; Mordecai et al. 2019). It has been suggested that this homologous gap represents possible genome segmentation in this clade (Edgar et al. 2020). We were, unfortunately, unable to determine the complete genomic structure of the viruses identified here and so cannot directly address this issue. It is noteworthy, however, that the S protein of kanakana letovirus shared more sequence similarity with toroviruses rather than coronaviruses, while trout perch letovirus and *Pacific salmon nidovirus* consistently fell within the *Coronaviridae*. This pattern

Table 2. Novel viruses identified in this study by mining the Sequence Read Archive (SRA), their detection method and sampling locations.

Virus	Host	NCBI accession of SRA data	Sampling location	Contig length (nt)	Closest amino acid match; GenBank accession
Blenny letovirus	<i>Acanthemblemaria</i> sp.	SRA search (SRR5997671)	Caribbean	4,995	Pacific salmon nidovirus ORF1b (56%); QEG08237.1
Mud hopper letovirus	<i>Boleophthalmus pectinirostris</i>	SRA search (SRR5012117)	China	4,170	Pacific salmon nidovirus ORF1b (56%); QEG08237.1
Waltoni letovirus	<i>Schizothorax waltoni</i>	SRA search (SRR6233344)	China	906	Pacific salmon nidovirus ORF1b (52%); QEG08237.1
Broadnosed pipefish letovirus	<i>Syngnathus typhle</i>	SRA search (SRR663248)	Europe	225	Pacific salmon nidovirus ORF1b (55%); QOQ03052.1
Trout perch letovirus	<i>Percopsis omiscomaycus</i>	SRA search (SRR5997814)	North America	2,046	Shrew coronavirus ORF1b (39%); YP_009755838.1
				3,942	Bat coronavirus spike protein (20%); QBP43256.1
Giant mudskipper bafinivirus	<i>Periophthalmodon schlosseri</i>	SRA search (SRR5012122)	China	6,786	White bream virus ORF1b (53%); YP_803213.1
				3,519	Fathead minnow nidovirus spike protein (40%); YP_009505583
Golden Chinese loach bafinivirus	<i>Botia supercilialis</i>	SRA search (SRR5997844)	China	6,996	White bream virus ORF1b (57%); YP_803213.1
				3,618	White bream virus Spike protein (42%); YP_803215.1

of phylogenetic incongruence between these genomic regions is indicative of past recombination, or reassortment, within the *Nidovirales*. At present, the occurrence of this process seems to be limited to only some aquatic vertebrate hosts, although given the propensity for coronaviruses to recombine (Denison et al. 2011), it is likely that this will be observed in additional species.

First recorded in 2011, lamprey populations in New Zealand's Southland region have been affected by LRS, causing reddening along the length of the body with increased mortalities (Brosnahan et al. 2019). While we cannot assign the novel kanakana letovirus discovered here as the causative agent of this syndrome, similar disease symptoms have been observed in salmon infected with *Pacific salmon nidovirus* (Mordecai et al. 2019). Since lamprey are classed as a nationally vulnerable threatened species (Dunn et al. 2018), it is important that further investigation is undertaken to determine the true prevalence of coronavirus in lamprey populations, whether the virus is associated with LRS and if the virus has made a recent jump from a reservoir host species. The increased deliberate movement of fish around the world for aquaculture, ornamental fish trade, and bio-fouling on vessels all are pathways for the potential transmission of disease internationally (Cano et al. 2020).

That *Letoviruses* form a sister clade to all other coronaviruses suggests that the backbone of the *Coronaviridae* phylogeny overall mirrors that of their vertebrate hosts, with different viral genera associated with different vertebrate classes, and the clear transition from aquatic to land (Zhang et al. 2018). Nevertheless, lamprey represent an ancient lineage of jawless fish that fall sister to all other vertebrates, including jawed fish, amphibians, and reptiles. In this context, it is particularly notable that the novel coronavirus identified here in the pouched lamprey does not fall basal to other fish coronaviruses, in turn revealing that cross-species transmission events have played a major role in the evolutionary history of these viruses and which is also apparent

from our co-phylogenetic reconciliation analysis. Hence, as is the case with many families of RNA viruses (Geoghegan, Duchêne, and Holmes 2017), the broad association between different coronaviruses and their vertebrate host classes suggests that the evolutionary history of this group likely dates to at least the origin of the vertebrates some 530 million years ago. However, on top of this overall pattern of long-term virus-host co-divergence, there have also been frequent instances of cross-species virus transmission. As the case of the pouched lamprey demonstrates, such host jumping events may be particularly commonplace in hosts that reside in aquatic environments, perhaps reflecting a combination of evolutionary antiquity (and hence more long-term opportunity for cross-species transmission) and greater physical interactions between multi-species assemblages, although this clearly needs to be investigated in greater detail. Indeed, the parasitic behaviour of lamprey might facilitate viral transmission.

4. Materials and methods

4.1 Lamprey sample collection in Aotearoa New Zealand and RNA sequencing

LRS describes pouched lamprey from the South Island of New Zealand that display red lesions on their gills, fins, and body walls (Brosnahan et al. 2019). From 2011 to 2013, the New Zealand MPI received specimens exhibiting possible signs of LRS for diagnostic testing, including bacteriology, histology, and molecular analysis targeting specific known diseases to determine if an infectious agent was present. No infectious agent attributed to LRS was identified through this testing. Samples that MPI received in 2012 were used in the current study. These comprised 40 pouched lamprey native to New Zealand. Lamprey skin and muscle tissues displaying reddening, heart, kidney, liver, gut, and visceral fat tissues were analysed for the presence of pathogens in 2012 by MPI,

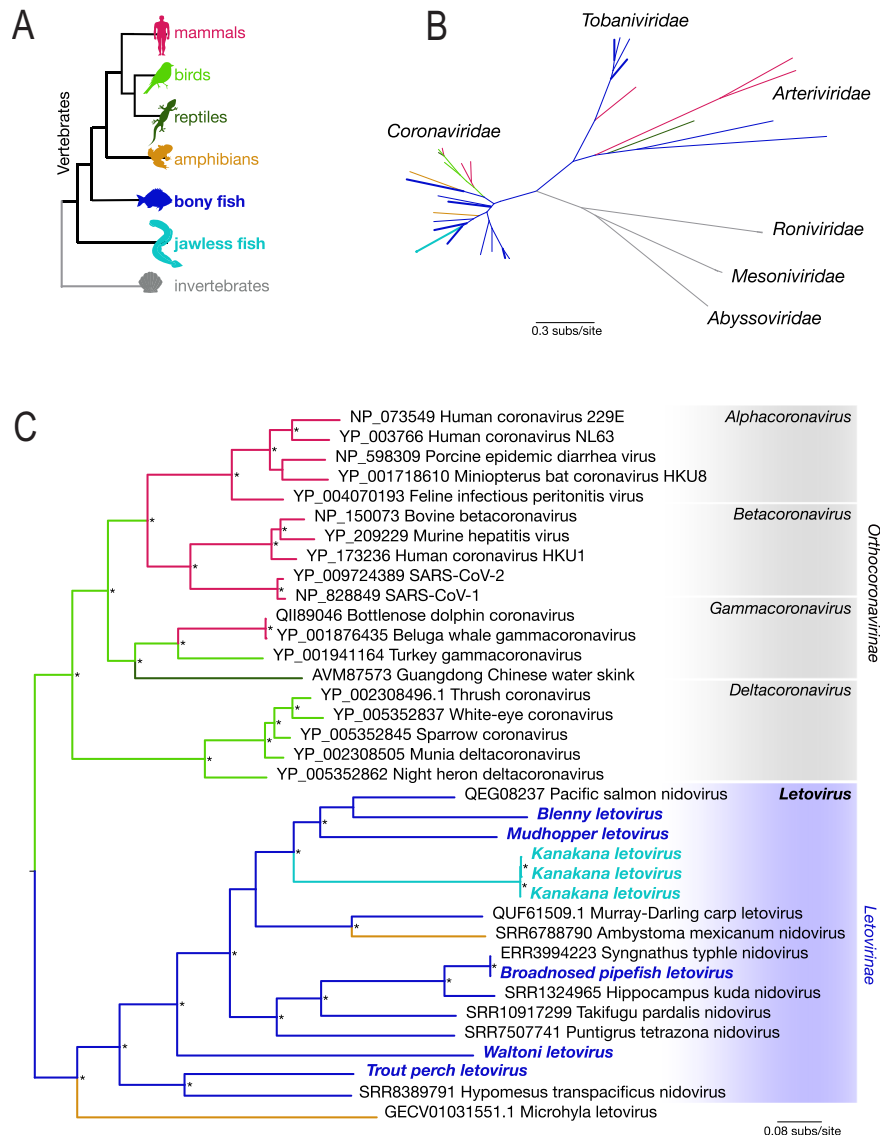


Figure 3. (A) Schematic phylogram showing where bony and jawless fish fall in the greater animal phylogeny and colour-coded to signify the host class in (B) and (C), informed by the relevant literature on vertebrate evolution. (B) An order-level phylogenetic tree showing the relationships between families within the Nidovirales. Branches are colour-coded to signify the host class shown in part a. Branches in bold represent those viruses identified in this study. All branches are scaled according to the number of amino acid substitutions per site. Supplementary Table S3 provides the accession numbers for each sequence used in the tree. Amino acid alignments are available at https://github.com/jemmageoghegan/Fish_coronaviridae. (C) Maximum likelihood phylogenetic tree of the open reading frame 1b (ORF1b), containing the viral RdRp, showing the topology of the nine newly discovered coronaviruses (bold font; blue and cyan) that fall within Letovirinae. The order-level tree in (B) was used to root the tree. All branches are scaled according to the number of amino acid substitutions per site. Percentage node labels with bootstrap support >70 per cent are indicated with an asterisk (*).

and these tissues were then archived in formalin-fixed paraffin-embedded (FFPE) blocks and stored at room temperature until they were sent to the University of Otago for RNA sequencing in 2020. Fresh tissues were obtained from four additional *Geotria australis* individuals, collected on separate field expeditions from the Mokau River, Aria, in 2015 ($n = 2$) and Okuti River, Banks Peninsula, in 2019 ($n = 2$), both of which are separate catchments areas to locations sampled in 2012. These individuals displayed no signs of LRS that has never been recorded from the Mokau or

Okuti rivers. These samples were used as controls. Muscle and skin subsamples from the control individuals were taken in the field and were preserved in RNALater (Thermo Fisher Scientific) and in a -80°C freezer. Sampling was conducted in consultation with local Māori (rūnanga and kaumatua).

Skin tissue was dissected from 26 FFPE blocks, representing 26 individuals. Total RNA was extracted using a Macherey-Nagel NucleoSpin® totalRNA FFPE XS kit following the NucleoSpin total-RNA FFPE XS protocol. Control samples from the Mokau and Okuti

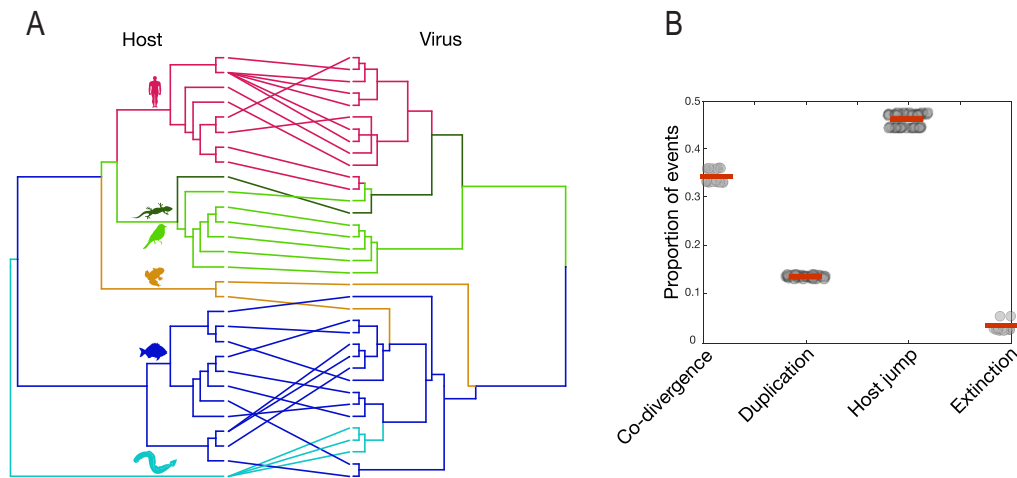


Figure 4. (A) Tanglegram of rooted phylogenetic trees for the *Coronaviridae* and their hosts. Lines and branches are coloured to represent host class, as in Fig. 2. The ‘untangle’ function was used in TreeMap3.0 to maximise the congruence between the host (left) and virus (right) phylogenies. Supplementary Fig. S2 provides the names of the hosts and viruses. The virus tree is that shown in Fig. 3C. (B) Reconciliation analysis of each virus family using eMPress (Santichaivekin et al. 2020) illustrates the proportion of specific evolutionary events, with the mean indicated by a red horizontal line. The ‘event costs’ associated with incongruences between trees were conservative towards co-divergence and defined here as follows: 0 for co-divergence, 1 for duplication, 1 for host-jumping and 1 for extinction.

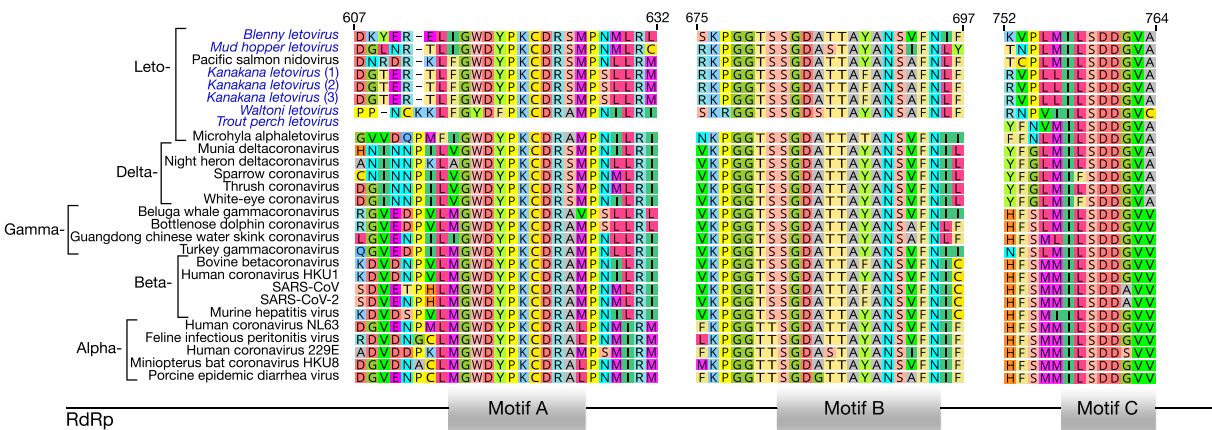


Figure 5. Multiple virus amino acid sequence alignment of the RdRp showing three motifs (A, B and C) conserved throughout all known genera within the *Coronaviridae*, including the newly discovered viruses within the *Letovirinae* for which virus transcripts in this region were identified.

ivers were extracted separately using a Zymo Research Direct-zol RNA Miniprep Plus kit. The purity of the extracted RNA from all sample types was evaluated (NanoDrop and Qubit Fluorometer), and 12 of the highest scoring samples (eight LRS and four controls) were sent to Otago Genomics Facility for sequencing using the Truseq Total RNA Library Preparation Protocol (Illumina). Host ribosomal RNA was depleted using the Ribo-Zero-Gold Kit (Illumina) to enhance viral discovery. Libraries were quality checked and sequenced in one lane on the HiSeq 2500 V4 Illumina platform.

4.2 Data mining the Sequence Read Archive for coronaviruses

To identify additional coronavirus-like sequences, we screened bony fish transcriptomes present in the SRA. We focused on a subset of available SRA libraries, particularly those RNA-seq

libraries without sequencing selection (e.g. poly(A)⁺ -selected). We excluded zebrafish (*Danio rerio*) runs from the screens due to the large number of experimental data sets. To reduce redundancy, we selected the first six runs for a given species in each experimental data set. In total, 1,279 libraries (both single- and paired-end reads) were selected. SRA runs were obtained through the European Nucleotide Archive (<https://www.ebi.ac.uk/ena/browser/home>) that mirrors the SRA but allows for reads to be downloaded directly in a FASTQ. Reads were then analysed using our coronavirus discovery pipeline. The pipeline consisted of an initial screen that aligned the reads against a custom database of *Coronaviridae* and other protein sequences within the *Nidovirales* using Diamond BLASTx (v2.0.2) (Buchfink, Xie, and Huson 2015) with an e-value threshold of 1×10^{-5} . Results were manually inspected in Geneious Prime (www.geneious.com) to exclude spurious alignments to low-complexity regions. Where coronavirus-like reads were detected ($n = 32$), we quality trimmed

and assembled *de novo* raw reads into contigs using Trinity RNA-Seq (v2.9.1) (Haas et al. 2013). The assembled contigs were then compared to NCBI non-redundant protein database using Diamond BLASTx with an e-value threshold of 1×10^{-5} . To identify highly divergent sequences, we regularly updated our custom *Coronaviridae* protein database with the novel viruses identified. Where a coronavirus-like sequence was found, we examined all SRA runs in the experimental data set for the given species.

4.3 Transcript sequence similarity searching for novel coronaviruses

Sequencing reads were first quality trimmed and then assembled *de novo* using Trinity RNA-Seq (v2.11.0) (Haas et al. 2013). The assembled contigs were annotated based on similarity searches against the NCBI nucleotide (nt) and non-redundant protein (nr) databases using BLASTn (Altschul et al. 1990) and Diamond BLASTx (v2.0.2) (Buchfink, Xie, and Huson 2015).

4.4 Phylogenetic analysis

To infer the evolutionary relationships of the coronaviruses, newly discovered translated viral contigs were combined with representative amino acid sequences from both the *Coronaviridae* and the order *Nidovirales*. All these sequences were obtained from NCBI GenBank. The sequences retrieved were then aligned with those generated here using MAFFT (v7.4) (Kato et al. 2002) employing the E-INS-i algorithm. Ambiguously aligned regions were removed using trimAl (v1.2) (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009). To estimate phylogenetic trees, we utilised the maximum likelihood approach available in IQ-TREE (v1.6.8) (Nguyen et al. 2015), selecting the best-fit model of amino acid substitution, LG, with ModelFinder (Kalyaanamoorthy et al. 2017), and using 1,000 bootstrap replicates. Phylogenetic trees were annotated with FigTree (v1.4.2).

To visualise the occurrence of cross-species transmission and virus-host co-divergence in the *Coronaviridae*, we reconciled the co-phylogenetic relationship between coronaviruses and their hosts. A host phylogeny was manually constructed using topologies available in the appropriate literature, and a tanglegram connecting the host and virus trees was inferred using TreeMap v3.0 (Charleston 2011): lines between the trees connect the host with its virus. We utilised the 'untangle' function, which rotates the branches of one tree, to minimise the number of line crosses (i.e. cross-species transmission events). To quantify the relative frequencies of cross-species transmission versus virus-host co-divergence, we reconciled the co-phylogenetic relationship between viruses and their hosts using eMPress (Santichaivekin et al. 2020). This approach employs a maximum parsimony approach to determine the best 'map' of the virus phylogeny onto the host phylogeny. We set the duplication, host-jump, and extinction event types to have a cost of one, while that of host-virus co-divergence was considered a 'null event' and therefore had a cost of zero. Following the parsimony principle, the reconciliation proceeds by minimising the total event cost.

4.5 Estimation of virus abundance

Transcriptomes were quantified using RNA-Seq by Expectation-Maximisation as implemented within Trinity (Li and Dewey 2011) and standardised by the number of paired reads in a given library. This enabled us to estimate the relative abundance of each virus transcript in these data. For comparison, and to assess the sequencing depth across libraries, we also estimated the relative abundance of a stably expressed host reference gene, ribosomal protein S13 (RSP13).

4.6 PCR confirmation

To further confirm the presence of the novel coronavirus in the pouched lamprey, total RNA was re-extracted from the 12 HiSeq-sequenced samples using the methods described above. The total RNA was then reverse transcribed (RT) using Maxima Reverse Transcriptase kit (Thermo Scientific) and ReadyMade Random Hexamers (Integrated DNA Technologies). The Maxima Reverse Transcriptase (Thermo Scientific) protocol was followed and did not include the GC-rich template options: 1 µl template total RNA, 1 µl ReadyMade Random Hexamers (10 mM), 1 µl dNTP mix (10 mM), 11.5 µl nuclease-free water, 4 µl 5X RT Buffer, 0.5 µl (20U) RiboLock RNase Inhibitor (Thermo Scientific), and 1 µl Maxima Reverse Transcriptase. The final volume (20 µl) was stored at -20°C until used in PCRs. See Supplementary Table S1 for the primer set used.

4.7 Etymology of virus names

New viruses discovered in this study were tentatively named by drawing from their host species' common names. The virus found in pouched lamprey used the Māori name, kanakana, since they were identified in the South Island of New Zealand and following discussions with representatives for iwi (Māori tribes). We suggest that all of these viruses be grouped within the subfamily *Letovirinae*.

Data availability

Raw sequencing reads from pouched lamprey samples are available online by request via the Genomics Aotearoa database to maintain Māori data sovereignty of native (taonga) species (<https://repo.data.nesi.org.nz/TAONGA-LAMPREY>). All virus sequences are available on NCBI (accession numbers: MZ203498-MZ203506).

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

The authors thank Peter Stockman (Ngati Maniapoto, Te Rūnanga o Wairua) and Keith Bradshaw and Duncan Ryan (Te Runaka o Awarua) for sample collection. We are grateful to the iwi representatives for providing insights and for their collaboration and to the New Zealand Ministry for Primary Industries (MPI) for access to samples.

Funding

This study was supported by funding from the New Zealand Ministry of Business Innovation and Employment (MBIE) (CO1X1615). A.K.M. is supported by a University of Otago Postgraduate Scholarship and the NIH Minority Health International Research Training Programme. J.L.G. is funded by a New Zealand Royal Society Rutherford Discovery Fellowship (RDF-20-UOO-007) and a Marsden Fast Start grant (20-UOO-105). E.C.H. is funded by an ARC Australian Laureate Fellowship (FL170100022).

Conflict of interest: None declared.

References

- Altschul, S. F. et al. (1990) 'Basic Local Alignment Search Tool', *Journal of Molecular Biology*, 215: 403–10.
- Babaian, A., and Edgar, R. C. (2021) 'Ribovirus Classification by a Polymerase Barcode Sequence', *bioRxiv*. [10.1101/2021.03.02.433648](https://doi.org/10.1101/2021.03.02.433648)
- Brosnahan, C. L. et al. (2019) 'Lamprey (*Geotria australis*; Agnatha) Reddening Syndrome in Southland Rivers, New Zealand 2011–2013: Laboratory Findings and Epidemiology, Including the Incidental Detection of an Atypical *Aeromonas salmonicida*', *New Zealand Journal of Marine and Freshwater Research*, 53: 416–36.
- Buchfink, B., Xie, C., and Huson, D. H. (2015) 'Fast and Sensitive Protein Alignment Using DIAMOND', *Nature Methods*, 12: 59–60.
- Bukhari, K. et al. (2018) 'Description and Initial Characterization of Metatranscriptomic Nidovirus-Like Genomes from the Proposed New Family Abysoviridae, and from a Sister Group to the Coronavirinae, the Proposed Genus Alphaletovirus', *Virology*, 524: 160–71.
- Cano, I. et al. (2020) 'Isolation of a Chinook Salmon Bafinivirus (CSBV) in Imported Goldfish *Carassius auratus* L. in the United Kingdom and Evaluation of Its Virulence in Resident Fish Species', *Viruses*, 12: 578.
- Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009) 'trimAl: A Tool For Automated Alignment Trimming in Large-Scale Phylogenetic Analyses', *Bioinformatics*, 25: 1972–3.
- Charleston, M. A. (2011), *TreeMap* <<https://sites.google.com/site/cophylogeny/home>> accessed 18 Mar 2021.
- Costa, V. A. et al. (2021) 'Metagenomic Sequencing Reveals a Lack of Virus Exchange between Native and Invasive Freshwater Fish across the Murray-Darling Basin, Australia', *Virus Evolution*, 7(1): veab034.
- Denison, M. R. et al. (2011) 'Coronaviruses: An RNA Proofreading Machine Regulates Replication Fidelity and Diversity', *RNA Biology*, 8: 270–9.
- Dunn, N. R. et al. (2018) *Conservation Status of New Zealand Freshwater Fishes, 2017. New Zealand Threat Classification Series 24*, Department of Conservation, New Zealand.
- Edgar, R. C. et al. (2020) 'PetaBase-Scale Sequence Alignment Catalyzes Viral Discovery', *bioRxiv*. [10.1101/2020.08.07.241729](https://doi.org/10.1101/2020.08.07.241729)
- Geoghegan, J. L., Duchêne, S., and Holmes, E. C. (2017) 'Comparative Analysis Estimates the Relative Frequencies of Co-divergence and Cross-species Transmission within Viral Families', *PLoS Pathogens*, 13(2): e1006215.
- Gess, R. W., Coates, M. I., and Rubidge, B. S. (2006) 'A Lamprey from the Devonian Period of South Africa', *Nature*, 443: 981–4.
- Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, Gorbalenya, A. E. et al. (2020) 'The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-nCoV and Naming It SARS-CoV-2', *Nature Microbiology*, 5: 536–44.
- Haas, B. J. et al. (2013) 'De Novo Transcript Sequence Reconstruction from RNA-seq Using the Trinity Platform for Reference Generation and Analysis', *Nature Protocols*, 8: 1494–512.
- Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods*, 14: 587–9.
- 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.
- Li, B., and Dewey, C. N. (2011) 'RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome', *BMC Bioinformatics*, 12: 323.
- Miyashita, T. et al. (2021) 'Non-ammocoete Larvae of Palaeozoic Stem Lampreys', *Nature*, 591: 408–12.
- Mordecai, G. J. et al. (2019) 'Endangered Wild Salmon Infected by Newly Discovered Viruses', *eLife*, 8: e47615.
- Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.
- Santichaivekin, S. et al. (2020) 'eMPress: A Systematic Cophylogeny Reconciliation Tool', *Bioinformatics*, btaa978.
- Shi, M. et al. (2018) 'The Evolutionary History of Vertebrate RNA Viruses', *Nature*, 556: 197–202.
- Snijder, E. J. et al. (2003) 'Unique and Conserved Features of Genome and Proteome of SARS-coronavirus, an Early Split-off from the Coronavirus Group 2 Lineage', *Journal of Molecular Biology*, 331: 991–1004.
- Wu, F. et al. (2020) 'A New Coronavirus Associated with Human Respiratory Disease in China', *Nature*, 579: 265–9.
- Zhang, Y.-Z. et al. (2018) 'The Diversity, Evolution and Origins of Vertebrate RNA Viruses', *Current Opinion in Virology*, 31: 9–16.

Limited cross-species virus transmission in a spatially restricted coral reef fish community

Vincenzo A. Costa,¹ David R. Bellwood,² Jonathon C.O. Mifsud,¹ Kate Van Brussel,¹ Jemma L. Geoghegan,^{3,4,†} Edward C. Holmes,^{1,‡} and Erin Harvey^{1,*,§}

¹Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia, ²Research Hub for Coral Reef Ecosystem Functions, College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia, ³Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand and ⁴Institute of Environmental Science and Research, Wellington 5022, New Zealand

[†]<https://orcid.org/0000-0003-0970-0153>

[‡]<https://orcid.org/0000-0001-9596-3552>

[§]<https://orcid.org/0000-0002-9379-2064>

*Corresponding author: E-mail: erin.harvey@sydney.edu.au

Abstract

The Great Barrier Reef (GBR)—the largest coral reef ecosystem in the world—supports over 1,200 fish species with some of the highest population densities and diversities observed in vertebrates, offering a high potential for virus transmission among species. As such, the GBR represents an exceptional natural ecosystem to determine the impact of host community diversity on virus evolution and emergence. In recent decades, the GBR has also experienced significant threats of extinction, making it one of the most vulnerable ecosystems on the planet. Despite the global importance of the GBR, our understanding of virus diversity and connectivity in tropical reef fishes remains poor. Here, we employed metatranscriptomic sequencing to reveal the viromes of sixty-one reef fish species. This identified transcripts representing 132 putative viral sequences, 38 of which exhibited strong phylogenetic relationships with known vertebrate-associated viral genera, including a novel *Santee-Cooper ranavirus* (*Iridoviridae*). We found little evidence for virus transmission between fish species living within a very restricted geographical space—a 100-m² coral reef ecosystem—suggesting that there might be important host barriers to successful cross-species transmission despite regular exposure. We also identified differences in virome composition among reef fish families, such that cryptobenthic reef fishes—characterized by small body sizes and short life spans—exhibited greater virome richness compared to large reef fishes. This study suggests that there are important barriers to cross-species virus transmission and that successful emergence in a reef fish community likely requires active host adaptation, even among closely related host species.

Key words: virome; Great Barrier Reef; virus; metatranscriptomics; fish; virus discovery.

Introduction

Despite their long evolutionary history, extensive diversity, and complex ecological interactions, fish are greatly under-sampled in studies of viral ecology and evolution. Economically, fish contribute approximately 401 billion USD to the global economy and supply a yearly average of 20.5 kg per capita for consumption (FAO 2020). Yet fish face continual and potentially ruinous threats from emerging viral infections (Kibenge 2019). Climate-associated changes in species interactions are likely to have devastating ecological and economic consequences, particularly in the context of emerging infectious diseases (Burge et al. 2014). It was recently estimated that climate change may be responsible for the extinction of almost half of the economically important fish species in the tropical Pacific region by 2100 (Lam et al. 2020).

Tropical coral reefs are particularly vulnerable to biodiversity loss. In recent decades, global warming has caused numerous coral bleaching events worldwide, with detrimental cascading effects on reef ecosystem functioning and biodiversity

(De'ath et al. 2012; Heron et al. 2016; Stuart-Smith et al. 2018). While tropical coral reefs make up only a small fraction of the marine environment, they support enormous biodiversity, accounting for approximately one-third of all currently described marine fish species (Spalding and Grenfell 1997). Many are 'cryptobenthic reef fishes', characterized by small body sizes (i.e. adult sizes of approximately 5 cm), short lifespans, cryptic behaviour, and benthic positioning on coral reefs (Depczynski and Bellwood 2003; Brandl et al. 2018). These diverse reef fish assemblages are of significant economic and cultural value to humans through aquaculture, fisheries, tourism, and the aquarium trade (Moberg and Folke 1999; Brandl et al. 2018; Woodhead et al. 2019).

Despite the economic and socioecological importance of reef fishes, little is known about the composition of their viromes or how ecological and phylogenetic variability within a reef fish community impacts cross-species virus transmission. Host community diversity likely plays a central role in virus emergence, as contact between donor and recipient hosts is a prerequisite

for virus transmission (Parrish et al. 2008). Yet, revealing the exact nature of that role has proven challenging (Ostfeld and Keesing 2012). A high diversity of hosts could provide more transmission opportunities for viruses, elevating the risk of disease emergence. Conversely, increased host community diversity may reduce the probability of disease emergence through the 'dilution effect', in which species richness provides more primary hosts for pathogens, in turn reducing disease occurrence in some other species (Schmidt and Ostfeld 2001; Ostfeld and Keesing 2012). As tropical coral reefs are biodiversity hotspots, they serve as an ideal natural ecosystem for exploring the impact of host diversity on the extent, pattern, and evolution of virus diversity. Indeed, reef fishes display some of the highest densities and highest diversities of potential vertebrate hosts on the planet. In our study location, a standard sampling area of 3.5 m² consistently supports between 50 and 150 fishes belonging to 15–25 species (Bellwood et al. 2006). These values equate to a mean density of approximately twenty-eight fishes per m². In terrestrial systems, vertebrates are typically at densities in the order of 0.001 per m², rarely, if ever, exceeding 1 per m² (Santini et al. 2018). Furthermore, these fish densities and richness are sustained year-round (Lefèvre et al. 2016), offering a continual potential for cross-species transmission in a readily transmitting aquatic medium.

Much of our knowledge on the tropical reef virosphere is skewed towards those viruses associated with coral species and their symbionts (Thurber et al. 2017). While epizootic infections have previously been reported in tropical reef fishes across the Western Atlantic and the Gulf of Mexico (PANEK 2005), there are minimal data on viral diversity in fish from the Great Barrier Reef (GBR), Australia, although this is the largest and among the most threatened reef ecosystems in the world (Bellwood et al. 2004; De'ath et al. 2012; Stuart-Smith et al. 2018). A recent metatranscriptomic analysis of the pygmy goby (*Eviota zebolina*) identified three novel viruses (families *Arenaviridae*, *Hantaviridae*, and *Paramyxoviridae*) despite *Eviota* exhibiting maximum lifespans of between 60 and 100 days (including a 24–26-day pelagic larval phase) (Depczynski and Bellwood 2006; Geoghegan et al. 2021).

The rate of infectious disease emergence is expected to increase in marine environments, particularly in the context of climate change (Burge et al. 2014). As ocean temperatures continue to rise, many eastern Australian tropical fishes have begun to shift their distribution poleward to temperate reefs situated at higher latitudes through tropicalization (Vergés et al. 2014; Booth et al. 2018). While the broad-scale ecological impacts of tropicalization are becoming increasingly apparent in native temperate fishes (Vergés et al. 2014), it is unclear how this will impact virus ecology and infectious disease emergence.

Tropical ornamental fishes may also act as viral vectors of disease in economically important farmed species. For example, in Australia, dwarf gourami (*Trichopodus lalius*) can transmit infectious kidney and spleen necrosis virus (genus *Megalocytivirus* and *Iridoviridae*) to domestic fishes, often with detrimental impacts including disease outbreaks in iconic species such as Murray cod (*Maccullochella peelii*) (Lancaster, Williamson, and Schroen 2003; Go et al. 2006). Moreover, tropical wrasses (Labridae) have been considered effective biological control agents in aquaculture for their natural ability to consume pests (Barton et al. 2020). However, temperate cleaner wrasses have been reported to be important drivers of outbreaks of viral haemorrhagic septicaemia virus (*Rhabdoviridae*) in farmed salmonids (Murray 2016). As such, revealing viral diversity in ornamental tropical reef fishes is imperative to understanding the risk of disease emergence in both wild and domestic fish populations.

Revealing the nature of viral transmission within and among species in natural environments is central to understanding disease emergence. With exceptionally high species diversity and highly variable individual abundances and ecologies, coral reefs offer an opportunity to explore cross-species virus transmission in a complex natural high-diversity ecosystem. Our goal, therefore, was to employ total RNA sequencing (i.e. metatranscriptomics) to reveal the diversity, abundance, and composition of viruses infecting reef fishes from the GBR, utilizing a fish community from a 100-m² coral reef ecosystem. In particular, we aimed to (1) reveal viral diversity in tropical reef fishes, (2) determine the frequency of virus transmission within a spatially restricted reef fish community, (3) determine whether there are differences in virome composition between reef fish families, as well as between cryptobenthic reef fishes and large reef fishes, that differ in size, metabolic rate, lifespan, and fecundity (Brandl et al. 2018), and (4) identify novel viruses that may pose an emerging threat to Australian fisheries, aquaculture, and the aquarium trade.

Materials and methods

Animal ethics

Fish were collected under a Great Barrier Reef Marine Park Authority permit (G16/37684.1) and James Cook University Animal Ethics permit A2752.

Tropical reef fish sample collection

Fishes ($n=192$) were collected in early April 2021 at Orpheus Island, GBR (18°36'44.3"S 146°28'59.4"E). All were "healthy" in that they were intact with good colour and no visible signs of disease, stress, or wasting. The vast majority were adult individuals. These included sixty-one species across sixteen reef fish families: Gobiidae (gobies) ($n=30$ species), Labridae (wrasses) ($n=6$), Pomacentridae (damselfishes) ($n=5$), Blenniidae (blennies) ($n=5$), Acanthuridae (surgefishes) ($n=3$), Apogonidae (cardinalfishes) ($n=2$), Monacanthidae (filefishes) ($n=2$), Tetraodontidae (pufferfishes) ($n=1$), Pseudochromidae (dottybacks) ($n=1$), Chaetodontidae (butterflyfishes) ($n=1$), Atherinidae (silversides, hardyheads) ($n=1$), Serranidae (groupers) ($n=1$), Tripterygiidae (triplefin blennies) ($n=1$), Muraenidae (moray eels) ($n=1$), Bythitidae (brotulas) ($n=1$), and Ophichthidae (snake eels) ($n=1$) (Supplementary Table S1). We sampled one to twelve individuals per species (± 2.7 SD) (Supplementary Table S1). Of these species, forty-three (thirty-two cryptobenthic reef fishes and eleven large reef fishes) ($n=148$ individuals) were collected from a reef fish community within a 100-m² sampling area (along the northern margin of Pioneer Bay). The other eighteen (seven cryptobenthic reef fishes and eleven large reef fishes) ($n=44$ individuals) were collected from similar habitats (i.e. shallow fringing reefs) around Orpheus Island (Supplementary Table S1). Fishes within the focal sampling area were collected using an enclosed clove oil method, in which a small portion of the reef is enclosed within a fine net, and all fishes within the net were anaesthetized using clove oil (Ackerman and Bellwood 2002). Additional fish were collected using nets and/or dilute clove oil. All fish caught were placed either dissected (liver and gills) or whole in RNAlater and then transported to the lab on ice (Supplementary Table S1). Specimens were then stored at -80°C until RNA extraction.

RNA extraction, library preparation, and metagenomic next-generation sequencing

Tissue specimens (e.g. liver and gills or whole fish) were processed together as a single extraction for each individual fish sample.

The combined tissues were placed in 600 µl of lysis buffer containing 0.5 per cent foaming reagent (Reagent DX, Qiagen) and 1 per cent of β-mercaptoethanol (Sigma-Aldrich). Tissue samples were homogenized with a TissueRuptor (Qiagen) for up to 1 minute at 5,000 rpm. To remove tissue residues, the homogenate was centrifuged at full speed for 3 minutes. RNA from the clear supernatant was extracted using the RNeasy Plus Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol.

RNA was quantified using a UV-Vis cuvette spectrophotometer (DeNovix, Delaware, USA) and a parallel capillary electrophoresis instrument (Fragment Analyzer; Agilent, CA, USA). RNA from each individual fish was then pooled within each species (Supplementary Table S1), resulting in a total of sixty-one RNA sequencing libraries. These libraries were constructed using the Truseq Total RNA Library Preparation Protocol (Illumina). Host ribosomal RNA was depleted with the Ribo-Zero Plus Kit (Illumina), and paired-end sequencing (150 bp) was carried out on the NovaSeq 6000 platform (Illumina). To reduce the impact of index hopping and false virus-host assignments, each library was sequenced on two different lanes. Library construction and metatranscriptomic sequencing were performed by the Australian Genome Research Facility.

Virome assembly

Sequencing reads were first quality trimmed using Trimmomatic v.0.38 (Bolger, Lohse, and Usadel 2014) and then *de novo* assembled into contigs using MEGAHIT v.1.2.9, with default parameter settings (Li et al. 2015b). Assembled contigs for each library were used as queries against the NCBI nucleotide (nt) and non-redundant protein (nr) databases using BLASTn and Diamond (BLASTx) (Buchfink, Xie, and Huson 2015). To enable the identification of divergent viral sequences, we used an e-value search threshold of 1×10^{-5} (Shi et al. 2016, 2018; Geoghegan et al. 2021). All contigs with matches to viral sequences (NCBI/GenBank taxid: 10239) were predicted as open reading frames (ORFs) using Geneious Prime (v.2022.0) (Kearse et al. 2012) (www.geneious.com). To remove any false positives, each putative viral ORF was translated into amino acid sequences and used as a query to perform a second BLAST search against the NCBI nt and nr databases. From this second search, ORFs with matches to host genes (e.g. from fish or bacteria) were deemed as false positives and removed. ORFs with top hits to viruses from the initial search were selected and utilized for further analysis. Viral contig completion and contamination were determined using CheckV (Nayfach et al. 2021). The coverage was assessed by mapping using BOWTIE2 v 2.3.3.1 (Langmead and Salzberg 2012).

Virus taxonomic assignment and phylogenetic analysis

We used a phylogenetic approach to taxonomically assign our putative viral sequences and infer their evolutionary histories. First, we aligned our putative viral sequences with complete sequences of related viruses available on NCBI/GenBank (August 2021) using MAFFT v.7.450, with the E-INS-i algorithm (Katoh and Standley 2013). While some of these sequences were partial, they are adequate for phylogenetic analysis (Geoghegan et al. 2021). We used background sequences from the similarity search as well as additional sequences listed by the International Committee of Viral Taxonomy (ICTV) (https://talk.ictvonline.org) for each viral family. These included the conserved RNA-dependent RNA polymerase (RdRp) for RNA viruses and the DNA polymerase and major capsid protein for DNA viruses. The amino acid sequence

alignment was trimmed using TrimAl v.1.2 to remove ambiguously aligned regions with a gap threshold of 0.9 and a variable conserve value (Capella-Gutierrez, Silla-Martinez, and Gabaldon 2009). The best-fit model of amino acid substitution was estimated with the 'ModelFinder Plus' (-m MFP) flag in IQ-TREE (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017). Using these data, we estimated phylogenetic trees using a maximum likelihood approach with 1,000 bootstrap replicates using IQ-TREE. Trees were annotated using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Phylogenetic position and level of sequence similarity were used to determine whether a viral sequence was likely infecting fish (i.e. 'vertebrate-associated') or derived from diet or environment (i.e. 'non-vertebrate'): the latter often exhibits considerable genetic divergence and hence are phylogenetically distinct (Shi et al. 2018; Zhang et al. 2018; Costa et al. 2021; Geoghegan et al. 2021). Vertebrate-associated viral sequences were classified as novel species according to similarity thresholds for each viral family as specified by the ICTV.

Viral genome annotation

Viral genomes were annotated using the NCBI conserved domain (CDD) search tool and InterProScan with the TIGRFAMs (v15.0), SFLD (v4.0), PANTHER (v15.0), SuperFamily (v1.75), PROSITE (v2022_01), CDD (v3.18), Pfam (v34.0), SMART (v7.1), PRINTS (v42.0), and CATH-Gene3D databases (v4.3.0) (Jones et al. 2014).

Removal of endogenous viral elements

To identify whether our putative viral contigs were expressed endogenous viral elements, we initially searched for disrupted ORFs using Geneious Prime (v.2022.0). Next, we screened for flanking host regions using CheckV and BLASTn. We also searched for flanking long terminal repeats as these are associated with integrated proviruses (Johnson 2019). Contigs with intact ORFs that were not flanked by host genes or repeats were subsequently used as a query against all vertebrate genomes available on NCBI using tblastn, with an e-value threshold of 1×10^{-20} (Shi et al. 2018). We used all published vertebrate genomes as no genomes are currently available for the reef fish studied here. Viral sequences identified as endogenous were removed from subsequent analyses.

RT-PCR validation of vertebrate-associated viruses

We used RT-PCR to validate the presence of the putative exogenous vertebrate-associated viruses detected here. Primers for each virus were designed using Primer3 v.0.4.0 (Untergasser et al. 2012). SuperScript IV One-Step RT-PCR (Invitrogen) was used to amplify viral genes from the total RNA of all sequencing library pools, with annealing temperatures ranging from 50°C to 55°C (Supplementary Table S2).

Abundance estimation and virome statistical analysis

We employed RNA-seq by Expectation-Maximization (v1.2.28) to quantify the relative abundance of transcripts within each fish species transcriptome (Haas et al. 2013). These included both viral genes and the stably expressed host reference gene, ribosomal protein S13 (RPS13), which was used to assess sequencing depth across libraries (Geoghegan et al. 2018, 2021). Abundance measures were standardized by dividing values against the total reads for each library. We calculated both alpha and beta diver-

sity to compare virome composition between reef fish families, as well as between cryptobenthic reef fishes ($n = 39$ species) and large reef fishes ($n = 22$ species). We also analysed the composition of the non-vertebrate virome as a form of internal control, as these viruses are not impacted by aspects of fish biology. Accordingly, we used Rhea scripts to calculate alpha diversity, including viral abundance, observed virome richness, and Shannon diversity (Lagkouravos et al. 2017). Statistical comparisons of alpha diversity were modelled using generalized linear models and tested using a likelihood-ratio test (χ^2) and Tukey's post hoc analysis with the *multcomp* package (Hothorn, Bretz, and Westfall 2008). To compare viral communities between reef fish assemblages, we calculated beta diversity using a Bray–Curtis distance matrix with the *phyloseq* package (McMurdie and Holmes 2013). These data were then tested using permutational multivariate analysis of variance (permanova) with the *vegan* package (adonis) (Dixon 2003). All plots were constructed using *ggplot2* (Valero-Mora 2010).

Results

Total diversity and abundance of viruses discovered in tropical reef fish

We sequenced sixty-one metatranscriptomes of tropical reef fish species for virus discovery, resulting in a median of 99,658,615 (range 45,583,181–121,905,422) reads and 371,944 (range 161,447–626,467) contigs with an average N50 of 1,049

bp per library (Supplementary Table S1). From these data, we identified transcripts representing thirty-eight putative exogenous viral sequences that exhibited strong phylogenetic relationships with viral families and genera known to infect vertebrate species, comprising eleven negative-sense single-stranded RNA viruses (-ssRNA) (Hantaviridae, Chuviridae, Orthomyxoviridae, Rhabdoviridae, and Paramyxoviridae), eighteen positive-sense single-stranded RNA viruses (+ssRNA) (Astroviridae, Hepeviridae, Picornaviridae, Flaviviridae, and Coronaviridae), one double-stranded RNA virus (dsRNA) (Reoviridae), six single-stranded DNA viruses (ssDNA) (Parvoviridae and Circoviridae), and two double-stranded DNA viruses (dsDNA) (Iridoviridae) (Table 1; Fig. 1). Using PCR, we confirmed the presence of thirty-six of the thirty-eight vertebrate-associated viruses from our reef fish RNA libraries (Table 1). Two libraries contained an insufficient volume of RNA for virus detection. In total, thirty-two libraries (seventeen cryptobenthic reef fishes and fifteen large reef fishes) contained no vertebrate-associated viral sequences.

Our metatranscriptomic analysis also identified ninety-four viral sequences that shared sequence similarity and phylogenetically clustered with viruses associated with invertebrates (arthropods, crustaceans, molluscs, platyhelminthes, myriapods, and nematodes) and fungal hosts (i.e. the 'non-vertebrate' group; Supplementary Figs S1–S4). These viruses were classified within the Flaviviridae (27 per cent of total non-vertebrate virome abundance), Narnaviridae (24 per cent), Picornaviridae (20 per cent), Nodaviridae (10 per cent), Hepeviridae (7 per cent), Solemoviridae

Table 1. Description of vertebrate-associated viruses identified in this study.

Virus name	Reads	Virus family	Host	ORFs/ segments	Length (nt)	Complete genome	Novel	PCR
Goby astrovirus 2	24,100	Astroviridae	<i>Cabillus tongarevae</i>	3	7,226	Y	Y	+
Goby astrovirus 2	2,767	Astroviridae	<i>Istigobius decoratus</i>	3	7,085	Y	Y	+
Goby astrovirus 2	9,023	Astroviridae	<i>Istigobius</i>	3	7,155	Y	Y	+
			<i>nigroocellatus</i>					
Goby astrovirus 2	1,184	Astroviridae	<i>Asterropteryx</i>	3	7,045	Y	Y	+
			<i>semipunctatus</i>					
Goby astrovirus 1	2,863	Astroviridae	<i>Istigobius goldmanni</i>	3	7,087	Y	Y	+
Goby astrovirus 3	6,709	Astroviridae	<i>Luposicya lupus</i>	2	7,355	Y	Y	+
Blenniella astrovirus	59	Astroviridae	<i>Blenniella</i> sp.	1	744	N	Y	+
<i>Salaria guttatus</i>	40,334	Chuviridae	<i>S. guttatus</i>	3	10,801	Y	Y	+
<i>piscichuvirus</i>								
<i>Abudefduf bengalensis</i>	54	Circoviridae	<i>A. bengalensis</i>	1	873	N	Y	+
<i>circovirus</i>								
<i>Enneapterygius</i>	26	Coronaviridae	<i>E. tutuillae</i>	1	856	N	Y	+
<i>tutuillae letovirus</i>								
<i>Chaetodon aureofaciatus</i>	1,029	Flaviviridae	<i>C. aureofaciatus</i>	1	8,826	Y	Y	+
<i>tus hepacivirus</i>								
<i>Luposicya lupus</i>	6,144	Hantaviridae	<i>L. lupus</i>	3	11,533	Y	Y	+
<i>actinovirus</i>								
Pygmy goby	972	Hantaviridae	<i>Eviota zebrina</i>	3	5,894	N	N	+
<i>hantavirus</i>								
<i>Trimma fangi</i>	106	Hantaviridae	<i>T. fangi</i>	2	1,319	N	Y	+
<i>actinovirus</i>								
<i>Pleurosicya acti-</i>	123	Hantaviridae	<i>Pleurosicya</i> sp.	2	2,103	N	Y	+
<i>novirus</i>								
<i>Cephalopholis boenak</i>	9	Hantaviridae	<i>C. boenak</i>	1	1,029	N	Y	+
<i>actinovirus</i>								
<i>Istigobius decoratus</i>	16	Hepeviridae	<i>I. decoratus</i>	1	393	N	Y	+
<i>hepevirus</i>								
<i>Halichoeres melanurus</i>	9,835	Iridoviridae	<i>H. melanurus</i>	9	10,821	N	N	+
<i>ranavirus</i>								
<i>Enneapterygius</i>	35,545	Iridoviridae	<i>E. tutuillae</i>	9	20,970	N	Y	+
<i>tutuillae iridovirus</i>								

(continued)

Table 1. (Continued)

Virus name	Reads	Virus family	Host	ORFs/ segments	Length (nt)	Complete genome	Novel	PCR
<i>Cheilodipterus quinquelineatus</i> orthomyxovirus	979	Orthomyxoviridae	<i>C. quinquelineatus</i>	5	9,090	N	Y	+
Pygmy goby paramyxovirus	238	Paramyxoviridae	<i>E. zebrina</i>	1	1,314	N	N	+
<i>Enneapterygius tutuillae</i> paramyxovirus	322	Paramyxoviridae	<i>E. tutuillae</i>	1	618	N	Y	+
<i>Istigobius ornatus</i> parvovirus	1,746	Parvoviridae	<i>I. ornatus</i>	3	4,284	Y	Y	+
<i>Istigobius rigillus</i> parvovirus	266	Parvoviridae	<i>I. rigillus</i>	1	1,656	N	Y	+
<i>Ecsenius stictus</i> parvovirus	2,707	Parvoviridae	<i>E. stictus</i>	2	4,286	Y	Y	+
<i>Dinematichthys</i> parvovirus	122	Parvoviridae	<i>Dinematichthys</i> sp.	1	1,236	N	Y	+
<i>Luposicya lupus</i> ichthama-parvovirus	12,328	Parvoviridae	<i>L. lupus</i>	1	1,299	N	Y	+
<i>Blenniella picornavirus</i>	2	Picornaviridae	<i>Blenniella</i> sp.	1	347	N	Y	+
<i>Enneapterygius tutuillae</i> picornavirus	225	Picornaviridae	<i>E. tutuillae</i>	1	3,840	N	Y	+
<i>Exyrias</i> picornavirus	44	Picornaviridae	<i>Exyrias</i> sp.	1	1,373	N	Y	+
<i>Halichoeres marginatus</i> picornavirus	27,506	Picornaviridae	<i>H. marginatus</i>	1	8,344	Y	Y	+
<i>Neopomacentrus bankieri</i> picornavirus	2,955	Picornaviridae	<i>N. bankieri</i>	1	8,408	Y	Y	+
<i>Asterropteryx spinosa</i> picornavirus	8,307	Picornaviridae	<i>A. spinosa</i>	1	8,405	Y	Y	+
<i>Asterropteryx semipunctatus</i> picornavirus	210	Picornaviridae	<i>A. semipunctatus</i>	1	992	N	Y	+
<i>Pomacentrus moluccensis</i> picornavirus	12	Picornaviridae	<i>P. moluccensis</i>	1	321	N	Y	+
<i>Trimma okinawae</i> rhabdovirus	13	Rhabdoviridae	<i>T. okinawae</i>	2	945	N	Y	+
<i>Trimma striatum</i> rhabdovirus	2	Rhabdoviridae	<i>T. striatum</i>	1	488	N	Y	-
<i>Istigobius nigroocellatus</i> reovirus	282	Reoviridae	<i>I. nigroocellatus</i>	6	6,966	N	Y	-

(4 per cent), Totiviridae (1 per cent), Partitiviridae, Iflaviridae, Picobirnaviridae, Phenuiviridae, Qinviridae, Rhabdoviridae, Tombusviridae, Mimiviridae, Natoreviridae, and Jingchuvirales (all less than 1 per cent) (Li et al. 2015a; Shi et al. 2016). We also identified viruses related to negevirus (2 per cent of non-vertebrate virus abundance) (Nunes et al. 2017) and quenyaviruses (<1 per cent) (Obbard et al. 2020) (Supplementary Fig. S1). While many of these families include viruses that are known to infect fishes (e.g. Flaviviridae, Picornaviridae, Nodaviridae, Hepeviridae, Totiviridae, and Rhabdoviridae), the viruses identified here were highly divergent and clustered with those that infect a broad range of invertebrate and fungal species (Supplementary Figs S1–S4). We therefore focused our analysis on vertebrate-associated viruses.

Genome organization and phylogenetic relationships of vertebrate-associated viruses -ssRNA viruses

Five actinoviruses were identified (subfamily Actantavirinae and Hantaviridae) (Fig. 2E). All were novel, with the exception of pygmy

goby hantavirus previously identified in *E. zebrina* (Geoghegan et al. 2021). We identified complete and partial actinovirus genomes in *Trimma fangi*, *Luposicya lupus*, *Cephalopholis boenak*, and *Pleurostictus* sp. (Fig. 2E; Supplementary Table S3). The genome of *L. lupus* actinovirus contained all three expected actinovirus genomic regions (Fig. 2E). The S segment possessed an additional ORF upstream of the nucleoprotein in antisense orientation, also seen in the genome of perch actinovirus, but not in other actinoviruses (Hierweger et al. 2021).

Two partial genomes of paramyxoviruses were identified, one in *E. zebrina* (i.e. pygmy goby paramyxovirus) (Geoghegan et al. 2021) and the other in *Enneapterygius tutuillae*. This virus was highly divergent from other classified ray-finned fish paramyxoviruses, including the genus *Aquaparamyxovirus* (Fig. 2A). We also identified partial genomes (L protein and nucleocapsid sequences) of two novel rhabdoviruses in *Trimma okinawae* and *Trimma striatum*. The *T. okinawae* rhabdovirus fell within the monophyletic *Tupavirus* group that infects birds and mammals (Kurz et al. 1986; Allison et al. 2011), whereas *T. striatum* rhabdovirus clustered with the

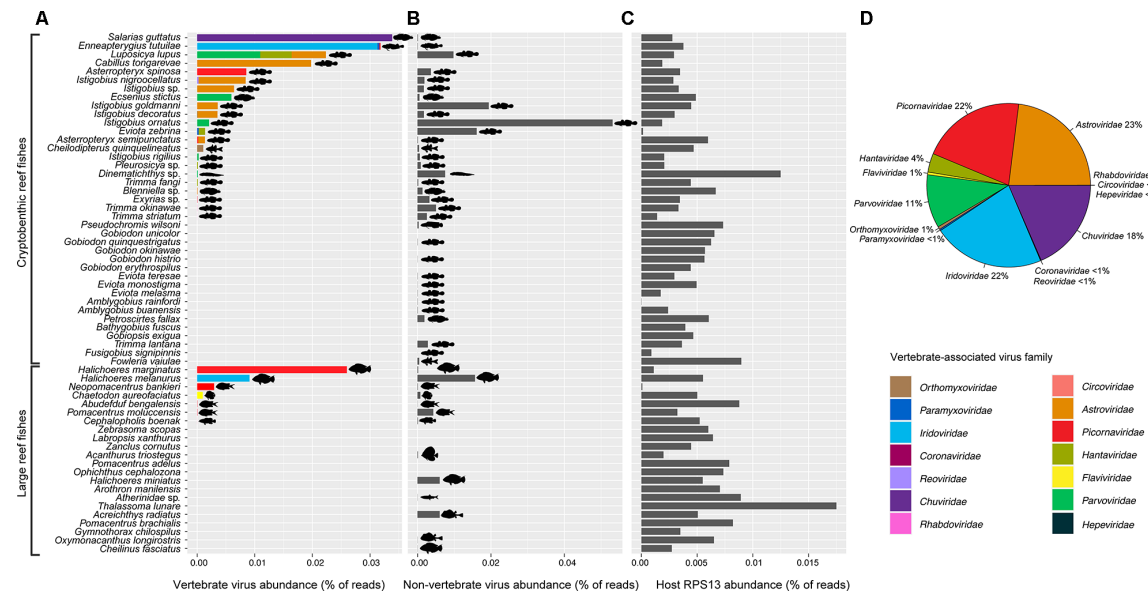


Figure 1. The standardized abundance of viral and host reads across reef fish libraries. The abundance (i.e. percentage of reads) of (A) likely vertebrate-associated viruses; (B) non-vertebrate-associated viruses (i.e. those from algae, fungi, coral, arthropods, crustaceans, and protists); and (C) host reference gene RPS13. Silhouettes represent viruses associated with fish species. (D) The total standardized abundance of vertebrate-associated viral families.

dimarhabdoviruses (Shi et al. 2018) (Fig. 2B). Among other -ssRNA viruses, we identified a novel pischichuvirus in *Salarias guttatus* and orthomyxovirus in *Cheilodipterus quinquelineatus* that showed sequence similarity to quaranjaviruses (Fig. 2D–E; Supplementary Table S3).

+ssRNA viruses

Astroviruses were detected in seven fish species. All were members of the Gobiidae, except for *Blenniella* sp. (Blenniidae). The astroviruses identified here clustered with those that infect a broad range of ray-finned and jawless fish species (Fig. 3D). Similarly, we identified a divergent virus from this group in *L. lupus*, tentatively named goby astrovirus 3. This virus had a genome of 7,355 nt with a predicted 5' untranslated region (UTR) of 347 nt, two main ORFs, and a 3' UTR of 347 nt. One ORF encoded the astrovirus ORF1a protein (protease), while the other encoded a polyprotein containing the RdRp and capsid protein. This arrangement is also seen in the genome of Wenling rattails astrovirus 5 but not in other currently described fish astroviruses that exhibit three main ORFs (Shi et al. 2018).

The complete genomes of three novel picornaviruses were identified in *Halichoeres marginatus*, *Neopomacentrus bankieri*, and *Asterropteryx spinosa*, and partial sequences of five novel picornaviruses were identified in *Asterropteryx semipunctatus*, *Pomacentrus moluccensis*, *E. tutuilae*, *Exyrias*, and *Blenniella* sp. (Fig. 3A). Although these fish species were members of the same community, all eight viruses were highly divergent. Phylogenetic analysis of RdRp revealed clustering with other fish picornaviruses, with the exception of *E. tutuilae* picornavirus and *Exyrias* picornavirus that clustered with *Ampivir* identified in the common newt (*Lissotriton vulgaris*) (Reuter et al. 2015) (Fig. 3A). We also identified the full genome of a novel hepacivirus (Flaviviridae) in *Chaetodon aureofaciatus* with conserved genomic regions—Core C,

Envelope E1, NS2, NS3, NS4, and NS5B—that fell within a distinct group of aquatic hepaciviruses (Shi et al. 2018) (Fig. 3B).

dsRNA viruses

The near-complete genome of a novel reovirus was identified in *Istigobius nigroocellatus*. This included six segments encoding the VP1 (guanyl transferase), VP2 (RdRp), VP3 (helicase), VP4 (NTPase), VP5 (outer capsid protein), and VP6 (core protein) proteins. Phylogenetic analysis of the VP2 gene revealed clustering within the genus *Aquareovirus*, sharing 48–60 per cent amino acid similarity with its closest relatives (Supplementary Table S3; Supplementary Fig. S5).

ssDNA viruses

We discovered a basal group of four parvoviruses that fell within the subfamily *Parvovirinae* (Parvoviridae) (Fig. 4A). The genome of *Ecsenius stictus* parvovirus was a single contig of 4,286 nt, containing two main ORFs. The left ORF was 1,374 nt and encoded the conserved non-structural protein NS1 that has DNA helicase and ATPase function (Pénzes et al. 2020). The right ORF encoded the structural VP1 protein (2,571 bp). *Istigobius ornatus* parvovirus had a genome of 4,284 nt and contained three main ORFs, one encoding NS1 (1,878 nt) and the other two encoding putative structural proteins (Fig. 4B). We also identified the partial genome of a novel ichthamaparvovirus in *L. lupus* and a novel circovirus in *Abudefduf bengalensis* that clustered within a distinct clade of fish circoviruses (Fig. 4C).

dsDNA viruses

The analysis of the *Halichoeres melanurus* meta-transcriptome identified nine conserved proteins sharing 99–100 per cent similarity with all currently described Santee-Cooper ranavirus isolates,

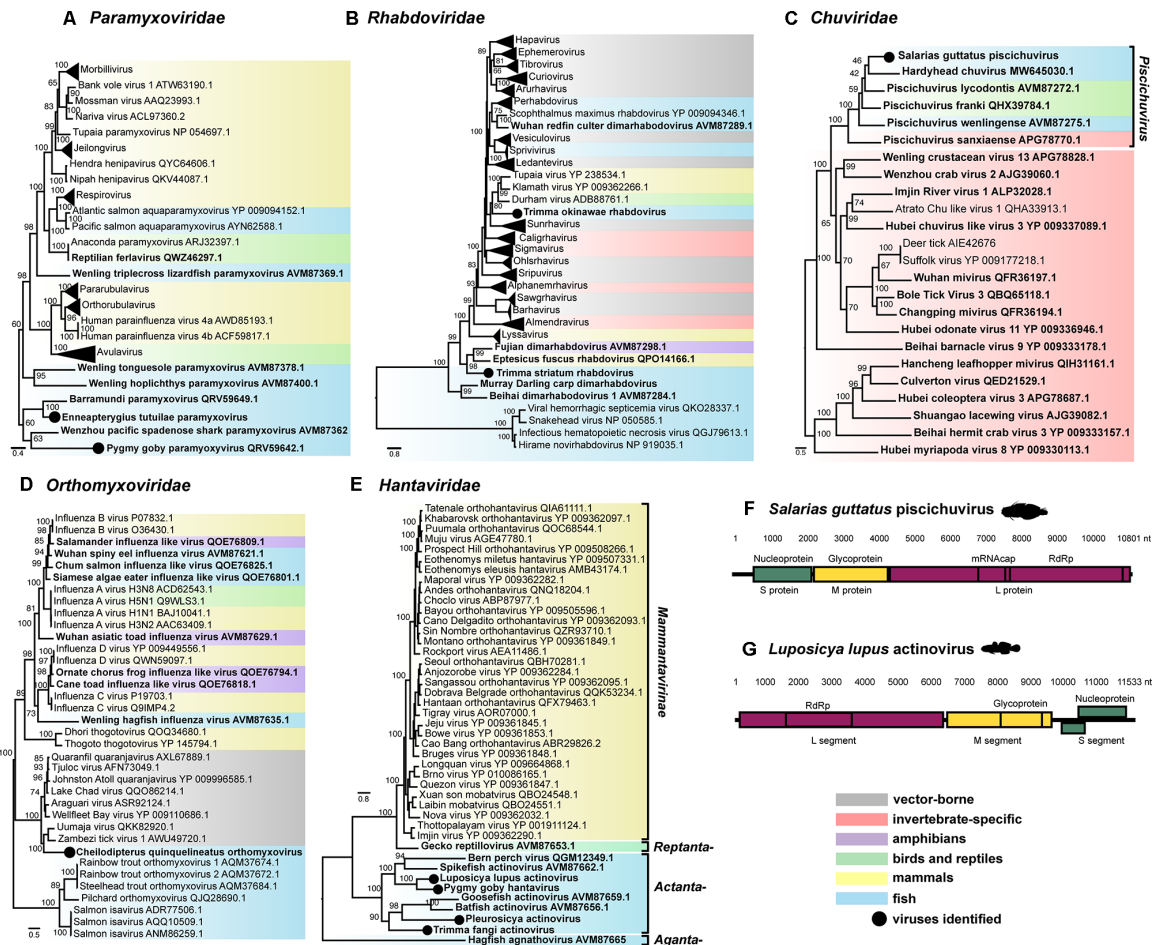


Figure 2. The genome structure and maximum likelihood phylogenies of the RdRp gene of -ssRNA viruses identified in this study. (A) The phylogeny of novel and related virus species of the *Paramyxoviridae*. (B) The phylogeny of the *Rhabdoviridae*. (C) The phylogeny of the *Chuviridae*. The genus *Piscichuvirus* is labelled to illustrate vertebrate-associated viruses. (D) The phylogeny of the *Orthomyxoviridae*. (E) The phylogeny of the *Hantaviridae*. The subfamilies *Mammantavirinae*, *Reptantavirinae*, *Actantavirinae*, and *Agantavirinae* are labelled to illustrate virus-host co-divergence. (F) The schematic genome of *Salaria guttatus piscichuvirus*. (G) The schematic genome of *Lupiscyca lupus actinovirus*. Trees were midpoint rooted for clarity. The scale bar represents the number of amino acid substitutions per site. Tip labels represent virus name with NCBI/GenBank accession. Viruses in the bold text represent sequences identified through metagenomics. Tree branches are highlighted to broadly represent host taxonomy. Schematic genome diagrams illustrate genome orientation, length, predicted ORFs, and gene products.

such as largemouth bass virus, mandarin fish ranavirus, koi ranavirus, doctor fish virus, and guppy virus 6 (*Iridoviridae*; Supplementary Table S3). We used the highly conserved major capsid protein and DNA polymerase for phylogenetic analysis (Whittington, Becker, and Dennis 2010), which further confirmed a novel Santee-Cooper ranavirus isolate, tentatively named *H. melanurus* ranavirus (Fig. 5). Largemouth bass virus and mandarin fish ranavirus are highly lethal in farmed populations, causing 95–100 per cent mortality (Zhang et al. 2017; Zhao et al. 2020). Our metatranscriptomic analysis also identified a novel iridovirus in *E. tuuila* that fell sister to erythrocytic necrosis virus (ENV) (Pagowski et al. 2019) and clustered with other erythrocytic-like viruses identified in amphibians and reptiles (Wellehan et al. 2008; Alves de Matos et al. 2010; Grosset et al. 2014; Russo et al. 2021). ENV was the closest relative across all proteins identified (Supplementary Table S3).

Cross-species virus transmission in a reef fish community

Despite the large number of viruses identified, we only found evidence for one cross-species transmission within our GBR ecosystem. This involved astroviruses found in five different fish species that clustered together and exhibited high levels of amino acid sequence similarity (Fig. 3D). Specifically, phylogenetic comparisons of ORF1b (RdRp) revealed two viral species: goby astrovirus 1, identified in *Istigobius goldmanni*, and goby astrovirus 2, identified in *I. nigroocellatus*, *Istigobius decoratus*, *A. semipunctatus*, and *Cabillus tongarevae* (Fig. 3D). These two viruses exhibit 82.5 per cent amino acid sequence similarity across the virus genome, while the four sequences of goby astrovirus 2 had 96.8 per cent amino acid similarity (Fig. 3D).

We also detected related viruses (i.e. those from the same virus family) in several different fish species, including the *Hantaviridae*,

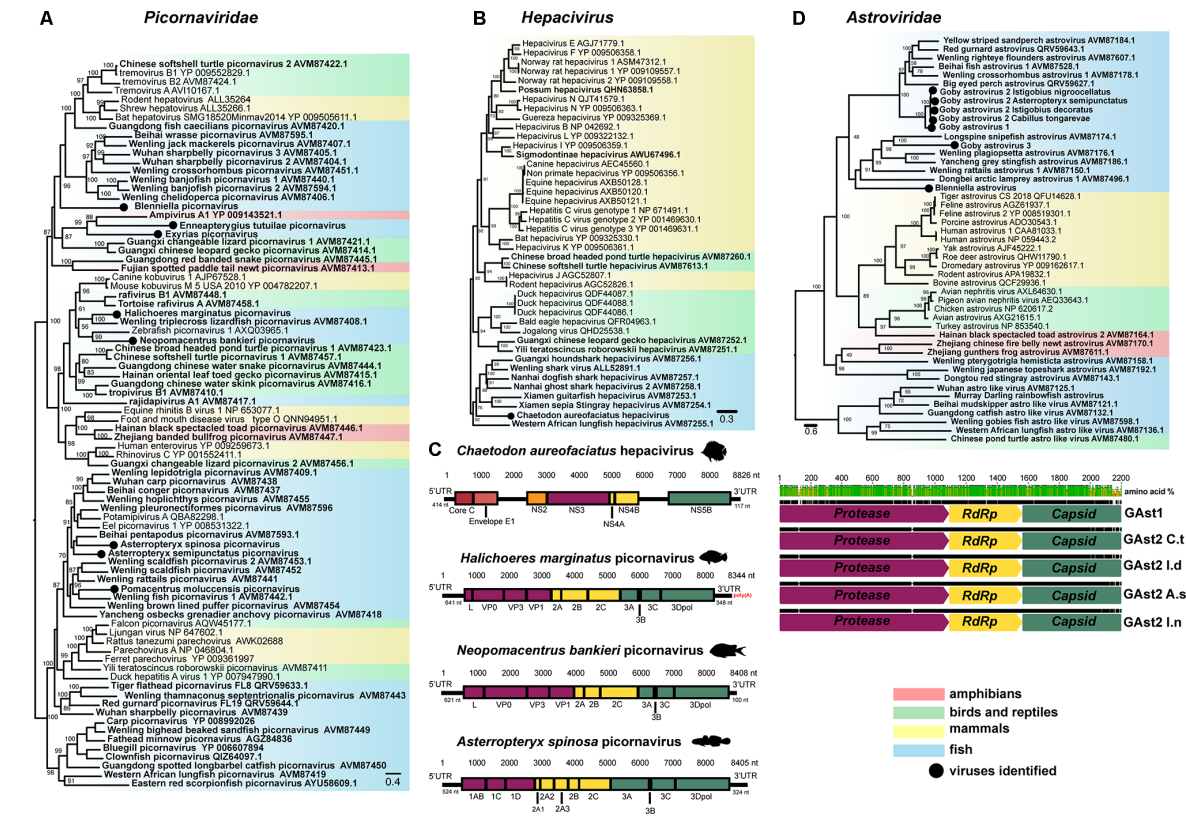


Figure 3. The genome organization and phylogenies of the RdRp gene of +ssRNA viruses identified in this study. (A) The maximum likelihood phylogeny of the Picornaviridae. (B) The phylogeny of the genus Hepacivirus (Flaviiviridae). (C) Schematic genomes illustrate genome orientation, length, and gene products. (D) The phylogeny of the Astroviridae and amino acid alignment of goby astrovirus 2 (GAST2) isolates against goby astrovirus 1 (GAST1). GAST2 are labelled to represent host species: GAST2 C.t = *C. tongarvae*, I.d = *I. decoratus*, A.s = *A. semipunctatus*, I.n = *I. nigroocellatus*. Trees were midpoint rooted for clarity. The scale bar represents the number of amino acid substitutions per site. Tip labels represent virus name with NCBI/GenBank accession. Viruses in the bold text represent sequences identified through metagenomics. Trees are highlighted to broadly illustrate host taxonomy.

Rhabdoviridae, Paramyxoviridae, and Picornaviridae (Fig. 6A). However, most of these were sufficiently divergent in sequence that they likely reflect common ancestry rather than direct host jumping in the reef ecosystem. For example, of the eight picornaviruses identified in this study, the closest relatives were *A. spinosa* picornavirus and *A. semipunctatus* picornavirus that shared only 44 per cent amino acid similarity.

In addition, we identified viral species not directly infecting the fish themselves but rather associated with the local environment, diet, or microbiome (i.e. non-vertebrate viruses) that were also transmitted among reef fish assemblages. These were as follows: quenyaviruses (95 per cent amino acid similarity between *P. moluccensis* and *A. bengalensis*), flavi-like viruses (91.2 per cent between *I. goldmanni* and *I. ornatus*), narnaviruses (97.7 per cent between *I. goldmanni*, *I. nigroocellatus*, and *Istigobius rigillus*) and totiviruses (93.5 per cent between *Amblygobius buanesis* and *Amblygobius rainfordi*) (Supplementary Figs S1, S2, and S4; Fig. 6C). That there were more instances of cross-species transmission of non-vertebrate viruses compared to those viruses that actively replicate in fish suggests that the latter group are subject to strong host barriers, even among closely related species.

Comparisons of viral alpha and beta diversity between reef fish families

We compared vertebrate virome composition between reef fish families, as well as between cryptobenthic reef fishes and large reef fishes (i.e. that differ in size). In our data set, cryptobenthic reef fish families included the Gobiidae, Apogonidae, Blenniidae, Tripterygiidae, Bythitidae, and Pseudochromidae, while the large reef fish families included the Pomacentridae, Acanthuridae, Tetraodontidae, Atherinidae, Serranidae, Monacanthidae, Chaetodontidae, Labridae, Muraeidae, and Ophichthidae (Brandl et al. 2018).

Three statistical measures were used to assess alpha diversity: viral abundance (i.e. the standardized number of viral reads), observed viral richness (i.e. the number of viruses), and the Shannon index. We found an association between fish size and observed viral richness, with cryptobenthic reef fishes harbouring more viruses than large reef fishes ($\chi^2 = 2.795$, $df = 1$, $P = 0.028$) (Fig. 7). Among cryptobenthic reef fishes, *E. tutuila* contained the most vertebrate-associated viruses with four viral species, followed by *L. lupus* with three viral species (Fig. 6A). To assess whether this difference might be driven by these two outliers, we

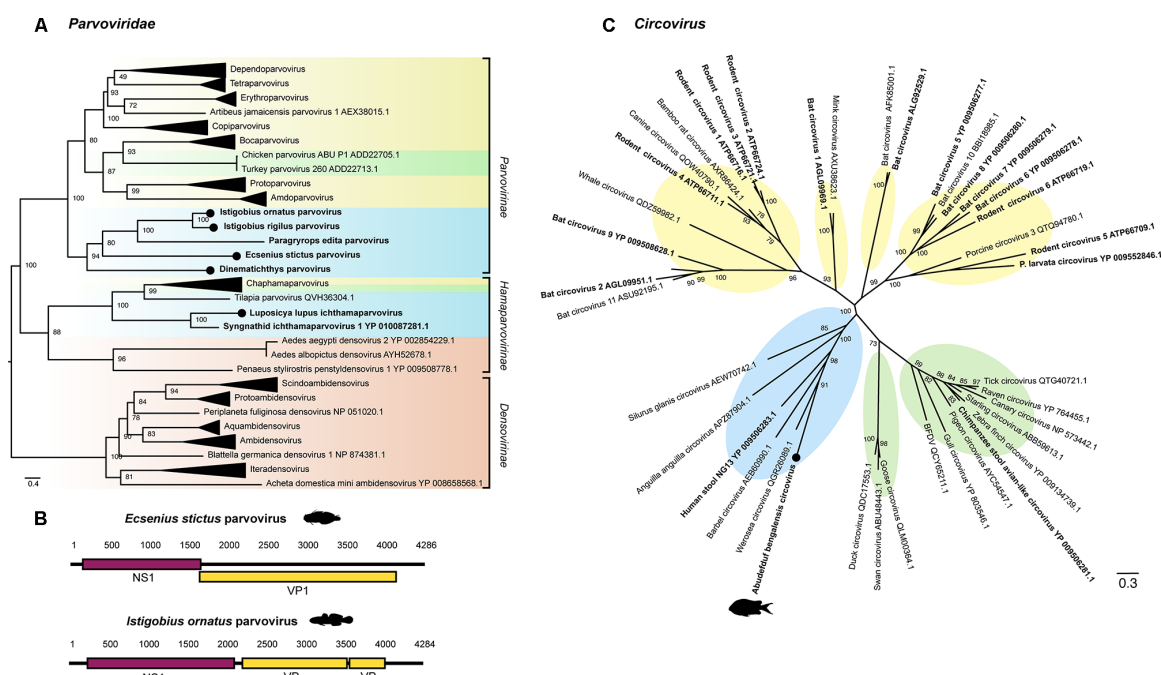


Figure 4. Phylogenetic relationships of the ssDNA viruses identified in this study. (A) The phylogenetic analysis of the NS1 gene of the Parvoviridae. Viruses discovered here are represented as black circles. The tree is midpoint rooted for clarity. The scale bar represents the number of amino acid substitutions per site. (B) Schematic genomes (nt) of discovered parvoviruses. Coloured boxes represent both structural (NS1) and non-structural (VP) ORFs. (C) Unrooted phylogeny of the replication-associated protein of the genus Circovirus (Circoviridae). The circovirus identified is represented as a fish symbol. Branches of both trees are highlighted to represent the host class: blue, fish; red, invertebrates; green, birds; and yellow, mammals. Viruses in the bold text represent sequences identified through metagenomics.

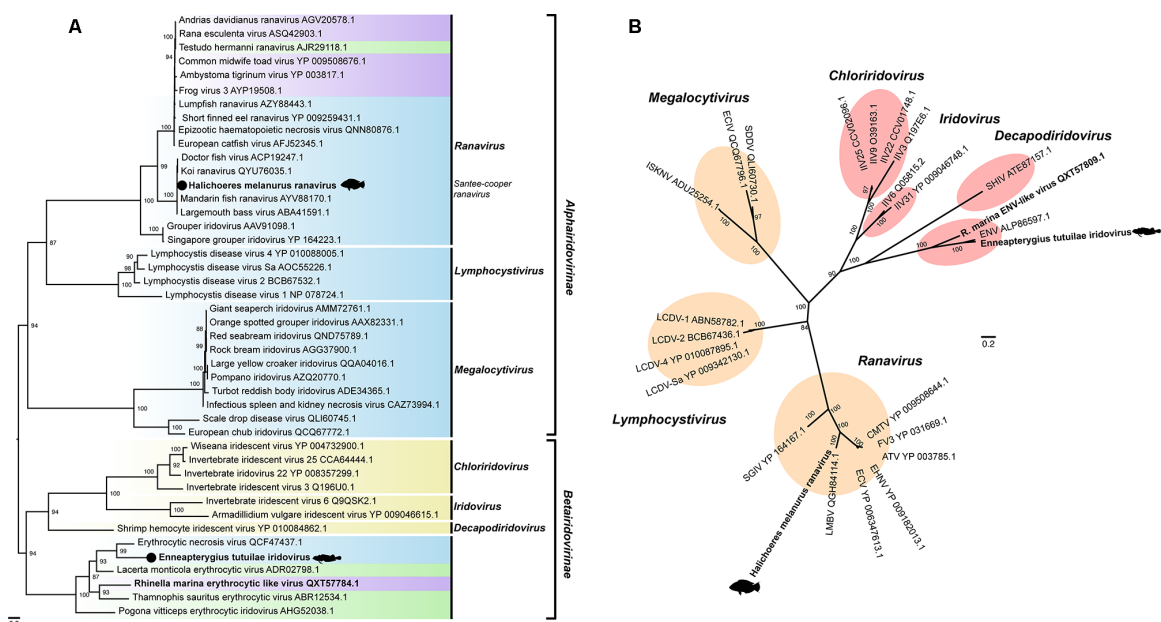


Figure 5. (A) The phylogenetic analysis of the DNA polymerase gene among the Iridoviridae. The scale bar represents the number of amino acid substitutions per site. Tip labels represent virus name with NCBI/GenBank accession. Viruses discovered here are represented as a black circle and fish symbol. Branches are highlighted to illustrate host class: blue, fish; purple, amphibians; green, reptiles; and yellow, invertebrates. (B) Unrooted phylogeny of the major capsid protein of the Iridoviridae. Branches are coloured to represent both subfamilies: orange, Alphaviridae; red, Betaviridae. Viruses in the bold text represent sequences identified through metagenomics.

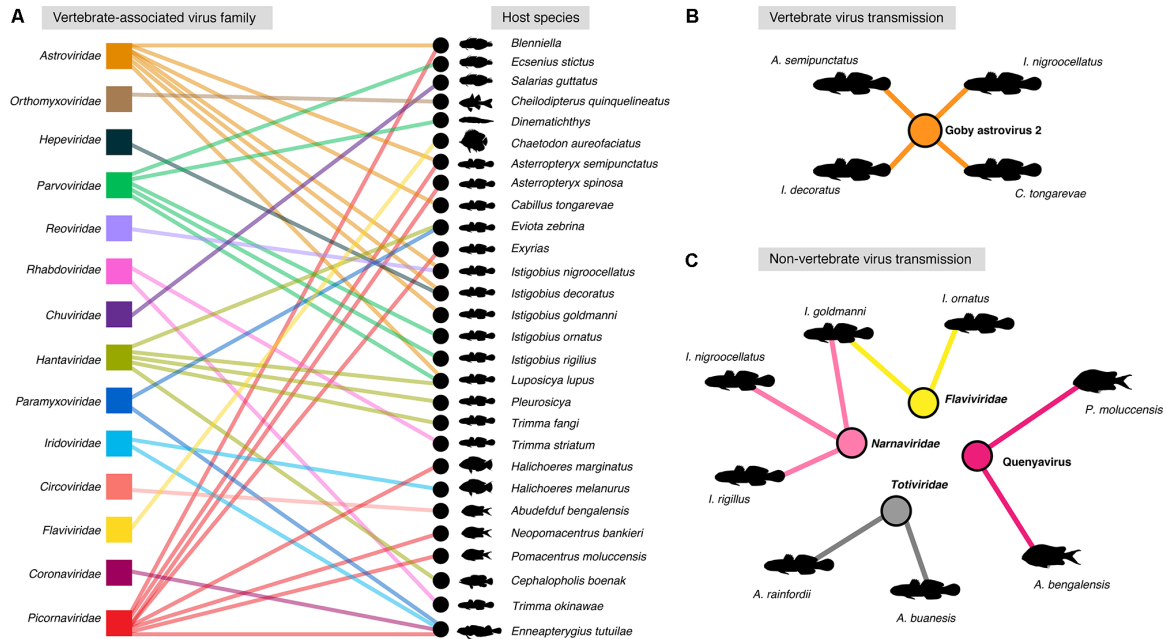


Figure 6. (A) The bipartite network illustrating viral families shared between fish taxa. Viral families are represented as a coloured box, with viruses connected between host species as a coloured line. (B) Network diagram illustrating vertebrate-associated viral species shared between fish. Coloured circle and lines illustrate viral species. (C) Network diagram revealing non-vertebrate viral species shared between fish libraries. All fish silhouettes broadly represent host species at the family level.

omitted these libraries and repeated the analysis. This likewise revealed a significant difference with cryptobenthics exhibiting greater observed viral richness than large reef fish ($\chi^2 = 1.507$, $df = 1$, $P = 0.044$).

We identified an association between reef fish family and observed viral richness, with the Tripterygiidae exhibiting greater viral diversity than all other reef fish families ($\chi^2 = 16.678$, $df = 15$, $P = 0.007$). However, we found no association between reef fish family and viral abundance ($P = 0.153$), Shannon diversity ($P = 0.901$), or beta diversity ($R^2 = 0.334$, $P = 0.064$). Likewise, we identified no significant differences in viral abundance ($P = 0.271$), Shannon diversity ($P = 0.142$), nor beta diversity ($R^2 = 0.048$, $P = 0.121$) between cryptobenthic reef fishes and large reef fishes.

As a form of internal control, we repeated our analyses of viral abundance and diversity on the non-vertebrate viruses identified here. This analysis revealed no significant differences in viral abundance between reef fish families ($P = 0.994$) nor between cryptobenthic reef fishes and larger reef fishes ($P = 0.355$). Similarly, we found no significant difference in observed viral richness between fish families ($P = 0.733$) although cryptobenthic reef fishes exhibited higher observed non-vertebrate viral richness than large reef fish families ($\chi^2 = 10.805$, $df = 1$, $P = 0.016$). We observed no difference in Shannon diversity among fish families ($P = 0.453$), as well as between cryptobenthic reef fishes and larger reef fishes ($P = 0.070$). Similarly, there was no association between beta diversity and reef fish families ($R^2 = 0.279$, $P = 0.126$) nor between cryptobenthic reef fishes and large reef fishes ($R^2 = 0.335$, $P = 0.058$).

We tested for any differences in alpha diversity between dissected and whole fish libraries. This identified no significant differences in vertebrate-associated viral abundance ($P = 0.903$),

observed viral richness ($P = 0.058$), or Shannon diversity ($P = 0.156$). Similarly, we detected no significant differences in non-vertebrate viral abundance ($P = 0.055$), observed viral richness ($P = 0.063$), and Shannon diversity ($P = 0.364$). Finally, we tested whether the number of fish per library impacted viral alpha diversity. As expected, this revealed a positive relationship between the number of replicates per fish library and vertebrate viral richness ($R^2 = 0.115$, $P = 0.001$).

Discussion

The GBR supports over 1,200 species of fish and is the largest coral reef ecosystem in the world, comprising 2,500 reefs across approximately 344,400 km² (De'ath et al. 2012). Despite the global importance of the GBR, little is known about the natural diversity of viruses that infect tropical reef fishes, as well as the ecological and evolutionary processes that allow such viruses to spread within a reef fish community. To this end, we employed metatranscriptomic sequencing to characterize the viromes of sixty-one tropical reef fish species, including those occupying a 100-m² reef fish community and several from the broader surrounding ecosystem. This identified transcripts representing 132 viral sequences, including 38 that exhibited strong phylogenetic relationships with viral genera known to infect vertebrates.

While we sampled coral reef fishes from a small spatial area, we identified a marked absence of cross-species virus transmission, with the only instance of host jumping being the presence of a single viral species (Astroviridae) in four goby species (Figs 3C and 6B). However, the observed level of genetic variation (96.8 per cent amino acid similarity) is indicative of past host switching during fish evolution, which may span millions of years, rather than direct host jumping within the ecosystem sampled here.

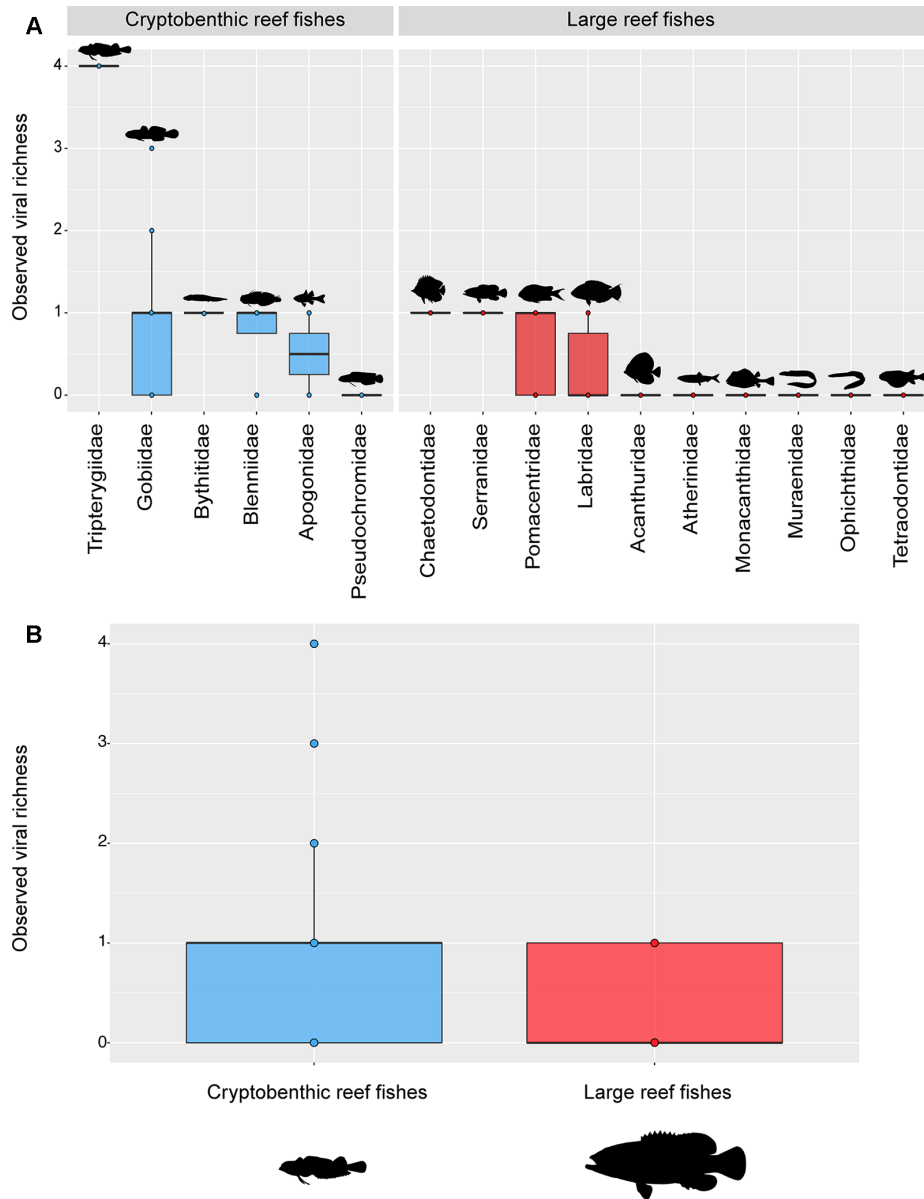


Figure 7. (A) The vertebrate-associated viral richness observed in all tropical reef fish families. (B) The comparison of observed vertebrate-associated viral richness between cryptobenthic reef fishes and large reef fishes. Gobiidae ($n = 30$ species), Labridae ($n = 6$), Pomacentridae ($n = 5$), Blenniidae ($n = 5$), Acanthuridae ($n = 3$), Apogonidae ($n = 2$), Monacanthidae ($n = 2$), Tetraodontidae ($n = 1$), Pseudochromidae ($n = 1$), Chaetodontidae ($n = 1$), Atherinidae ($n = 1$), Serranidae ($n = 1$), Tripterygiidae ($n = 1$), Muraenidae ($n = 1$), Bythitidae ($n = 1$), and Ophichthidae ($n = 1$).

Given the restricted spatial area considered, and the density of potential hosts, such a lack of cross-species transmission suggests that there may be important host barriers (genetic, immune, and physiological) to virus switching among the reef fishes sampled herein. This is supported by the observation that non-vertebrate viruses—that we assume are not impacted by aspects of fish genetics—were characterized by higher levels of cross-species virus transmission. Although more research is needed to understand the nature of these barriers in fish, they may in part reflect differences in host cell receptor binding (Longdon et al. 2014). For instance, although the betacoronavirus RaTG13 sampled from

Rhinolophus affinis bats is closely related (~96 per cent sequence similarity) to SARS-CoV-2, it is unable to bind to the human ACE2 receptor (Wrobel et al. 2020).

These results also suggest that the reef ecosystem might select for specialist viruses, such that there is no fitness advantage in infecting multiple fish species despite ample exposure. Most fishes within this reef fish community are typically found in high abundances that are sustained year-round (Bellwood et al. 2006; Lefèvre et al. 2016; Brandl et al. 2018). The apparent lack of seasonal variation in fish densities as well as continuously high contact rates and rapid life cycles implies that most of the viruses detected here

have evolved to specialize within a reef fish community and maximize fitness in a single host type. Indeed, specialist-generalist theory suggests that in most host–parasite systems, evolution should select for specialist viruses, particularly when host populations remain stable over space and time (Elena, Agudelo-Romero, and Lalic 2009). Moreover, the evolution of generalist viruses is often restricted by antagonistic pleiotropy, in which mutations that are beneficial in one host species are deleterious in others (Lefevre et al. 2019). That we observed very limited evidence for virus generalization across multiple viral families implies that the ecological stability of this reef environment facilitates virus specialization.

We detected differences in vertebrate virome composition between reef fish families, with cryptobenthic reef fishes harbouring more viruses than large reef fishes (Fig. 7). Due to their small body size, cryptobenthic reef fishes exhibit significantly higher rates of metabolism than large reef fishes, resulting in high energy demands with a low tolerance for starvation (Brandl et al. 2018). Given their extremely short lifespans, there are high energy demands for their rapid growth and complex reproductive strategies (e.g. sex change), such that there might be important energy trade-offs between growth, reproduction, and immunity, particularly as immune systems require a substantial allocation of resources for recognizing and eliminating pathogens (Lochmiller and Deerenberg 2000). While it is unclear if and how such energy demands impact immunity in cryptobenthic reef fishes, it is notable that this lifestyle is associated with a more diverse virome than large reef fishes, suggesting that these species may be more susceptible to infection. Indeed, with a daily mortality rate of 8–9 per cent (Depczynski and Bellwood 2005, 2006), investing in immunity may be a poor strategy, and it is possible that these energy-saving strategies are key to the spectacular reproductive success of these cryptobenthic fishes (Brandl et al. 2019). Whether as simple larvae in the pelagic realm, or on the reef, cryptobenthic fishes may be able to minimize energetically expensive activities such as immunity and swimming, freeing up energy to maintain adequate fecundity, despite exceptionally high mortality rates and infection risks.

Among reef fish families, the Tripterygiidae exhibited the greatest observed viral richness, with *E. tutuilae* harbouring four viral species. *Enneapterygius tutuilae* is a generalist on coral reefs, utilizing sand and rubble, soft coral, cave, and open reef microhabitats (Depczynski and Bellwood 2004). It is therefore possible that habitat generalism may increase interactions with other fish species and hence increase the likelihood of being infected by a larger number of viruses although only one example of relatively recent cross-species virus transmission was detected in our system. In contrast, *Gobiodon* species are extreme habitat specialists and may often only occupy a single coral species as habitat for its entire life with minimal interactions with other fish species (Munday, Jones, and Caley 1997). This lifestyle may explain why all five *Gobiodon* species examined here harboured no vertebrate-associated viruses, as well as few non-vertebrate-associated viruses (Fig. 1A–B).

Given the phylogenetic distance between fish and other vertebrate classes, as well as long-term associations through virus–host co-divergence (Zhang et al. 2018), it was not unexpected that 95 per cent of the vertebrate-associated viruses discovered here clustered with other fish viruses. A case in point comes from the phylogeny of the *Hantaviridae* that clearly reflects the broad evolutionary history of vertebrates, with the *Agantavirinae* (jawless fish) falling basal to the *Actantavirinae* (ray-finned fish), *Reptantavirinae* (reptiles), and *Mammantavirinae* (mammals),

respectively (Fig. 2). Despite their likely long-term presence in ray-finned fish, actinoviruses have only recently been discovered and have been associated with disease in farmed species (Shi et al. 2018; Geoghegan et al. 2021; Hierweiger et al. 2021). For instance, a novel actinovirus—perch actinovirus—was identified in diseased European perch (*Perca fluviatilis*) with high concentrations of viral RNA in gill endothelial cells and macrophages (Hierweiger et al. 2021). Our discovery of four actinoviruses in gobies suggests that these viruses may be widespread in this family and hence should be monitored closely if interacting with farmed populations.

Another virus of concern identified in our study is a novel isolate of *Santee-Cooper ranavirus* in *H. melanurus*. To the best of our knowledge, this is the first discovery of this virus in Australia. Our detection of this virus in seemingly healthy wrasses, as well as its natural presence in cleaner wrasses (*Labroides dimidiatus*), suggests that this family of tropical reef fishes may be important reservoir hosts for the *Santee-Cooper ranavirus* group (Hedricks and McDowell 1995). For instance, disease outbreaks with high mortality rates have only been observed in farmed freshwater fish species including largemouth bass (*Micropterus salmoides*), mandarin fish (*Siniperca chuatsi*), and koi (*Cyprinus carpio*) (Plumb et al. 1996; Grizzle et al. 2002; George et al. 2015; Zhang et al. 2020; Zhao et al. 2020).

Despite the long-known presence of viral erythrocytic necrosis in various fish species across the North Atlantic and Pacific Oceans, the genome of the causative virus—ENV—was only recently characterized (Pagowski et al. 2019). We identified a closely related virus in *E. tutuilae*. Viral erythrocytic necrosis in a juvenile triggerfish (*Rhinecanthus aculeatus*) has been reported nearby at Lizard Island (Davies et al. 2009), suggesting that this virus may be circulating between several reef locations along the GBR. Phylogenetic comparisons of DNA polymerase and major capsid protein revealed a distinct ‘ectothermic vertebrate’ clade that could be clearly classified as a novel genus within the *Betairidovirinae*, *Iridoviridae* (Fig. 5).

While we observed significant differences in virome composition between coral reef fishes, there were necessary limitations in our sampling. For instance, cryptobenthic fishes made up 62 per cent of the fish diversity in this study. In addition, twenty-three libraries contained only one individual. Such unequal sample sizes likely impacted the statistical power of our analyses. Accordingly, where possible, future work should balance the number of samples as well as increase the number of reef sampling sites. Another limitation was the difficulty in detecting highly divergent viral sequences with the similarity-based methods used here. Moreover, given the RNA-based nature of metatranscriptomics, we were only able to detect DNA viruses that were actively generating transcripts, in turn reducing the chances of identifying viruses that may be expressing genes at lower frequencies.

In sum, our study identified a large diversity of viruses in tropical reef fish assemblages. We identified a marked absence of virus exchange within a reef fish community, suggesting that there may be important host barriers for successful cross-species virus transmission. Our discovery of a novel *Santee-Cooper ranavirus* isolate in seemingly healthy wrasses highlights the importance of virological surveillance in marine wildlife, particularly as this virus can cause significantly high mortality in farmed fishes. As such, these species should be considered in biosecurity risk assessments and screened if utilized for aquaculture or aquarium operations (Rimmer et al. 2015). Accordingly, future studies should also investigate its susceptibility in Australian food fish to fully assess its threat of emergence (Hedricks and McDowell 1995). Overall, this study increases our knowledge on the severely

understudied coral reef fish virome and provides the first data on virus–host interactions in a reef fish community.

Data availability

All sequence reads are available on the NCBI Sequence Read Archive under BioProject PRJNA841039, and all generated virus sequences have been deposited in NCBI/GenBank under accessions ON595936–ON596019.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Funding

This work was funded by the Australian Research Council (ARC) Australian Laureate Fellowships to E.C.H. (FL170100022) and D.R.B. (FL190100062) and an ARC Discovery Project grant to E.C.H. and J.L.G. (DP200102351). J.L.G. is funded by a New Zealand Royal Society Rutherford Discovery Fellowship (RDF-20-UOO-007) and a Marsden Fast Start grant (20-UOO-105).

Conflict of interest: The authors declare no competing interests.

References

Ackerman, J. L., and Bellwood, D. R. (2002) 'Comparative Efficiency of Clove Oil and Rotenone for Sampling Tropical Reef Fish Assemblages', *Journal of Fish Biology*, 60: 893–901.

Allison, A. B. et al. (2011) 'Characterization of Durham Virus, a Novel Rhabdovirus that Encodes Both a C and SH Protein', *Virus Research*, 155: 112–22.

Alves de Matos, A. P. et al. (2010) 'New Viruses from *Lacerta monticola* (Serra da Estrela, Portugal): Further Evidence for a New Group of Nucleo-Cytoplasmic Large Deoxyriboviruses', *Microscopy and Microanalysis*, 17: 101–8.

Barton, J. A. et al. (2020) 'Biological Controls to Manage Acropora-eating Flatworms in Coral Aquaculture', *Aquaculture Environment Interactions*, 12: 61–6.

Bellwood, D. R. et al. (2004) 'Confronting the Coral Reef Crisis', *Nature*, 429: 827–33.

Bellwood, D. R. et al. (2006) 'Coral Bleaching, Reef Fish Community Phase Shifts and the Resilience of Coral Reefs', *Global Change Biology*, 12: 1587–94.

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

Booth, D. J. et al. (2018) 'Predicting Success of Range-expanding Coral Reef Fish in Temperate Habitats Using Temperature-abundance Relationships', *Frontiers in Marine Science*, 5: 31.

Brandl, S. J. et al. (2018) 'The Hidden Half: Ecology and Evolution of Cryptobenthic Fishes on Coral Reefs', *Biological Reviews of the Cambridge Philosophical Society*, 93: 1846–73.

Brandl, S. J. et al. (2019) 'Demographic Dynamics of the Smallest Marine Vertebrates Fuel Coral Reef Ecosystem Functioning', *Science*, 364: 1189–92.

Buchfink, B., Xie, C., and Huson, D. H. (2015) 'Fast and Sensitive Protein Alignment Using DIAMOND', *Nature Methods*, 12: 59–60.

Burge, C. A. et al. (2014) 'Climate Change Influences on Marine Infectious Diseases: Implications for Management and Society', *Annual Review of Marine Science*, 6: 249–77.

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009) 'trimAl: A Tool for Automated Alignment Trimming in Large-scale Phylogenetic Analyses', *Bioinformatics*, 25: 1972–3.

Costa, V. A. et al. (2021) 'Metagenomic Sequencing Reveals a Lack of Virus Exchange between Native and Invasive Freshwater Fish across the Murray–Darling Basin, Australia', *Virus Evolution*, 7: veab034.

Davies, A. J. et al. (2009) 'Suspected Viral Erythrocytic Necrosis (VEN) in a Juvenile Blackbar Triggerfish, *Rhinecanthus aculeatus*, from Lizard Island, Great Barrier Reef, Australia', *Marine Biodiversity Records*, 2: E149.

De'ath, G. et al. (2012) 'The 27-year Decline of Coral Cover on the Great Barrier Reef and Its Causes', *Proceedings of the National Academy of Sciences*, 109: 17995–9.

Depczynski, M., and Bellwood, D. R. (2003) 'The Role of Cryptobenthic Reef Fishes in Coral Reef Trophodynamics', *Marine Ecology Progress Series*, 256: 183–91.

——— (2004) 'Microhabitat Utilisation Patterns in Cryptobenthic Coral Reef Fish Communities', *Marine Biology*, 145: 455–63.

——— (2005) 'Shortest Recorded Vertebrate Lifespan Found in a Coral Reef Fish', *Current Biology*, 15: 288–9.

——— (2006) 'Extremes, Plasticity, and Invariance in Vertebrate Life History Traits: Insights from Coral Reef Fishes', *Ecology*, 87: 3119–27.

Dixon, P. (2003) 'VEGAN, a Package of R Functions for Community Ecology', *Journal of Vegetation Science*, 14: 927–30.

Elena, S. F., Agudelo-Romero, P., and Lalic, J. (2009) 'The Evolution of Viruses in Multi-host Fitness Landscapes', *The Open Virology Journal*, 3: 1–6.

FAO. (2020), *The State of World Fisheries and Aquaculture. Sustainability in Action* <<https://www.fao.org/documents/card/en/c/ca9229en/>> accessed 9 Feb 2022.

Geoghegan, J. L. et al. (2018) 'Hidden diversity and evolution of viruses in market fish', *Virus Evolution*, 4: vey031.

——— (2021) 'Virome Composition in Marine Fish Revealed by Metatranscriptomics', *Virus Evolution*, 7: veab005.

George, M. R. et al. (2015) 'Isolation and Characterization of a Ranavirus from Koi, *Cyprinus carpio* L., Experiencing Mass Mortalities in India', *Journal of Fish Diseases*, 38: 389–403.

Go, J. et al. (2006) 'The Molecular Epidemiology of Iridovirus in Murray Cod (*Maccullochella peelii peelii*) and Dwarf Gourami (*Colisa lalia*) from Distant Biogeographical Regions Suggests a Link between Trade in Ornamental Fish and Emerging Iridoviral Diseases', *Molecular and Cellular Probes*, 20: 212–22.

Grizzle, J. M. et al. (2002) 'First Isolation of Largemouth Bass Virus', *Diseases of Aquatic Organisms*, 50: 233–5.

Grosset, C. et al. (2014) 'Intraerythrocytic Iridovirus in Central Bearded Dragons (*Pogona vitticeps*)', *Journal of Veterinary Diagnostic Investigation*, 26: 354–64.

Haas, B. J. et al. (2013) 'De Novo Transcript Sequence Reconstruction from RNA-seq Using the Trinity Platform for Reference Generation and Analysis', *Nature Protocols*, 8: 1494–512.

Hedricks, R. P., and McDowell, T. S. (1995) 'Properties of Iridoviruses from Ornamental Fish', *Veterinary Research*, 26: 423–7.

Heron, S. F. et al. (2016) 'Warming Trends and Bleaching Stress of the World's Coral Reefs 1985–2012', *Scientific Reports*, 6: 38402.

Hierweger, M. M. et al. (2021) 'Novel Filoviruses, Hantavirus, and Rhabdovirus in Freshwater Fish, Switzerland, 2017', *Emerging Infectious Diseases*, 27: 3082–91.

Hothorn, T., Bretz, F., and Westfall, P. (2008) 'Simultaneous Inference in General Parametric Models', *Biometrical Journal*, 50: 346–63.

Johnson, W. E. (2019) 'Origins and Evolutionary Consequences of Ancient Endogenous Retroviruses', *Nature Reviews Microbiology*, 17: 355–70.

Jones, P. et al. (2014) 'InterProScan 5: Genome-scale Protein Function Classification', *Bioinformatics*, 30: 1236–40.

- Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods*, 14: 587–9.
- Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.
- Kearse, M. et al. (2012) 'Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data', *Bioinformatics*, 28: 1647–9.
- Kibenge, F. S. B. (2019) 'Emerging Viruses in Aquaculture', *Current Opinion in Virology*, 34: 97–103.
- Kurz, W. et al. (1986) 'Isolation and Characterization of a Tupaia Rhabdovirus', *Intervirology*, 25: 88–96.
- Lagkouvardos, I. et al. (2017) 'Rhea: A Transparent and Modular R Pipeline for Microbial Profiling Based on 16S rRNA Gene Amplicons', *PeerJ*, 5: e2836.
- Lam, V. W. Y. et al. (2020) 'Climate Change, Tropical Fisheries and Prospects for Sustainable Development', *Nature Reviews Earth & Environment*, 1: 440–54.
- Lancaster, M. J., Williamson, M. M., and Schroen, C. J. (2003) 'Iridovirus-associated Mortality in Farmed Murray Cod (*Maccullochella peelii peelii*)', *Australian Veterinary Journal*, 81: 633–4.
- Langmead, B., and Salzberg, S. L. (2012) 'Fast Gapped-read Alignment with Bowtie 2', *Nature Methods*, 9: 357–9.
- Lefeuivre, P. et al. (2019) 'Evolution and Ecology of Plant Viruses', *Nature Reviews Microbiology*, 17: 632–44.
- Lefèvre, C. D. et al. (2016) 'Consequences of Extreme Life History Traits on Population Persistence: Do Short-lived Gobies Face Demographic Bottlenecks?', *Coral Reefs*, 35: 399–409.
- Li, C. X. et al. (2015a) 'Unprecedented Genomic Diversity of RNA Viruses in Arthropods Reveals the Ancestry of Negative-sense RNA Viruses', *eLife*, 4: e05378.
- Li, D. et al. (2015b) 'MEGAHIT: An Ultra-fast Single-node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph', *Bioinformatics*, 31: 1674–6.
- Lochmiller, R. L., and Deerenberg, C. (2000) 'Trade-offs in Evolutionary Immunology: Just What Is the Cost of Immunity?', *Oikos*, 88: 87–98.
- Longdon, B. et al. (2014) 'The Evolution and Genetics of Virus Host Shifts', *PLoS Pathogens*, 10: e1004395.
- McMurdie, P. J., and Holmes, S. (2013) 'Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data', *PLoS One*, 8: e61217.
- Moberg, F., and Folke, C. (1999) 'Ecological Goods and Services of Coral Reef Ecosystems', *Ecological Economics*, 29: 215–33.
- Munday, P. L., Jones, G. P., and Caley, M. J. (1997) 'Habitat Specialisation and the Distribution and Abundance of Coral-dwelling Gobies', *Marine Ecology Progress Series*, 152: 227–39.
- Murray, A. G. (2016) 'A Modelling Framework for Assessing the Risk of Emerging Diseases Associated with the Use of Cleaner Fish to Control Parasitic Sea Lice on Salmon Farms', *Transboundary and Emerging Diseases*, 63: 270–7.
- Nayfach, S. et al. (2021) 'CheckV Assesses the Quality and Completeness of Metagenome-assembled Viral Genomes', *Nature Biotechnology*, 39: 578–85.
- Nguyen, L. T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.
- Nunes, M. R. T. et al. (2017) 'Genetic Characterization, Molecular Epidemiology, and Phylogenetic Relationships of Insect-specific Viruses in the Taxon Negevirus', *Virology*, 504: 152–67.
- Obbard, D. J. et al. (2020) 'A New Lineage of Segmented RNA Viruses Infecting Animals', *Virus Evolution*, 6: vez061.
- Ostfeld, R. S., and Keesing, F. (2012) 'Effects of Host Diversity on Infectious Disease', *Annual Review of Ecology, Evolution, and Systematics*, 43: 157–82.
- Pagowski, V. A. et al. (2019) 'Distribution and Phylogeny of Erythrocytic Necrosis Virus (ENV) in Salmon Suggests Marine Origin', *Viruses*, 11: 358.
- Panek, F. M. (2005) 'Epizootics and Disease of Coral Reef Fish in the Tropical Western Atlantic and Gulf of Mexico', *Reviews in Fisheries Science*, 13: 1–21.
- Parrish, C. R. et al. (2008) 'Cross-species Virus Transmission and the Emergence of New Epidemic Diseases', *Microbiology and Molecular Biology Reviews*, 72: 457–70.
- Pénzes, J. J. et al. (2020) 'Reorganizing the Family Parvoviridae: A Revised Taxonomy Independent of the Canonical Approach Based on Host Association', *Archives of Virology*, 165: 2133–46.
- Plumb, J. A. et al. (1996) 'An Iridovirus Isolated from Wild Largemouth Bass', *Journal of Aquatic Animal Health*, 8: 265–70.
- Reuter, G. et al. (2015) 'A Highly Divergent Picornavirus in an Amphibian, the Smooth Newt (*Lissotriton vulgaris*)', *Journal of General Virology*, 96: 2607–13.
- Rimmer, A. E. et al. (2015) 'Detection of Dwarf Gourami Iridovirus (Infectious Spleen and Kidney Necrosis Virus) in Populations of Ornamental Fish Prior to and after Importation into Australia, with the First Evidence of Infection in Domestically Farmed Platy (*Xiphophorus maculatus*)', *Preventive Veterinary Medicine*, 122: 181–94.
- Russo, A. G. et al. (2021) 'Discovery of Novel Viruses Associated with the Invasive Cane Toad (*Rhinella marina*) in Its Native and Introduced Ranges', *Frontiers in Microbiology*, 12: 733631.
- Santini, L. et al. (2018) 'Global Drivers of Population Density in Terrestrial Vertebrates', *Global Ecology and Biogeography*, 27: 968–79.
- Schmidt, K. A., and Ostfeld, R. S. (2001) 'Biodiversity and the Dilution Effect in Disease Ecology', *Ecology*, 82: 609–19.
- Shi, M. et al. (2016) 'Redefining the Invertebrate RNA Virosphere', *Nature*, 540: 539–43.
- Shi, M. et al. (2018) 'The Evolutionary History of Vertebrate RNA Viruses', *Nature*, 556: 197–202.
- Spalding, M. D., and Grenfell, A. M. (1997) 'New Estimates of Global and Regional Coral Reef Areas', *Coral Reefs*, 16: 225–30.
- Stuart-Smith, R. D. et al. (2018) 'Ecosystem Restructuring along the Great Barrier Reef following Mass Coral Bleaching', *Nature*, 560: 92–6.
- Thurber, R. V. et al. (2017) 'Virus-host Interactions and Their Roles in Coral Reef Health and Disease', *Nature Reviews Microbiology*, 15: 205–16.
- Untergasser, A. et al. (2012) 'Primer3—New Capabilities and Interfaces', *Nucleic Acids Research*, 40: e115.
- Valero-Mora, P. M. (2010) 'ggplot2: Elegant Graphics for Data Analysis', *Journal of Statistical Software*, 35: 1–3.
- Vergés, A. et al. (2014) 'The Tropicalization of Temperate Marine Ecosystems: Climate-mediated Changes in Herbivory and Community Phase Shifts', *Proceedings of the Royal Society B*, 281: 20140846.
- Wellehan, J. F. X. et al. (2008) 'Characterization of an Erythrocytic Virus in the Family Iridoviridae from a Peninsula Ribbon Snake (*Thamnophis sauritus sackeri*)', *Veterinary Microbiology*, 131: 115–22.
- Whittington, R. J., Becker, J. A., and Dennis, M. M. (2010) 'Iridovirus Infections in Finfish—Critical Review with Emphasis on Ranaviruses', *Journal of Fish Diseases*, 33: 95–122.
- Woodhead, A. J. et al. (2019) 'Coral Reef Ecosystem Services in the Anthropocene', *Functional Ecology*, 33: 1023–34.

- Wrobel, A. G. et al. (2020) 'SARS-CoV-2 and Bat RaTG13 Spike Glycoprotein Structures Inform on Virus Evolution and Furin-cleavage Effects', *Nature Structural & Molecular Biology*, 27: 763–7.
- Zhang, Y. Z. et al. (2018) 'The Diversity, Evolution and Origins of Vertebrate RNA Viruses', *Current Opinion in Virology*, 31: 9–16.
- Zhang, W. et al. (2020) 'Widespread Outbreaks of the Emerging Mandarinfish Ranavirus (MRV) Both in Natural and ISKNV-FKC Vaccinated Mandarinfish *Siniperca chuatsi* in Guangdong, South China, 2017', *Aquaculture*, 520: 734989.
- Zhao, R. et al. (2020) 'A New Ranavirus of the Santee-Cooper Group Invades Largemouth Bass (*Micropterus salmoides*) Culture in Southwest China', *Aquaculture*, 526: 735363.



Evidence for an ancient aquatic origin of the RNA viral order *Articulavirales*

Mary E. Petrone^{a,b}, Rhys Parry^c, Jonathon C. O. Mifsud^a, Kate Van Brussel^a, Ian Vorhees^d, Zoe T. Richards^{e,f}, and Edward C. Holmes^{a,b,1}

Edited by Harmit Malik, Fred Hutchinson Cancer Center, Seattle, WA; received June 22, 2023; accepted October 3, 2023

The emergence of previously unknown disease-causing viruses in mammals is in part the result of a long-term evolutionary process. Reconstructing the deep phylogenetic histories of viruses helps identify major evolutionary transitions and contextualizes the emergence of viruses in new hosts. We used a combination of total RNA sequencing and transcriptome data mining to extend the diversity and evolutionary history of the RNA virus order *Articulavirales*, which includes the influenza viruses. We identified instances of *Articulavirales* in the invertebrate phylum Cnidaria (including corals), constituting a novel and divergent family that we provisionally named the “Cnidenomoviridae.” We further extended the evolutionary history of the influenza virus lineage by identifying four divergent, fish-associated influenza-like viruses, thereby supporting the hypothesis that fish were among the first hosts of influenza viruses. In addition, we substantially expanded the phylogenetic diversity of quaranjaviruses and proposed that this genus be reclassified as a family—the “Quaranjaviridae.” Within this putative family, we identified a novel arachnid-infecting genus, provisionally named “Cheliceravirus.” Notably, we observed a close phylogenetic relationship between the Crustacea- and Chelicerata-infecting “Quaranjaviridae” that is inconsistent with virus–host codivergence. Together, these data suggest that the *Articulavirales* has evolved over at least 600 million years, first emerging in aquatic animals. Importantly, the evolution of the *Articulavirales* was likely shaped by multiple aquatic–terrestrial transitions and substantial host jumps, some of which are still observable today.

Articulavirales | influenza | Cnidaria | evolution | virosphere

Zoonotic viruses transmitted from other animals to humans have caused multiple epidemics in recent decades (1, 2), and the frequency of these events is projected to increase as a consequence of a myriad of reasons including climate change (3). To combat this threat research on zoonotic risk aims to identify viruses that are primed to spill over into humans. It is commonly assumed that the viruses most likely to cause future epidemics are genetically similar to those that have previously caused outbreaks (4, 5). A range of field-based and bioinformatic methods aimed at detecting potentially zoonotic viruses are predicated on this assumption (6–8). Although these approaches help construct a picture of virus diversity, the process of cross-species virus transmission that drives disease emergence in the short term also plays a key role in virus speciation in the long term, shaping virus–host associations that likely date back to the existence of single-celled organisms (9). In addition to revealing their antiquity, elucidating the deep evolutionary history of known disease-causing viruses provides important context for understanding the true rate at which viruses jump species boundaries to infect new hosts.

Because they reveal the viromes of a diverse array of organisms, metagenomic data are a powerful tool for tracing long evolutionary histories (10). These data have already shown that the global virosphere is vast and largely unexplored (10), and the recent exploration of marine environments has demonstrated that the ocean is a rich source of virus diversity (11). Metagenomic research has also shown that viruses once thought to be restricted to mammals in reality exist in a wide variety of other vertebrates (12), suggesting that they have long evolutionary histories in animals. Similarly, disease-causing lineages of vertebrate viruses can have origins in invertebrate hosts. For example, flavi-like viruses (single-strand positive-sense RNA viruses) were recently identified in ancient invertebrates (Cnidaria) (13), basal chordates (Ascidia) (13), and in diverse aquatic environments (14). Collectively, this suggests that flavi-like viruses and their relatives, which include such human pathogens as dengue virus and yellow fever virus, emerged concurrently with the origin of the Metazoa (i.e., animals) some 750 to 800 Mya (15). Indeed, that aquatic ecosystems may have harbored the ancestors of terrestrial viruses speaks to their antiquity and the biological diversity that these environments support.

Significance

Viruses that cause disease in humans comprise a tiny fraction of global virus diversity. Determining the long-term evolutionary history of disease-causing virus lineages places the events that drive virus emergence in an evolutionary context. Understanding when and from where viral lineages originate, and how often they jump species boundaries to emerge in new hosts, is central to this process. We addressed these questions with respect to the RNA virus order *Articulavirales* which contains the vertebrate influenza viruses. Our metagenomic and phylogenetic study doubled the number of known *Articulavirales* families, revealed a complex interplay of virus–host codivergence and cross-species virus transmission over deep evolutionary time, and traced the order's likely origins to ancient aquatic animals at least 600 Mya.

Author contributions: M.E.P. and E.C.H. designed research; M.E.P., J.C.O.M., K.V.B., I.V., and Z.T.R. performed research; M.E.P. and R.P. analyzed data; and M.E.P., R.P., J.C.O.M., K.V.B., I.V., Z.T.R., and E.C.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

¹To whom correspondence may be addressed. Email: Edward.holmes@sydney.edu.au.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2310529120/-DCSupplemental>.

Published October 31, 2023.

The order *Articulavirales* is of special importance in exploring the intersection of marine life and the evolution of disease-causing viruses. Viruses within the *Articulavirales* have segmented, negative-sense RNA genomes. This order is currently organized into two families—the *Orthomyxoviridae* and the *Amnoonviridae*. The former receives frequent public health attention because it contains three genera that can cause disease in humans: *Thogotovirus*, *Quarantavirus*, and *Influenzavirus*. Thogotoviruses were first isolated in the 1960s (16), while discoveries of disease-causing quarantaviruses are ongoing [e.g., the discovery of Wellfleet Bay virus as the cause of avian mortality in 2014 (17)]. Both thogotoviruses and quarantaviruses periodically spill over into humans, other mammals, and birds through tick-mediated transmission. In contrast, influenza viruses cause seasonal epidemics in human populations, as well as occasional pandemics.

Aquatic animals serve as hosts for viruses in both families. The *Amnoonviridae* is a family of primarily fish-infecting viruses (18), with two species also found in amphibians (19, 20). Tilapia lake virus, a genus within this family, has important implications for the global aquaculture industry because it is associated with severe disease in tilapia (21). Similarly, salmon isavirus (*Orthomyxoviridae*) causes overt disease in salmon (22), and influenza-like viruses have been identified in fish and amphibians (12, 23), although with unknown disease associations. Given their host distribution, the *Amnoonviridae* likely emerged in aquatic animals. However, the long branch lengths associated with this family in current the *Articulavirales* phylogeny suggest substantial unsampled diversity.

Genomic architecture varies within and between the families of the *Articulavirales*, but all contain three polymerase segments: polymerase basic 1 (PB1), polymerase basic 2 (PB2), and polymerase basic 3/polymerase acidic (PB3/PA). These proteins form a heterotrimer comprising the canonical RNA-dependent RNA polymerase (RdRp) (24) that is routinely used as a phylogenetic marker for RNA viruses (25). PB1 contains the palm domain, defined by the SDD amino acid motif, and is therefore the most highly conserved of the subunits. For this reason, PB1 is the segment that has the highest probability of being identified from metagenomic data.

We applied a combination of total RNA sequencing and data mining to revisit the evolutionary history of the *Articulavirales*. To date, no *Articulavirales* genera have been identified in nonbilaterian (i.e., lacking body symmetry) invertebrates of the Cnidaria, which includes corals, jellyfish, and hydra. To address this, we analyzed the RNA viromes of Hexacorals and Octocorals from Australia. Concurrently, we screened the NCBI Transcriptome Shotgun Assemblies (TSA) database for novel orthomyxo-like viruses associated with aquatic animals: Phylum Arthropoda (e.g., shrimp and crabs); Phylum Chordata, Subphylum Vertebrata (fish); Subphylum Tunicata (sea squirts and tunicates); Phylum Mollusca (e.g., squid and whelk); Phylum Porifera (sponges); and Phylum Bryozoa. We combined the resulting datasets to characterize the role of ancient aquatic animals in the evolution of the *Articulavirales* and determine when this order may have emerged in relation to the evolution of the Metazoa.

Results

Discovery of Highly Divergent *Articulavirales* in Ancient Invertebrates. Influenza viruses are overrepresented in publicly available data for viruses in the order *Articulavirales*. Of the 76,887 *Articulavirales* PB1 segments available on NCBI Virus in January 2023, only 291 (0.38%) were not influenza or influenza-like viruses (Fig. 1A). Thus, our first objective was to expand the known diversity of noninfluenza viruses in this order.

To search for highly divergent *Articulavirales*, we performed total RNA sequencing on 128 corals (Cnidaria) collected in Western Australia. From these, we identified *Articulavirales* PB1 segments in two libraries—*Heliopora coerulea* (blue coral) (GenBank: OQ939982.1) and *Acropora samoensis* (stony coral) (GenBank: OQ939981.1)—using Diamond BLASTx. Although PB1 is the most conserved polymerase segment, both segments were highly divergent, sharing $\leq 25\%$ amino acid identity with those of known viruses (*SI Appendix, Table S1*). Neither contained premature stop codons nor returned BLAST hits to known coral genes, suggesting they are not endogenous virus elements (EVEs). However, they were present at very low abundance (both $<0.001\%$ non-rRNA reads). The segment recovered from the *Acropora*-associated library

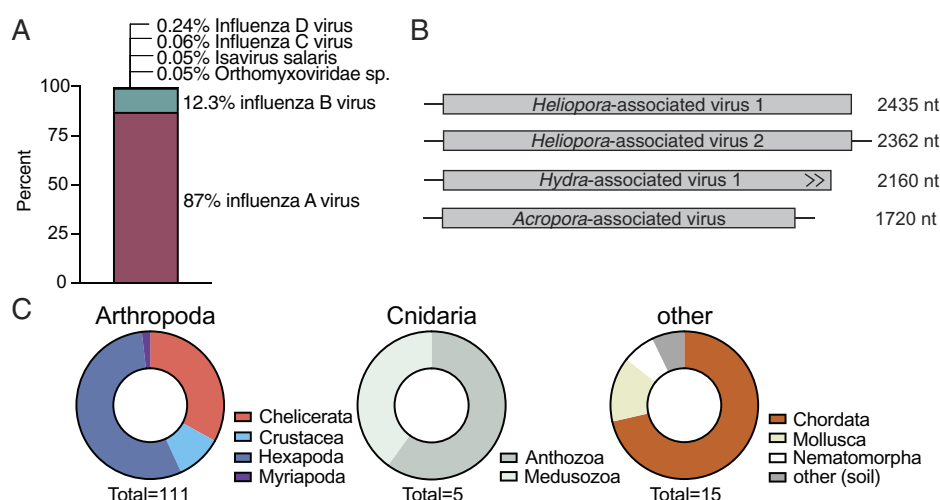


Fig. 1. Expanding the known diversity of the *Articulavirales*. (A) Overrepresentation of influenza A and B viruses in publicly available genomic data. Distribution of 99.7% of publicly available PB1 segments on NCBI Virus as of January 2023 by virus species. Those virus species not shown each comprise $<0.05\%$ of available data. (B) PB1 segments of novel *Articulavirales* identified in Cnidaria hosts. *Heliopora*-associated virus 1 was identified through sequencing while *Heliopora*-associated virus 2 was identified through data mining. (C) Distribution of host species and classes associated with the *Articulavirales* detected through total RNA sequencing and data mining.

(1720nt) was shorter than the *Heliopora*-associated segment (2435nt) (Fig. 1B), and the two segments shared minimal sequence similarity to each other (17.8% amino acid pairwise identity). This substantial genetic distance is consistent with the phylogenetic relationship of *Heliopora* and *Acropora* corals. The former are commonly known as blue corals belonging to the Subclass Octocorallia, Order Scleractyonacea, and estimated to have diverged from other Anthozoa about 500 Mya (26). *Acropora*, commonly known as staghorn corals, are from the Subclass Hexacorallia, Order Scleractinia, which emerged more recently, likely in the Permian around 260 Mya (26). Analysis of the composition of each library using CCMetagen showed that the majority of eukaryotic genetic material (98%) in the *Acropora* library was associated with Cnidaria (SI Appendix, Fig. S1A).

In contrast, 73% of contigs assembled from the *Heliopora* library were Suessiales (dinoflagellate symbionts of corals) (SI Appendix, Fig. S1B). Of the contigs associated with Anthozoa in this library, most (76%) were assigned to Scleractinia rather than *Helioporidae*. This suggested that *Helioporidae* reads were misassigned due to overrepresentation of Hexacorallia in the database (SI Appendix, Fig. S1B). To test this, we assembled the rRNA contigs identified using sortMeRNA and assessed the host composition against the NCBI nt database. A total of 84.0% of rRNA reads that corresponded to an entry in the nt database were assigned to Octocorallia, with only 4.98% assigned to Hexacorallia (Dataset S1). Assuming that any *Articulavirales* that infect protists such as dinoflagellates would share less than 25% sequence identity to known animal-infecting viruses, the high proportion of reads assigned to Octocorallia suggest that *Heliopora* (Octocorallia) was the most likely host of the virus identified in this library.

We used these segments to screen the Cnidaria assemblies available in the NCBI TSA database (n = 50). This yielded two additional novel PB1 segments: one in a *Heliopora coerulea* coral library (GFVH01041936.1) and another in a *Hydra vulgaris* hydra library (HAAC01043982.1). The second *Heliopora*-associated PB1 segment (2362nt) was slightly shorter than the first (2435nt) and shared 47.6% amino acid pairwise identity. The open reading frame (ORF) of the hydra-associated segment was uninterrupted but incomplete (Fig. 1B). PB2 and PB3 segments could not be identified in either.

Having identified *Articulavirales* in basal invertebrates, we reasoned that this viral order likely contains unrealized diversity in other aquatic invertebrate hosts. To address this, we screened TSA libraries for Arthropoda, Porifera, Ascidiacea, Bryozoa, Mollusca, Cnidaria, and fish using an iterative process. Our first input sequence was the PB1 segment of the Wenling hagfish influenza-like virus (AVM87635.1) because this was identified in an aquatic host—a jawless fish—and is highly divergent among the *Orthomyxoviridae* in the current *Articulavirales* phylogeny. This search yielded 30 novel (i.e., sharing ≤80% amino acid similarity to the closest BLASTp hit) PB1 segments of at least 1000nt in length. We then used each of these segments as input into the same query, and continued until no additional new sequences could be recovered. Through this process, we detected 94 novel PB1 segments of at least 1000nt in length. To supplement this dataset, we evaluated all RdRp contigs assembled from the Serratus project with an assigned “unknown” *Articulavirales* RdRp palmprint as per the palmDB database (27). After removing known viruses, duplicated sequences, and fragments less than 800nt in length, we identified 35 additional novel viruses.

In total, we discovered 131 novel *Articulavirales* PB1 segments. Most of these viral sequences were associated with arthropod hosts (n = 111), which likely reflects a sampling bias toward Arthropoda

rather than the true host distribution of *Articulavirales* (Fig. 1C and Dataset S2).

Previous studies have presented evidence for the endogenization of *Orthomyxoviridae* polymerase segments (28). We therefore assessed whether the contigs identified here could be EVEs. PB1 candidates that included premature stop codons were excluded from our dataset. We screened each segment for the presence of host genes using CheckV (29), and none were identified. Because the presence of multiple virus segments in a library would be strongly indicative of exogenous expression, we screened each TSA library in which we had found a PB1 segment for additional virus segments. We recovered three polymerase segments for 91 of the novel viruses we identified, and for 73 and 81 of these, we successfully retrieved nucleoprotein and hemagglutinin segments, respectively (Dataset S2). Our inability to detect additional polymerase segments in the invertebrate libraries using both sequence- and structure-based approaches was likely due to very low similarity of these segments to those of any known virus.

The *Articulavirales* Comprises at least Four Families and Likely Originated in Ancient, Aquatic Animals. Given the low sequence similarity of the PB1 segments we identified in the Cnidaria libraries, we hypothesized that these viruses would be phylogenetically distinct from other families in the order. To test this, we aligned the novel 131 sequences identified in this study with a representative sample of publicly available *Articulavirales* PB1 sequences (Dataset S3) using MUSCLE (30) and inferred a maximum likelihood phylogeny using IQ-Tree (31).

The inclusion of the viruses identified here had a substantial impact on the phylogenetic structure of this order. The Cnidaria-associated viruses formed a putative new family currently comprising nonarthropod invertebrate hosts. We have provisionally named this family the “Cnidenomoviridae.” Although there is generally strong support within the “Cnidenomoviridae” (SI Appendix, Fig. S2), the phylogenetic position of this family within the *Articulavirales* is uncertain (ufboot = 56). The known hosts of this family are currently coral, hydra, tunicate (Urochordata), sea cucumber (Echinodermata), sea slug (Mollusca), and mussel (Mollusca) (32) (Fig. 2A, “Cnidenomoviridae”). Viruses associated with *Heliopora coerulea* corals were more closely related to each other than to the virus identified in the *Acropora* coral, tentatively suggesting class specificity (SI Appendix, Fig. S2). This relationship also suggests that *Heliopora* is the host of the *Heliopora*-associated virus we identified through sequencing despite the library composition.

The addition of the arthropod-associated viruses identified through our TSA and palmDB screens suggested that the genus *Quaranjavirus* and the quaranja-like viruses should be reclassified as a new family that we have provisionally denoted as the “Quaranjaviridae” (Fig. 2A). We also identified a clade of quaranja-like viruses associated with spiders (Arthropoda, Chelicerata) (Fig. 2A, “Cheliceravirus”). Although the PB1 segments of this putative new genus were arguably sufficiently divergent from the “Quaranjaviridae” to constitute a distinct family, this was not true for the other segments, and we discuss this in more detail below. The new tree topology also supported the addition of a genus (Isavirus) to the *Amnoonviridae* through the reclassification of salmon isavirus, which is currently classified as *Orthomyxoviridae* (Fig. 2A, red arrow). In sum, we propose that *Articulavirales* comprises at least four families: *Orthomyxoviridae*, *Amnoonviridae*, “Quaranjaviridae”, and “Cnidenomoviridae”. We adopt this new nomenclature for the remainder of the manuscript.

Putative host associations throughout the order are inconsistent with strict virus–host codivergence. For example, viruses associated with Mollusca and Cnidaria hosts did not belong exclusively to the

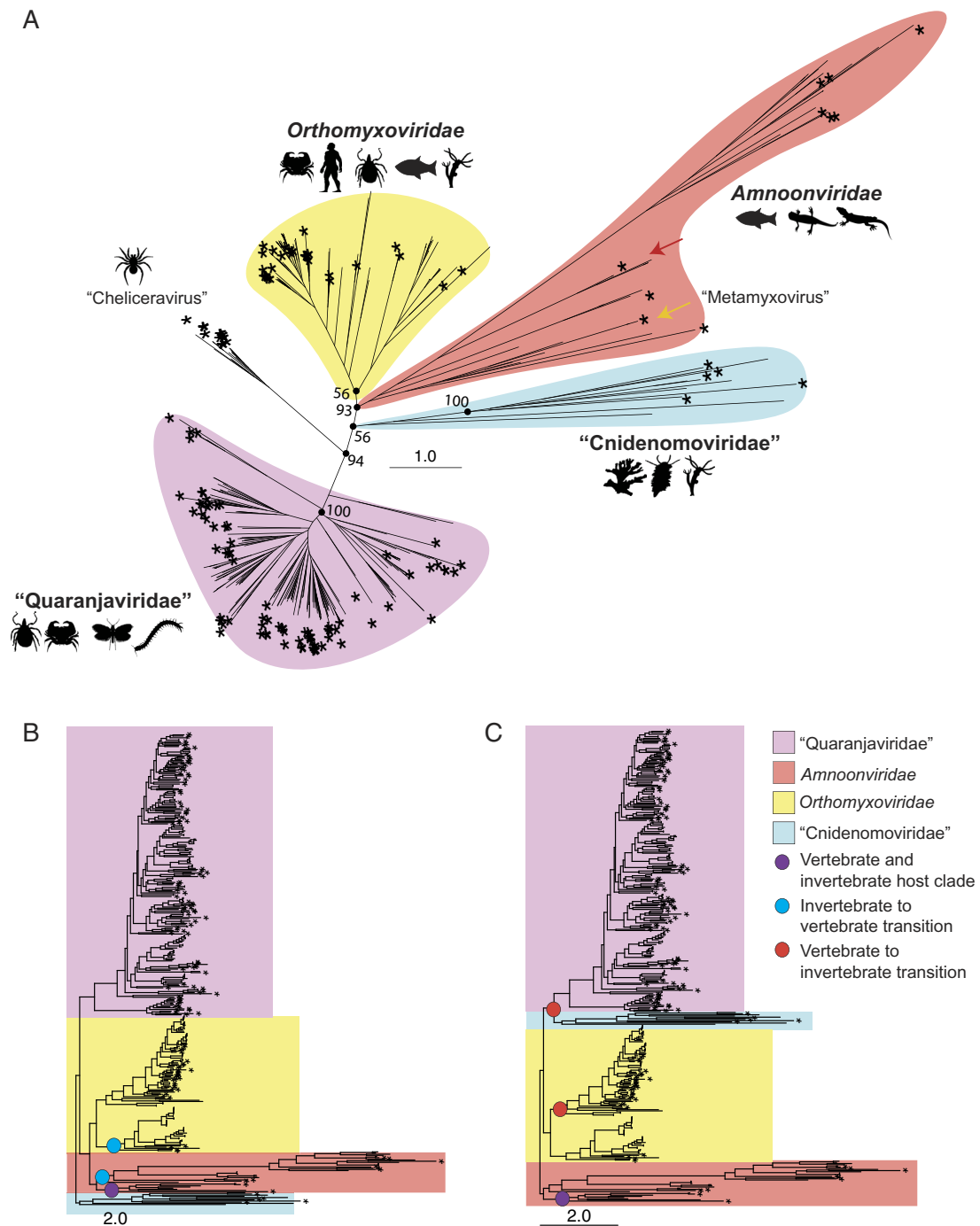


Fig. 2. Phylogenetic evidence for four virus families within the order *Articulavirales*. (A) Unrooted maximum likelihood tree of the *Articulavirales* PB1 segment, with branch lengths scaled to the number of amino acid substitutions per site. *Novel viruses identified in this study. The red arrow marks Salmon isavirus while the gold arrow marks an aquatic amnoon-like virus clade associated with Mollusca and fish hosts ("Metamyxovirus"). (B) PB1 maximum likelihood tree rooted on basal invertebrate family "Cnidenomoviridae". (C) PB1 maximum likelihood tree rooted on the *Amnoonviridae*. In all cases, branch lengths are scaled to the number of amino acid substitutions per site.

"Cnidenomoviridae". In particular, we identified amnoon-like viruses associated with a squid (*Sepioloidea lineolata*) and a hydra (GHUC-01003666.1) that formed a clade with viruses associated with fish and whelk (Mollusca) (Fig. 2A, gold arrow, "Metamyxovirus"). The phylogenetic position of this clade, which comprises exclusively

aquatic hosts (salmon louse, marine flatworm, lizardfish, squid, whelk, crustacean, gobie, and hydra), had poor support within the *Amnoonviridae* (ufboot = 37) such that the true placement of this putative new genus—"Metamyxovirus"—is uncertain. In the metamyxoviruses there is no clear separation of vertebrate and invertebrate

hosts as observed between *Thogotovirus* and *Influenzavirus* and within other families in the *Articulavirales*, although true host associations cannot be verified with metagenomic data alone.

To interpret the evolutionary history of the *Articulavirales*, we considered various possible rooting positions on the phylogeny. The structure of the tree presents ambiguities because both the *Amnoonviridae* and the “*Cnidenomoviridae*” are defined by long branch lengths, and both are plausible roots. We first assessed whether the branching patterns were an artifact of an alignment error by comparing the topology of trees inferred using the MUSCLE (30) and MAFFT (33) aligners that utilize different algorithms (*SI Appendix, Fig. S3*). No substantive differences in tree topology were observed, thereby suggesting that large sampling gaps persist in these families and account for the branch lengths. Rooting the phylogeny on the novel basal invertebrate family “*Cnidenomoviridae*” is compatible with long-term virus–host codivergence (Fig. 2*B*). This phylogenetic history means that the *Articulavirales* likely originated in aquatic invertebrates, potentially around 600 Mya when Medusozoa and Anthozoa diverged (26). In this scenario, the *Amnoonviridae* form a sister clade to the *Orthomyxoviridae* and the “*Quaranjaviridae*”. Applying the parsimony criterion, a transition to vertebrate hosts would have occurred twice independently (Fig. 2*B*, blue circles), while a partial transition to include vertebrate hosts would have occurred with the evolution of “*Metamyxovirus*” (Fig. 2*B*, purple circle).

In contrast, placing the *Amnoonviridae* as the root introduces vertebrate-to-invertebrate transitions, such that the evolution of the order was not shaped by virus–host codivergence (Fig. 2*C*). In this representation of phylogenetic history, the “*Cnidenomoviridae*” are a sister clade to the “*Quaranjaviridae*”, and there is a clear division between the invertebrate-only and vertebrate host sections of the tree (the “*Metamyxoviruses*” notwithstanding). There is also a division between arthropod-associated (the “*Quaranjaviridae*”) and non-arthropod-associated (the “*Cnidenomoviridae*”) viruses, although this partitioning may eventually change with additional sampling. Midpoint rooting is relatively consistent with placing the root on the branch leading to the *Amnoonviridae*, except the monophyly of this family is lost (*SI Appendix, Fig. S4*). Despite these important differences, in all three rooting scenarios, the tree structure indicates that the *Articulavirales* first emerged in ancient, aquatic hosts (either fish or marine invertebrates) before spreading to terrestrial animals.

Fish Were among the Earliest Hosts of Influenza-Like Viruses.

Tracing the evolution of influenza and influenza-like viruses could shed light on the mechanisms by which these viruses emerged in mammals. Through data mining, we identified the PB1 segment of four divergent influenza-like viruses associated with the transcriptomes of fish: a Siberian sturgeon (*Acipenser baerii*, GISE01045129.1), a flounder (*Paralichthys olivaceus*, SRR8334112), a seahorse (*Hippocampus guttulatus*, SRR1324961), and a grass carp (*Ctenopharyngodon idella*, SRR6475468). In addition to identifying PB1, we successfully recovered four segments (PB2, PB3, NP, and HA) from the sturgeon library, seven segments (PB2, PB3, NP, HA, NA, M1/M2, and NS1/NEP) from the flounder library, five segments from the seahorse library (PB2, PB3, NP, M1/M2, and HEP), and three segments from the grass carp (PB2, PB3, and NP) (*SI Appendix, Fig. S5*). All segments contained complete ORFs.

The genomes of the sturgeon- and carp-associated viruses differed from those of canonical influenza viruses, including those discovered in other fish, in two notable ways. First, the predicted protein sequences of the PB2 segment of these viruses are approximately 100 amino acids longer than other segments found in the viruses from this clade (*SI Appendix, Fig. S5 A and C*). These

additional amino acids do not appear to be the result of an insertion as they occur at the 3′ end of the segment. Second, while there is a conserved VGG motif in the palm domain of the PB1 segment of known influenza and thogotoviruses, the PB1 segment of the sturgeon- and carp-associated virus motif is IGG which is common among the “*Quaranjaviridae*”, found in some members of the “*Cnidenomoviridae*”, and entirely absent in the *Amnoonviridae* (Fig. 3*A*). The flounder- and seahorse-associated PB1 segments have a VGG motif. Polymerase motifs I–IV were well conserved (Fig. 3*A*). In Motif I, “KWNE” is conserved throughout the clade. The hagfish and our novel influenza-like viruses share identical sequences in motif II. While the SDD sequence in motif III is requisite for functional *Articulavirales* PB1 segments, influenza and influenza-like viruses have a second serine (SSDD) in this motif. Interestingly, motif IV (canonically GINMS) is most divergent in the hagfish virus compared to the sturgeon and carp viruses, but all are characterized by asparagine and serine in the third and fifth positions, respectively.

As the sturgeon transcriptome was generated from the animal’s intestine, we hypothesized that this virus may have been derived from the animal’s diet. To test this, we analyzed the library composition using KMA v1.3.9a (34) and CCMetagen v1.1.3 (35). Interestingly, 64% of all contigs were associated with the host (Actinopteri), and 29.5% aligned to other Chordates (14% Primates, 16.5% Birds, 0.5% *Felidae*) (*SI Appendix, Fig. S6A*). Although sturgeon consume other animals in nature, the authors of the study that generated this sequencing library reported that sturgeon were fed commercial fish feed prior to collection (36), rendering the introduction of a competent RNA virus from a different animal species unlikely. The close and consistent phylogenetic relationship of the sturgeon- and carp-associated segments (Fig. 3*B–D*) further supports fish as the bona fide host of these viruses. Thus, given that we were able to recover multiple segments and in the absence of another plausible host, we concluded that the sturgeon was the likely host of this virus.

Phylogenetic analysis of the three polymerase segments revealed that the sturgeon- and carp-associated viruses consistently fall within the influenza clade but are more divergent than all known influenza-like viruses, while the placement of the flounder- and seahorse-associated viruses is less certain (Fig. 3*B–D*). Previously, the most divergent virus in the influenza clade was an influenza-like virus associated with a hagfish (12). However, our analysis found that this placement is dependent on the alignment method. Alignment with MAFFT using the BLOSUM45 and BLOSUM62 scoring matrices recapitulated the reported placement of the hagfish-associated polymerase segments when the novel influenza-like viruses were added to the dataset (*SI Appendix, Fig. S7*). For all three segments, the flounder-, seahorse-, carp-, hagfish-, and sturgeon-associated polymerase segments clustered outside of established species of influenza viruses, with the sturgeon and carp segments consistently comprising the most divergent lineage. Alignment with MUSCLE affected the placement of the hagfish-, seahorse-, and flounder-associated PB1 segments, placing them in the same clade albeit with low support (ufboot = 37) (Fig. 3*B*). However, this relationship was better resolved when inferred in the context of the entire order (Fig. 2). Here, the sturgeon- and carp-associated viruses were the most divergent in the clade, followed by the hagfish-associated virus and then the flounder- and seahorse-associated viruses, all with ufboot >95 (*SI Appendix, Fig. S8*). Alignment with MUSCLE also placed the PB2 and PB3 segments of the hagfish-, seahorse-, and flounder-associated viruses as sister lineages to influenza A and B viruses (Fig. 3*C and D*). The polymerase segments of the sturgeon- and carp-associated viruses were consistently the most divergent relative to known influenza-like viruses when either alignment

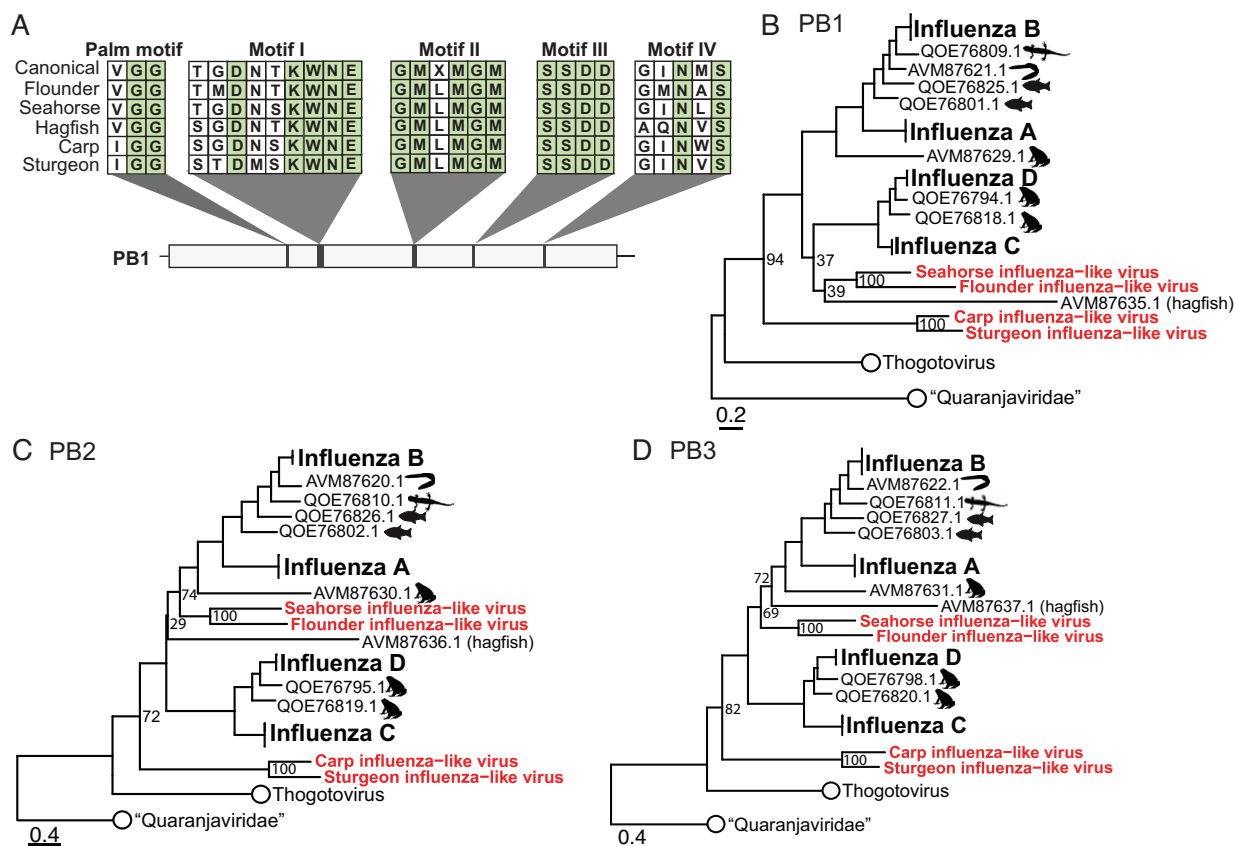


Fig. 3. Characterization of a divergent influenza-like viruses identified in a flounder, seahorse, carp, and sturgeon. (A) Conserved motifs of the PB1 segment of flounder-, seahorse-, sturgeon-, carp-, and hagfish-associated influenza-like viruses. Maximum likelihood phylogenetic trees of the influenza virus PB1 (B), PB2 (C), and PB3 (D) segments. Putative aquatic hosts are indicated with icons. Segments from viruses identified in this study are shown in red. All trees were rooted using the “Quaranjviridae” as outgroups with branch lengths scaled to the number of amino acid substitutions per site.

method was used. Moreover, the placement of the three polymerase segments was relatively well supported at the base of the influenza virus clade (ufboot = 94, 72, 82 Fig. 3 B–D).

To confirm that the unstable placement of the seahorse-, flounder-, and hagfish-associated virus segments was not due to infection by multiple influenza-like strains, we rescreened each library for additional influenza-like segments. All segments presented in Fig. 3 were used as input. No additional segments were found in the libraries, suggesting that the lack of robust support is a consequence of the divergent nature of the viruses rather than to coinfections.

These findings suggest that influenza-like viruses can infect all classes of fish such that these animals may have served as early, if not the first, hosts of influenza-like viruses before their later emergence in birds and mammals.

Phylogenetic Relationships of “Quaranjviridae” Segments Suggest a History of Cross-Species Transmission. Having reasoned that aquatic animals were likely central to the evolution of the *Articulavirales*, we next investigated their role in the evolution of the “Quaranjviridae”. Nearly all viruses in this family, including those identified in this study, are associated with arthropods as their primary host. While we identified a quaranja-like virus in a horsehair worm (Nematomorpha), as this animal parasitizes arthropods, we assumed that the respective virus was more likely

to be associated with an arthropod than with the worm. Previous studies have also identified “Quaranjviridae” in sediment (14) and feces samples (37), and host associations cannot be determined in these instances. Among the 87 novel “Quaranjviridae” we identified here, nine were associated with crustaceans (three amphipoda, two crab, and one each for shrimp, squat lobster, copepod, and tanaid).

To evaluate the relationship of these aquatic viruses within the family, we inferred maximum likelihood phylogenetic trees for the three polymerase segments, the nucleoprotein (NP), and hemagglutinin (HA). The composition of these trees necessarily differed as not all segments (particularly NP and HA) were publicly available for known viruses, preventing direct topological comparisons. PB1 segments were available for all viruses, arguably rendering this tree the most reliable. The topology of the PB1 phylogeny revealed both virus–host codivergence and cross-species virus transmission. This phylogeny can be divided into two clades, albeit with limited support (Fig. 4A, ufboot = 58). Clade 1 comprises predominantly Hexapoda-associated viruses. However, all four Arthropoda subphyla are represented in this clade, and their relative placement is consistent with the host phylogeny with the exception of one spider-associated virus. In the current host phylogeny, Chelicerata (i.e., ticks, mites, and scorpions) are thought to have diverged first, while Crustacea form a sister clade to the Hexapoda (38) (Fig. 4A,

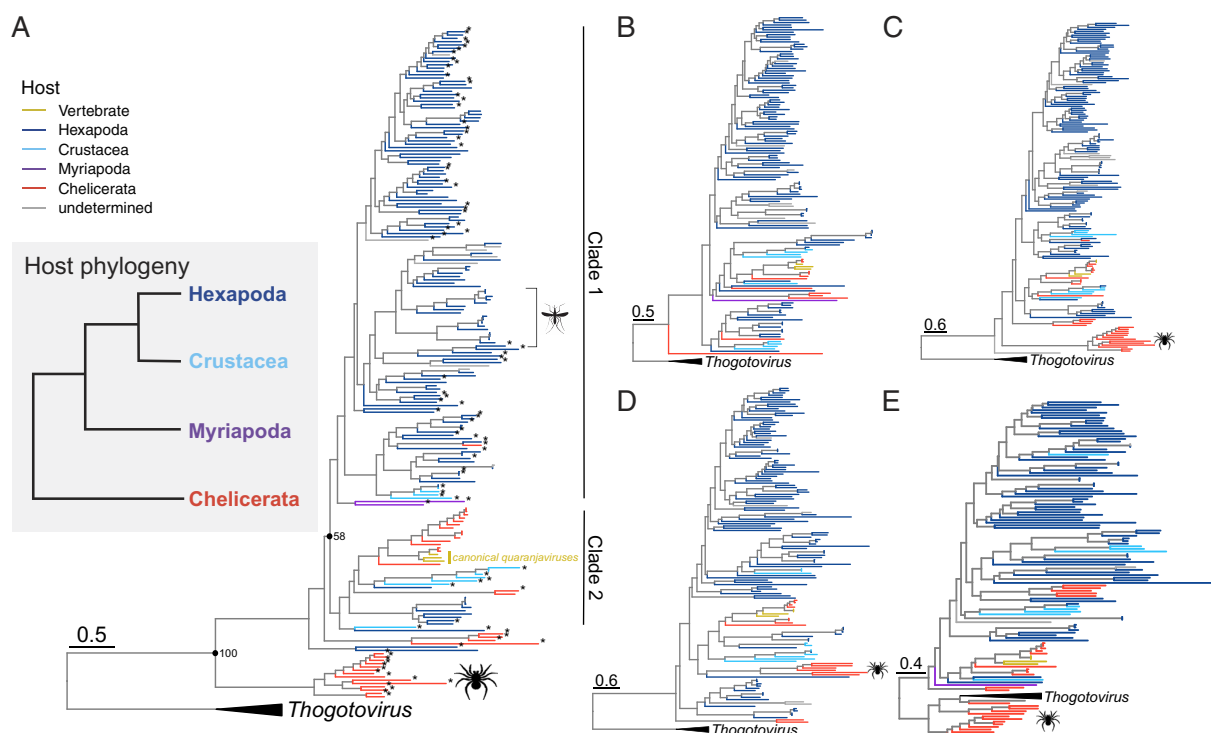


Fig. 4. Evolution of the “Quaranjaviridae” is shaped by virus–host codivergence and cross-species transmission. (A–E) Maximum likelihood phylogenies of the “Quaranjaviridae” rooted with *Thogotovirus*. “Cheliceravirus” are denoted by a spider icon. (A), PB2 (B), PB3 (C), NP (D), and HA (E). Branch color indicates putative host. Due to limitations in data availability differing numbers of segment were used for each tree (PB1: $n = 208$; PB2: $n = 144$; PB3: $n = 168$; NP: $n = 149$; HA: $n = 128$) (A) Mosquito-borne viruses are denoted with a mosquito icon. Host phylogeny adapted from Thomas et al. (38) *Novel viruses identified in this study. Branch lengths are scaled according to the number of amino acid substitutions per site.

Inset). We identified two viruses associated with centipedes (Myriapoda) that fell as sister taxa to Hexapoda- and Crustacea-associated viruses (Fig. 4A). Viruses associated with the same host genera within the Hexapoda were generally more closely related, suggesting limited recent cross-species transmission. For example, mosquito-associated viruses largely clustered together (Fig. 4A, mosquito icon).

In contrast, the organization of clade 2 is consistent with a history of host jumping. Segments from five Crustacea-associated viruses (two crab-associated, and one each associated with lobster, shrimp and Amphipoda) formed a sister clade to Chelicerata-associated quaranjaviruses, and this relationship was observed in multiple segments (Fig. 4 A–D). The only exception was the HA phylogeny, in which the Crustacea-associated viruses were dispersed throughout the tree (Fig. 4E). However, due to the limited availability of HA segments in public datasets, this placement is likely to be an artifact of undersampling. The topology of this clade is also inconsistent with codivergence: Rather, it is suggestive of substantial host jumping. The ability to jump between diverse host species is apparent in the canonical quaranjaviruses (Quaranfil quaranjavirus, Johnston Atoll quaranjavirus), as evidenced by the ability of these viruses to infect vertebrate and invertebrate hosts (Fig. 4A, “canonical quaranjaviruses”).

The placement of “Cheliceravirus” was inconsistent among segments. Both the PB1 and PB3 segments form a sister clade to the rest of the family with high support (Fig. 4 A and C). Few PB2 segments could be recovered from this putative genus. In contrast, the NP segments were more closely related to the Crustacea- and other Chelicerata-associated segments, while most of the HA

segments cluster with the *Thogotovirus* outgroup. However, this phylogeny may change with the addition of more sequences.

Discussion

We present evidence that the order *Articulavirales* originated in aquatic ecosystems and should be reorganized into four families. We document evidence of Cnidaria-associated *Articulavirales*, demonstrating that viruses in this order are likely able to infect ancient invertebrates. Our study also identifies an arachnid-associated genus in the “Quaranjaviridae”, which we have termed “Cheliceravirus”, and supports the addition of a new genus, denoted the “Metamyxovirus”, that appears to infect both vertebrate and invertebrate hosts. Finally, we greatly expanded the number of known quaranja-like viruses and propose that this genus be reclassified as a family—the “Quaranjaviridae”.

Although accurately rooting the phylogenetic tree of the *Articulavirales* is challenging, it is possible that ancient invertebrate lineages rather than fish are the first hosts of the *Articulavirales* for two reasons. First, this topology avoids introducing arguably unlikely vertebrate-to-invertebrate transitions. Second, rooting the tree with the *Amnoonviridae* contradicts some theories of virus genomic architectural evolution (9). If the *Amnoonviridae* is the root of the tree the most parsimonious host progression places fish as the first hosts of the *Orthomyxoviridae*. The characterized segments of Tilapia lake virus (PB1) and of salmon isavirus (PB1, PB2, and PB3) are approximately 50nt to 300nt shorter than their counterparts in the *Orthomyxoviridae*. This relationship implies that *Articulavirales* segments have increased in length since the genesis of the order. However, it has

been hypothesized that viral segments become shorter and more streamlined over time (9). Moreover, the PB2 segment of the novel sturgeon- and carp-associated influenza-like viruses are longer than the PB2 segments from descendant taxa in the clade, implying a lengthening and then shortening of this segment. Thus, we argue that the currently available *Articulavirales* phylogeny should be rooted on the “Cnidenomoviridae”. Importantly, however, both tree topologies support an aquatic origin of the *Articulavirales* and indicate that this order may have persisted since the split of Anthozoa and Medusozoa Cnidaria in the Ediacaran (~640 Mya) or the evolution of fish, which is thought to have occurred at approximately the same time.

Consideration of the *Articulavirales* as a whole highlights that viruses within this order utilize a large repertoire of transmission routes. Influenza virus spreads via respiratory droplets in mammals and feces among birds. Fish-to-fish influenza transmission may also be respiratory, as fish- and amphibian-associated influenza-like viruses have been detected in gill tissue (23). There is evidence that both salmon isavirus and TiLV can be transmitted vertically (39, 40) as well as horizontally, potentially through an oral-fecal route (41). Characterising the transmission dynamics of the “Cnidenomoviridae” between basal invertebrates is beyond the scope of this study, but previous studies have suggested that crustaceans could serve as vectors for aquatic RNA viruses (42). Invertebrate-to-vertebrate transmission facilitated by terrestrial arthropods has evolved independently at least twice within the *Articulavirales* (“Quaranjaviridae” and *Thogotovirus*), and motile animals would have been necessary if the *Articulavirales* first emerged in sessile invertebrates such as corals. Crustaceans could play such a role, but fish are also plausible vector candidates because they can travel long distances and serve as hosts for at least two *Articulavirales* families. Answering this question and elucidating the mechanism by which the “Quaranjaviridae” spread among nonvertebrate biting arthropods will require experimental data but could shed light on mechanisms of cross-species transmission.

Our findings support the hypothesis that fish were early, if not the first, hosts of influenza-like viruses. This hypothesis was originally put forward after the discovery of a divergent influenza-like virus in a hagfish, although based on an assumption of virus–host codivergence (12, 43). The presence of divergent influenza-like viruses in a hagfish, carp, and sturgeon suggests that these viruses have likely existed throughout the entire evolutionary history of fish. Cyclostomata, to which Myxini (including hagfish) belong, is thought to have diverged from other vertebrate phyla around 600 Mya (44), suggesting that *Articulavirales* adapted to vertebrate hosts early in its evolutionary history. No influenza-like viruses have been detected in invertebrates. Although substantial sampling gaps persist, the discovery of other *Articulavirales* in a wide range of invertebrate hosts suggests that influenza viruses, as defined as a distinct phylogenetic group, are limited to vertebrate hosts, and implies that fish were the first. A similar evolutionary history has been proposed for the vertebrate-infecting family *Coronaviridae* (45). Amphibians form an evolutionary lineage between fish and reptiles and bridge aquatic and terrestrial ecosystems. The relative placement of amphibians is consistent with both virus–host codivergence and an aquatic-to-terrestrial transition. We find that the latter is more strongly supported by the phylogeny of the influenza clade. Taken together, the evolution of influenza viruses may therefore have been shaped by cross-class host jumps. Some events of this kind are still observable today [e.g., bird-to-mammal transmission of H5N1 (46)], while others, like the hypothetical transmission of influenza-like viruses between fish and other aquatic vertebrates, represent deep evolutionary relationships.

Our study places vector-borne virus emergence into an evolutionary context. In the family “Quaranjaviridae”, only arachnid-associated viruses are known to spill over into vertebrates. While this could be due in part to the biting behavior of ticks, the “Quaranjaviridae” have also been detected in mosquito genera known to transmit RNA viruses to humans (*Aedes* and *Culex*). Mosquito-borne transmission of quaranjaviruses to humans has not been reported. Although it may be that cryptic transmission does occur, as vertebrates and particularly humans are disproportionately oversampled, these events are likely to be captured if they are associated with overt disease. Thus, of the possible explanations for this ostensible absence, we consider two to be of particular interest: Spill over of mosquito-borne quaranjaviruses does occur but does not cause disease, or spill over of these viruses into vertebrates does not occur because of genetic differences between mosquito- and tick-borne “Quaranjaviridae”. If the former, it may be that these viruses are unable to evade the vertebrate adaptive immune system. If the latter, they may not be able to adapt to vertebrate host receptors. In either case, the absence of mosquito-borne quaranjavirus conferred disease suggests that there are key biological differences between Hexapoda-borne and Chelicerata-borne quaranjaviruses, and the latter are better able to adapt to vertebrate hosts. Elucidating these differences will require experimental data and could have important implications for understanding vector-borne virus emergence in vertebrates. The close relationship between arachnid- and crustacean-associated “Quaranjaviridae” again indicates that the evolutionary history of the *Articulavirales* has not followed strict virus–host codivergence. Chelicerata diverged prior to both Hexapoda and Crustacea (38), yet Crustacea-associated viruses are more closely related to Chelicerata-associated viruses and are basal to the rest of the family. Sampling horseshoe crabs, which are ancient, aquatic Chelicerata, could provide insight into how this transition occurred.

This study was not without limitations. We utilized publicly available data for most of our analyses, introducing sampling biases (e.g., overrepresentation of Arthropoda hosts) and limiting the size of our dataset. As such, the range of hosts of the *Articulavirales* is likely far larger than what we have described. Without experimental data, we cannot definitively establish the true hosts of the viruses we have identified; however, there is sufficient contextual data to draw reasonable inferences. We were unable to obtain access to the raw sequencing reads of the sturgeon library (TSA ID GIPE01): The reported SRAs did not match the assembled contigs available through the TSA database. As a consequence, we could not assess the abundance of the sturgeon influenza-like virus. Additionally, the use of cross-sectional metagenomic data (i.e., data collected at a single time point) precludes the assignment of true virus–host associations. Despite this, the host associations we assumed were supported by phylogenetic analysis. For example, viruses identified in *Heliopora* corals clustered together, as do most mosquito-associated “Quaranjaviridae”. Finally, while the overall topology of the phylogenies we inferred had strong support, internal tree nodes were less robust. This may reflect substantial undersampling within the order, and the addition of more viruses when discovered should improve the veracity of the placement of internal nodes. For this reason, we avoided overly interpreting internal relationships, which limited the depth of our conclusions.

We have demonstrated that metagenomic data can be used to elucidate the origins of respiratory and arthropod-borne viruses that are the focus of modern public health initiatives. Continuing to explore virus diversity in a wide range of animal hosts will help to answer key evolutionary questions. For example, how did the evolution of the adaptive immune system influence virus evolution,

particularly when ancient invertebrate viruses first appeared in vertebrates? Answering this and related questions will be especially useful for improving methods to estimate zoonotic risk. Importantly, our findings show that, as one of the first animals to evolve on Earth and therefore one of the likely first animal hosts of RNA viruses, aquatic Metazoa have played a central role in shaping the modern virome. Further exploration of marine virus diversity will undoubtedly yield deeper insights into the drivers of virus diversity and shed light on how these simple replicating entities have successfully parasitized all known forms of life.

Methods

Sample Collection. All (n = 128) coral samples were collected and identified by Zoe Richards. Location details for the two individual corals in which we identified novel *Articulavirales* sequences are as follows: *Heliopora coerulea* #210, WAM Z42167, subtidal 12m, collected 20.09.2016, NW Montelivet Island, Bonaparte Archipelago, inshore Kimberley, Western Australia. Site 196; S14.28778, E125.2127. *Acropora samoensis*, WAM Z421003, #158, intertidal, collected 18.09.2016, North Patricia Island, Bonaparte Archipelago, inshore Kimberley, Western Australia. Site 189S14.266603, E125.29700.

RNA Extraction and Sequencing. The first coral (*Heliopora coerulea*) was processed in 2018. The sample was thawed on ice, removed from RNA later, and rinsed gently with sterile RNA and DNA-free 1 × PBS solution (GIBCO). Sterile tools were then used to remove a small fragment (~1 g), with care taken to include both soft and hard tissues. The tissue was homogenized using a sterile drill, and total RNA was extracted with the RNeasy Plus Mini Kit (Qiagen). RNA quality was determined using Bioanalyzer fragment analysis, followed by Ribo-Zero Plus library preparation and next-generation sequencing (Illumina HiSeq 2500) performed at the Australian Genome Research Facility (AGRF).

The second coral (*Acropora samoensis*) was processed in 2022. This sample was homogenized in liquid nitrogen using a mortar and pestle. RNA was extracted using the RNeasy Plus Mini Kit (Qiagen). The quality and concentration of the RNA were assessed with Qubit. Ribo-Zero Plus library preparation and next-generation sequencing (Illumina NovaSeq S4) were performed at AGRF. Negative extraction controls were used to assess potential contaminants. No relevant contaminants were identified. The extracted RNA from each sample was not pooled prior to sequencing.

Virus Discovery. Raw sequencing reads were trimmed using Trimmomatic v0.38 (47). Reads mapping to rRNA were removed using sortMeRNA v4.3.3 (48). Contigs were assembled from trimmed and filtered reads using MEGAHIT v1.2.9 (49). Assembled reads were screened against the NCBI nonredundant (nr) protein database (as of June 2022) and a custom RdRp database using Diamond BLASTx v2.0.9. The custom RdRp database contains all currently identified RdRp sequences. In both cases, we used an e-value cutoff of 1×10^{-5} . Contigs with hits to known viruses with less than 80% amino acid identity were considered new virus species. All contigs with hits to the RdRp database were cross-checked against the nr results, and virus candidates that hit to host genes were excluded from further analysis. Contigs identified as *Articulavirales* segments were translated using the ExPASy online translation tool (<https://web.expasy.org/translate/>), and each translation was visually inspected in Geneious. Read abundance was estimated using RSEM v1.3.0 (50) implemented in Trinity v2.5.1 (51).

A structure-based approach was adopted to search for additional polymerase segments in the coral libraries. All contigs of at least 1200nt in length that did not correspond to an entry in the RdRp or nr databases were translated using EMBOSS. Contigs with open reading frames between 600 and 900 amino acids were retained. This range was selected according to the range in lengths of known *Articulavirales* polymerase segments. The remaining contigs were analyzed using Phyre2 to identify structural similarities to known *Articulavirales* polymerase segments.

Assessment of Endogenous Virus Elements (EVEs). All virus candidates were compared to the nr database (as of June 2022) to confirm that they did not share similarity to known host genes. We also assessed the presence of host gene contamination using the contamination function implemented in CheckV (29). No virus candidates were found to be EVEs.

NCBI TSA Database Screen. We performed tBLASTn searches limited to the following organisms: Ascidiacea, Porifera, Cnidaria, Arthropoda, Bryozoa, Mollusca, fish (Agnatha, Chondrichthyes, Actinopterygii, Osteichthyes actinopterygian, Osteichthyes coelacanthiform, Osteichthyes dipnoan, Hyperoartia, Gnathostomata, Petromyzontidae, marine lamprey, Arctic lamprey), and Amphibia. These organisms encompass the majority of aquatic animals with available TSA libraries. To identify divergent PB1 segments, we began by using the Wenling hagfish influenza-like virus (AVM87635), Beihai orthomyxo-like virus 2 (APG77864) as input, as well as both PB1 segments identified in corals. Divergent hits (less than 80% amino acid identity) were confirmed to be novel using BLASTx (nr database). We used these sequences as input for tBLASTn against the TSA projects described above. Each novel sequence was translated using ExPASy and visually inspected in Geneious. In all cases, only translations containing the RdRp SDD motif that defines the *Articulavirales* were considered. A preliminary phylogenetic analysis was performed to remove identical sequences. In this case, sequences were aligned in MAFFT v7.490 (33), and the phylogenetic tree was inferred using the maximum likelihood approach in IQ-TREE v1.6.12 (31) with ModelFinder, which selected LG+F+R10 as the best-fit model.

Serratus Screen. To identify contigs assembled as part of the Serratus project (27) containing “novel” or unassigned *Orthomyxoviridae* PB1 segments, we downloaded and manually queried the master RdRp contig and metadata file for *Orthomyxoviridae*-like PALMdb palmpoint (or barcode) accessions (n = 336). These PALMdb palmpoints are assigned based on taxonomy and clustered based on 90% identity (52). Putative novel Orthomyxo-like PB1 sequences were manually validated using BLASTx against the NCBI nr virus database. SRA accessions from identified novel viral-like sequences were assembled using MEGAHIT, and additional segments were extracted from assemblies using tBLASTn.

Phylogenetic Tree Estimation. To compile a representative dataset of publicly available *Articulavirales* PB1 segments, we downloaded all noninfluenza and influenza-like *Articulavirales* PB1 segments available on NCBI “Virus” that were at least 500 amino acids in length (n = 291). For completeness, we used all available spellings of “PB1” and “polymerase basic 1.” Sequences associated with a laboratory host were excluded. We included all influenza-like viruses identified in nonmammalian and nonreptilian hosts (n = 8) and a representative sample of influenza A, B, C, and D viruses (n = 22). Identical sequences were removed using CD-HIT v4.6.1 with a threshold of 0.99. The remaining sequences were aligned with MUSCLE v5.1 (30). After trimming the aligned sequences to the conserved motifs in Geneious and removing gaps with trimAl v1.4 (53), identical sequences were again removed according to genetic distance. The final dataset contained 288 sequences.

Sequences for all other phylogenies presented in this study were aligned using MUSCLE v5.1 unless otherwise specified. In all cases, gaps were removed using trimAl v1.4., and maximum likelihood phylogenetic trees based on amino acid sequences were inferred in IQ-TREE v1.6.12 using 1000 ultrafast bootstraps and ModelFinder (Table 1). Trees were rendered with ggtree (54) implemented in R v4.1.2 and visualized with Adobe Illustrator.

Analysis of Host Gene Distribution in Sequencing Libraries. We analyzed the host gene assignment and distribution of the sequencing libraries associated with the Siberian sturgeon (*Acipenser baerii*, TSA: GIPE01, BioProject:

Table 1. Substitution models selected by Model Finder during phylogenetic inference

Phylogenetic tree	Figure	Substitution model
<i>Articulavirales</i> PB1	2 A–C	LG+F+R10
Influenza clade PB1	3A	LG+F+R10
Influenza clade PB2	3B	LG+F+R3
Influenza clade PB3/PA	3C	LG+F+Γ4
Quarjviridae PB1	4A	LG+F+R10
Quarjviridae PB2	4B	LG+F+R6
Quarjviridae PB3	4C	LG+F+R9
Quarjviridae NP	4D	LG+F+R10
Quarjviridae HA	4E	WAG+F+R6

PRJNA591120) as well as the corals from which we extracted RNA (*Acropora samoensis* and *Heliopora coerulea*). For the sturgeon library, the transcriptome shotgun contigs preassembled and associated with the BioProject ID were used as input. In all three cases, we used KMA v1.3.9a (34) and CCMetagen v1.1.3 (35) and visualized the results with Prism v9.5.0 and Adobe Illustrator.

Data, Materials, and Software Availability. Sequence data for the two viruses identified through metatranscriptomic sequencing (Heliopora cnidenomovirus 1 and Acropora cnidenomovirus 1) have been submitted to both the SRA (BioProject PRJNA966762) (55) and GenBank (accessions [OQ939982.1](#) and [OQ939981.1](#), respectively). The sequence accessions for the viruses identified here through data mining are provided in [SI Appendix](#). The multiple sequence alignments used to generate all of the phylogenetic trees presented in this study are freely available at https://github.com/mary-petrone/ancient_articulavirales (56).

ACKNOWLEDGMENTS. We thank Prof. Alexander Sens for his suggestion of the name Cnidenomoviridae. The name derives from the Greek Κνίδη δ

(nettle, or sea nettle, and the root of Cnidaria) and -νῶμος (“dwelling in” or “pasturing on”). We also thank J.-S. Eden and C. Le Lay for their assistance extracting RNA from the *Heliopora* coral. This work was funded by a National Health and Medical Research Council (NHMRC) Investigator award and by AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong Special Administrative Region, China. Coral samples were collected on fieldwork funded by the Western Australian Museum, Woodside Energy, the Net Conservation Benefits Fund and ARC Linkage Project LP160101508.

Author affiliations: ^aSydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia; ^bLaboratory of Data Discovery for Health Limited, Hong Kong Special Administrative Region, China; ^cSchool of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4067, Australia; ^dJames A. Baker Institute for Animal Health, Department of Microbiology and Immunology, College of Veterinary Medicine, Cornell University, Ithaca, NY 14850; ^eCoral Conservation and Research Group, Trace and Environmental DNA Laboratory, School of Molecular and Life Sciences, Curtin University, Perth, WA 6102, Australia; and ^fCollections and Research, Western Australian Museum, Welshpool, WA 6106, Australia

1. E. de Wit, N. van Doremalen, D. Falzarano, V. J. Munster, SARS and MERS: Recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.* **14**, 523–534 (2016).
2. M. Worobey *et al.*, The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. *Science* **377**, 951–959 (2022).
3. R. A. Weiss, A. J. McMichael, Social and environmental risk factors in the emergence of infectious diseases. *Nat. Med.* **10**, S70–S76 (2004).
4. M. E. Petrone, E. C. Holmes, E. Harvey, Through an ecological lens: An ecosystem-based approach to zoonotic risk assessment: An ecosystem-based approach to zoonotic risk assessment. *EMBO Rep.* **24**, e56578 (2023).
5. C. J. Carlson *et al.*, The future of zoonotic risk prediction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **376**, 20200358 (2021).
6. N. Mollentze, S. A. Babayan, D. G. Strecker, Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS Biol.* **19**, e3001390 (2021).
7. L. Brierley, A. Fowler, Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. *PLoS Pathog.* **17**, e1009149 (2021).
8. T. Allen *et al.*, Global hotspots and correlates of emerging zoonotic diseases. *Nat. Commun.* **8**, 1124 (2017).
9. E. C. Holmes, *The Evolution and Emergence of RNA Viruses* (Oxford University Press, 2009).
10. E. Harvey, E. C. Holmes, Diversity and evolution of the animal virome. *Nat. Rev. Microbiol.* **20**, 321–334 (2022).
11. A. A. Zayed *et al.*, Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
12. M. Shi *et al.*, The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).
13. J. C. O. Mifsud *et al.*, Transcriptome mining extends the host range of the *Flaviviridae* to non-bilaterians. *Virus Evol.* **9**, veac124 (2023).
14. Y.-M. Chen *et al.*, RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nat. Microbiol.* **7**, 1312–1323 (2022).
15. C. G. G. Bamford, W. M. de Souza, R. Parry, R. J. Gifford, Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of flavivirids (family *Flaviviridae*). *Virus Evol.* **8**, veac085 (2022).
16. D. A. Haig, J. P. Woodall, D. Danskin, Thogoto virus: A hitherto undescribed agent isolated from ticks in Kenya. *J. Gen. Microbiol.* **38**, 389–394 (1965).
17. A. B. Allison *et al.*, Cyclic avian mass mortality in the northeastern United States is associated with a novel orthomyxovirus. *J. Virol.* **89**, 1389–1403 (2015).
18. O. M. H. Turnbull *et al.*, Meta-transcriptomic identification of divergent *Amnooviridae* in fish. *Viruses* **12**, 1254 (2020).
19. A. S. Ortiz-Baez, J.-S. Eden, C. Moritz, E. C. Holmes, A divergent *Articulavirus* in an Australian Gecko identified using meta-transcriptomics and protein structure comparisons. *Viruses* **12**, (2020).
20. I. Olendraitte, K. Brown, A. E. Firth, Identification of RNA virus-derived RdRp sequences in publicly available transcriptomic datasets. *Mol. Biol. Evol.* **40**, msad060 (2023).
21. K. Fitzsimmons, Global Tilapia Market Update 2015 (2016).
22. D. Bouchard *et al.*, Isolation of infectious salmon anemia virus (ISAV) from Atlantic salmon in New Brunswick, Canada. *Dis. Aquat. Organ.* **35**, 131–137 (1999).
23. R. Parry, M. Wille, O. M. H. Turnbull, J. L. Geoghegan, E. C. Holmes, Divergent influenza-like viruses of Amphibians and fish support an ancient evolutionary association. *Viruses* **12**, (2020).
24. H. Fan *et al.*, Structures of influenza A virus RNA polymerase offer insight into viral genome replication. *Nature* **573**, 287–290 (2019).
25. J. Charon, J. P. Buchmann, S. Sadiq, E. C. Holmes, RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *Virus Evol.* **8**, veac082 (2022).
26. A. D. Rogers, “Cnidarians (Cnidaria)” in *The Timetree of Life*, S. B. Hedges, S. Kumar, Eds. (Oxford University Press, 2009), pp. 233–238.
27. R. C. Edgar *et al.*, PetaBase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022).
28. A. G. Russo, A. G. Kelly, D. Enosi Tupulotu, M. M. Tanaka, P. A. White, Novel insights into endogenous RNA viral elements in bodes scapularis and other arbovirus vector genomes. *Virus Evol.* **5**, vez010 (2019).
29. S. Nayfach *et al.*, CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
30. R. C. Edgar, Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).
31. B. Q. Minh *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
32. M. Shi *et al.*, Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
33. K. Katoh, J. Rozewicki, K. D. Yamada, MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).
34. P. T. L. C. Clausen, F. M. Aarestrup, O. Lund, Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).
35. V. R. Marcelino *et al.*, CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biol.* **21**, 103 (2020).
36. D. Chen *et al.*, Low lethal doses of *Streptococcus iniae* caused enteritis in Siberian sturgeon (*Acipenser baerii*). *Fish Shellfish Immunol.* **104**, 654–662 (2020).
37. K. Van Brussel *et al.*, Faecal virome of the Australian grey-headed flying fox from urban/suburban environments contains novel coronaviruses, retroviruses and sapoviruses. *Virology* **576**, 42–51 (2022).
38. G. W. C. Thomas *et al.*, Gene content evolution in the arthropods. *Genome Biol.* **21**, 15 (2020).
39. S. H. Marshall, R. Ramirez, A. Labra, M. Carmona, C. Muñoz, Bona fide evidence for natural vertical transmission of infectious salmon anemia virus in freshwater brood stocks of farmed Atlantic salmon (*Salmo salar*) in Southern Chile. *J. Virol.* **88**, 6012–6018 (2014).
40. J. Yankasem, P. Tattiyapong, A. Kamlangdee, W. Surachetpong, Evidence of potential vertical transmission of tilapia lake virus. *J. Fish Dis.* **42**, 1293–1300 (2019).
41. F. Piercezan, S. Yun, W. Surachetpong, E. Soto, Intragastric and intracoeleomic injection challenge models of tilapia lake virus infection in Nile tilapia (*Oreochromis niloticus* L.) fingerlings. *J. Fish Dis.* **42**, 1301–1307 (2019).
42. R. Parry, S. Asgari, Discovery of novel crustacean and cephalopod flaviviruses: Insights into the evolution and circulation of flaviviruses between marine invertebrate and vertebrate hosts. *J. Virol.* **93**, e00432–19 (2019).
43. M. Wille, E. C. Holmes, The ecology and evolution of influenza viruses. *Cold Spring Harb. Perspect. Med.* **10**, a038489 (2020).
44. S. B. Hedges, “Vertebrates (Vertebrata)” in *The Timetree of Life*, S. B. Hedges, S. Kumar, Eds. (Oxford University Press, 2009), pp. 309–314.
45. A. K. Miller *et al.*, Slippery when wet: Cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the *Coronaviridae*. *Virus Evol.* **7**, veab050 (2021).
46. M. Agüero *et al.*, Highly pathogenic avian influenza A(H5N1) virus infection in farmed minks, Spain, October 2022. *Euro Surveill.* **28**, (2023).
47. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
48. E. Kopylova, L. Noé, H. Touzet, SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
49. D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
50. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
51. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
52. A. Babaian, R. Edgar, Ribovirus classification by a polymerase barcode sequence. *PeerJ* **10**, e14055 (2022).
53. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
54. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, Gttree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
55. M. E. Petrone *et al.*, BioProject PRJNA966762. NCBI1. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA966762>. Accessed 3 March 2023.
56. M. E. Petrone *et al.*, Ancient Articulavirales. GitHub. https://github.com/mary-petrone/ancient_articulavirales. Accessed 28 August 2023.

Divergent hepaciviruses, delta-like viruses, and a chu-like virus in Australian marsupial carnivores (dasyurids)

Erin Harvey,^{*,†} Jonathon C.O. Mifsud,[‡] Edward C. Holmes,[§] and Jackie E. Mahar^{*,¶}

Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia

[†]<https://orcid.org/0000-0002-9379-2064>

[‡]<https://orcid.org/0000-0001-8934-9193>

[§]<https://orcid.org/0000-0001-9596-3552>

[¶]<https://orcid.org/0000-0003-2687-9810>

*Corresponding author: E-mail: erin.harvey@sydney.edu.au

Abstract

Although Australian marsupials are characterised by unique biology and geographic isolation, little is known about the viruses present in these iconic wildlife species. The Dasyuromorphia are an order of marsupial carnivores found only in Australia that include both the extinct Tasmanian tiger (thylacine) and the highly threatened Tasmanian devil. Several other members of the order are similarly under threat of extinction due to habitat loss, hunting, disease, and competition and predation by introduced species such as feral cats. We utilised publicly available RNA-seq data from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database to document the viral diversity within four Dasyuromorph species. Accordingly, we identified fifteen novel virus sequences from five DNA virus families (Adenoviridae, Anelloviridae, Gammaherpesvirinae, Papillomaviridae, and Polyomaviridae) and three RNA virus taxa: the order Jingchuvirales, the genus *Hepacivirus*, and the delta-like virus group. Of particular note was the identification of a marsupial-specific clade of delta-like viruses that may indicate an association of deltaviruses with marsupial species. In addition, we identified a highly divergent hepacivirus in a numbat liver transcriptome that falls outside of the larger mammalian clade. We also detect what may be the first Jingchuvirales virus in a mammalian host—a chu-like virus in Tasmanian devils—thereby expanding the host range beyond invertebrates and ectothermic vertebrates. As many of these Dasyuromorphia species are currently being used in translocation efforts to reseed populations across Australia, understanding their virome is of key importance to prevent the spread of viruses to naive populations.

Keywords: virus discovery; metatranscriptomics; hepacivirus; deltavirus; marsupial; chuvirus.

1. Introduction

Australian wildlife has evolved in isolation for approximately 45 million years, resulting in a unique mammalian fauna, of which 87 per cent are endemic (Chapman 2009). This includes carnivorous marsupials of the order Dasyuromorphia (Kealy and Beck 2017). Dasyuromorphs have experienced extinction as a direct result of human activity, with the Tasmanian tiger (*Thylacinus cynocephalus*) an iconic symbol of human-mediated mammalian extinction (Feigin, Frankenberg, and Pask 2022). Indeed, Australia is experiencing one of the highest rates of mammalian extinction globally due to changing land use, disease, and competition and predation from introduced species (Woinarski et al. 2011). Despite these threats, we know little about the viruses that infect these unique and at-risk species and how these viruses might impact population health.

Marsupials are an infraclass of mammals, characterised by their distinctive pouch in which they carry live young. The order

Dasyuromorphia represent all carnivorous marsupials in Australia apart from the omnivorous bandicoots (Peramelemorphia) (Zemann et al. 2013) and are found only on the mainland of Australia and its surrounding islands, such as the Australian island state of Tasmania and Papua New Guinea (Kealy and Beck 2017). The largest marsupial carnivore was the extinct thylacine, followed by the now-endangered Tasmanian devil (*Sarcophilus harrisii*) that has replaced the tiger as the apex predator in Tasmania. Through a combination of hunting, habitat loss, competition for resources with introduced species and, potentially, the introduction of exotic pathogens, the last known thylacine died in 1936 and the species was declared officially extinct in 1986 (Paddle 2000). The Tasmanian devil is also threatened by a contagious cancer that has contributed to an average population decline of 77 per cent in the past 20 years (Lazenby et al. 2018).

Beyond these two iconic species, lesser-known marsupial carnivores are also at risk of extinction. The numbat (*Myrmecobius*

fasciatus) is the last remaining member of the family Myrmecobiidae, one of the few diurnal marsupials, and the only one that feeds exclusively on termites (Zemann et al. 2013). Due to habitat loss and predation by introduced species such as cats and foxes, the numbat only persists in two natural locations in southwestern Australia and it is estimated that there are less than 1,000 numbats remaining in the wild (Peel et al. 2022). In addition, other species that had been believed to be of little concern, such as the fat-tailed dunnart (*Sminthopsis crassicaudata*), have recently been listed as threatened by the state of Victoria due to population decline (Scicluna, Gill, and Robert 2021). To combat these dramatic declines in populations of marsupial carnivores across Australia, programmes have been developed to relocate individuals from thriving populations in protected locations, such as the quoll populations in Tasmania, as well as from captive breeding populations (Portas et al. 2020). Similarly, the yellow-footed antechinus (*Antechinus flavipes*) has been proposed as a candidate species for translocation as part of a rewilding project in an area near Canberra, Australia (Manning 2021). These programmes have experienced varied success, and an often-overlooked risk with the translocation of wildlife is the introduction of pathogens into naive populations already struggling under the burden of anthropogenic activities and introduced species (French et al. 2022).

We know little about the viruses circulating within dasyuromorphs. To date, only a single virome project has been performed on these animals—in this case, the faecal virome of Tasmanian devils (Chong et al. 2019). There have been a small number of studies of marsupial carnivores experiencing overt signs of disease, including the characterisation of Dasyurid herpesvirus 1 in *Antechinus* (Barker, Carbonell, and Bradley 1981), and a chimeric papilloma-polyomavirus has been identified in bandicoots (Woolford et al. 2007). There have also been a small number of serological and bioinformatic screening studies targeting or including marsupial carnivores. For example, the molecular and serological study of marsupial tissues for the identification of herpesviruses led to the identification of a novel herpesvirus in a Tasmanian devil (Stalder et al. 2015), while an analysis of available transcriptomic data and reference host genomes characterised the endogenous viral elements (EVEs) of marsupial carnivores (Harding et al. 2021).

Herein, we present the first unbiased virus discovery analysis of marsupial carnivores. Because of the inherent challenges in acquiring samples from dasyuromorphs that are generally protected species, we instead mined the available transcriptomes present on the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), followed by genomic and phylogenetic analysis.

2. Methods

2.1. Identification of virus contigs in transcriptome data

A custom virus detection pipeline was used to screen all available Dasyuromorphia (NCBI taxonomic identifier: taxid 2759) RNA-seq datasets available on the SRA, excluding faecal samples. Raw FASTQ files for all libraries were downloaded using Kingfisher (<https://github.com/wwood/kingfisher-download>). Sequencing reads first underwent quality trimming and adapter removal using Trimmomatic (v0.38) with parameters SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5, and MINLEN:25, prior to assembly (Bolger, Lohse, and Usadel 2014).

De novo assembly was conducted using MEGAHIT with default parameters (v1.2.9) (Li et al. 2015). The assembled contigs were

then compared to the RdRp-scan RNA-dependent RNA polymerase (RdRp) core protein sequence database (v0.90) (Charon et al. 2022) and the protein version of the Reference Viral Databases (v23.0) (Goodacre et al. 2018; Bigot et al. 2019) using Diamond BlastX (v2.0.9) with an e-value cut-off of 1×10^{-5} (Buchfink, Reuter, and Drost 2021). To remove potential false positives, contigs with hits to virus sequences were used as a query against the NCBI nucleotide database (as of May 2023) using Blastn, and all contigs with sequence identity to non-virus nucleotide sequences were removed from the query set (Camacho et al. 2009). The remaining contigs were then aligned against the NCBI non-redundant protein database (as of March 2023) using Diamond BlastX, and contigs with hits to virus proteins were further examined. The *de novo* assembler and FindORFs tool, available in Geneious (Kearse et al. 2012), were used to further assemble genomes where necessary and to identify ORFs in potential virus sequences, respectively. EVEs were identified by referencing Harding et al. (2021) and removed manually. NCBI Web BLAST (<https://www.ncbi.nlm.nih.gov/BLAST>) was then used to check for false positives, disrupted open reading frames (ORFs), and to manually assess alignment to virus motifs.

2.3. Virus abundance

The abundance of virus contigs was measured using the RNA-seq by Expectation Maximisation (v1.3.0) program (Li and Dewey 2011). The expected count of viral contigs was used to calculate viral abundance as a percentage of total reads. All calculations and graphing were performed using R (R Core Team 2021).

2.4. Phylogenetic analysis

Sequences representing novel virus species, defined as those with <95 per cent nucleotide sequence identity to their closest relative, were assigned to a taxonomic family based on their identity to previously characterised virus species. For each family, a reference dataset was downloaded from NCBI Virus (Hatcher et al. 2017) and the completeness of this dataset was assessed by comparing it to the International Committee on Virus Taxonomy (ICTV)-recognised species for each family. Unclassified species related to the viruses discovered here were added using a Web BLAST search using the novel virus sequence as the query. Due to the high level of divergence within the reference dataset, amino acid alignments were estimated for all families except the *Kolmioviridae* (Deltavirus), in which case a nucleotide alignment was used. Multiple sequence alignments were inferred using MAFFT (v 7.402) (Katoh and Standley 2013) with local pair alignment for amino acid sequences and global pair alignment for the deltavirus nucleotide sequences. All alignments were then trimmed to remove ambiguous regions using trimAl (v1.4.1) (Capella-Gutierrez, Silla-Martinez, and Gabaldon 2009) with a gap threshold of 0.8 and a similarity threshold of 0.005 and then manually assessed using AliView (Larsson 2014). Phylogenetic trees were estimated using IQ-TREE 2 (v2.2.2) (Minh et al. 2020) with the appropriate substitution model selected using ModelFinder (Kalyaanamoorthy et al. 2017). Branch support was assessed with 1,000 bootstrap replicates using ultrafast bootstrapping (Hoang et al. 2017).

2.5. Mapping sequences to herpesvirus

As only one reference sequence for a single gene (DNA-dependent DNA polymerase) was available for the species of *Antechinus* herpesvirus previously identified, Bowtie2 (v2.2.5) (Langmead and Salzberg 2012) was used to align the non-host reads to the

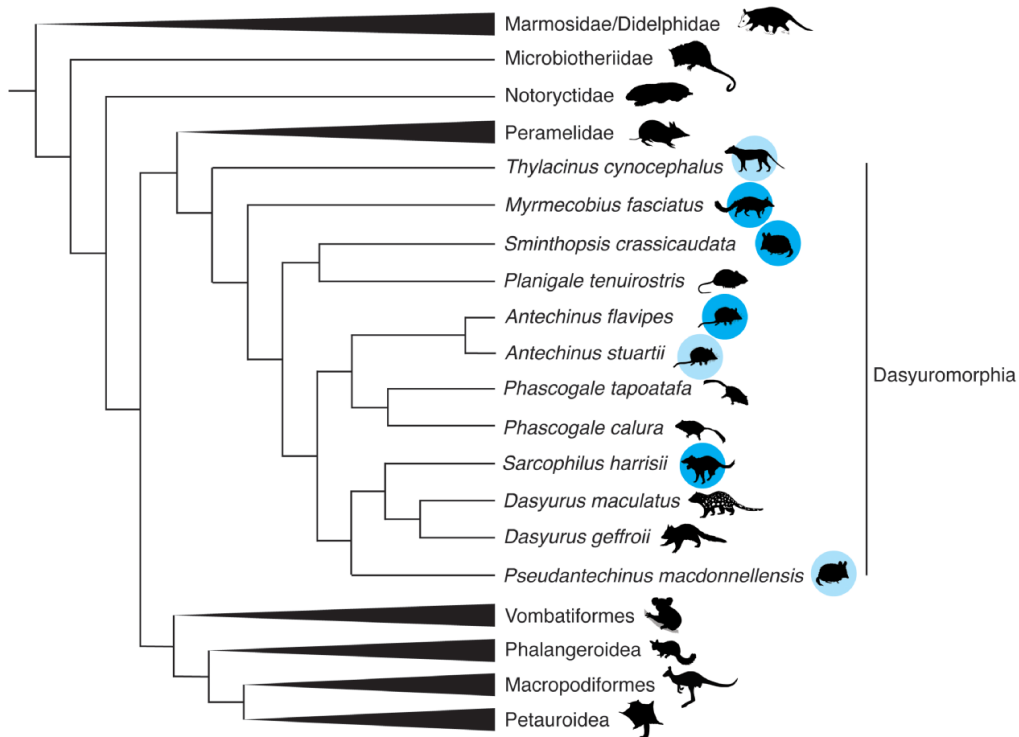


Figure 1. A cladogram depicting the phylogenetic relationships of marsupials. Adapted from Duchêne et al. (2017). Species with SRA RNA-seq data available are indicated by a blue circle, species in which viruses were identified in this study are indicated with dark blue circles, while species in which no viruses were found are indicated with a light blue circle. Animal silhouettes are sourced from PhyloPic (<https://www.phylopic.org/>) produced by Sarah Werning, Gabriela Palomo-Munoz. Creative Commons licence is found at <https://creativecommons.org/licenses/by/3.0/>.

Dasyurid herpesvirus 1 DNA-dependent DNA polymerase reference sequence (MF576269.1). The resulting alignment was viewed in Geneious (Kearse et al. 2012) and used to assess the percentage sequence identity between the novel sequence and the available reference.

2.6. Library composition assessment

To assess the taxonomic composition of each library, contigs were aligned to a custom NCBI nucleotide database without environmental and artificial sequences (<https://researchdata.edu.au/indexed-reference-databases-kma-ccmetagen/1371207>) using the KMA aligner (v1.3.9a) and the CCMetagen program (v1.1.3) (Clausen et al. 2018; Marcelino et al. 2020). The abundance of each taxonomic group was determined by counting the number of nucleotides that matched the reference sequence, with an additional correction for template length using the default parameter in KMA. For data visualisation, CCMetagen was used to generate Krona graphs, which were subsequently edited in Adobe Illustrator (<https://www.adobe.com>).

3. Results

3.1. Screening Dasyuromorphia for viruses

As of March 2023, the NCBI SRA database contained 446 RNA-seq libraries from Dasyuromorphia (Supplementary Table S1), comprising seven species: *S. crassicaudata* (fat-tailed dunnart) (21 libraries), *A. flavipes* (yellow-footed antechinus) (113 libraries), *A. stuartii* (brown antechinus) (10 libraries), *S. harrisii* (Tasmanian devil) (292 libraries), *T. cynocephalus* (Tasmanian tiger) (6 libraries), *M.*

fasciatus (numbat) (3 libraries), and *Pseudantechinus macdonnellensis* (fat-tailed false antechinus) (1 library).

Virus sequence similarity screening of the available libraries resulted in potential positive hits in 206 of the screened libraries, including Tasmanian devil, yellow-footed antechinus, numbat, and fat-tailed dunnart. Following secondary assembly and removal of contigs with hits to endogenous viruses, twenty-two partial or full virus genomes were identified in forty-three libraries including four host species (Fig. 1). These virus sequences were taxonomically assigned to five DNA virus families: Adenoviridae, Anelloviridae, Herpesviridae, Papillomaviridae, and Polyomaviridae, and three RNA virus taxa: the order Jingchuvirales, the genus Hepacivirus, and the delta-like virus group (Table 1). Of these, two were partial sequences of novel herpesviruses in the Tasmanian devil and the fat-tailed dunnart for which no gene that could be used for phylogenetic confirmation could be assembled, and were thus excluded from further study. There was also a fragment of the ORF2 of an anellovirus present in a Tasmanian devil library, but as the ORF1 fragment that is used for phylogenetic classification of these viruses was absent, and ORF2 is rarely published for this virus species, it was similarly excluded.

Within the virus-positive dataset, further analysis revealed that four *Antechinus* libraries from the same study (SRR11306636, SRR11306642, SRR11306648, and SRR11306672) appeared to contain viruses identified as likely contaminants. Specifically, the human-associated viruses Gammapapillomavirus 19 and Betapapillomavirus 1 were identified with over 99 per cent nucleotide sequence identity, and human reads were found in three of these libraries (Supplementary Fig. S1). These viruses were excluded

Table 1.

Virus name	Closest blast hit (no.)	Percentage identity to top blast hit (aa)	Number of nucleotides recovered	Percentage of expected genome recovered (using closest relative as reference)	Virus taxonomy (to the lowest determinable taxonomic level)	Host	Tissue	BioProject	Library ID
Fat-tailed dunnart deltavirus	Rodent deltavirus (QJD13562.1)	71	642	38	Deltavirus	<i>Smynthopsis crassicaudata</i>	Adult eye	PRJNA554238	SRR9673767
Tasmanian devil deltavirus	Rodent deltavirus (QJD13562.1)	71	741	44	Deltavirus	<i>Sarcophilus harrisi</i>	Tasmanian devil facial tumour	PRJNA693818	SRR13765777
Tasmanian devil chu-like virus	Lishi spider virus 1 (A/G39051.1)	24	12,118	100	Jingchuvirales	<i>Sarcophilus harrisi</i>	Tasmanian devil facial tumour (DFTD1) cell lines	PRJNA422607	SRR6380970
Antechinus hepaticus	Northern treeshrew hepatitis (CA15760841.1)	48	9,094	100	Hepacivirus	<i>Antechinus flavipes</i>	Spleen, kidney, liver, stomach, and cerebrum	PRJNA565840	SRR10127601 SRR10127603 SRR11306583 SRR11306585 SRR11306588 SRR11306615 SRR11306617 SRR11306621 SRR11306628 SRR11306634 SRR11306639 SRR11306641 SRR11306643 SRR11306667 SRR11306686
Numbat hepatitis	Duck hepatitis (QKT21547.1)	29	10,486	100	Hepacivirus	<i>Myrmecobius fasciatus</i>	Lung and liver	PRJNA786364	SRR17244188
Dasyurid herpesvirus 4	Ursid gamma-herpesvirus 2 (AZP55492.1)	37	841	0.50	Gammaherpesvirinae	<i>Antechinus flavipes</i>	Male prostate	PRJNA565840	SRR11306664
Dasyurid herpesvirus 5	Ursid gamma-herpesvirus 2 (AZP55492.1)	38	20,464	12	Gammaherpesvirinae	<i>Antechinus flavipes</i>	Male stomach	PRJNA565840	SRR11306624

(continued)

Table 1. (Continued)

Virus name	Closest blast hit (no.)	Percentage identity to top blast hit (aa)	Number of nucleotides recovered	Percentage of expected genome recovered (using closest relative as reference)	Virus taxonomy (to the lowest determinable taxonomic level)	Host	Tissue	BioProject	Library ID
Tasmanian devil polyomavirus 3	Goose haemorrhagic polyomavirus (UJT42146.1)	43	2,513	46	Polyomaviridae	<i>Sarcophilus harrisi</i>	Lip tissue	PRJNA693818	SRR13765816 SRR13765794
Antechinus	Torque teno virus (QBA84075.1)	28	1,745	79	Anelloviridae	<i>Antechinus flavipes</i>	Spleen	PRJNA565840	SRR1306585 SRR1306684
Antechinus adenovirus	Bat mastadenovirus WIV13 (YP_009272931.1)	69	9,474	32	Adenoviridae	<i>Antechinus flavipes</i>	Kidney, prostate, and spleen	PRJNA565840	SRR1306638 SRR1306640 SRR1306641 SRR1306643 SRR1306644 SRR1306646 SRR1306647 SRR1306648 SRR1306664 SRR1306685
Tasmanian devil papillomavirus 3	Human papillomavirus type 229 (UQF78855.1)	45	4,378	60	Papillomaviridae	<i>Sarcophilus harrisi</i>	Lip tissue	PRJNA693818	SRR13765825
Tasmanian devil papillomavirus 4	Capra hircus papillomavirus type 2 (QH12244.1)	40	5,443	75	Papillomaviridae	<i>Sarcophilus harrisi</i>	Lip tissue	PRJNA693818	SRR13765801
Tasmanian devil papillomavirus 5	Human papillomavirus (QAB13979.1)	43	2,105	29	Papillomaviridae	<i>Sarcophilus harrisi</i>	Lip tissue	PRJNA693818	SRR13765803 SRR13765816
Tasmanian devil papillomavirus 6	Human papillomavirus (QAB13979.1)	60	5,469	75	Papillomaviridae	<i>Sarcophilus harrisi</i>	Lip tissue	PRJNA693818	SRR13765808
Tasmanian devil papillomavirus 7	Bos taurus papillomavirus 20 (YP_009272609.1)	40	3,198	45	Papillomaviridae	<i>Sarcophilus harrisi</i>	Lip tissue	PRJNA693818	SRR13765811



aligning to the contig (Fig. 2). Tasmanian devil deltavirus was identified in a sample of Tasmanian devil facial tumour disease (DFTD) taken from a female Tasmanian devil. The assembled contig was again fragmentary, at 741 nucleotides in length, and the abundance was also low with 3.98×10^{-7} per cent of reads (expected count: twenty-seven) aligning to the contig. In both cases, no other viruses were identified in the sequencing libraries.

3.3. RNA viruses

We identified three species of novel RNA viruses in this study: two hepaciviruses (positive-sense RNA, family Flaviviridae), one in antechinus and one in a numbat, and a chu-like virus (negative-sense RNA virus, order Jingchuvirales) here named Tasmanian devil chu-like virus. The chu-like virus, which has an unsegmented, non-circular genome, was identified in a DFTD cell line that was cultured from primary tissue (Kozakiewicz et al. 2021). The genome contained four predicted ORFs, one of these being the RdRp, while the remaining three had no sequence identity to any reference sequence. This was at the highest abundance of all viruses identified in this study, representing 0.2 per cent of total reads (expected read count: 253,672) aligning to the viral

370

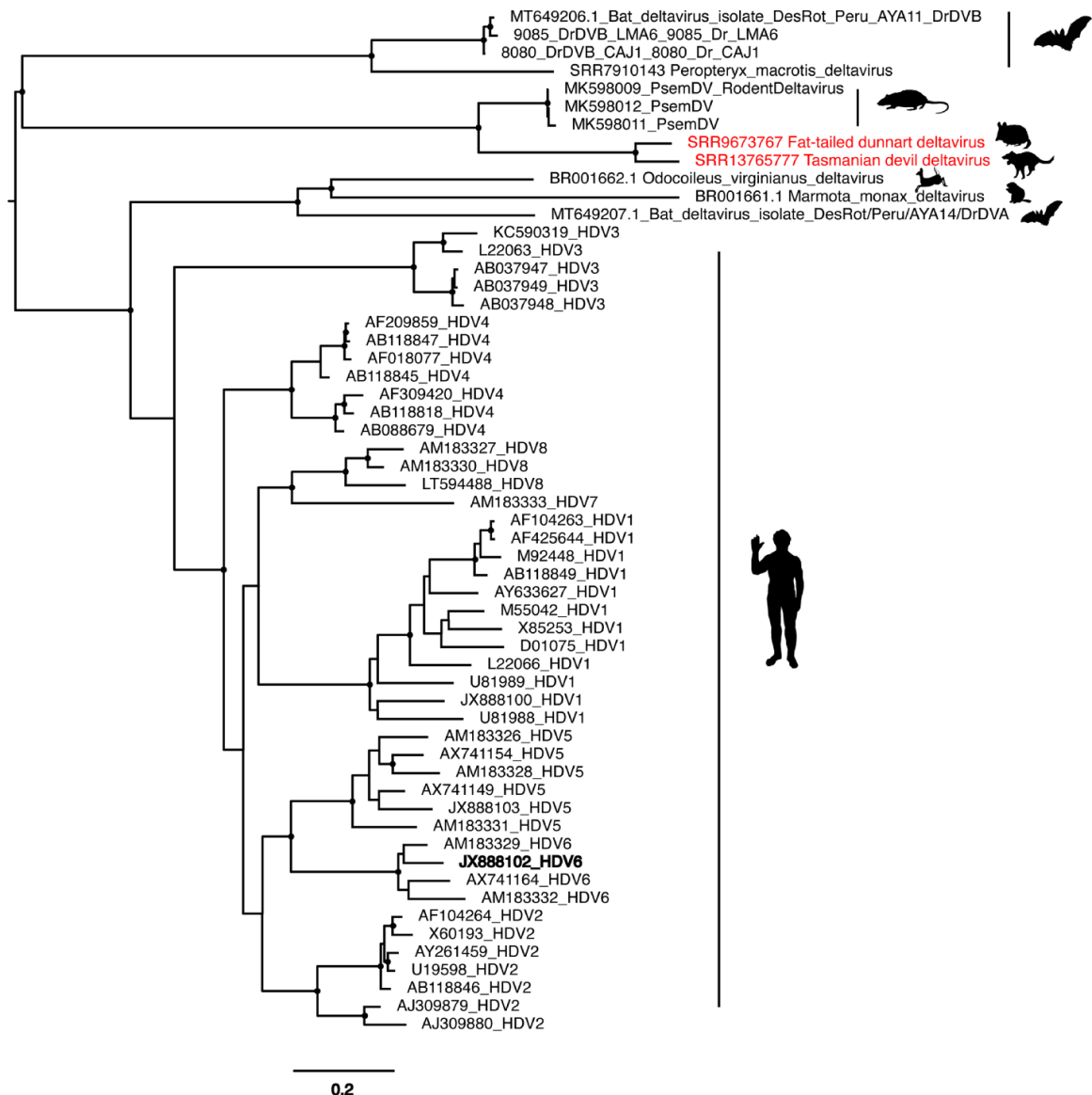


Figure 3. Phylogeny of mammalian deltavirus small delta antigen-like protein nucleotide sequence. The phylogeny was estimated using the GTR+I+4 nucleotide substitution model. The nucleotide alignment used to generate this tree was 679 positions in length. Sequences identified in this study are shown in red text and the library ID is indicated in the taxa name. Branch support >90 per cent is indicated with a black dot at the node. The scale bar indicates the number of nucleotide substitutions per site and the tree is midpoint rooted for clarity. This alignment is based on that provided by Bergner et al. (2021). Animal silhouettes indicate the host species.

genome (Fig. 2) of 12,118 nucleotides. A phylogenetic analysis of the RdRp revealed that Tasmanian devil chu-like virus fell outside of the Chuviridae and was highly divergent from known viruses including the only known vertebrate clade of Chuviridae, the genus *Piscichuvirus*, although it did cluster within the order Jingchuvirales (Fig. 4A). The high abundance in the cultured Tasmanian devil tissues and the absence of any sequencing reads of a non-target source in this library (BioProject SRR6380970) tentatively suggest that it does replicate in these cells and was not the result of contamination (Supplementary Fig. S1).

Antechinus hepaticivirus was detected in multiple libraries from a single BioProject (PRJNA565840) in which multiple tissues of thirteen antechinus individuals were sequenced. The assembled virus genome was 9,094 nucleotides in length and was detected in five of the thirteen individuals and in at least two tissue types from each positive individual, with libraries of liver tissue consistently having the highest abundance (1.18×10^{-4} – 6.44×10^{-4} per cent of total reads). The genome structure was a single polyprotein, consistent with the genus *Hepaticivirus*. The full virus genome was assembled from library SRR11306639, a library of male liver tissue and the

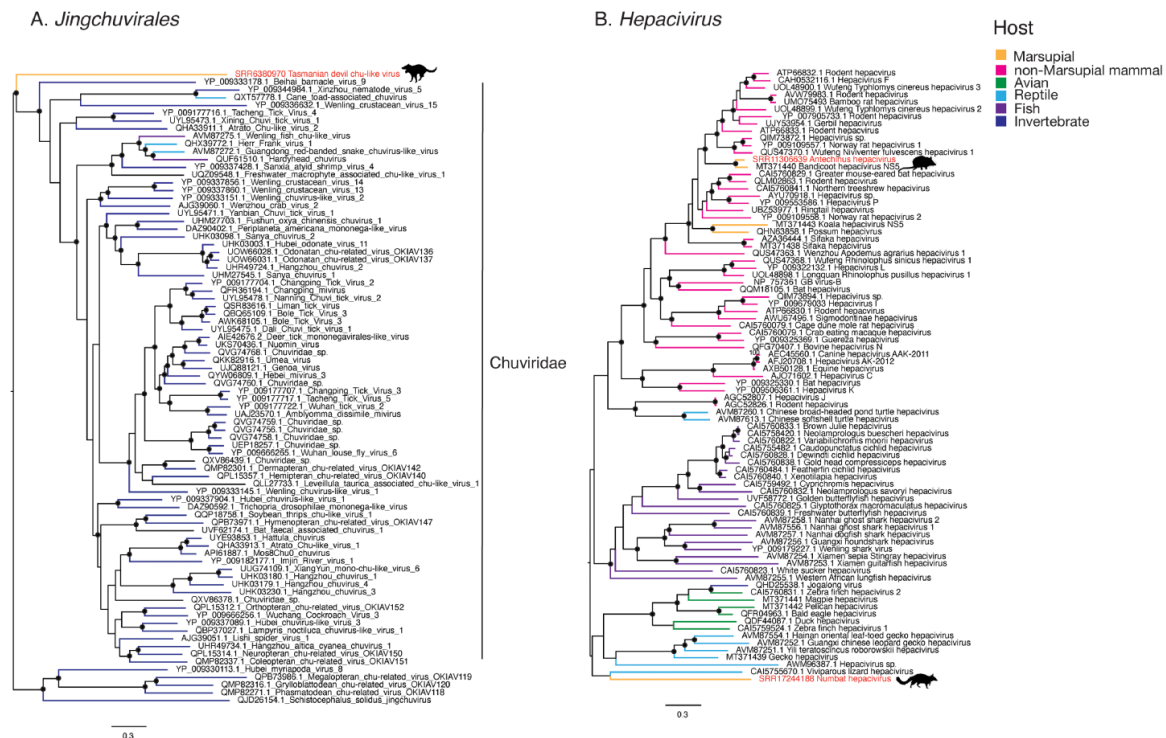


Figure 4. Phylogenetic trees of the RNA viruses identified in this study. (A) Phylogeny of the order Jingchuvirales based on the RdRp amino acid sequence using the LG+G4+I+4 substitution model. The alignment used to generate this tree was 1,160 positions in length. (B) Phylogeny of the hepacivirus NS5 protein sequence estimated using the LG+G4+I+4 substitution model. The alignment used to generate this tree was 594 positions in length. Scale bars indicate the number of amino acid substitutions per site. Trees are midpoint rooted for clarity. Branch support values >90 per cent are indicated with a black dot at the node. Viruses identified in this study are highlighted with red taxa labels. Animal silhouettes indicate the host species of the viruses identified here and tip branches are coloured according to broader host species.

sequence identity across the other five individuals varied between 93 and 98 per cent. The virus was also identified in spleen, kidney, stomach, and cerebrum. Based on the NS5 (RdRp) protein, this virus was most closely related to a hepacivirus identified in an *Ixodes holocyclus* tick engorged with the blood of a long-nosed bandicoot in Australia (Harvey et al. 2019; Porter et al. 2020). In turn, these two viruses were related to rodent hepacivirus (Fig. 4B).

Finally, a highly divergent hepacivirus was identified in a transcriptome of numbat liver tissue (Peel et al. 2022) with an abundance of 9.9e-5 per cent of total reads and a length of 10,486 nucleotides with a single polyprotein consistent. This virus, here named numbat hepacivirus, had 37 per cent amino acid sequence identity to the closest blast hit (Norway rat hepacivirus 2, YP_009325411.1) over the NS5 protein and 29 per cent amino acid sequence identity across the whole genome to the closest blast hit (Duck hepacivirus, QKT21547.1). Numbat hepacivirus clustered phylogenetically with a clade of avian- and reptile-associated hepaciviruses based on the NS5 protein (Fig. 4B), but on a long branch characterised by low bootstrap support (63 per cent), suggesting that its phylogenetic position is uncertain due to high-sequence divergence.

3.4. DNA viruses

Although RNA-seq data were analysed here, our dataset contained a surprising diversity of novel DNA virus transcripts. Herpesviruses were present in three species—antechinus, Tasmanian devil, and fat-tailed dunnart—but only two antechinus libraries

contained sufficient gene sequences for phylogenetic analysis and therefore only these species were analysed further. A very small fifty-five nucleotide fragment of glycoprotein H of the herpesvirus identified in Tasmanian devils exhibited 100 per cent nucleotide sequence identity to Dasyurid herpesvirus 3 (Chong et al. 2019).

Orthoherpesvirus sequences were identified in four Antechinus libraries—two from female spleen libraries containing two contigs each—although none were appropriate for phylogenetic analysis and did not overlap with the genes present in the other two libraries. These sequences were not investigated further. Within the remaining two libraries, one library from male stomach tissue (SRR11306624) contained two short contigs that included the glycoprotein B gene (as well as one hypothetical protein gene sequence), which could be utilised in phylogenetic analysis, here named Dasyurid herpesvirus 5. The other, a library of prostate tissue, contained twenty-four contigs with sequence identity to gammaherpesvirus genes, also including the glycoprotein B sequence, here named Dasyurid herpesvirus 4. A comparison of these two glycoprotein B sequences revealed that they shared only 25 per cent amino acid sequence identity. We used Bowtie2 to align reads from the herpesvirus-positive libraries to the Dasyurid herpesvirus 1 reference sequence (MF576269.1), for which only the DNA-dependent DNA polymerase is currently available. A single read from SRR11306664 (containing Dasyurid herpesvirus 4) aligned with 92 per cent nucleotide sequence identity to the Dasyurid herpesvirus 1 reference, while no reads aligned from the other three Antechinus libraries containing

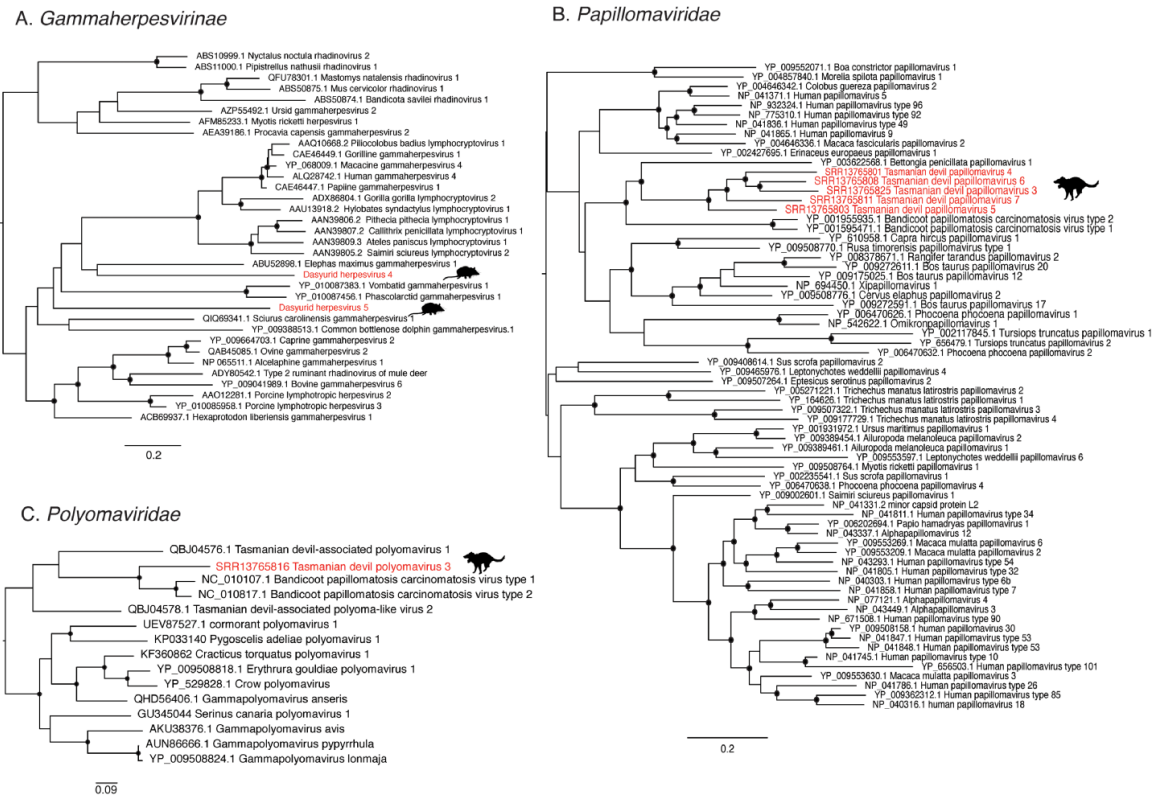


Figure 5. Phylogenies of *Papillomaviridae*, *Polyomaviridae*, and *Gammaherpesvirinae*. (A) Phylogeny of the family *Gammaherpesvirinae* estimated using the LG+I+4 amino acid substitution model. The alignment used to generate this tree was 318 positions in length. (B) Phylogeny of the *Papillomaviridae* L1 and L2 proteins estimated using the LG+I+4 substitution model. The alignment used to generate this tree was 809 positions in length. (C) Phylogeny of the *Polyomaviridae* large T antigen protein estimated using the LG+I+4 substitution model. The alignment used to generate this tree was 446 positions in length. Trees are midpoint rooted for clarity. Scale bars indicate the number of amino acid substitutions per site. Viruses identified in this study are highlighted in red text and animal silhouettes indicate the host species. Branch support values >90 per cent are indicated with a black dot at the node.

herpesvirus contigs. Based on phylogenetic analysis of glycoprotein B, Dasyurid herpesvirus 4 was most closely related to an elephant herpesvirus (Fig. 5A). Dasyurid herpesvirus 5 fell as a sister lineage to this group, although the sequence is very short and the node is characterised by low bootstrap support. Without the DNA polymerase, the most commonly sequenced orthoherpesvirus gene used for taxonomic demarcation, the taxonomic relationship of these viruses was uncertain.

The most diverse viral family identified was the family *Papillomaviridae*. Specifically, we discovered transcripts from five species of novel papillomavirus in six libraries of Tasmanian devil lip tissue from a study of DFTD (Kozakiewicz et al. 2021). These viruses were named Tasmanian devil papillomavirus 3–7 to remain consistent with previously identified Tasmanian devil-associated papillomaviruses (Chong et al. 2019). Partial gene transcripts were identified in all six libraries, although they were not consistent across all libraries. The L2 gene (minor capsid protein) was found in all six libraries and thus used for comparison. Five of these were distinct papillomavirus L2 genes, with amino acid sequence identity ranging from 36 to 59 per cent. As the ICTV species demarcation for papillomavirus genera is <70 per cent, these were deemed to represent five distinct virus species. As the L1 (major capsid) protein was present in five libraries with distinct L2 transcripts, L1 and L2 were used for phylogenetic analysis. These sequences

formed a distinct clade of marsupial papillomaviruses with a betong papillomavirus and the bandicoot papillomavirus that is the sole member of the genus *Dyolambdapapillomavirus* (Fig. 5B). The presence of papillomaviruses did not correlate with the DFTD status of the animal as both DFTD positive and negative samples contained papillomaviruses and there was no phylogenetic distinction based on the DFTD status.

As the study for which these data were generated was designed to determine geographical patterns in DFTD expression profiles (Kozakiewicz et al. 2021), we assessed whether geographic location affected the phylogenetic clustering by manual inspection of the tree. No obvious pattern was observed. Two papillomavirus species were previously identified in Tasmanian devil faecal samples—Tasmanian devil-associated papillomavirus 1 and Tasmanian devil-associated papillomavirus 2—although only the E1 protein was published for these viruses. Of the virus sequences discovered here, only a small fragment (317 nt) of the E1 region could be recovered for Tasmanian devil papillomavirus 6. This exhibited 48 per cent amino acid (aa) identity to Tasmanian devil-associated papillomavirus 1, 55 per cent aa identity to Tasmanian devil-associated papillomavirus 2, and 28 per cent aa identity to *Bettongia penicillata* papillomavirus 1.

Of interest, a library that contained the E2 protein of Tasmanian devil papillomavirus 5 (SRR13765816) also contained a contig

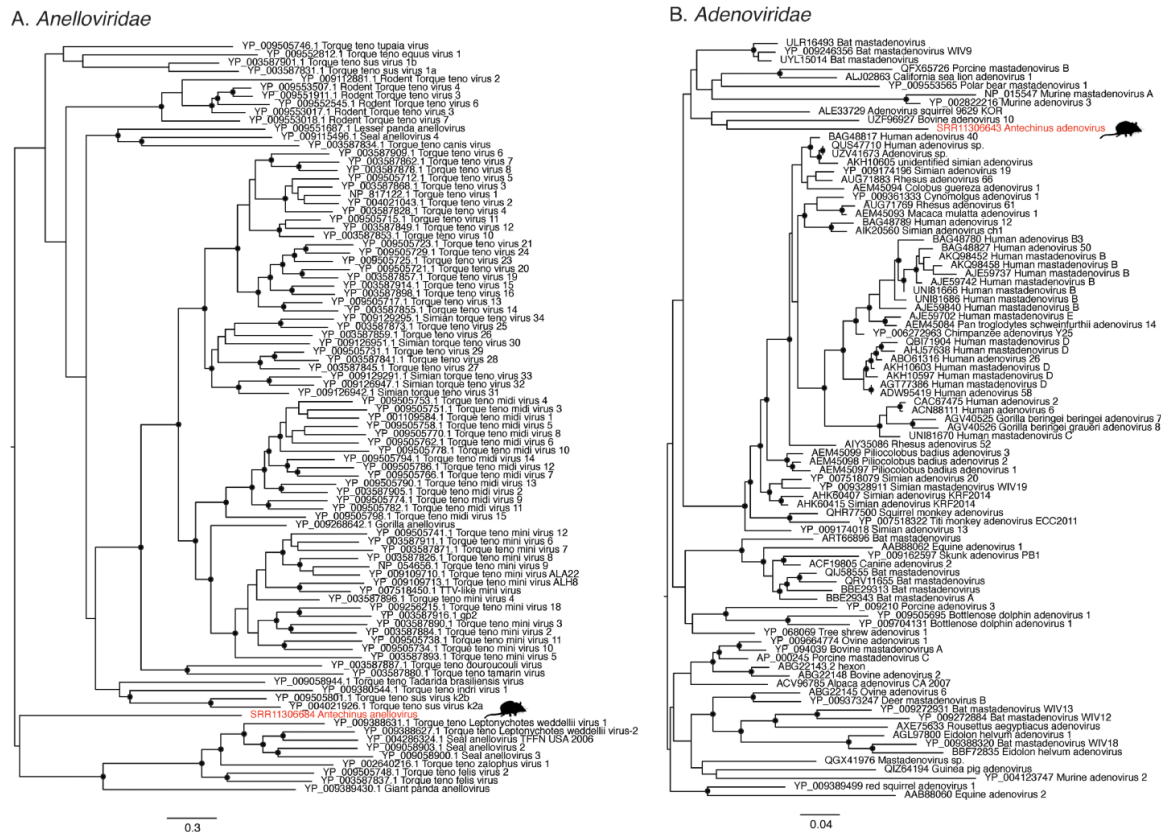


Figure 6. Phylogenies of Adenoviridae and Anelloviridae. (A) Phylogeny of the Anelloviridae hexon protein estimated using the LG+I+4 substitution model. The alignment used to generate this tree was 440 positions in length. (B) Phylogeny of the Adenoviridae hexon protein estimated using the LG+I+4 substitution model. The alignment used to generate this tree was 545 positions in length. Viruses identified in this study are indicated with red text and animal silhouettes indicate the host species. Trees are midpoint rooted for clarity. Branch support values >90 per cent are indicated with black dot at the node. Scale bars indicate the number of amino acid substitutions per site.

showing sequence identity to the large T antigen of a novel polyomavirus. A second library, SRR13765794, contained an identical polyomavirus large T antigen but no other virus sequences. No other polyomavirus genes were detected in these libraries. As this large T antigen sequence was related to but distinct from Tasmanian devil polyomavirus 1 and 2 (Fig. 5C), we named this novel virus Tasmanian devil polyomavirus 3.

Transcripts of two novel anelloviruses were detected, although only one of these—here named Antechinus anellovirus—contained the ORF1 sequence and hence was analysed further. Phylogenetically, Antechinus anellovirus was extremely divergent in the ORF1 protein from the current diversity of Anelloviridae and was closely related to a group of anelloviruses detected in seals, cats, and a giant panda (Fig. 6A). Antechinus anellovirus sequences with 100 per cent sequence identity were detected in two libraries of spleen, co-occurring with Antechinus hepacivirus in one of these libraries.

Similarly, the adenovirus transcripts identified in this study were also found in the same samples as Antechinus hepacivirus: in the kidney tissue of two individuals, as well as in eight other libraries (seven individuals) where it was the only virus identified. Across libraries, the virus sequences were 99 per cent similar at the nucleotide level. The virus was mainly found in libraries

of kidney tissue (8/10), as well as in one library of spleen and one of prostate tissue. Antechinus adenovirus was most closely related to members of the genus *Mastadenovirus*, but due to the high degree of divergence, the phylogenetic position of this virus is still unclear (Fig. 6B).

4. Discussion

We identified twenty-two full or partial virus genomes from eight virus taxa within 446 SRA transcriptome libraries of tissue samples from species within the order Dasyuromorphia. Of these virus genomes, fifteen were subjected to phylogenetic analysis. The remaining seven virus contigs either were viruses identified as contaminants (in four cases discussed later) or were too short for meaningful analysis; this was the case with two herpesviruses identified in a fat-tailed dunnart library and a Tasmanian devil library, as well as the ORF2 segment of a novel anellovirus in a Tasmanian devil. Human papillomavirus sequences were identified in four antechinus libraries, three of which also contained contaminating human sequences. Although this is a risk when analysing SRA data, the use of taxonomic classification software such as CCMetagen (Marcelino et al. 2020) can be used to identify contaminating host sequences. We can also be confident that the viruses identified here are exogenous, rather than endogenous,

as high-quality host genomes are available for three of the four Dasyurid species studied here, and a comprehensive analysis of EVEs in marsupials has previously been undertaken (Harding et al. 2021).

Until 2018, deltaviruses were believed to be exclusive to humans. However, this was disproven by the identification of delta-like viruses in birds and snakes (Wille et al. 2018; Hetzel et al. 2019). Since this time, additional deltaviruses have been identified through SRA screening (Bergner et al. 2021). Of note, we describe the first marsupial deltaviruses that also form what appears to be a marsupial-specific clade. Specifically, Tasmanian devil deltavirus and fat-tailed dunnart deltavirus cluster in a sister clade to a group of rodent-associated deltaviruses also identified in an SRA-screening study (Bergner et al. 2021). Further exploration of the marsupial virome is needed to determine if this putative marsupial cluster of deltaviruses holds true, such that these viruses have associated with marsupial hosts for their entire 160-million-year history (Luo et al. 2011), or if there has been frequent cross-species transmission on more recent timescales as in other deltaviruses (Bergner et al. 2021).

Two distinct marsupial carnivore hepaciviruses were also identified in this study. Hepaciviruses are associated with liver disease in humans, although in most animals their pathogenicity is unknown. Three marsupial hepaciviruses have previously been identified: Koala hepacivirus, Possum hepacivirus, and Collins beach virus, all of which were identified metagenomically (Chang et al. 2019; Harvey et al. 2019; Porter et al. 2020). Antechinus hepacivirus was most closely related to a suspected bandicoot-associated hepacivirus, denoted Collins beach virus, while Koala hepacivirus and Possum hepacivirus cluster together. These two marsupial clades are paraphyletic, with the Antechinus hepacivirus and Collins beach virus grouping with rodent hepaciviruses and the Possum and Koala hepaciviruses clustering with a broader range of mammals. This phylogenetic pattern may be a result of bandicoots and antechinus sharing a similar ecological niche—ground-dwelling foragers (along with rodents)—thereby enabling cross-species transmission, while Koala and Possum hepaciviruses are mainly tree-dwelling herbivores. The lack of a marsupial-specific clade, as well as the grouping with rodent-associated hepaciviruses, is also indicative of multiple introductions of hepaciviruses to marsupials, perhaps via rodent species, although this will again need to be resolved through more extensive sampling. Of note, Antechinus hepacivirus was identified in a relatively large proportion of individuals (5/13), which could suggest a high prevalence in the population and hence be of particular concern if populations are being used for translocation in population reseedling programmes (Manning 2021). In addition, we identified a highly divergent hepacivirus in a numbat liver transcriptome. This virus shared only 29 per cent amino acid sequence identity to the closest blast hit and did not fall within the larger mammalian/marsupial clade, suggesting the presence of a third, highly divergent lineage of marsupial hepaciviruses. Further study of marsupial-associated hepaciviruses is needed to shed light on the emergence and evolution of these viruses in Australian marsupials and to determine what effect these infections may have on their hosts.

A highly divergent chu-like virus was identified in a transcriptome of a DFTD cell line (SRR6380970) grown from primary tissues (Kosack et al. 2019). The Chuviridae and chu-like viruses belong to the order Jingchuvirales, which was until recently believed to be invertebrate specific (Di Paola et al. 2022). The discovery of a chu-like virus in the brain of a snake with neurological disease (Argenta et al. 2020) and in a meta-transcriptomic study of reptiles (Shi et al.

2018) led to the identification of a fish-reptile-associated genus (*Piscichuvirus*) within the Chuviridae. More recently, three novel *Piscichuvirus* species were identified and associated with encephalitis in turtles (Laovechprasit et al. 2023), suggesting that this genus could be associated with neurological disease. Tasmanian devil chu-like virus is the sister group to the Chuviridae within the order Jingchuvirales and is highly divergent from the characterised species. Interestingly, it has been suggested that DFTD is of Schwann cell origin (Murchison et al. 2010), a cell type associated with the peripheral nervous system, such that the novel chu-like virus identified here could similarly be associated with the nervous system. However, as this virus was identified from SRA data and from a library of cultured cells, we cannot be certain of the true host of this virus. Further work is needed to determine if this virus is capable of infecting DFTD cells and what effect this infection has on the cancer cells.

Transcripts from five families of DNA viruses were identified in these samples. Papilloma-, polyoma-, and herpesviruses are often associated with disease and, in some cases cancer, while anellovirus and adenovirus have less definitive links to disease. We identified a novel anellovirus in antechinus spleen transcriptomes, which co-occurred with Antechinus hepacivirus in one library. These viruses are ubiquitous and their disease association is still debated, although it has been suggested that they may play a role as co-infecting viruses (Webb, Rakibuzzaman, and Ramamoorthy 2020). Similarly, Antechinus adenovirus was identified in kidney, prostate, and spleen tissue transcriptomes and co-occurred with hepacivirus in two of ten libraries. This virus clustered with a group of mastadenoviruses identified in mice, cows, pigs, bats, and a polar bear. Mastadenoviruses are in some cases associated with diseases including encephalitis, respiratory disease (Chen and Tian 2018), and gastrointestinal symptoms (Dayaram et al. 2018). In this case, no disease was reported in the individuals sequenced, but this is difficult to assess in SRA-screening studies.

Sequences from two distinct species of the family Gammaherpesvirinae were identified in antechinus transcriptomes of stomach and prostate tissue, along with partial herpesvirus sequences in Tasmanian devil peripheral blood mononuclear cells and fat-tailed dunnart mitogen-stimulated splenocytes. Herpesviruses are relatively well characterised in marsupials, with Dasyurid herpesvirus 1 identified in Antechinus in 2014 (Amery-Gale et al. 2014) and subsequent studies revealing two Tasmanian devil-associated herpesviruses (Dasyurid herpesvirus 2 and Dasyurid herpesvirus 3) (Stalder et al. 2015; Chong et al. 2019). However, the species demarcation of viruses within the family now termed Orthoherpesviridae is unclear, with no defined genetic distance or biological features used for classification (Gatherer et al. 2021). Given that very few genomic sequences are available for the related viruses, we suggest 92 per cent nucleotide identity across the single read of the DNA-dependent DNA polymerase (the only gene available for Dasyurid herpesvirus 1) as sufficient genetic distance to represent a novel species. DNA sequencing would likely lead to the recovery of the full genome, including gene sequences necessary to determine the relatedness of these species to each other and to other Dasyurid herpesviruses, but this is beyond the scope of this study.

We identified five distinct but related species of papillomavirus in Tasmanian devil lip tissue transcriptomes within the libraries of a single SRA BioProject. The DFTD status of the sampled individual did not appear to be associated with the presence of papillomavirus, nor did the DFTD status correlate with any clustering in the phylogeny of these species. Two Tasmanian devil-associated

papillomaviruses were previously identified in a faecal meta-transcriptome, but as these animals are carnivores, it was impossible to determine their true host (Chong et al. 2019). Phylogenetic analysis revealed that all these sequences were most closely related to each other and more broadly to *Bettongia penicillata* papillomavirus 1, isolated from papillomatous lesions on a brush-tailed bettong in Australia (Bennett et al. 2010), and more distantly with Bandicoot papillomatosis carcinomatosis virus types 1 and 2. Bandicoot papillomatosis carcinomatosis virus uniquely exhibits genomic features of both papillomavirus and polyomavirus proteins with the L1 and L2 proteins of a papillomavirus and the large and small T antigen proteins of a polyomavirus (Woolford et al. 2007). Interestingly, the large T antigen protein sequence of a novel polyomavirus was identified in one library in this study, which also contained the E2 protein of Tasmanian devil papillomavirus 5, although this is not consistent with the genomic structure of the previously identified hybrid virus. Phylogenetically, the large T antigen sequence of Tasmanian devil polyomavirus 3 was most closely related to the large T antigen sequence of the hybrid Bandicoot papillomatosis carcinomatosis virus, with Tasmanian devil-associated polyomavirus 1 and Tasmanian devil-associated polyoma-like virus 2 being more divergent. Interestingly, Tasmanian devil-associated polyomavirus 1 and Tasmanian devil-associated polyoma-like virus 2 were identified in the same faecal meta-transcriptome study as Tasmanian devil-associated papillomaviruses 1 and 2. DNA sequencing will be required to confirm the genomic structure and phylogenetic relationships of these viruses.

Although our study identified evidence of a relatively high diversity of DNA viruses within transcriptomic data, the lack of complete genomes limits the value of these sequences in phylogenetic comparison and genomic characterisation of these species. In contrast, there was a relatively limited diversity of RNA viruses in these data, which may be a product of the sequencing methodology used. For example, poly-A selection during library preparation will remove many virus sequences (Visser et al. 2016). However, despite being prepared with a poly-A selection, we identified a high diversity of viruses within the yellow-footed *Antechinus* libraries, with 10/13 sequenced individuals being associated with at least one virus. *Antechinus* hepacivirus was particularly prevalent, and as these species are being proposed for species reintroduction programmes, it is of particular importance that the viromes of these individuals are investigated to prevent the spread of these viruses to naive populations. From the data provided here, PCR primers could be designed to screen for these viruses in individuals being used in translocation programmes and in cases of undiagnosed disease in marsupial species to determine if there is some disease association.

In light of the diversity of marsupial carnivore viruses identified in this small study of transcriptomic data not generated for the purposes of virus discovery, it is evident that considerable viral diversity exists in these species and that further study is required to understand the viromes and viral evolutionary history of these unique species. Given the threat that these species face as a result of human activity, it is imperative that we understand the risk that viral disease poses to what are often small, isolated populations.

Data availability

All viral genomes and corresponding sequences assembled in this study and the alignments generated in this study for phylogenetic analysis are available at <https://github.com/erinhunter4/>

[dasyurid_virome](https://github.com/JonathonMifsud/BatchArtemisSRAMiner). The SRA virus discovery pipeline is available at <https://github.com/JonathonMifsud/BatchArtemisSRAMiner>.

Supplementary data

Supplementary data is available at Virus Evolution online.

Acknowledgements

We acknowledge the University of Sydney's high-performance computing cluster, Artemis, for providing computing. This work would not be possible without the sequencing data that have been generously shared by the research community, and we are grateful to the NCBI SRA team for their management of this resource.

Funding

This work was funded by a National Health and Medical Research Council Investigator Grant to E.C.H. (no. GNT201719) and AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong Special Administrative Region, China.

Conflict of interest: None declared.

References

- Amery-Gale, J. et al. (2014) 'Detection and Identification of a Gamma-herpesvirus in *Antechinus* spp. in Australia', *Journal of Wildlife Diseases*, 50: 334–9.
- Argenta, F. F. et al. (2020) 'Identification of Reptarenaviruses, Hartmanviruses, and a Novel Chuvirus in Captive Native Brazilian Boa Constrictors with Boid Inclusion Body Disease', *Journal of Virology*, 94: e00001-20.
- Barker, I. K., Carbonell, P. L., and Bradley, A. J. (1981) 'Cytomegalovirus Infection of the Prostate in the Dasyurid Marsupials, *Phascogale tapoatafa* and *Antechinus stuartii*', *Journal of Wildlife Diseases*, 17: 433–41.
- Bennett, M. D. et al. (2010) 'The First Complete Papillomavirus Genome Characterized from a Marsupial Host: A Novel Isolate from *Bettongia penicillata*', *Journal of Virology*, 84: 5448–53.
- Bergner, L. M. et al. (2021) 'Diversification of Mammalian Deltaviruses by Host Shifting', *Proceedings of the National Academy of Sciences*, 118: e2019907118.
- Bigot, T. et al. (2019) 'RVDB-prot, a Reference Viral Protein Database and Its HMM Profiles', *F1000Research*, 8: 530.
- Bolger, A. M., Lohse, M., and Usadel, B. (2014) 'Trimmomatic: A Flexible Trimmer for Illumina Sequence Data', *Bioinformatics*, 30: 2114–20.
- Buchfink, B., Reuter, K., and Drost, H. G. (2021) 'Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND', *Nature Methods*, 18: 366–8.
- Camacho, C. et al. (2009) 'BLAST+: Architecture and Applications', *BMC Bioinformatics*, 10: 421.
- Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009) 'trimAl: A Tool for Automated Alignment Trimming in Large-scale Phylogenetic Analyses', *Bioinformatics*, 25: 1972–3.
- Chang, W.-S. et al. (2019) 'Metagenomic Discovery and Co-infection of Diverse Wobbly Possum Disease Viruses and a Novel Hepacivirus in Australian Brushtail Possums', *One Health Outlook*, 1: 5.
- Chapman, A. (2009) 'Numbers of living species in Australia and the world', Report for the Australian Biological Resources Study, pp.

- 1–78. Canberra, Australia: Department of the Environment, Water, Heritage and the Arts.
- Charon, J. et al. (2022) 'RdRp-scan: A Bioinformatic Resource to Identify and Annotate Divergent RNA Viruses in Metagenomic Sequence Data', *Virus Evolution*, 8: veac082.
- Chen, S., and Tian, X. (2018) 'Vaccine Development for Human Mastadenovirus', *Journal of Thoracic Disease*, 10: S2280–94.
- Chong, R. et al. (2019) 'Fecal Viral Diversity of Captive and Wild Tasmanian Devils Characterized Using Virion-Enriched Metagenomics and Metatranscriptomics', *Journal of Virology*, 93: e00205–19.
- Clausen, P. T. L. C. et al. (2018) 'Rapid and precise alignment of raw reads against redundant databases with KMA', *BMC Bioinformatics*, 19: 307.
- Dayaram, A. et al. (2018) 'Novel Divergent Polar Bear-Associated Mastadenovirus Recovered from a Deceased Juvenile Polar Bear', *mSphere*, 3: e00171-18.
- Di Paola, N. et al. (2022) 'Jingchuvirales: A New Taxonomical Framework for a Rapidly Expanding Order of Unusual Monjiviricete Viruses Broadly Distributed among Arthropod Subphyla', *Applied and Environmental Microbiology*, 88: e0195421.
- Duchêne, D. A. et al. (2017) 'Analysis of Phylogenomic Tree Space Resolves Relationships among Marsupial Families', *Systematic Biology*, 67: 400–12.
- Feigin, C., Frankenberg, S., and Pask, A. (2022) 'A Chromosome-Scale Hybrid Genome Assembly of the Extinct Tasmanian Tiger (*Thylacinus cynocephalus*)', *Genome Biology and Evolution*, 14: evac048.
- French, R. K. et al. (2022) 'Novel Viral and Microbial Species in a Translocated Toutouwai (*Petroica longipes*) Population from Aotearoa/New Zealand', *One Health Outlook*, 4: 16.
- Gatherer, D. et al. (2021) 'ICTV Virus Taxonomy Profile: Herpesviridae 2021', *Journal of General Virology*, 102: 001673.
- Goodacre, N. et al. (2018) 'A Reference Viral Database (RVDB) to Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection', *mSphere*, 3: e00069-18.
- Harding, E. F. et al. (2021) 'Ancient Viral Integrations in Marsupials: A Potential Antiviral Defence', *Virus Evolution*, 7: veab076.
- Harvey, E. et al. (2019) 'Extensive Diversity of RNA Viruses in Australian Ticks', *Journal of Virology*, 93: 3.
- Hatcher, E. L. et al. (2017) 'Virus Variation Resource—Improved Response to Emergent Viral Outbreaks', *Nucleic Acids Research*, 45: D482–90.
- Hetzel, U. et al. (2019) 'Identification of a Novel Deltavirus in Boa Constrictors', *mBio*, 10: 2.
- Hoang, D. T. et al. (2017) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution*, 35: 518–22.
- Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods*, 14: 587–9.
- Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.
- Kealy, S., and Beck, R. (2017) 'Total Evidence Phylogeny and Evolutionary Timescale for Australian Faunivorous Marsupials (*Dasyuromorphia*)', *BMC Evolutionary Biology*, 17: 240.
- Kearse, M. et al. (2012) 'Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data', *Bioinformatics*, 28: 1647–9.
- Kosack, L. et al. (2019) 'The ERBB-STAT3 Axis Drives Tasmanian Devil Facial Tumor Disease', *Cancer Cell*, 35: 125–139.e9.
- Kozakiewicz, C. P. et al. (2021) 'Spatial Variation in Gene Expression of Tasmanian Devil Facial Tumors despite Minimal Host Transcriptomic Response to Infection', *BMC Genomics*, 22: 698.
- Langmead, B., and Salzberg, S. L. (2012) 'Fast Gapped-Read Alignment with Bowtie 2', *Nature Methods*, 9: 357–9.
- Laovechprasit, W. et al. (2023) 'Piscichuviral Encephalitis in Marine and Freshwater Chelonians: First Evidence of Jingchuviral Disease', *bioRxiv*, 2023.02.24.528524.
- Larsson, A. (2014) 'AliView: A Fast and Lightweight Alignment Viewer and Editor for Large Datasets', *Bioinformatics*, 30: 3276–8.
- Lazenby, B. T. et al. (2018) 'Density Trends and Demographic Signals Uncover the Long-term Impact of Transmissible Cancer in Tasmanian Devils', *Journal of Applied Ecology*, 55: 1368–79.
- 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, B., and Dewey, C. N. (2011) 'RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome', *BMC Bioinformatics*, 12: 323.
- Luo, Z.-X. et al. (2011) 'A Jurassic Eutherian Mammal and Divergence of Marsupials and Placentals', *Nature*, 476: 442–5.
- Manning, A. (2021) 'NESP Threatened Species Recovery Hub. Selecting Faunal Species for Reintroduction to Mulligans Flat Woodland Sanctuary', ACT, Project 4.1.2 Research Findings Factsheet, pp. 1–3.
- Marcelino, V. R. et al. (2020) 'CCMetagen: Comprehensive and Accurate Identification of Eukaryotes and Prokaryotes in Metagenomic Data', *Genome Biology*, 21: 103.
- Minh, B. Q. et al. (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37: 1530–4.
- Murchison, E. P. et al. (2010) 'The Tasmanian Devil Transcriptome Reveals Schwann Cell Origins of a Clonally Transmissible Cancer', *Science*, 327: 84–7.
- Paddle, R. (2000) *The Last Tasmanian Tiger: The History and Extinction of the Thylacine*, p. 273. Cambridge University Press: New York.
- Peel, E. et al. (2022) 'Genome Assembly of the Numbat (*Myrmecobius fasciatus*), the Only Termitivorous Marsupial', *GigaByte*, 2022: gigabyte47.
- Portas, T. J. et al. (2020) 'Baseline Health and Disease Assessment of Founder Eastern Quolls (*Dasyurus viverrinus*) during a Conservation Translocation to Mainland Australia', *Journal of Wildlife Diseases*, 56: 547–59.
- Porter, A. F. et al. (2020) 'Novel Hepaci- and Pegi-like Viruses in Native Australian Wildlife and Non-human Primates', *Virus Evolution*, 6: veaa064.
- R Core Team (2021) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing <<https://www.R-project.org/>>.
- Royall, A. H. et al. (2019) 'Of Eyes and Embryos: Subfunctionalization of the CRX Homeobox Gene in Mammalian Evolution', *Proceedings of the Royal Society B: Biological Sciences*, 286: 20190830.
- Sciicluna, E. L., Gill, B. P., and Robert, K. A. (2021) 'Fat-tailed Dunnarts (*Sminthopsis crassicaudata*) of the Werribee Grasslands: A Case Study of a Species in Decline', *Australian Journal of Zoology*, 69: 27–32.
- Shi, M. et al. (2018) 'The Evolutionary History of Vertebrate RNA Viruses', *Nature*, 556: 197–202.
- Stalder, K. et al. (2015) 'Prevalence and Clinical Significance of Herpesvirus Infection in Populations of Australian Marsupials', *PLoS One*, 10: e0133807.
- Visser, M. et al. (2016) 'Next-generation Sequencing for Virus Detection: Covering All the Bases', *Virology Journal*, 13: 85.

- Webb, B., Rakibuzzaman, A. G. M., and Ramamoorthy, S. (2020) 'Torque Teno Viruses in Health and Disease', *Virus Research*, 285: 198013.
- Wille, M. et al. (2018) 'A Divergent Hepatitis D-Like Agent in Birds', *Viruses*, 10: 720.
- Woinarski, J. C. Z. et al. (2011) 'The Disappearing Mammal Fauna of Northern Australia: Context, Cause, and Response', *Conservation Letters*, 4: 192–201.
- Woolford, L. et al. (2007) 'A Novel Virus Detected in Papillomas and Carcinomas of the Endangered Western Barred Bandicoot (*Perameles bougainville*) Exhibits Genomic Features of Both the Papillomaviridae and Polyomaviridae', *Journal of Virology*, 81: 13280–90.
- Zemann, A. et al. (2013) 'Ancestry of the Australian Termitivorous Numbat', *Molecular Biology and Evolution*, 30: 1041–5.

1 **Limited transmission of microbial species among coral** 2 **reef fishes from the Great Barrier Reef, Australia**

3

4 Vincenzo A. Costa¹, David R. Bellwood², Jonathon C.O. Mifsud¹, Jemma L. Geoghegan^{3,4},
5 Erin Harvey¹, Edward C. Holmes^{1*}

6

7

8 ¹Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of
9 Sydney, Sydney, NSW 2006, Australia.

10 ²Research Hub for Coral Reef Ecosystem Functions, College of Science and Engineering,
11 James Cook University, Townsville, QLD 4811, Australia.

12 ³Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand.

13 ⁴Institute of Environmental Science and Research, Kenepuru, Porirua, 5022, New Zealand.

14

15 *Corresponding author:

16 Prof. Edward C. Holmes,

17 Sydney Institute for Infectious Diseases, School of Medical Sciences,

18 The University of Sydney, Sydney, NSW 2006, Australia.

19 Email: edward.holmes@sydney.edu.au

20 **Abstract**

21 Reef fishes account for one-third of all extant marine fishes and exhibit enormous
 22 biodiversity within a highly interactive ecosystem. Yet relatively little is known about the
 23 diversity and evolution of microbial species (bacteria, viruses, and eukaryotes) associated
 24 with reef fish, even though this may provide valuable insights into the factors that shape
 25 microbial communities within vertebrate hosts as well as the extent and pattern of cross-
 26 species transmission. Through metatranscriptomic sequencing we characterised the viruses,
 27 bacteria, and single-celled eukaryotes from 128 reef fish species inhabiting Lizard Island and
 28 Orpheus Island on the Great Barrier Reef, Australia. We assessed whether microbial
 29 communities differed between islands that are separated by approximately 450 kilometres,
 30 and to what extent viruses were able to emerge in new hosts. Notably, despite strong
 31 ecological interactions in the reef environment, and the presence of the same families and
 32 subfamilies of viruses and bacteria on both islands, there was minimal evidence for the cross-
 33 species transmission of individual microorganisms among fish species. An exception was the
 34 high prevalence of the bacterial pathogen *Photobacterium damsela* among apparently
 35 healthy cardinalfishes from both islands, indicating that these fish species are natural
 36 reservoirs within the reef system. Overall, these data suggest that reef fishes have microbial-
 37 host associations that arose prior to the formation of the Great Barrier Reef, leading to strong
 38 host barriers to cross-species microbial transmission even within a highly interactive and
 39 species-rich environment.

40 **Introduction**

41 Symbiotic interactions are ubiquitous in nature and play an important role in animal, plant,
 42 and microbial evolution. In reef systems, the interdependence of corals, fishes, and their
 43 microbial symbionts supports fundamental ecological processes, and their dissociation may
 44 cause devastating declines in species abundance and reef functioning (1). Healthy tropical
 45 coral reefs are characterised by remarkable biodiversity and exceptionally complex
 46 interactions. This provides an ideal forum for investigating the evolutionary and ecological
 47 factors shaping microbial communities, especially among teleost fishes. Teleost fishes are the
 48 most speciose group of vertebrates on coral reefs and rank among the most phylogenetically
 49 and ecologically diverse group of vertebrates, accounting for one-third of all currently
 50 described marine fishes (2, 3). Reef fishes exhibit exceptional dispersal capabilities. Most
 51 have geographic ranges spanning thousands of kilometres, with some species spanning
 52 approximately two thirds of the global tropics (4). This level of interconnectivity is also
 53 exhibited on individual reefs, with highly complex food webs. Reef fishes display diverse
 54 trophic guilds (e.g., carnivores, mobile invertivores, omnivores, planktivores, sessile
 55 invertivores, herbivores/detritivores), occur in a range of habitats (e.g., coral, sand, rubble,
 56 caves), and commonly live in exceptionally close proximity (5-7).

57 This high degree of interaction is exemplified in the cryptobenthic reef fishes: behaviourally
 58 cryptic species with adult body sizes of approximately 5 cm or less that typically occupy the
 59 benthic zone (6). Cryptobenthic reef fishes engage with larger reef fishes through extensive
 60 predator-prey interactions, and their frequent consumption is an integral component of coral
 61 reef food webs via the transfer of energy from microscopic prey to large predators (8). For
 62 example, the dwarf goby (*Eviota sigillata*) has a maximum lifespan of just 59 days and
 63 experiences mortality rates of 7.8% per day (9). In addition to predation, fishes interact
 64 through cleaning (including the removal of blood-sucking parasites) (10) and extensive food

65 webs of coprophagy (consuming the faeces of other fishes) (11). The potential for microbial
66 transmission on reefs is therefore considerable.

67 Although reef fish co-exist in a highly diverse and interactive ecosystem, relatively little is
68 known about whether and how microbial composition (i.e., of viruses, bacteria, eukaryotes)
69 differs across fish groups within and among communities. Studies of microbial ecology in
70 teleosts have largely focused on animals utilised in aquaculture and in laboratory model
71 species, with very few investigations of wild ecosystems (12-15). Fish-microbe interactions
72 are highly beneficial for fish nutrition and immunity and are shaped by host and
73 environmental factors including trophic level, age, water quality, and host phylogeny (12, 16-
74 20).

75 The extent of phylogenetic divergence between animal species has an important impact on
76 virus ecology and evolution, particularly the likelihood of successful cross-species virus
77 transmission, and it is therefore a key determinant of infectious disease emergence (21-25).

78 The “phylogenetic distance” theory posits that microorganisms are more likely to be
79 transmitted between closely related species that have conserved cellular properties, such as
80 cell receptors (26). This idea is supported by numerous studies across a broad spectrum of
81 host taxa (e.g., vertebrates, invertebrates) and pathogen groups including virus-host
82 interactions in reef fishes (16, 22). For example, recent work has shown that reef fishes from
83 a spatially restricted (100 square metre) community from Orpheus Island in the Great Barrier
84 Reef (GBR), Australia, harbour diverse viral assemblages that are highly host-specific despite
85 ample opportunity for cross-species transmission (16).

86 With approximately 2500 coral reefs and 900 islands, fishes of the GBR constitute a natural
87 model system to investigate spatial patterns of microbial evolution and diversity in vertebrate
88 hosts. Molecular and fossil evidence indicates that the majority of reef fish families
89 originated during the Paleocene and Eocene, approximately 66 to 50 million years ago (Ma).

90 Subsequently, a notable acceleration in lineage diversification took place during the
91 Oligocene and Miocene (34–5.3 Ma), with this shift occurring in the Indo-Australia
92 Archipelago (IAA) (i.e., the IAA biodiversity hotspot) (27). During the Miocene (23–5.3 Ma),
93 the reciprocal diversification of fish and coral species likely led to the development of the
94 functional reef ecosystems that are observable today. By the early stages of the Pleistocene
95 (5.3–0 Ma), almost all reef fish taxonomic groups were established and began their settlement
96 on reefs across all tropical oceans (27), with the formation of the GBR fish communities
97 likely occurring within the last 10,000 years following the stabilization of sea level to its
98 current height ~6000–8000 years ago (5, 28). Whether and how these colonization events
99 have shaped microbial evolution and diversity is unknown. For example, many reef fishes—
100 particularly cryptobenthics that have limited dispersal—exhibit strong site fidelity,
101 maintaining the same community composition year-round (29). The seemingly consistent
102 community composition might therefore lead to the generation of distinct microbial
103 communities in different geographic areas, that will be most pronounced for rapidly evolving
104 RNA viruses. Conversely, it is possible that the high ecological similarities among disjunct
105 reef locations might result in broadly similar microbial compositions among fish
106 communities.

107 We characterised the total assemblage of viruses, bacteria, and single-celled eukaryotes from
108 128 reef fish species spread across Lizard Island and Orpheus Island in the GBR. This study
109 comprised 28 reef fish families, making it one of the largest investigations of microbial
110 diversity and evolution in reef fish undertaken to date. Using metatranscriptomics, we aimed
111 to determine the relationship between host community diversity and microbial diversity and
112 identify whether these communities differ between fish species from two islands separated by
113 approximately 450 kilometres. We also aimed to determine the impact of host ecology on
114 microbial diversity and evolution and to identify whether particular fish groups are potential

115 reservoirs for important viral or bacterial pathogens. This is of particular importance given
116 the high utilisation of reef fish in aquaculture as well as the ongoing threats of biodiversity
117 loss on coral reefs from anthropogenic climate change, itself a significant contributor to the
118 emergence of infectious diseases (30).

119

120 **Results**

121 **Composition of sequence reads**

122 We sequenced a total of 10.7 billion RNA reads, including 4.7 billion reads newly generated
123 from Lizard Island fishes. The remaining reads (Orpheus Island) are available on NCBI
124 Sequence Read Archive (SRA) under BioProject PRJNA841039 (16). These data were
125 generated from a total of 140 sequencing libraries, representing 128 reef fish species and 28
126 families (Figure 1) (mean 54,617,146 reads per library). Fish RNA accounted for 93% of the
127 total reads, followed by RNA associated with cnidarians (5.7%), bacteria (0.31%), single-
128 celled eukaryotes (0.29%), arthropods (0.23%), platyhelminths (0.18%), molluscs (0.04%),
129 poriferans (0.03%), annelids (0.02%), nematodes (0.01%), fungi (0.009%), and viruses
130 (0.009%) (Supplementary Table 1).

131

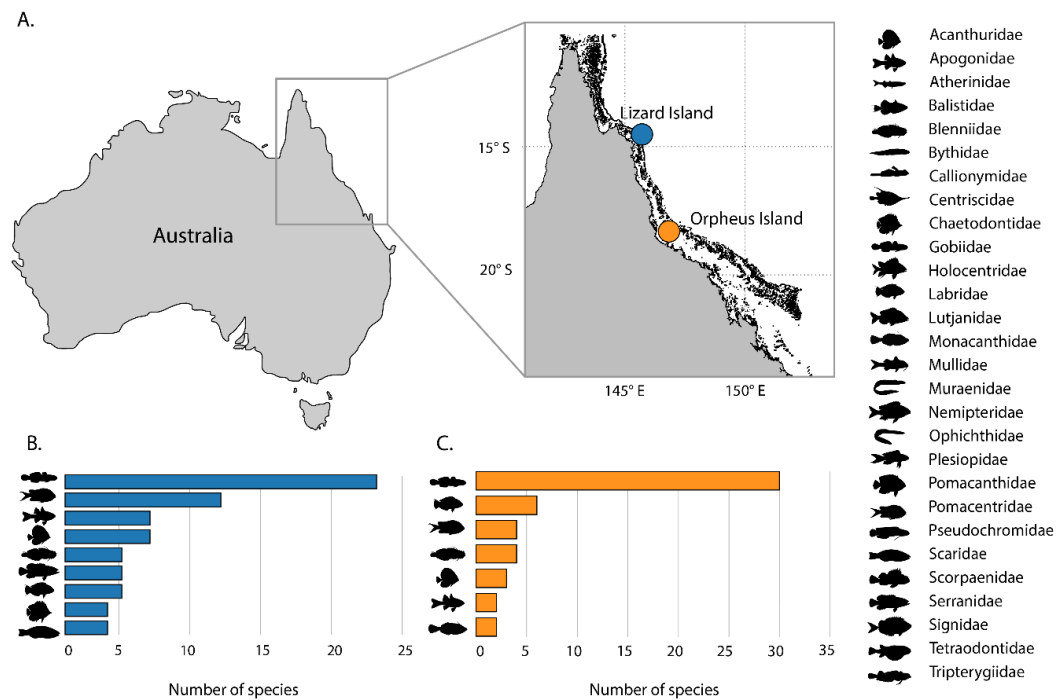


Figure 1. Sampling locations and taxonomic diversity of reef fish. (A) Location of Lizard Island and Orpheus Island in the GBR. (B) Taxonomy of samples collected from Lizard Island. (C) Taxonomy of samples collected from Orpheus Island. All other families with less than one species sampled are omitted from panels B and C.

Diversity and abundance of the reef fish virome

We identified sequences representing 64 vertebrate-associated viruses (i.e., those likely infecting fish tissues), including 27 newly discovered from Lizard Island fishes (Supplementary Table 2). The *Astroviridae* comprised 32.4% of vertebrate-associated viral reads, followed by the *Iridoviridae* (21.8%), *Picornaviridae* (18.6%), *Chuviridae* (15.9%), *Parvoviridae* (6.8%), *Hantaviridae* (2.9%) with all other groups representing <1% of the total viral reads: *Flaviviridae*, *Orthomyxoviridae* (order *Articulavirales*), *Poxviridae*,

145 *Paramyxoviridae*, *Reoviridae*, *Circoviridae*, *Coronaviridae*, *Hepeviridae*, *Rhabdoviridae*,
146 and *Caliciviridae* (Figure 2).

147 As well as vertebrate-associated viruses, we identified 194 viruses that were likely infecting
148 porifera, arthropods, molluscs, fungi, plants, and microbial eukaryotes (e.g., dinoflagellates),
149 including 100 that were newly discovered at Lizard Island. Because these viruses were likely
150 associated with fish diet and not infecting the fish themselves, we assume that their presence
151 does not reflect key aspects of fish biology (i.e., immunity or receptor binding) and hence
152 effectively serve as a “negative control” in comparison to the vertebrate-associated viruses.
153 We refer to this group as “non-vertebrate” viruses. The most abundant viral groups in this
154 category were the *Flaviviridae* (24% of non-vertebrate viral reads), unclassified
155 picornaviruses (i.e. “picorna-like” viruses) (22.5%) *Narnaviridae* (21.9%), *Nodaviridae*
156 (8.9%), *Hepeviridae* (6.2%), *Partitiviridae* (4%), *Solemoviridae* (3.2%), *Negevirus* (2.1%)
157 and *Totiviridae* (2%), with all other groups comprising <1%: *Reovirales*, *Weivirus*,
158 *Qinviridae*, *Iflaviridae*, *Dicistroviridae*, *Rhabdoviridae*, *Picobirnaviridae*, *Quenyavirus*,
159 *Bunyavirales*, *Chuviridae*, and *Tombusviridae*.

160 **Spatial comparisons of the reef fish virome**

161 To determine whether viral composition (i.e., virus families or subfamilies) differed between
162 Orpheus and Lizard islands, we analysed beta diversity using permutational multivariate
163 analysis of variance (PERMANOVA) with the Bray–Curtis dissimilarity matrix. This
164 revealed no significant difference in vertebrate-associated viral communities on both islands
165 ($F = 1.42$, $p = 0.06$), with overlapping viromes at the viral family/subfamily level (Figure 2b).
166 Accordingly, both islands contained viruses assigned to the *Picornaviridae*, *Astroviridae*,
167 *Parvoviridae*, *Hantaviridae*, *Orthomyxoviridae*, *Coronaviridae*, *Hepeviridae* and *Iridoviridae*
168 (Figure 2d-e). A similar pattern was observed when analysing the non-vertebrate virome ($F =$

1.53, $p = 0.07$) (Figure 2c) with both islands dominated by unclassified picornaviruses, *Narnaviridae*, *Totiviridae*, *Partitiviridae*, and *Nodaviridae*.

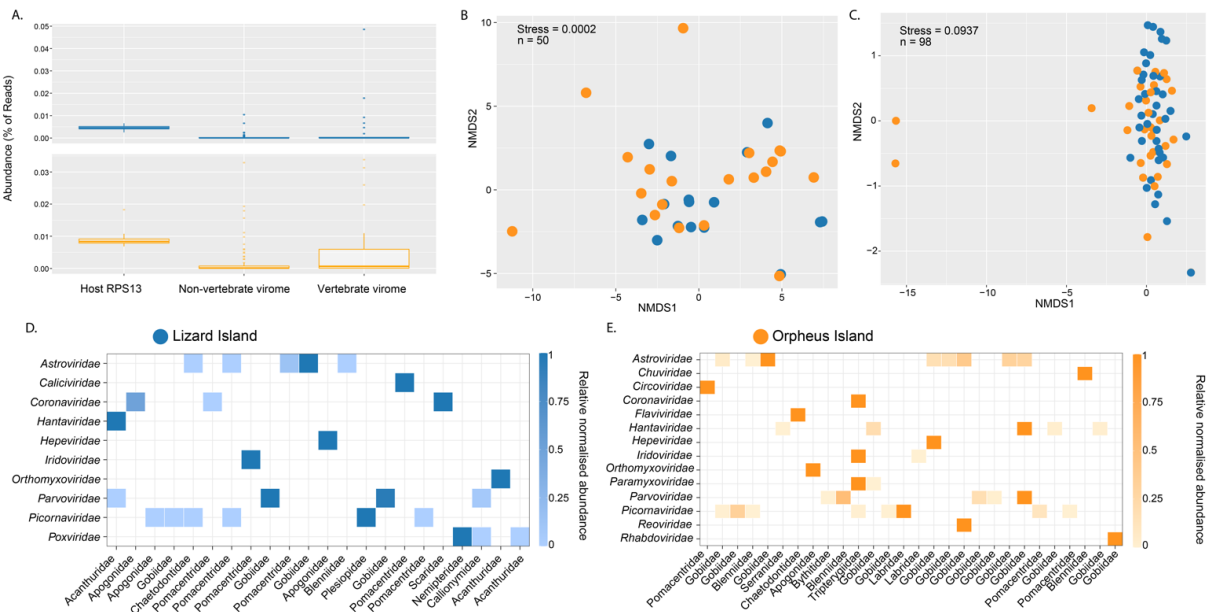


Figure 2. (A) Abundance of host gene marker (RPS13), non-vertebrate virome and vertebrate virome. (B) Non-metric multi-dimensional scaling (NMDS) plot (Bray-Curtis dissimilarity matrix) for vertebrate-associated viruses for each fish species. (C) NMDS plot for non-vertebrate-associated viruses for each fish species. (D-E) Relative normalised abundance of vertebrate-associated viral families from each island.

Evolutionary history and biogeographical patterns of vertebrate-associated viruses

While our analysis revealed similarities at the level of virus family/subfamily, almost all of the viruses identified exhibited levels of genetic divergence that reflected long-term virus-host associations, rather than recent cross-species transmission within each reef ecosystem. The only instance of the same virus being shared between fish species was the presence of

184 highly similar astroviruses (~96 similarity across the entire genome) in gobies from Orpheus
185 Island (see ref. 16 for full description of these viruses). No viruses were shared among the
186 fishes sampled from Lizard Island.

187 We now describe the phylogenetic relationships of the shared viral groups in turn. We focus
188 primarily on the relationships of viruses between both islands as well as those newly
189 discovered at Lizard Island.

190 ***Positive-sense single-stranded RNA viruses (+ssRNA): Picornaviridae, Astroviridae,***
191 ***Hepeviridae, Caliciviridae and, Coronaviridae***

192 The *Picornaviridae* were the most common viral group in our data set with 14 viruses: six
193 from Lizard Island and eight from Orpheus Island. We discovered a group of four novel
194 viruses that formed a distinct clade with the newly formed genus *Danipivirus*, represented by
195 a single virus that is commonly detected in model zebrafishes (31) (Figure 2). In this group it
196 was notable that we identified two relatively closely related viruses (70.6% sequence
197 similarity across the entire polyprotein) in damselfishes (Pomacentridae) from both islands. A
198 more distantly related virus was identified in *Chaetodon baronessa* (Chaetodontidae; 48-49%
199 RdRp similarity with both damselfish picornaviruses) suggesting that these viruses
200 diversified within reef fish, although on an unknown time scale. Evidence of reef
201 diversification was also identified in *Pomacentrus nagasakiensis* picornavirus and *Blenniella*
202 picornavirus from Orpheus Island that exhibited 66.1% similarity and were related to
203 fipiviruses found only at Lizard Island (Figure 3).

204 We identified 12 astroviruses, including five novel viruses from Lizard Island. These were
205 spread across three major clades: clade I, represented exclusively by fish viruses including
206 five goby viruses from Orpheus Island (see above); clade II, similarly represented by fish
207 viruses including those from both Lizard and Orpheus Island; and clade III that fell sister to

the genera *Mamastrovirus*, found in mammals and *Avastrovirus*, exclusively in birds (Figure 3). Notably, both *Plagiotremus tapeinosoma* astrovirus (Lizard Island) and *Blenniella* astrovirus (Orpheus Island) fell within clade III, exhibiting 75.8% sequence similarity in the RdRp gene. Similarly, both *Eviota* astrovirus and *Chaetodon baronessa* astrovirus clustered together within clade II, while the other reef fish viruses were more divergent.

Among other +ssRNA viruses, we identified two hepeviruses – *Istigobius decoratus* hepevirus (Orpheus Island) and *Fowleria viaulae* hepevirus (Lizard Island) – that grouped with other fish hepeviruses in phylogenetic trees, as well as a novel calicivirus in *Pomacentrus brachialis* that similarly fell within a broad group of fish and amphibian caliciviruses (Figure 3).

Both letoviruses (*Coronaviridae*, subfamily *Letovirinae*) —*Chromis atripectoralis* letovirus and *Scarus psittacus* letovirus—from Lizard Island formed a basal clade with *Microhyla* letovirus and three other viruses identified in African cichlids (32, 33). Together, this group likely formed a novel genus within the *Letovirinae* that includes diverse ectothermic hosts such as amphibians, jawless and ray-finned fishes (32).

Negative-sense single-stranded RNA viruses (-ssRNA): Hantaviridae and Articulavirales

A notable observation from our previous sampling of reef fish was the detection of hantaviruses in four different goby species at Orpheus Island. In contrast, at Lizard Island we only detected one hantavirus in a surgeonfish (*Acanthurus nigrofusus*) that was sister to Wenling red spikefish hantavirus (NCBI/GenBank accession: AVM87662.1) and was highly divergent to those found at Orpheus Island (~29% RdRp similarity). Overall, these viruses fell within the subfamily *Actantavirinae*, that exclusively infects ray-finned fishes. Similarly, we detected two viruses with high levels of divergence (40% similarity) from both islands that were related to quaranjaviruses (order *Articulavirales*) (Figure 3).

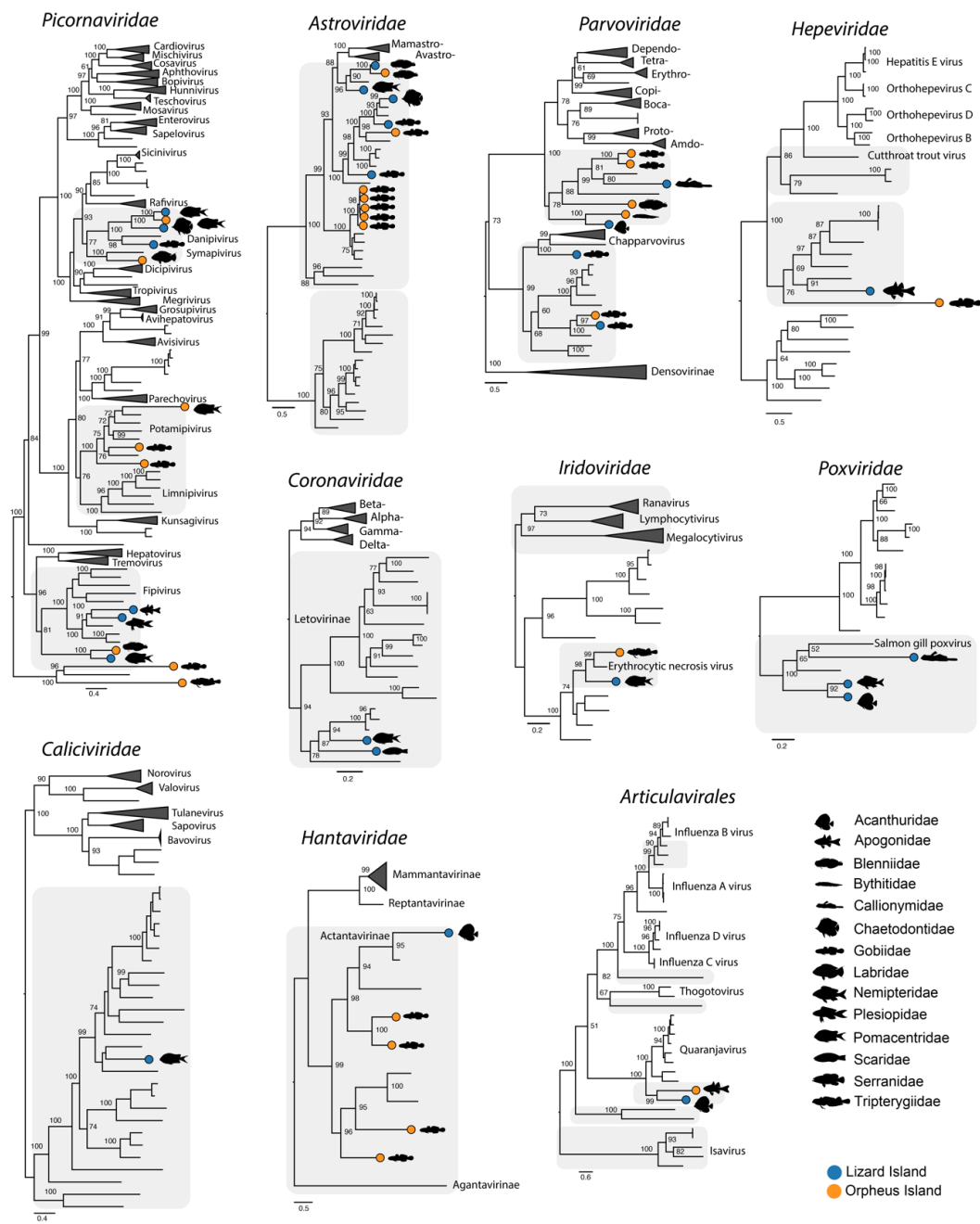
232 **DNA viruses**

233 We identified reef fish from three lineages within the *Parvoviridae*: the *Parvovirinae*,
 234 *Chapparvovirus*, and *Ichthamaparvovirus* groups. Reef fish *Parvovirinae* fell as a sister
 235 lineage to mammalian and avian viruses (e.g. *Dependoparvovirus*, *Aveparvovirus*), again
 236 with high sequence divergence between islands. For example, the closest relatives among
 237 both islands were *Acanthurus nigrofusus* parvovirus (Lizard Island) and *Dinematichthys*
 238 parvovirus (Orpheus Island) that shared 62% NS1 gene sequence similarity (Figure 3).
 239 Notably, we identified a novel chapparvovirus in *Cryptocentrus strigilliceps* that was related
 240 to tilapia parvovirus—a pathogen in farmed Tilapia in China—making it the second fish virus
 241 identified in this group (34). Among the genus *Ichthamaparovirus*, both *pleurosicya*
 242 *ichthamaparvovirus* (Lizard Island) and *Luposicya lupus ichthamaparvovirus* (Orpheus Island)
 243 shared a common ancestor (57.9% NSI similarity) and formed a distinct clade with other fish-
 244 infecting parvoviruses (16, 33, 35).

245 Similar patterns were observed in reef fish *Betairidovirinae* (*Iridoviridae*), between both
 246 *Chrysiptera rollandi* iridovirus (Lizard Island) and *Enneapterygius tutuilae* iridovirus
 247 (Orpheus Island). These viruses exhibited 82.2% similarity in the conserved major capsid
 248 protein (MCP) and clustered with erythrocytic necrosis virus (ENV).

249 Poxviruses were only identified at Lizard Island, with phylogenetic analysis placing them
 250 with other fish viruses—Salmon gill pox virus and carp edema virus—within the subfamily
 251 *Chordopoxvirinae*. Notably, *Zebrasoma veliferum* poxvirus and *Scolopsis bilineata* poxvirus
 252 shared a common ancestor (80.3% similarity), strongly suggestive of an origin in reef fish
 253 (Figure 3).

254



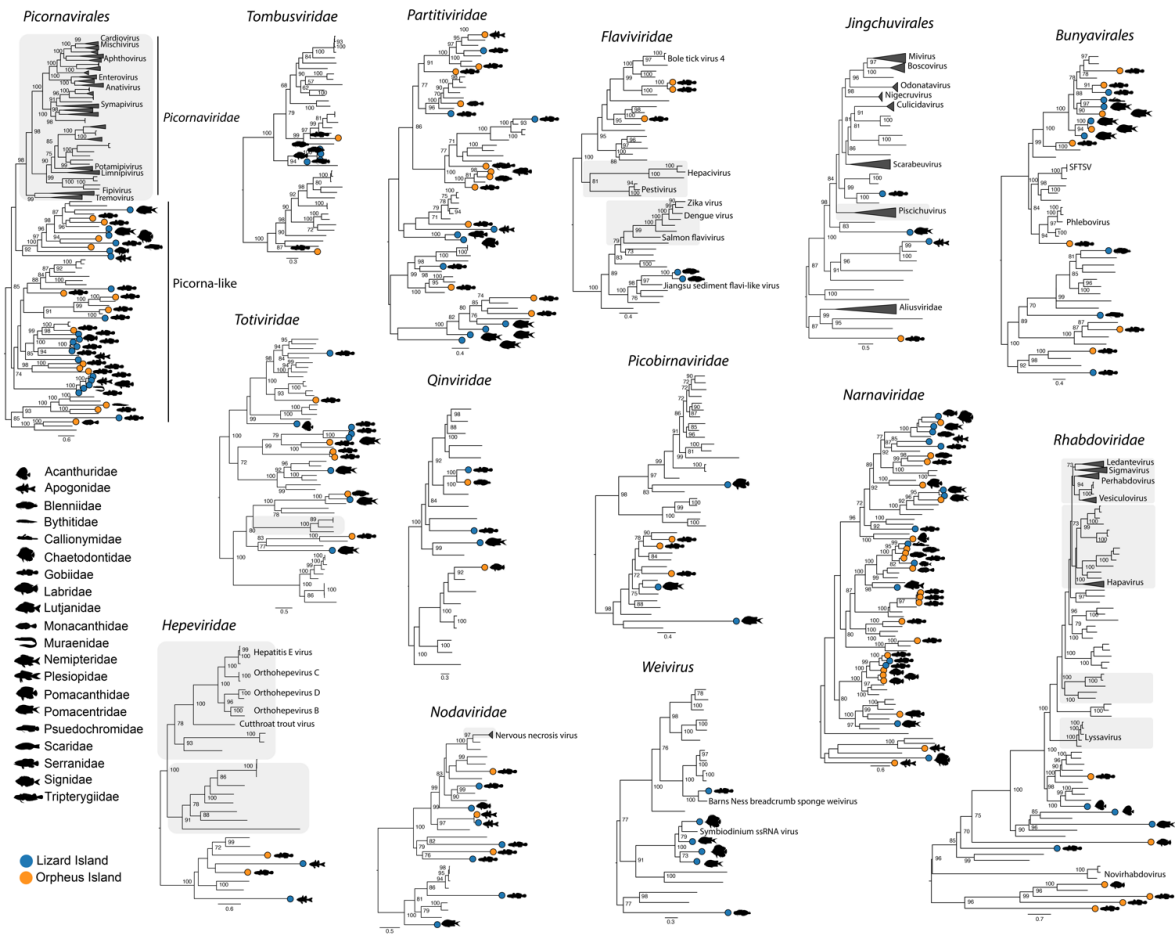
255

256 **Figure 3.** Phylogenetic analysis reveals relationships among reef fish viruses from both
257 islands. Phylogenies were estimated using the RdRp gene for RNA viruses (*Picornaviridae*,
258 *Astroviridae*, *Hepeviridae*, *Hantaviridae*, *Caliciviridae*, *Coronaviridae*, *Articulavirales*), NS1
259 gene for parvoviruses and DNA polymerase for iridoviruses and poxviruses. Coloured circles
260 on branch tips represent viruses identified in this study. The scale bar represents the number

261 of amino acid substitutions per site. Shaded branches represent fish viruses. Trees were
262 midpoint rooted for clarity only.

263 **Close phylogenetic relationships of non-vertebrate-associated viruses**

264 A common observation in our data set was that genetically diverse reef fish exhibited very
265 similar assemblages of non-vertebrate-associated viruses (Figure 2). This pattern sits in
266 marked contrast to the vertebrate-associated viruses, that were rarely transmitted among
267 species. In particular, we identified the same virus (i.e., 98-99% similarity) in multiple host
268 species within the following groups: unclassified picornaviruses, *Tombusviridae*, *Totiviridae*,
269 *Nodaviridae*, *Narnaviridae*, *Bunyavirales* and (Figure 4). As expected, most of these cases of
270 virus sharing occurred within each island. For example, we identified the same tombusvirus
271 in two blennies and one wrasse from Lizard Island as well as the same narnavirus in three
272 gobies from Orpheus Island (Figure 4). A notable exception was the presence of two
273 nodaviruses identified in *Fowleria vaiulae* from both islands that exhibited 98% similarity in
274 the RdRp gene. While nodaviruses are capable of infecting fish species (e.g. nervous necrosis
275 virus) we detected reads associated with decapods in both libraries. These reads, in
276 combination with the phylogenetic positions of these viruses (i.e., highly divergent from
277 nervous necrosis virus and more related to crustacean viruses) strongly implies that these
278 viruses are of dietary origin.



279

280 **Figure 4.** Close phylogenetic relationships among non-vertebrate-associated viruses sampled
281 from reef fish. Phylogenies were estimated using amino acid sequences of the RdRp gene.
282 Coloured circles on branch tips represent viruses identified in this study. The scale bar
283 represents the number of amino acid substitutions per site. Shaded branches represent
284 vertebrate-associated viruses. Trees were midpoint rooted for clarity only.

285

286 Diversity and abundance of bacteria in reef fish assemblages

287 After the removal of cyanobacteria, the phylum *Proteobacteria* accounted for 73.5% of the
288 total sequence reads from bacteria, followed by *Firmicutes* (9%), *Actinobacteria* (8%),
289 *Bacteroidetes* (2.9%), *Spirochaetes* (2.3%), *Fusobacteria* (1.7%), with all other phyla each

290 representing <1% (Figure 5a). At the family level, the *Vibrionaceae* (42.7%) and
 291 *Enterobacteriaceae* (11.6%) were present at the highest frequencies followed by the
 292 *Endozoicomonadaceae* (5.9%), *Comamonadaceae* (5.9%), *Shewanellaceae* (5.8%),
 293 *Propionibacteriaceae* (5%), *Clostridiaceae* (4.9%), *Micrococcaceae* (2.3%),
 294 *Fusobacteriaceae* (2%), *Pseudomonadaceae* (1.3%), and *Mycoplasmataceae* (1.3%) (Figure
 295 5).

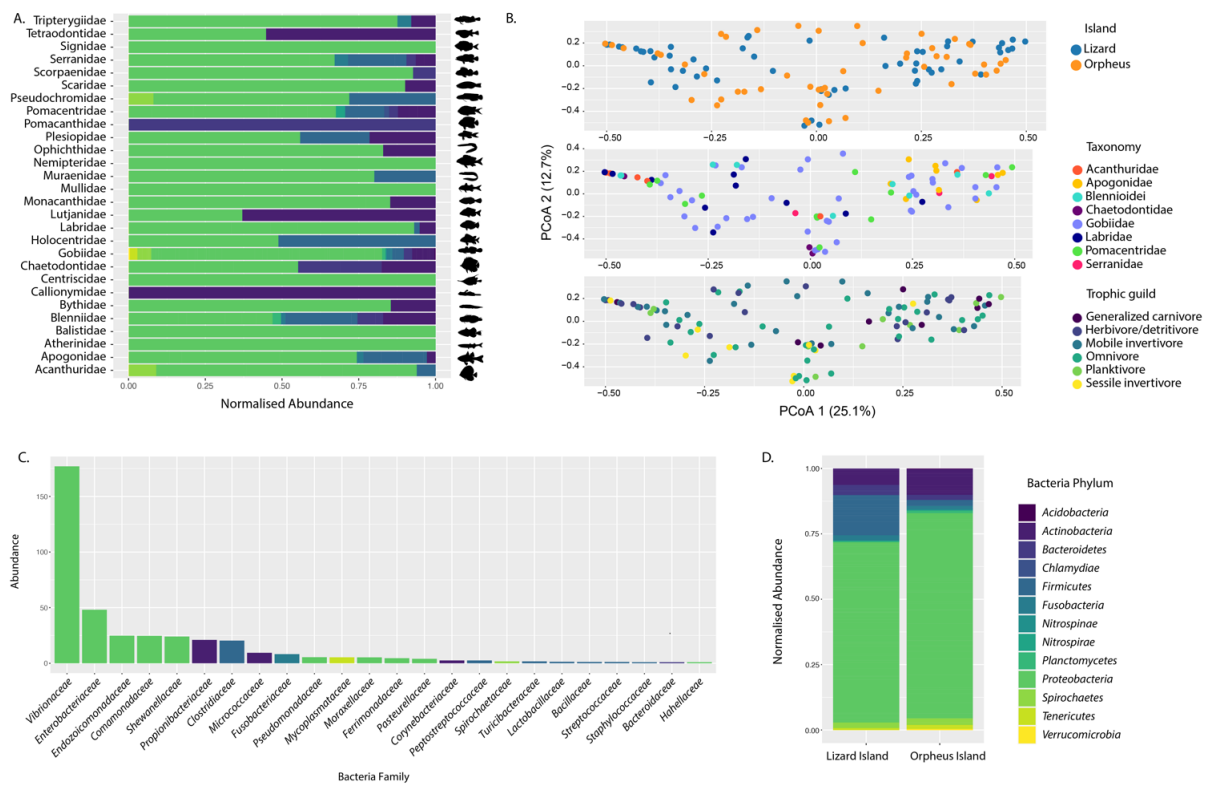
296 To explore patterns of bacterial diversity between both islands and among reef fish groups,
 297 we performed a principal coordinate analysis (PCoA). This revealed no partitioning
 298 according to reef location with largely overlapping bacterial communities between the two
 299 islands ($F = 1.711$, $R^2 = 0.015$, $p = 0.082$) (Figure 6). PERMANOVA revealed significant
 300 differences in bacterial composition between fish taxonomic groups
 301 ($F = 2.42$, $R^2 = 0.153$, $p = 0.001$). However, there was overlap among all fish species,
 302 particularly the gobies, which may be explained by their high involvement in coral reef food
 303 webs (5, 6) (Figure 5b).

304 Despite the high degree of overlap in bacterial families among the reef fishes, there was
 305 limited evidence for the sharing of individual bacterial species between fish species. Notably,
 306 however, we identified *Photobacterium damsela*—an important pathogen in aquaculture—
 307 in 88% of the cardinalfish species examined from both islands (Figure 6). We also detected
 308 other potentially pathogenic *Vibrionaceae*, including four that were related to those within
 309 the *Vibrio harveyi* clade: *V. parahaemolyticus*, *V. campbellii*, and *V. owensii*. (Supplementary
 310 Figure 1). In addition, we identified *V. fortis*—an opportunistic pathogen of coral—in the
 311 surgeonfish, *Ctenochaetus binotatus* (36) (Supplementary Figure 1).

312 To assess the impact of fish ecology on bacterial composition, we grouped fish species into
 313 six trophic guilds: carnivores, mobile invertivores, omnivores, planktivores, sessile
 314 invertivores, herbivores/detritivores (7). While we detected significant differences in bacterial

composition between these groups ($F = 1.485$, $R^2 = 0.074$, $p = 0.035$), there was similarly high overlap with no clear partitioning according to host trophic guild (Figure 5b).

317



318

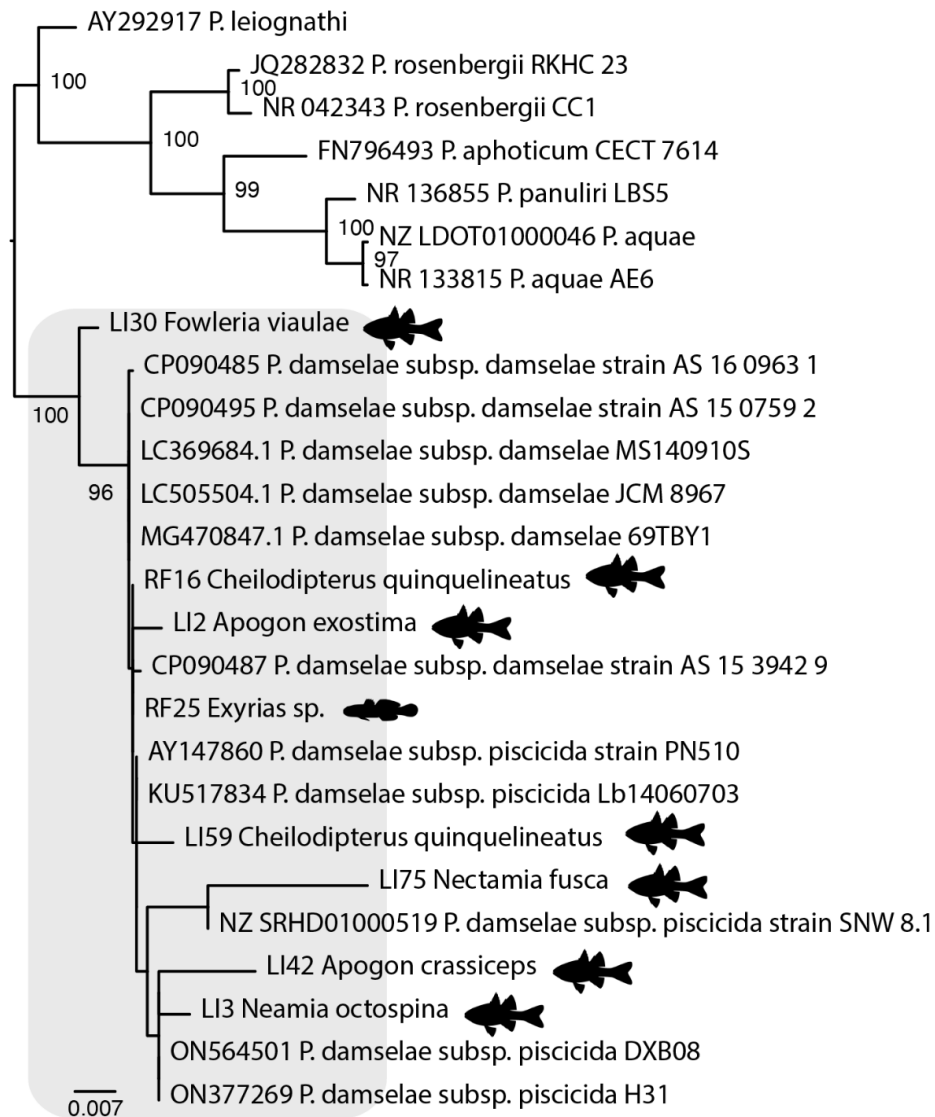
319 **Figure 5.** (A) Normalised abundance of bacterial phyla for each reef fish family. (B)

320 Principal coordinate analysis (PCoA) plots of bacterial communities for island, host

321 taxonomy and trophic guild using Bray-Curtis dissimilarity matrix. (C) Abundance of

322 bacterial families coloured by phylum. (D) Normalised abundance of bacterial phyla for each

323 island.



324

325

326 **Figure 6.** Maximum likelihood phylogeny of the genus *Photobacterium*, estimated using
327 nucleotide sequences of the 16S gene. Fish silhouettes represent individuals identified in this
328 study. The scale bar represents the number of nucleotide substitutions per site. Tree was
329 midpoint rooted for clarity only.

330

331 **Composition of single-celled eukaryotes**

332 Finally, we identified transcripts representing single-celled eukaryotes from 15 phyla.
 333 Dinoflagellates were the most abundant group (46% of the total eukaryotic reads) followed
 334 by Bacillariophyta (21.6%), Foraminifera (12.6%), Cercozoa (5.8%), Euglenozoa (2.9%),
 335 Apicomplexa (2.7%), Ciliophora (2.5%), Endomyxa (1.1%), Parabasalia, Haptista, Fornicata,
 336 and Heterolobosea (all less than one percent). Fungi—Ascomycota, Basidiomycota,
 337 Microsporidia—were identified at much lower frequencies, representing only 3.3% of the
 338 total reads in this category.

339 Using these data, we performed a PCoA based on fish location, taxonomy and ecology. As
 340 with the analysis of virome composition, this revealed overlapping microbial communities
 341 between both islands ($F = 1.208$, $R^2 = 0.015$, $p = 0.181$). However, there were significant
 342 differences in microbial communities between fish taxonomic groups
 343 ($F = 1.384$, $R^2 = 0.119$, $p = 0.002$), which may be driven by the separation of pomacentrids
 344 from other groups, in turn reflecting the larger diversity of microorganisms identified in this
 345 group (Figure 7a-b). When assessing the impact of host ecology, we similarly identified
 346 substantial overlap with no clear partitioning (Figure 7). Apart from *Symbiodinium* spp.,
 347 which form symbiotic relationships with coral rather than fish and hence are dietary-
 348 associated, we found no evidence for the same microbial species present in multiple fishes.

349 Among the Apicomplexa, which are common parasites of vertebrates, we detected reads
 350 associated with *Goussia* spp. in *Chaetodon aureofasciatus*, *Halichores melanurus*, and *Eviota*
 351 *melasma*. Overall, the parasitic families, such as the Eimeriidae (Apicomplexa),
 352 Trypanosomatidae, and Ichthyobodonidae (both Euglenozoa) were found in eight libraries at
 353 low transcript abundances, representing an average of 0.19% of the reads in each library.
 354

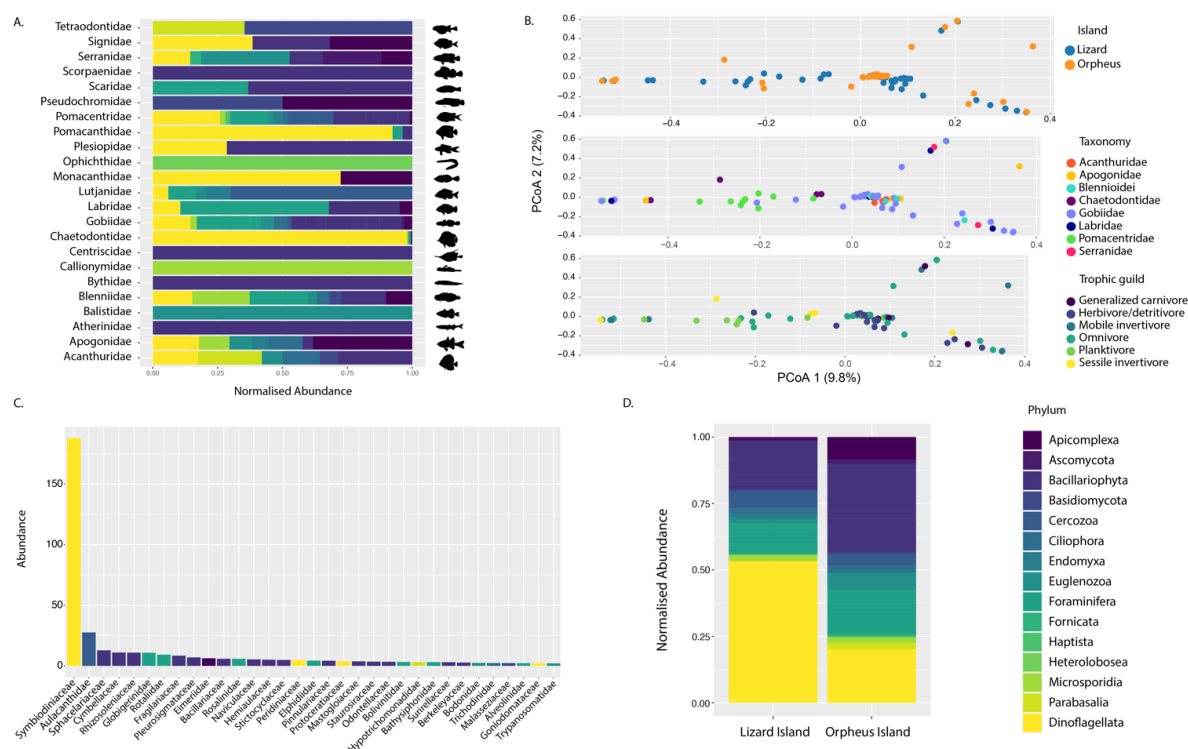


Figure 7. (A) Normalised abundance of single-celled eukaryotic phyla for each reef fish family. (B) Principal coordinate analysis (PCoA) plots of single-celled eukaryotic communities for island, host taxonomy and trophic guild using Bray-Curtis dissimilarity matrix. (C) Abundance of single-celled eukaryotic families coloured by phylum. (D) Normalised abundance of single-celled eukaryotic phyla for each island.

Discussion

Our metatranscriptomic analysis of 128 reef fish species revealed no significant differences in the composition of viral and microbial families/subfamilies between two islands located ~450 kilometres apart in the Australian Great Barrier Reef. Fish sampled from islands were associated with the presence of the *Picornaviridae*, *Astroviridae*, *Parvoviridae*, *Hantaviridae*, *Orthomyxoviridae*, *Coronaviridae*, *Hepeviridae* and *Iridoviridae*, as well as *Proteobacteria*

368 of the *Vibrionaceae*, *Enterobacteriaceae*, and *Endozoicomonadaceae*. Also of note was that
369 within each group of viruses we observed high levels of genetic diversity, with minimal
370 evidence for the same virus being shared among fish species (both within and between
371 islands), despite strong ecological interactions in the reef ecosystem and our spatially
372 restricted sampling.

373 These findings offer further support for recent studies of virus ecology in fish that show that
374 host phylogeny has a strong influence on virome composition (16, 33). For example, closely
375 related African cichlid species, that have evolved *in situ* within Lake Tanganyika over the last
376 10 million years, exhibit highly similar viromes with high levels of cross-species
377 transmission and viral generalism (33). African cichlids are members of a single family, the
378 Cichlidae, and exhibit some of the lowest pairwise genetic distances observed between
379 vertebrates (e.g., differences of 0.03% between some species). In contrast, reef fish
380 communities are composed of several divergent families, including 28 examined in this study
381 (7, 37), many of which were established around 66 Ma, occupying the biogeographic region –
382 Tethys – now covered by Europe and the Mediterranean Sea (27). During the Oligocene there
383 was a shift in species richness from Tethys to the IAA, eventually forming the IAA
384 biodiversity hotspot (27, 38). Most reef fish genera formed within the IAA around 30-15 Ma,
385 with accelerated speciation during the Miocene (27).

386 A notable difference between cichlids and reef fishes, which may explain their strikingly
387 contrasting levels of cross-species transmission, is that cichlids rapidly evolved within Lake
388 Tanganyika, while reef fishes diversified within the IAA, over many millions of years, prior
389 to their settlement on the GBR (27, 39). As the GBR that exists today formed after
390 Pleistocene sea level rises, reef fish communities only became established during the last
391 ~10,000 years, such that the genetic boundaries inhibiting cross-species transmission were
392 already established in the IAA before their settlement at reef locations. In marked contrast,

the adaptive radiation of the African cichlids would have provided a more favourable environment for cross-species transmission as their diversification occurred rapidly within the confinements of Lake Tanganyika, generating a large pool of closely related host species for infection. Indeed, a time-calibrated phylogeny of cichlid hepaciviruses showed that elevated rates of virus diversification coincided with a period of rapid cichlid speciation (33). These ecological and evolutionary patterns were similarly observed across the bacteriome, with *P. damsela* the only bacterium that was shared among multiple fish species. Indeed, *P. damsela* was primarily identified in cardinalfishes (Figure 6), further illustrating a phylogenetic effect. However, it is likely that this association represents a long-term symbiotic relationship between cardinalfishes and *P. damsela*, particularly as it was detected in *Cheilodipterus quinquelineatus* from both islands, as well as its presence in 88% of the cardinalfish libraries examined. This strongly suggests that *P. damsela* forms part of the natural microbiome in these species. It is noteworthy that most cardinalfish species are nocturnal on coral reefs, and the genus *Photobacterium* is renowned for its bioluminescence—via the expression of *lux* genes—including some strains of *P. damsela* (40). Indeed, *Photobacterium mandapamensis* is a bioluminescent symbiont of the urchin cardinalfish, *Siphamia tubifer*, where it provides light to attract prey (40, 41). Several cardinalfish species have evolved specialized “light organs” that harbour bioluminescent *Photobacterium*; however, these specialized organs are not present in all species (42). It is important to note that we did not observe the expression of *lux* genes in any of our libraries, and these species are not recognized for possessing a bioluminescent system. The presence of *P. damsela* in almost all Apogonidae libraries implies a longstanding relationship between *Photobacterium* and cardinalfishes on coral reefs, suggesting that this genus might have originated and diversified in the reef environment. Overall, these data show that cardinalfishes serve as natural hosts for *P. damsela* and should be monitored closely,

418 particularly if interacting with farmed populations that are often severely affected by *P.*
419 *damselae* infection (43).

420 The identification of two iridoviruses that were related to Erythrocytic necrosis virus (ENV;
421 83-89% MCP similarity) was also noteworthy. ENV is an important pathogen in the North
422 Atlantic and North Pacific oceans (44), and the presence of an ENV-like virus at Lizard
423 Island is consistent with a previous study that identified viral erythrocytic necrosis (VEN) in
424 a juvenile triggerfish (*Rhinecanthus aculeatus*) at this location (45). Moreover, this study
425 identified VEN-like bodies in *R. aculeatus* erythrocytes that were also found in the digestive
426 tract of associated gnathiid isopods that are common blood feeding parasites. Similarly, we
427 detected reads from gnathiids in *C. rollandi* from Lizard Island, implying that gnathiids could
428 act as vectors in the marine environment (45), although we did not detect any gnathiid reads
429 in *Enneapterygius tutuilae* from Orpheus Island. This association warrants further
430 investigation and has the potential to improve control measures against pathogenic ENV in
431 susceptible species such as pink (*Oncorhynchus gorboscha*) and chum (*Oncorhynchus keta*)
432 salmon (44).

433 It was notable that we identified a low number of protozoan parasites ($n = 8$), such as
434 apicomplexans and trypanosomes. In a similar manner, we detected low levels of fungal
435 reads (e.g., Ascomycota, Basidiomycota and Microsporidia), supporting the idea that the
436 marine environment, particularly coral reefs, contains low fungal biomass perhaps because of
437 the oligotrophic conditions compared to nutrient-rich terrestrial environments (46). The vast
438 majority of reads in this category came from dinoflagellates, diatoms, and foraminifera that
439 likely came from the reef environment, particularly dinoflagellates that form symbioses with
440 corals (1). Indeed, we discovered six viruses that fell within the “Weivirus” group—a group
441 of dinoflagellate, poriferan and mollusc viruses—grouping with a virus recently identified in
442 *Symbiodinium* spp. (47, 48) (Figure 4).

While our microbial profiling was able to detect diverse microbial communities, it is important to note that our sampling was initially performed for virological analysis (16), such that our methodology was optimized for the detection of viruses (e.g. ribosomal depletion). This may have limited our ability to detect bacteria and microbial eukaryotes. Moreover, there were necessary limitations in our sampling that impacted the power of our statistical analyses. For example, there was a considerably higher number of gobies sampled compared to other reef fish families (Figure 1). Moreover, these comparisons are based on combined tissues, such as liver and gills or whole fish (i.e. cryptobenthic reef fishes). Overall, given that metatranscriptomics is solely based on RNA-sequencing, we were only able to detect microbes and DNA viruses that were expressing genes during the time of sampling.

In summary, while our metatranscriptomic of reef fish communities from islands separated by ~450 kilometres revealed the same viral and bacterial families across both islands, there was strikingly little evidence for cross-species transmission within reef fish communities. As such, these data support the concept that fish are rich in microbial diversity, but that there are strong barriers to infection in host communities that display high levels of genetic diversity.

458

459 **Materials and methods**

460 **Ethics**

461 Fish were collected under a Great Barrier Reef Marine Park Authority permit (G16/37684.1)
462 and James Cook University Animal Ethics permit A2752.

463 **Fish sample collection**

464 Building on transcriptome data from our previous sampling of reef fish from Orpheus Island
465 (n = 192 fishes; 16 reef fish families) (16), an additional 163 individuals (24 families) were

collected at Lizard Island (14°40'08"S 145°27'34"E) during January 2022 (Supplementary Table 3). All fishes were intact with no visible signs of disease, and the vast majority were adults. These animals were captured using an enclosed clove oil method (16) from both Mermaid Cove and the Lagoon entrance, located on the north and south side of Lizard Island (Supplementary Table 1). All fish caught were placed either dissected (liver and gills) or whole in RNAlater and then transported to the lab on ice (Supplementary Table 1). Specimens were stored at -80°C until RNA extraction. Overall, we sampled 1-12 individuals per species (with a mean of 3 per species).

RNA extraction, metagenomic library preparation and next-generation sequencing

As described previously (16), tissue specimens (e.g. liver and gills or whole fish) were collectively processed as a single extraction for each individual fish sample. The combined tissues were submerged in 600 μ l of lysis buffer containing 30 μ l of foaming reagent (Reagent DX, Qiagen) and 60 μ l of β -mercaptoethanol (Sigma-Aldrich). Tissue samples were homogenized with a TissueRuptor (Qiagen) for up to 1 minute at 5,000 rpm. The homogenate was centrifuged at maximum speed for three minutes to remove tissue residues. The RNA was then extracted from the resulting clear supernatant using the RNeasy Plus Mini Kit (Qiagen, Hilden, Germany), following the manufacturer's guidelines.

RNA quantification was conducted utilizing a UV-Vis cuvette spectrophotometer (DeNovix, Delaware, USA) and a parallel capillary electrophoresis instrument (Fragment Analyzer; Agilent, CA, USA). RNA from individual fishes were pooled according to species, resulting in 79 RNA sequencing libraries newly generated from Lizard Island. All libraries were prepared using the TruSeq Total RNA Library Preparation Protocol (Illumina). Ribo-Zero Plus Kit (Illumina) was employed for host ribosomal RNA depletion, and paired-end sequencing (150 bp) was performed on the NovaSeq 6000 platform (Illumina). To mitigate

index hopping and minimize false virus–host assignments, each library was sequenced on two different lanes. Library construction and metatranscriptomic sequencing were performed by the Australian Genome Research Facility.

Assembly of reef fish viromes

We replicated the same methodology we used previously (16). Accordingly, raw RNA sequencing reads were quality trimmed using Trimmomatic v.0.38, employing the parameters SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5, and MINLEN:25, and assembled into contigs using MEGAHIT v.1.2.9, with default parameter settings (49, 50). Assembled contigs were compared against the NCBI non-redundant protein (nr) and nucleotide (nt) databases (August 2022) using DIAMOND (BLASTX) (v.2.0.9) and BLASTn (51). To enable the identification of divergent viral sequences, we used an e-value search threshold of 1×10^{-5} . Contigs with top matches to the kingdom “Viruses” (NCBI taxid: 10239) were predicted as open reading frames (ORFs) using Geneious Prime (v.2022.0) (www.geneious.com) (52). To remove false positives, all putative viral ORFs were translated into amino acid sequences and used as a query to perform a second search (BLASTP) against the NCBI nr database using Geneious Prime. ORFs with top matches to fish genes were deemed as false positives and removed from further analysis. To determine whether our putative viral contigs were expressed endogenous viral elements (EVEs) we screened for disrupted ORFs and flanking host regions using CheckV and BLASTn (16, 33). Viral contig contamination and completion was determined using CheckV (53). Transcript abundances of both host (RPS13 gene) and virus (16, 17, 54) were calculated using RNA-Seq by Expectation Maximization (RSEM) (v.1.3.0) and coverage was assessed by mapping using Bowtie2 (v.2.3.3.1) (55, 56).

Taxonomic assignment and genome annotation of reef fish viruses

We aligned the amino acid sequences of our putative viruses (partial or complete) with the complete sequences of related viruses available on NCBI/GenBank using the E-INS-i algorithm in MAFFT v.7.450 (57). To determine whether our viruses were novel species, we used levels of sequence similarity and phylogenetic relationships (see below) as specified by the International Committee of Viral Taxonomy (ICTV) (<https://talk.ictvonline.org>) for each viral genus/family. We used these criteria to determine whether a virus was likely infecting reef fishes (i.e. vertebrate-associated) or of “non-vertebrate” origin, such as those derived from fish diet, microbiome or environment (16, 17, 54). Viral genomes were annotated with the Live Annotate and Predict tool in Geneious using reference sequences from NCBI/GenBank, with a similarity threshold of 20%. We also used the NCBI conserved domain (CDD) search tool and InterProScan with the TIGRFAMs (v.15.0), SFLD (v.4.0), PANTHER (v.15.0), SuperFamily (v.1.75), PROSITE (v.2022_01), CDD (v.3.18), Pfam (v.34.0), SMART (v.7.1), PRINTS (v.42.0), and CATH-Gene3D databases (v.4.3.0) (58).

Viral phylogenetic analysis

To infer the evolutionary relationships of both the vertebrate and non-vertebrate associated viruses, we aligned the translated contigs with background protein sequences from each viral family/subfamily/genus selected from the ICTV classification and obtained from NCBI/GenBank. For RNA viruses, we used the conserved RNA-dependent RNA polymerase (RdRp), while for DNA viruses we used the DNA polymerase. Amino acid sequence alignments were trimmed using TrimAl (v.1.2) with a gap threshold of 0.9 and a variable conserve value (59). The best-fit model of amino acid substitution was estimated with the “ModelFinder Plus” (-m MFP) flag in IQ-TREE (v.1.6.12) (60, 61). We used a maximum likelihood approach to estimate phylogenetic trees using IQ-TREE, with 1000 bootstrap replicates. Trees were annotated using FigTree (v.1.4.4) (<http://tree.bio.ed.ac.uk/software%20figtree/>).

538 **Virus nomenclature**

539 Viruses were provisionally named (i.e., awaiting ICTV confirmation) according to host
540 species (e.g. *Halichoeres melanurus ranavirus*) as described previously (16).

541 **Microbial profiling**

542 To screen for transcripts associated with bacteria or single-celled eukaryotes, we aligned our
543 contigs to a custom database comprising all nucleotide sequences available on NCBI (with
544 the removal of environment or artificial sequences) using the KMA aligner and CCMetagen
545 (62, 63). We also used this output to assess metazoan reads—e.g., arthropod, mollusc,
546 platyhelminth, nematode—that may represent potential vectors for virus transmission. For
547 instances in which CCMetagen identified a microbe at the species level, we validated these
548 taxonomic assignments by: (i) performing an additional search (BLASTn) against a custom
549 16S (bacteria) or 18S (eukaryote) rRNA database, and (ii) analysing the BLASTX output (see
550 above) for hits to bacterial or eukaryotic proteins. The contigs from these BLAST hits were
551 predicted into ORFs, translated into amino acid sequences, and used as a query to perform a
552 second search against the NCBI using BLASTP for further validation. The 16S or 18S rRNA
553 gene was then utilised for phylogenetic analysis as a final validation.

554 **Analysis of Beta diversity**

555 To compare viral and microbial communities between reef fish assemblages, we calculated
556 beta diversity using a Bray–Curtis distance matrix with the phyloseq package in R (64). The
557 variables assessed were host taxonomy, location (i.e. island) and trophic guild. Accordingly,
558 fish species were categorised into six trophic guilds: carnivores, mobile invertivores,
559 omnivores, planktivores, sessile invertivores, herbivores/detritivores as described in (7). We
560 based our analysis on groups with three or more species. These data were then tested using

561 permutational multivariate analysis of variance (PERMANOVA) with the vegan package
562 (adonis) (65). All plots were constructed using ggplot2 in R (66).

563 **Data Availability**

564 Raw sequence reads have been deposited in the Sequence Read Archive (NCBI/SRA) under
565 BioProject PRJNA1078998. All viral sequences discovered have been deposited in
566 NCBI/GenBank under the accessions XXXX-XXXX. All phylogenetic trees, tables, and
567 chrona plots are available on GitHub under the repository
568 <https://github.com/vcosta16/reeffishvirome>.

570 **Acknowledgements**

571 This work was funded by an Australian Research Council Discovery Project (DP200102351)
572 and a National Health and Medical Research Council Investigator grant (GNT2017197) to
573 ECH, and an ARC Laureate Fellowship (FL 190100062) to DRB. JLG is funded by a New
574 Zealand Royal Society Rutherford Discovery Fellowship (RDF-20-UOO-007) and a Marsden
575 Fund Fast Start (20-UOO-105). We acknowledge the University of Sydney for providing the
576 high-performance computer cluster “Artemis”, that was used for this study.

577 **Author Contributions**

578 VAC and ECH conceptualised the study. VAC, DRB and JCOM performed the sampling.
579 VAC performed the analyses. VAC wrote and prepared the original draft. VAC, JCOM,
580 ECH, EH, DRB, and JLG edited and revised the manuscript. ECH and JLG funded the
581 project. EH and ECH supervised the project.

582 **Supplemental Information**

583 **Supplementary Figure 1.** Maximum likelihood phylogeny of the *Vibrio harveyi* clade
584 estimated using nucleotide sequences of the 16S gene. Fish silhouettes represent individuals
585 identified in this study. The scale bar represents the number of nucleotide substitutions per
586 site. Tree was midpoint rooted for clarity only.

587 **Supplementary Table 1.** Composition of host sequence reads.

588 **Supplementary Table 2.** Description of vertebrate-associated viruses.

589 **Supplementary Table 3.** Host library information.

590

591 **References**

- 592 1. Roth, MS. 2014. The engine of the reef: photobiology of the coral–algal
593 symbiosis. *Front Microbiol* 5:422.
- 594 2. Spalding, M, Grenfell, A. 1997. New estimates of global and regional coral reef
595 areas. *Coral Reefs* 16:225–230.
- 596 3. Eschmeyer WN, Fricke R, Fong JD, Polack DA. 2010. Marine fish diversity: history of
597 knowledge and discovery (Pisces). *Zootaxa* 2525:19-50.
- 598 4. Hughes TP, Bellwood DR, Connolly SR. 2002. Biodiversity hotspots, centres of
599 endemism, and the conservation of coral reefs. *Ecol Lett* 5:775-84.
- 600 5. Bellwood DR, Wainwright PC. 2002. The history and biogeography of fishes on coral
601 reefs. In *Coral reef fishes: dynamics and diversity in a complex ecosystem*. 5:32.
- 602 6. Brandl SJ, Goatley CH, Bellwood DR, Tornabene L. 2018. The hidden half: ecology and
603 evolution of cryptobenthic fishes on coral reefs. *Biol Rev* 93:1846-73.

- 604 7. Siqueira AC, Morais RA, Bellwood DR, Cowman PF. 2020. Trophic innovations fuel
605 reef fish diversification. *Nat Commun* 11:2669.
- 606 8. Mihalitsis M, Morais RA, Bellwood DR. 2022. Small predators dominate fish predation
607 in coral reef communities. *PLoS Biol* 20:e3001898.
- 608 9. Bellwood DR, Hoey AS, Ackerman JL, Depczynski M. 2006. Coral bleaching, reef fish
609 community phase shifts and the resilience of coral reefs. *Glob Change Biol* 12:1587-94.
- 610 10. Grutter AS. 1999. Cleaner fish really do clean. *Nature* 398:672-3.
- 611 11. Robertson DR. 1982. Fish feces as fish food on a Pacific coral reef. *Mar Ecol Prog*
612 7:253-65.
- 613 12. Clements KD, Angert ER, Montgomery WL, Choat JH. 2014. Intestinal microbiota in
614 fishes: what's known and what's not. *Mol Ecol* 23:1891-8.
- 615 13. Roeselers G, Mittge EK, Stephens WZ, Parichy DM, Cavanaugh CM, Guillemin K,
616 Rawls JF. 2011. Evidence for a core gut microbiota in the zebrafish. *ISME J* 5:1595-608.
- 617 14. Ghanbari M, Kneifel W, Domig KJ. 2015. A new view of the fish gut microbiome:
618 advances from next-generation sequencing. *Aquaculture* 448:464-75.
- 619 15. Pratte ZA, Besson M, Hollman RD, Stewart FJ. 2018. The gills of reef fish support a
620 distinct microbiome influenced by host-specific factors. *Appl Environ Microbiol*
621 84:e00063-18.
- 622 16. Costa VA, Bellwood DR, Mifsud JC, Van Brussel K, Geoghegan JL, Holmes EC,
623 Harvey E. 2023. Limited cross-species virus transmission in a spatially restricted coral
624 reef fish community *Virus Evol* 9:vead011.

- 625 17. Geoghegan JL, Di Giallonardo F, Wille M, Ortiz-Baez AS, Costa VA, Ghaly T, Mifsud
626 JC, Turnbull OM, Bellwood DR, Williamson JE, Holmes EC. 2021. Virome composition
627 in marine fish revealed by meta-transcriptomics. *Virus Evol* 7:veab005.
- 628 18. Wang AR, Ran C, Ringø E, Zhou ZG. 2018. Progress in fish gastrointestinal microbiota
629 research. *Rev Aquacult* 10:626-40.
- 630 19. Sullam KE, Essinger SD, Lozupone CA, O'Connor MP, Rosen GL, Knight RO, Kilham
631 SS, Russell JA. 2012. Environmental and ecological factors that shape the gut bacterial
632 communities of fish: a meta-analysis. *Mol Ecol* 21:3363-78.
- 633 20. Bolnick DI, Snowberg LK, Hirsch PE, Lauber CL, Knight R, Caporaso JG, Svanbäck R.
634 2014. Individuals' diet diversity influences gut microbial diversity in two freshwater fish
635 (threespine stickleback and Eurasian perch). *Ecol Lett* 17:979-87.
- 636 21. Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, Laughlin CA,
637 Saif LJ, Daszak P. 2008. Cross-species virus transmission and the emergence of new
638 epidemic diseases. *Microbiol Mol Biol* 72:457-70.
- 639 22. French RK, Anderson SH, Cain KE, Greene TC, Minor M, Miskelly CM, Montoya JM,
640 Wille M, Muller CG, Taylor MW, Digby A, Kakapo Recovery Team, Holmes EC. 2023.
641 Host phylogeny shapes viral transmission networks in an island ecosystem. *Nat Ecol*
642 *Evol* 7:1834-43.
- 643 23. Longdon B, Hadfield JD, Webster CL, Obbard DJ, Jiggins FM. 2011. Host phylogeny
644 determines viral persistence and replication in novel hosts. *PLoS Pathog* 7:e1002260.
- 645 24. Gupta P, Vishnudas CK, Robin VV, Dharmarajan G. 2020. Host phylogeny matters:
646 Examining sources of variation in infection risk by blood parasites across a tropical
647 montane bird community in India. *Parasit Vectors* 13:1-3.

- 648 25. Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C, Balloux F. 2020. The
649 phylogenetic range of bacterial and viral pathogens of vertebrates. *Mol Ecol* 29:3361-79.
- 650 26. Longdon B, Brockhurst MA, Russell CA, Welch JJ, Jiggins FM. 2014. The evolution
651 and genetics of virus host shifts. *PLoS Pathog* 10:e1004395.
- 652 27. Bellwood DR, Goatley CH, Bellwood O. 2017. The evolution of fishes and corals on
653 reefs: form, function and interdependence. *Biol Rev* 92:878-901.
- 654 28. Webster JM, Braga JC, Humblet M, Potts DC, Iryu Y, Yokoyama Y, Fujita K, Bourillot
655 R, Esat TM, Fallon S, Thompson WG. 2018. Response of the Great Barrier Reef to sea-
656 level and environmental changes over the past 30,000 years. *Nat Geosci* 11:426-32.
- 657 29. Lefèvre CD, Nash KL, González-Cabello A, Bellwood DR. 2016. Consequences of
658 extreme life history traits on population persistence: do short-lived gobies face
659 demographic bottlenecks? *Coral Reefs* 35:399-409.
- 660 30. Baker RE, Mahmud AS, Miller IF, Rajeev M, Rasambainarivo F, Rice BL, Takahashi S,
661 Tatem AJ, Wagner CE, Wang LF, Wesolowski A. 2022. Infectious disease in an era of
662 global change. *Nat Rev Microbiol* 20:193-205.
- 663 31. Altan E, Kubiski SV, Boros Á, Reuter G, Sadeghi M, Deng X, Creighton EK, Crim MJ,
664 Delwart E. 2019. A highly divergent picornavirus infecting the gut epithelia of zebrafish
665 (*Danio rerio*) in research institutions worldwide. *Zebrafish* 16:291-9.
- 666 32. Miller AK, Mifsud JC, Costa VA, Grimwood RM, Kitson J, Baker C, Brosnahan CL,
667 Pande A, Holmes EC, Gemmell NJ, Geoghegan JL. 2021. Slippery when wet: cross-
668 species transmission of divergent coronaviruses in bony and jawless fish and the
669 evolutionary history of the *Coronaviridae*. *Virus Evol* 7:veab050.

- 670 33. Costa VA, Ronco F, Mifsud JC, Harvey E, Salzburger W, Holmes EC. 2024. Host
671 adaptive radiation is associated with rapid virus diversification and cross-species
672 transmission in African cichlid fishes. *Curr Biol*, in press.
- 673 34. Liu W, Zhang Y, Ma J, Jiang N, Fan Y, Zhou Y, Cain K, Yi M, Jia K, Wen H, Liu W.
674 2020. Determination of a novel parvovirus pathogen associated with massive mortality in
675 adult tilapia. *PLoS Pathog* 16:e1008765.
- 676 35. Péntzes JJ, de Souza WM, Agbandje-McKenna M, Gifford RJ. 2019. An ancient lineage
677 of highly divergent parvoviruses infects both vertebrate and invertebrate hosts. *Viruses*
678 11:525.
- 679 36. Sun X, Li Y, Yang Q, Zhang H, Xu N, Tang Z, Wu S, Jiang Y, Mohamed HF, Ou D,
680 Zheng X. 2023. Identification of quorum sensing-regulated *Vibrio fortis* as potential
681 pathogenic bacteria for coral bleaching and the effects on the microbial shift. *Front*
682 *Microbiol* 14:1116737.
- 683 37. Svardal H, Salzburger W, Malinsky M. 2021. Genetic variation and hybridization in
684 evolutionary radiations of cichlid fishes. *Annu Rev Anim Biosci* 9:55-79.
- 685 38. Cowman PF, Bellwood DR 2013. The historical biogeography of coral reef fishes: global
686 patterns of origination and dispersal. *J Biogeogr* 40:209-24.
- 687 39. Ronco F, Matschiner M, Böhne A, Boila A, Büscher HH, El Taher A, Indermaur A,
688 Malinsky M, Ricci V, Kahmen A, Jentoft S., Salzburger, W. 2021. Drivers and dynamics
689 of a massive adaptive radiation in cichlid fishes. *Nature* 589:76-81.
- 690 40. Urbanczyk H, Ast JC, Dunlap PV. 2011. Phylogeny, genomics, and symbiosis of
691 *Photobacterium*. *FEMS Microbiol Rev* 35:324-42.

- 692 41. Gould AL, Henderson JB, Lam AW. 2022. Chromosome-Level Genome assembly of the
693 bioluminescent cardinalfish *Siphamia tubifer*: an emerging model for symbiosis research.
694 *Genome Biol Evol* 14:evac044.
- 695 42. Thacker CE, Roje DM. 2009. Phylogeny of cardinalfishes (Teleostei: Gobiiformes:
696 Apogonidae) and the evolution of visceral bioluminescence. *Mol Phylogenet Evol*
697 52:735-45.
- 698 43. Gouife M, Chen S, Huang K, Nawaz M, Jin S, Ma R, Wang Y, Xue L, Xie J. 2022.
699 *Photobacterium damsela* subsp. *damsela* in mariculture. *Aquac Int* 30:1453-80.
- 700 44. Pagowski VA, Mordecai GJ, Miller KM, Schulze AD, Kaukinen KH, Ming TJ, Li S,
701 Teffer AK, Tabata A, Suttle CA. 2019. Distribution and phylogeny of erythrocytic
702 necrosis virus (ENV) in salmon suggests marine origin. *Viruses* 11:358.
- 703 45. Davies AJ, Curtis L, Grutter AS, Smit NJ. 2009. Suspected viral erythrocytic necrosis
704 (VEN) in a juvenile blackbar triggerfish, *Rhinecanthus aculeatus*, from Lizard Island,
705 Great Barrier Reef, Australia. *Mar Biodivers Rec* 2:e149.
- 706 46. Roik A, Reverter M, Pogoreutz C. 2022. A roadmap to understanding diversity and
707 function of coral reef-associated fungi. *FEMS Microbiol Rev* 46:fuac028.
- 708 47. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS,
709 Buchmann J. 2016. Redefining the invertebrate RNA virosphere. *Nature* 540:539-43.
- 710 48. Benites LF, Stephens TG, Bhattacharya D. 2022. Multiple waves of viral invasions in
711 *Symbiodiniaceae* algal genomes. *Virus Evol* 8:veac101.
- 712 49. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
713 sequence data. *Bioinformatics* 30:2114-20.

- 714 50. Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-
715 node solution for large and complex metagenomics assembly via succinct de Bruijn
716 graph. *Bioinformatics* 31:1674–1676.
- 717 51. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using
718 DIAMOND. *Nat Methods* 12:59-60.
- 719 52. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper
720 A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012.
721 Geneious Basic: an integrated and extendable desktop software platform for the
722 organization and analysis of sequence data. *Bioinformatics* 28:1647–1649.
- 723 53. Nayfach S, Camargo AP, Schulz F, Eloie-Fadrosch E, Roux S, Kyrpides NC. 2021.
724 CheckV assesses the quality and completeness of metagenome-assembled viral
725 genomes. *Nat Biotechnol* 39:578–585.
- 726 54. Costa VA, Mifsud JC, Gilligan D, Williamson JE, Holmes EC, Geoghegan JL. 2021.
727 Metagenomic sequencing reveals a lack of virus exchange between native and invasive
728 freshwater fish across the Murray–Darling Basin, Australia. *Virus Evol* 7:veab034.
- 729 55. Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data
730 with or without a reference genome. *BMC Bioinformatics* 12:323.
- 731 56. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat*
732 *Methods* 9:357–359.
- 733 57. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
734 improvements in performance and usability. *Mol Biol Evol* 30:772-80.
- 735 58. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J,
736 Mitchell A, Nuka G, Pesseat S. 2014. InterProScan 5: genome-scale protein function
737 classification. *Bioinformatics* 30:1236-40.

- 738 59. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated
739 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.
- 740 60. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and
741 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol*
742 *Evol* 32:268–274.
- 743 61. Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermini LS. 2017.
744 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat*
745 *Methods* 14:587–589.
- 746 62. Clausen PT, Aarestrup FM, Lund O. 2018. Rapid and precise alignment of raw reads
747 against redundant databases with KMA. *BMC Bioinformatics* 19:1–8.
- 748 63. Marcelino VR, Clausen PTL, Buchmann JP, Wille M, Iredell JR, Meyer W, Lund O,
749 Sorrell TC, Holmes EC. 2020. CCMetagen: comprehensive and accurate identification of
750 eukaryotes and prokaryotes in metagenomic data. *Genome Biol* 21:103.
- 751 64. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive
752 analysis and graphics of microbiome census data. *PloS one* 8:e61217.
- 753 65. Dixon P. 2003. VEGAN, a package of R functions for community ecology. *J Veg Sci*
754 14:927-30.
- 755 66. Valero-Mora PM. 2010. ggplot2: elegant graphics for data analysis. *J Stat Softw* 35:1-3.



A ~40-kb flavi-like virus does not encode a known error-correcting mechanism

Mary E. Petrone^{a,b,1} , Joe Grove^c, Julien Mélade^a, Jonathon C. O. Mifsud^a , Rhys H. Parry^d , Ezequiel M. Marzinelli^e , and Edward C. Holmes^{a,b}

Edited by Harmit Malik, Fred Hutchinson Cancer Center, Seattle, WA; received February 22, 2024; accepted June 13, 2024

It is commonly held that there is a fundamental relationship between genome size and error rate, manifest as a notional “error threshold” that sets an upper limit on genome sizes. The genome sizes of RNA viruses, which have intrinsically high mutation rates due to a lack of mechanisms for error correction, must therefore be small to avoid accumulating an excessive number of deleterious mutations that will ultimately lead to population extinction. The proposed exceptions to this evolutionary rule are RNA viruses from the order *Nidovirales* (such as coronaviruses) that encode error-correcting exonucleases, enabling them to reach genome lengths greater than 40 kb. The recent discovery of large-genome flavi-like viruses (*Flaviviridae*), which comprise genomes up to 27 kb in length yet seemingly do not encode exonuclease domains, has led to the proposal that a proofreading mechanism is required to facilitate the expansion of nonsegmented RNA virus genomes above 30 kb. Herein, we describe a ~40 kb flavi-like virus identified in a *Haliclona* sponge metatranscriptome that does not encode a known exonuclease. Structural analysis revealed that this virus may have instead captured cellular domains associated with nucleic acid metabolism that have not been previously found in RNA viruses. Phylogenetic inference placed this virus as a divergent pesti-like lineage, such that we have provisionally termed it “Maximus pesti-like virus.” This virus represents an instance of a flavi-like virus achieving a genome size comparable to that of the *Nidovirales* and demonstrates that RNA viruses have evolved multiple solutions to overcome the error threshold.

evolution | virology | metatranscriptomics | error threshold | Flaviviridae

RNA viruses have traditionally been characterized by small, compact genomes, often close to 10 kb in length (1). The longest RNA virus genome documented to date is ~47 kb (2), considerably shorter than those seen in DNA viruses whose genomes can exceed 2.5 Mb (3). Despite the explosion in RNA virus discovery following the advent of metagenomic sequencing (4–7), the maximum genome sizes of most families of RNA viruses have remained relatively stable, with the longest nonsegmented genomes (i.e., >30 kb) consistently falling within families from the order *Nidovirales*. Some segmented viruses have been found to achieve comparable lengths, such as *Dendrolimus punctatus* cypovirus, which encodes a genome of ~33 kb across 16 segments ranging from 700 bp to 4,100 bp in length (8).

The most popular theory for the restricted genome sizes of RNA viruses is that it stems from the mutational burden associated with replication via an error-prone RNA polymerase [i.e., RNA-dependent RNA polymerase (RdRp) or reverse transcriptase] (9). Because these enzymes usually lack any form of error correction such as proofreading, longer genomes are expected to accumulate more mutations per genome replication. As most mutations are deleterious (or even lethal) (10), a high error rate will greatly reduce virus fitness, eventually leading to population extinction. Consequently, an upper cap on genome sizes reflects an “error threshold,” with high mutation rates necessarily resulting in smaller genomes.

The existence of a mutation-driven cap on virus genome size is supported by two lines of evidence. First, the treatment of experimental populations of RNA viruses with mutagenic agents, such as 5-azacytidine and 5-fluorouracil, results in major fitness losses as expected if these viruses reside close to the error threshold (11). Indeed, fitness losses resulting from the use of antiviral mutagenic agents such as ribavirin and molnupiravir (12) underpin viral therapies based on the induction of so-called “lethal mutagenesis” (13, 14). Second, members of the *Nidovirales*, which include RNA viruses with the largest known genomes (2), have acquired a unique proofreading mechanism. The *Coronaviridae* encode endo- and exonucleases (such as nsp14-ExoN) that are essential for replication (15) and that correct replication errors (16, 17). Similarly, members of the *Roniviridae* utilize NAD and ADP-ribose (NADAR) domains which are associated with RNA repair (18). Reducing the error rate per nucleotide should, in theory, have enabled nidovirus

Significance

A hallmark of RNA viruses is their small genome size, a feature attributed to the inability of their polymerase to correct errors. Because the accumulation of mutations results in the decline of fitness, RNA viruses must maintain small genomes to avoid extinction. Members of the *Nidovirales* are the sole exception because some have stolen error-correcting domains from cellular hosts. Here, we present an instance of a nonsegmented RNA virus outside of the *Nidovirales* that supports a genome size above 30 kb. This virus encodes a ~40 kb genome and falls within the *Amarillovirales*. Notably, we did not detect a known proofreading mechanism in its genome, suggesting that multiple evolutionary strategies are available for overcoming the error threshold.

Author affiliations: ^aSydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia; ^bLaboratory of Data Discovery for Health Limited, Hong Kong Special Administrative Region, China; ^cMRC-University of Glasgow Centre for Virus Research, Glasgow G61 1QH, United Kingdom; ^dSchool of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4067, Australia; and ^eSchool of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia

Author contributions: M.E.P., J.G., E.M.M., and E.C.H. designed research; M.E.P., J.G., J.M., J.C.O.M., R.H.P., and E.M.M. performed research; M.E.P., J.G., J.M., J.C.O.M., and R.H.P. analyzed data; and M.E.P., J.M., E.M.M., and E.C.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2024 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution License 4.0 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).

¹To whom correspondence may be addressed. Email: mary.petrone@sydney.edu.au.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2403805121/-/DCSupplemental>.

Published July 17, 2024.

genomes to expand in size. Indeed, the discovery of Nam Dinh virus, a ~20 kb nidovirus that is distantly related to the *Coronaviridae* and the *Roniviridae* and that encodes an exoribonuclease, further supports this hypothesis (19).

Although attractive, the error threshold theory for virus genome size was challenged by the discovery of the so-called “large-genome flavi-like” viruses (LGFs) that possess genomes up to 27 kb in length (20, 21) and hence comprise the second-longest group of nonsegmented RNA virus genomes documented to date. Importantly, however, the LGFs do not encode an exoribonuclease or similar domain, but rather a stabilizing methyltransferase domain. This domain has been proposed as a mechanism for promoting genome size expansion but is not directly involved in proofreading (22). Previously, no nonsegmented virus outside of the order *Nidovirales* had been identified with a genome exceeding 27 kb, and it was not known whether longer genomes could be supported in the absence of a known proofreading mechanism.

Herein, we describe a divergent, flavi-like virus with a genome of ~40 kb that does not encode an exoribonuclease or recognizable proofreading domain. This finding provides insights into the evolution of the *Flaviviridae* and demonstrates that exonuclease domains are not required to support RNA virus genomes above 27 kb.

Results

Identification of a Large and Highly Divergent Marine Flavi-Like Virus.

We identified a ~40 kb flavi-like virus in a sea sponge (order Haplosclerida) metatranscriptome. This sponge, along with 72 invertebrate specimens, was collected from Chowder Bay, Sydney Harbour, Sydney, Australia, on October 20, 2022. Assessment of the true host association of this virus was challenging due to the complexity of the sequencing library (*SI Appendix, Fig. S1A*), which in turn reflected the heterogeneous composition of the ecosystem from which it was collected (*SI Appendix, Fig. S1B*). Despite this, over 74% of rRNA reads in this library shared high sequence similarity with rRNA of sponge, particularly those belonging to the genus *Haliclona*, order Haplosclerida (*Dataset S1*). Given that the morphology of the specimen was consistent with *Haliclona* (*SI Appendix, Fig. S1 B, Inset*) and this genus has been previously documented in Sydney Harbour (ozcam.ala.org.au), we concluded that a *Haliclona* sponge was the most probable host organism of this virus.

This virus exhibited clear, yet minimal, sequence similarity with members of the *Flaviviridae*. It was most closely related to Fushun flavivirus 1 (length = 15 kb), with which it shared 25.7% sequence similarity (e-value = 1.31e-23, Reference Viral Database). Like other *Flaviviridae*, it comprised a single polyprotein (38 kb) and long 5' and 3' untranslated regions (UTRs) (488 nt and 1,277 nt, respectively) (Fig. 1A). The polyprotein sequence was complete, containing no premature stop codons and exhibited a strong AT bias (64%). Its abundance in the sequencing library was low, comprising less than 0.01% of non-rRNA reads (56,371 reads mapped to the flavi-like virus contig), but its sequencing coverage was robust, with a mean depth of 139 (SD = 13.2) (Fig. 1A). We recovered identical contigs with three independent assembly methods [MEGAHIT v1.2.9 (23), SPAdes v3.15.5 (24), and Trinity v2.8.6 (25)].

Fragments of this virus were detected in two of the three other libraries of samples collected from the same piling (Bryozoa, <0.0001% non-rRNA reads; sponge, 0.015% non-rRNA reads) [*SI Appendix, Fig. S2*, (28)]. To our knowledge, this is the longest known flavi-like virus, nearly 1.5 times the length of previously documented LGFs. It is also the longest known RNA virus outside of the order *Nidovirales* and the only known RNA virus over 30 kb to encode a single polyprotein.

To assess whether this putative virus was replicating in our environmental sample, we performed strand-specific PCR. Because the replication of RNA viruses involves the synthesis of negative-strand intermediates, both positive- and negative-sense strands would be produced during viral replication. Both strands could be amplified using RT-PCR (*SI Appendix, Fig. S3*) and quantified with qPCR (positive strand: 2.28×10^6 copies/mL, negative strand: 1.10×10^3 copies/mL).

Despite its length and low sequence similarity to the *Flaviviridae*, this virus possessed characteristics of canonical flavivirids. To explore genome organization and annotate protein function we used both sequence-based approaches (BLASTp, InterProScan, SuperFamily, NCBI CD-Search, and NCBIfam) and protein structure prediction followed by fold-based homology searches. For the latter, the polyprotein sequence was broken into overlapping 300 residue blocks for structure prediction, using ColabFold-AlphaFold2 (29, 30) and ESMFold (31), and queried using the Foldseek (32) server. The putative NS2/3 protein featured a consecutive serine protease and helicase pair (Fig. 1B). Manual annotation showed that the serine protease included the typical Ser (nucleophile), His (base), Asp (acid residue) catalytic triad (*SI Appendix, Fig. S4A*). Similarly, the putative NS5 RdRp protein was pesti-like in both sequence and predicted structure (Fig. 1C) and retained the highly conserved GDD motif (*SI Appendix, Fig. S4B*). Like other LGFs, a methyltransferase domain was present within the putative NS5. However, we found no regions that were similar to nido-like endo/exonucleases or NADAR domains (Fig. 1B).

An envelope protein surrounded by signal peptides and probable N-linked glycosylation sites was detected near the 5' terminal (Fig. 1B and D and *SI Appendix, Fig. S5*). ESMFold protein structure prediction yielded a domain of intermediate prediction confidence toward the C terminus of this region. This domain exhibited similarity to domain III of a class II fusion protein, with strongest homology to glycoproteins from bunya-(Gc) and alphaviruses (E1) rather than the E glycoprotein found in orthoflaviviruses (Phyre2: Rift Valley Fever virus, HHpred: Thrombocytopenia syndrome virus, Foldseek: Chikungunya virus, *Dataset S2*). This is consistent with the highly divergent nature of this protein, as this region exhibited only 16% identity with its closest homolog (Fig. 1D).

However, viral proteins are underrepresented in fold-based homology search servers such as Foldseek. To address this, we previously generated a protein foldome for the *Flaviviridae* composed of predicted structures for >450 viral species, all processed using the 300-residue sequence block strategy described above (27). We queried this database for similarity to the RdRp and glycoprotein folds identified in the sponge flavi-like virus using Foldseek (Fig. 1E and F). Being highly conserved, RdRp homologues were found in all species, with highest similarity residing in the pestiviruses and LGFs. Glycoprotein homologues were present across the large-genome flavi-, jingmen- and orthoflaviviruses, where similarity corresponded to E glycoproteins. Glycoprotein homologues could not be found in the pesti-, pegi- or hepaciviruses, which possess the E1E2 glycoproteins that are structurally distinct from the E glycoprotein (33). We concluded that this virus expresses an E-like glycoprotein homologous to other LGFs. Together, the quality of the sequence and the organization of the putative genome suggested that this unusually large virus was exogenous and related, albeit distantly, to the *Flaviviridae*.

Divergent, Cellular-Associated Domains in an RNA Virus Genome.

Given its size, we hypothesized that this virus would encode noncanonical flavivirus domains, particularly those that would help support the replication of such a large RNA genome. Accordingly, using both sequence- and structure-based approaches,

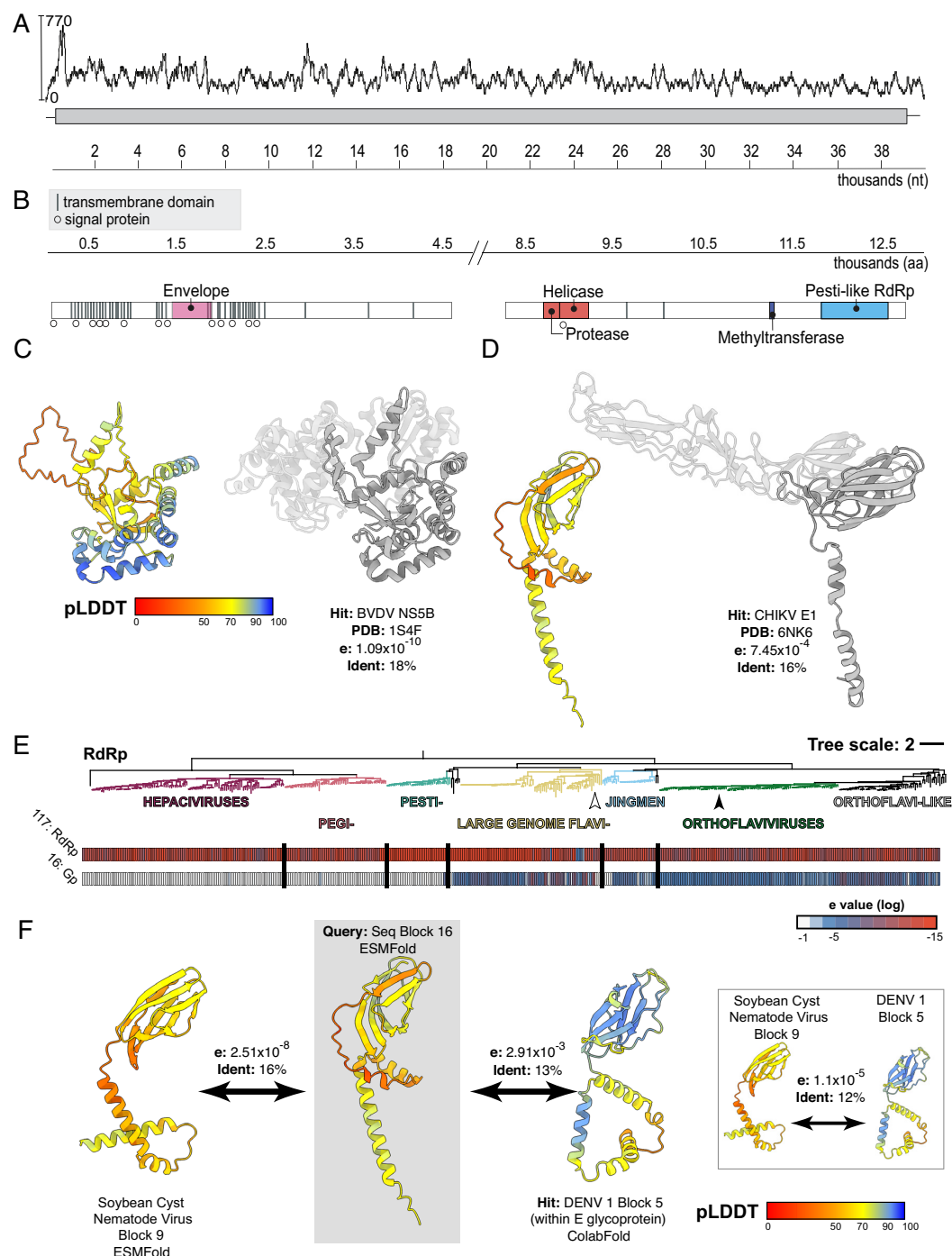


Fig. 1. Characterization of canonical *Flaviviridae* features in a highly divergent, large-genome flavi-like virus. (A) Sequencing coverage of nucleotide sequence. The y-axis measures sequencing depth. Reads were mapped using BMap (26) and the coverage plot was generated in Geneious Prime (www.geneious.com). (B) Functional domains identified using sequence-based and protein fold-based homology. (C) Structural prediction of the putative NS5 protein. Left: ESMFold structure prediction for polyprotein sequence block 117 (amino acid residues 11,700 to 12,000), color-coded by prediction confidence (predicted Local Distance Difference Test, pLDDT) as shown in the key. Right: homologous BVDV NS5B structure (PDB: 1S4F) returned by Foldseek as the top hit. Expected value (e-value) and percentage sequence identity are shown. Regions of structure not aligned by Foldseek are transparent. (D) Homology between sequence block 16 (with putative glycoprotein region) and Chikungunya virus (CHIKV) E1 structure (PDB: 6NK6), the top hit returned by Foldseek, displayed as in (C). (E) RdRp maximum likelihood phylogeny for the *Flaviviridae* as described by Mifsud et al. (27). The tree is labeled and colored to indicate major viral genera and groups. The heatmap displays log transformed e-values (color coded as shown in the key) demonstrating the distribution of homologues to block 117 (RdRp) and block 16 (glycoprotein) within the *Flaviviridae* protein foldome. (F) Example structural homologues to block 16 from a large-genome flavivirus (Soybean cyst nematode virus) and a canonical orthoflavivirus (Dengue virus 1 [DENV 1]); e-values and sequence identities are provided. The inset demonstrates structural similarity between these two hits. The location of the Soybean cyst nematode virus and DENV 1 are indicated on the *Flaviviridae* phylogeny (E) with white and black arrowheads, respectively.

we found three notable features: a divergent extended helical domain (amino acid pos. 4,346 to 4,475), a Tu elongation factor (pos. 6,419 to 6,586), and a nucleic acid metabolism cassette (pos. 7,286 to 8,206) (Fig. 2A).

The potential functions of these domains were uncertain. The role of the extended helical domain could not be determined. Both ESMFold and ColabFold structures were of intermediate to high confidence and were in excellent agreement, supporting their veracity. However, this feature did not share detectable homology with known proteins in both ESMFold and ColabFold predictions (Fig. 2B). Its position in the genome proximally downstream of the envelope protein could be consistent with host response antagonism like the canonical NS1 (35), although experimental analysis is required to test this. In cellular organisms, Tu elongation factors (Fig. 2C) principally facilitate translation (36). This may be particularly beneficial to a virus with a very long polyprotein, but Tu elongation factors are also thought to exhibit “moonlighting” (i.e., noncanonical) functions (36). In this instance the protein was highly divergent, exhibiting no detectable sequence similarity to any other protein in the NCBI nr database, including from cellular organisms, which may reflect functional divergence. We identified related domains in the genomes of some *Megaviricetes* and *Caudoviricetes* (i.e., DNA bacteriophage) when we screened the Reference Proteomes database with HMMER (37), although we found no homologues in RNA viruses. The predicted structure shared homology with eukaryotic and bacterial elongation factors, but it was most closely related (11.9% identity) to an ancestrally reconstructed bacterial Tu Elongation Factor (Fig. 2C). However, it should be noted that lower-confidence/low-frequency Foldseek hits suggest other potential identities for this fold (e.g., uridine phosphorylase; *SI Appendix, Fig. S6*).

Elements of the nucleic acid metabolism cassette were more conserved. This region comprised an adenylate kinase, a nucleoside

diphosphate kinase (NDPk), a phosphoprotein phosphatase, and a photolyase-related protein upstream of the NS2/3 protease-helicase pair (Fig. 2A and D–F). Among these domains, NDPks are the most highly conserved across all kingdoms of life (38) and have been found in some giant DNA viruses and bacteriophage (39). They are involved in RNA synthesis but have not previously been documented in RNA viruses. Interestingly, the flavi-associated NDPk domain was more closely related to those in cellular organisms than to those in other viruses, sharing up to 30% sequence similarity to homologous domains in bacteria and eukaryotes. To search for related domains in viruses, we screened the NCBI nonredundant (nr) protein database restricted to viruses (taxid: 10239) using the flavi-associated NDPk as input. This returned no results, indicating that the flavi-associated NDPk domain does not share detectable sequence similarity with its counterparts in DNA viruses and thus is not closely related. It also suggested that NDPk acquisition is a rare event in RNA virus evolution. Querying Reference Proteomes, UniProtKB, Swiss-Prot, PDB, and ColabFold databases using hmmsearch with the NDPk alignment as input also failed to detect this domain in any publicly available RNA virus genomes.

Phylogenetic analysis of the NDPk domain could not confidently resolve its evolutionary relationship to cellular organisms. Although it did not share detectable sequence similarity to homologous domains in DNA viruses, the flavi-like virus-associated domain shared <30% similarity to domains in Bacteria, Viridiplantae, Fungi, Archaea, Metazoa, and Hapista. Using the top 500 Blastp matches filtered for redundancy at a threshold of 90%, we inferred four phylogenetic trees from a combination of aligning and trimming methods (*SI Appendix, Fig. S7*). In three of the four phylogenies, the flavi-like virus-associated NDPk domain was most closely related to bacteria-associated domains, albeit with low support (*SI Appendix, Fig. S7 B–D*). Regardless of

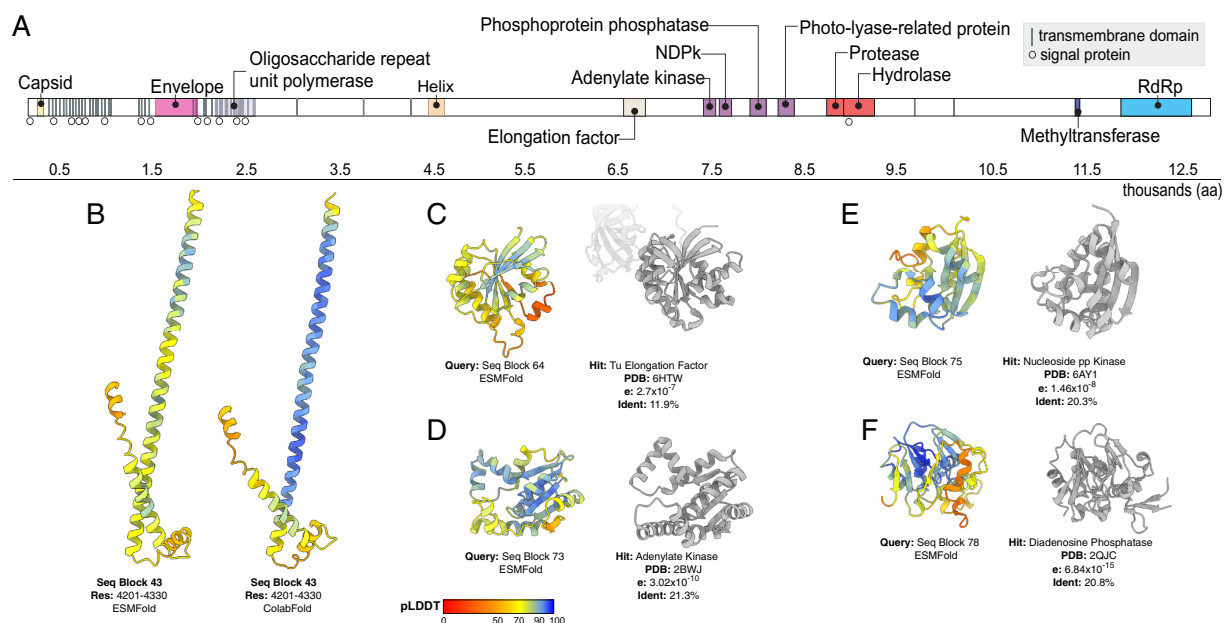


Fig. 2. Large-genome flavi-like virus polyprotein encodes a complement of noncanonical functional domains. (A) Annotation of domains detected using a combination of sequence and structure-based analyses: InterProScan, DeepTMHMM (transmembrane domains), SignalP (signal peptides), Pfam, CDD, and Foldseek. Domain annotations reflect domains sharing the closest structural homology to the putative viral domains. The polyprotein is scaled to thousand amino acids. NDPk: nucleoside diphosphate kinase. Helical domain: helical domain of unknown function. (B) Inferred ESMFold and ColabFold structures of a helical domain of unknown function [found in sequence block, (34)], color-coded by prediction confidence (pLDDT) as shown in the key (C–F) ESMFold structures of the stated sequence blocks alongside their cognate, top representative Foldseek hits, expected value, and percentage sequence identity are provided. Regions of structure not aligned by Foldseek are transparent.

its origins, this domain constitutes the only known example of such a feature in the characterized RNA virosphere.

Phylogenetic Support for Classification as a Large-Genome Flavi-Like Virus. The unique and divergent protein complement encoded by this virus suggested that it could provide insights into the evolutionary history of the *Flaviviridae*. A family-level phylogenetic analysis placed both the putative NS2/3 and NS5 proteins with pesti-like viruses in a deep phylogenetic position (Fig. 3 *A* and *B*). We herein refer to this section of the tree as the “*Pestivirus*-LGF clade” that comprises the genus *Pestivirus*, pesti-like viruses, and LGFs. The placement of the divergent NS5 sequence at the base of the LGF clade was well supported (ufboot = 92), and the size of the viral genome was consistent with other members of this group (Fig. 3 *A*, *Inset*). In contrast, the NS2/3 proteins did not form a monophyletic group, although this topology did not have robust support (Fig. 3 *B*). Here, the putative NS2/3 sequence from the divergent virus formed a sister group to two spider-associated pesti-like viruses (Xinzhou spider virus 3 and Shayang spider virus 4) (Fig. 3 *B*, ufboot = 64). These spider-associated viruses were previously reported to be the most divergent members of the *Pestivirus* and pesti-like virus clade (40), and this relationship was reflected in our NS5 tree (Fig. 3 *A*).

In an attempt to improve the accuracy of the phylogenetic position for both proteins, we realigned all NS2/3 and NS5 sequences belonging to the orthoflavivirus, pestivirus, pesti-like viruses, and LGFs using six combinations of sequence alignment and alignment trimming methods (Table 2). This did not yield more stable results for the placement of either region in the divergent virus (*SI Appendix*, Figs. S8 and S9). However, both

amino acid sequences from the virus newly identified here consistently fell at the base of the *Pestivirus*-LGF clade, suggesting that this virus represents part of the deep evolutionary history of this group. Both midpoint rooting and rooting on the *Orthoflavivirus* clade consistently placed our virus within the sister clade to containing the genus *Pestivirus*. Importantly, it shared a closer phylogenetic relationship to animal-associated viruses than to plant-associated viruses, which form a monophyletic group. The placement of the divergent NS2/3 sequence, which formed a sister group to the spider-associated pesti-like viruses, is consistent with a history of virus–host codivergence. However, the broad cross-kingdom range of hosts associated with the *Pestivirus*-LGF clade indicates that the evolutionary history of this clade with respect to virus–host interactions is complex.

Together, our analyses support the conclusion that this virus is a large-genome sponge-infecting pesti-like virus, which we have provisionally named “*Maximus* pesti-like virus.”

Discussion

We describe the largest known flavi-like virus and largest known RNA virus outside of the order *Nidovirales*. With a genome size of 39.8 kb, this virus surpassed the longest documented flavi-like virus to date by nearly 13 kb (20). Despite its large size, this virus possessed features that are characteristic of members of the *Flaviviridae*: an E-like glycoprotein, a protease/helicase pair, and a pesti-like RdRp. Although the E-like glycoprotein shared highest similarity to alpha- and bunyavirus glycoproteins, this is consistent with the notion that the class II membrane fusion machinery of the *Amarillovirales* (that contain the *Flaviviridae*), *Bunyavirales*,

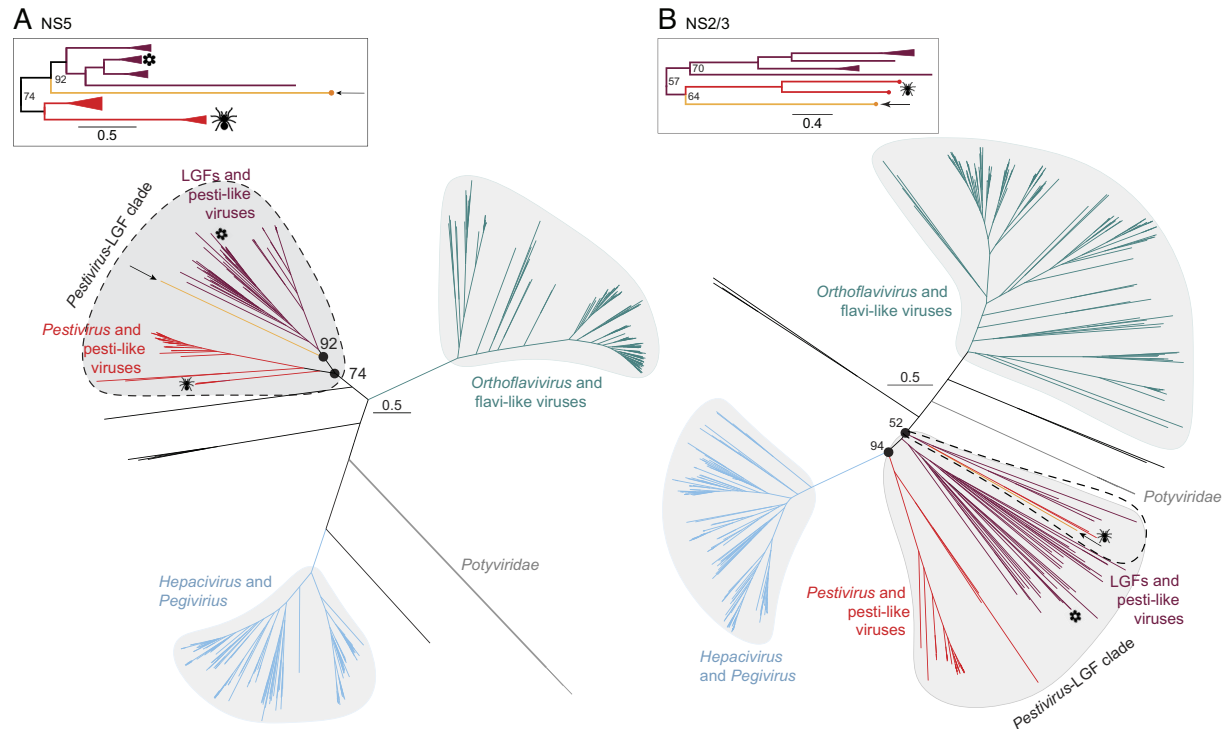


Fig. 3. Phylogenetic analysis of the divergent flavi-like virus places it among the pesti-like viruses. Unrooted phylogenetic trees of *Flaviviridae*. (A) NS5 and (B) NS2/3. Arrows indicate the divergent flavi-like virus identified here. Flower icon indicates plant-associated flavi-like viruses. Spider icon indicates spider-associated pesti-like viruses. Ultrafast bootstrap values are shown as numbers at select nodes. *Insets* show the groups containing the divergent flavi-like virus and correspond to the regions encircled by dotted lines in the main figure. The topology of the insets reflects midpoint rooting of each tree in the main figures. Tree branches are scaled according to amino acid substitutions.

Table 1. Primers used for the standard RNA transcripts, RT, and PCR for the strand-specific assay

RT reaction	Name	Sequence 5'–3'	Position (nt)
(+) strand	Tag Pos RT	<u>CGGGAAGGCGACTGGAGATCATATGAGCATTAGAGTACC</u>	4790
(–) strand	Tag Neg RT	<u>CCGTCATGGTGGCGAATAAGACTGGTGTCAATTAGAAGAATC</u>	4700
PCR and qPCR			
(+) strand	PCR Pos F	GACTGGTGTCAATTAGAAGAATC	4700
	PCR Pos R	<u>CGGGAAGGCGACTGGAG</u>	Tag
(–) strand	PCR Neg F	<u>CCGTCATGGTGGCGAATAA</u>	Tag
	PCR Neg R	ATCATATGAGCATTAGAGTACC	4790
Standard RNA			
(+) strand	T7 Pos F	GCGTAATACGACTCACTATAG GCTTCGATTCAAGATTACGTTTA	4586
	T7 Pos R	GATCCAAAGTGTCATGTATGA	5564
(–) strand	T7 Neg F	GCTTCGATTCAAGATTACGTTTA	4586
	T7 Neg R	GCGTAATACGACTCACTATAG TAGAGATCCAAAGTGTCATGTATGA	5564

Underlined nucleotides represent the unique nonviral tag sequence for the strand-specific RT reaction. Sequences in bold indicate the T7 promoter for the standard RNA generation. (+): positive; (–): negative; nt : nucleotides.

and *Martellivirales* evolved from a common ancestor (41). The polyprotein encoded a methyltransferase upstream of the RdRp, which is common throughout the *Flaviviridae* (22). The organization of the genetic components was also congruent with that of canonical *Flaviviridae*, a characteristic that differs from large nidoviruses, as these exhibit a flexible modularity in their genetic organization (2). Phylogenetic analysis placed the two most conserved proteins (NS2/3 and NS5) of this divergent virus deep within the *Pestivirus*-LGF clade, with a relatively close relationship to previously characterized LGFs. Although small portions of the genome shared sequence and structural homology to cellular proteins, we concluded that the virus was unlikely to be an endogenous viral element because the virus-associated sequences were highly divergent from their cellular homologues. Successful amplification and quantification of both positive- and negative-sense strands with PCR further supported this conclusion.

The identification of such a large flavi-like virus raises important questions. Does it possess a mechanism for correcting errors during replication? Or does it sustain a mutation rate similar to those seen in other flavivirids, and if so, how? We did not find evidence for nido-like endo- or exonucleases, the absence of which is consistent with other LGFs. Instead, the nucleic acid metabolism cassette encoded upstream of the putative NS2/3

could be a plausible proofreading candidate. One component of this cassette, deoxyribodipyrimidine photolyase is involved with nucleotide repair caused by UV damage (42). However, other components may serve different purposes. Adenylate kinases regulate adenosine phosphate concentrations (43), and, when present in dsDNA viruses, are thought to counterbalance the differential AT concentrations between the virus and its host along with NDPk (39). The genome of our divergent flavi-like virus exhibited a substantial AT bias (64%), and the nucleic acid metabolism cassette may in part function to sustain this. The only *Haliclona* genome available on NCBI (CAXHKC01) comprises ~58% AT. The lack of additional proofreading candidates reflected our inability to annotate or assess the role of large portions of the genome in silico because they did not share sequence or structural homology with known proteins. Isolation and culturing of this virus is needed to ascertain the function of these proteins. In vitro studies are also needed to estimate the intrinsic mutation rate of this virus so that it may be compared to those of canonical, ~10 kb vertebrate *Flaviviridae* (e.g., hepatitis C virus, dengue virus).

The genomic architecture of this virus may also be consequential in understanding the long-term evolutionary history of the *Flaviviridae*. In contrast to nidoviruses, which are thought to have

Table 2. Combinations of sequence alignment and trimming methods used in phylogenetic analysis of the *Pestivirus*-LGF clade

Protein	Alignment method	trimAl parameters	Alignment length (aa)	Substitution model selected
NS5	MAFFT	N/A	3,317	LG+F+R10
NS5	MUSCLE	N/A	6,085	LG+R10
NS5	MAFFT	-gappyout	621	LG+F+R10
NS5	MUSCLE	-gappyout	453	LG+F+R8
NS5	MAFFT	-gt 0.15 -cons 0.15	1,146	LG+F+R10
NS5	MUSCLE	-gt 0.15 -cons 0.15	1,178	LG+F+R10
NS2/3	MAFFT	N/A	3,321	LG+F+R6
NS2/3	MUSCLE	N/A	8,477	LG+F+R10
NS2/3	MAFFT	-gappyout	971	LG+F+R10
NS2/3	MUSCLE	-gappyout	1,016	LG+F+R10
NS2/3	MAFFT	-gt 0.15 -cons 0.15	1,525	LG+F+R10
NS2/3	MUSCLE	-gt 0.15 -cons 0.15	1,226	LG+F+R10

undergone genomic expansion after the acquisition of a proofreading mechanism (19, 22, 44), members of the *Pestivirus*-LGF clade may have undergone genome reduction. The phylogenetic placement of NS5 sequences near the base of the *Pestivirus*-LGF clade was moderately well supported (ufboot = 92), and this position was consistent across multiple combinations of sequence alignment and alignment trimming methods. This tree topology would be consistent with the phylogenetic position of its putative host, a sponge, which was among the first animals to evolve (34). Thus, one interpretation of the size of the *Maximus pesti*-like virus genome is that it is a relic of an ancestral state of the pestiviruses, and the canonical (i.e., 10 kb) *Pestivirus* genome resulted from a series of genome reduction events, perhaps concordant with their emergence in vertebrates where longer genomes would mean more potential immune targets for organisms that had evolved adaptive immunity (45). If large sections of the *Pestivirus* genome were lost over time, it follows that the proteins encoded in these sections were selected against. This could explain why we were unable to find similar protein domains in other RNA viruses. It also raises the interesting possibility that some RNA viruses previously encoded a proofreading mechanism and then lost it, if one is found in the divergent flavi-like virus described here. Such a pattern would again be opposite to what has been inferred from the evolutionary history of the *Nidovirales*. Robust sampling of basal marine invertebrates for the purposes of virus discovery is required to address this hypothesis more thoroughly.

Another explanation for the aberrant size of this virus is that it belongs to a select group of specialized viruses that are successful “genetic pirates.” The acquisition of the nucleic acid metabolism cassette and the other functional domains could have been the result of relatively recent genetic piracy wherein an ancestral virus captured genetic material from a cellular organism. The divergent flavi-like virus appears to have acquired these domains in at least two separate events because its putative elongation factor does not share detectable sequence similarity with known cellular homologues, whereas its NDPk domain does. This suggests that the ancestors of this virus may have been adept at stealing domains from cellular organisms, a characteristic that may not be representative of most ancestral *Flaviviridae*. However, no conclusions can be drawn on the origin of the large sections of the genome that could not be aligned to any known member of the *Flaviviridae*, were not present in any RNA virus searchable in publicly available databases, and did not share structural homology with known cellular proteins. It may be that these sections were derived from organisms that are underrepresented in publicly available databases. Alternatively, they may have been captured long before the identifiable domains such that they no longer share detectable structural similarity with their cellular homologs. An assessment of this “genetic piracy” hypothesis will require the discovery of additional LGFs of comparable size and a more rigorous exploration of the LGF “foldome.”

Regardless of when it occurred, the apparent acquisition of domains sharing sequence and structural similarity to those found in bacteria is puzzling. Bacterial NADAR and Nudix domains have been found in RNA viruses, including among the *Nidovirales*, but these are thought to have been acquired through horizontal gene transfer from a eukaryote (18, 46). Although they did not have robust statistical support, the topologies of our NDPk phylogenetic trees are inconsistent with this sequence of events because in three of four topologies, the flavi-associated NDPk was most closely related to bacteria-associated NDPk domains (*SI Appendix, Fig. S7*). One explanation is that ancestral viruses infected bacteria-eating protists. Because bacteria have evolved defenses that allow them to use their would-be predators as safe ecological niches (47), coinfection with an ancestral flavi-like virus could have afforded the virus an opportunity to steal bacterial domains. Alternatively, sponges

harbor many endosymbionts, including bacteria (48), which may have facilitated bacteria-virus gene transfer. It may also be that the flavi-associated NDPk shares closer sequence similarity to an unsampled eukaryote, and its position in the phylogenetic tree will change as new genetic data become available.

Filling sampling gaps in the RNA virosphere may shift paradigms in virus evolution if more exceptions to conventional wisdom are documented. Here, we present one such example by showing that flavi-like viruses can support genome sizes comparable to those observed in the *Nidovirales*. The absence of a detectable exonuclease domain demonstrates that other solutions are available to RNA viruses for overcoming theoretical error thresholds. Future studies may find that the *Nidovirales* and the *Flaviviridae* are not unique in their possession of large genomes. The continued exploration of RNA virus diversity is therefore key to revealing the patterns that shaped the long-term macroevolution of RNA viruses.

Materials and Methods

Sample Collection. All invertebrate samples analyzed here ($n = 72$) were collected on October 20, 2022, in Chowder Bay, Sydney, Australia. Samples were collected by divers wearing latex gloves and using forceps cleaned with 96% ethanol between sampling events. All samples were placed in RNA/DNA-free cryogenic tubes, which were snap-frozen and stored immediately in liquid Nitrogen. Samples were then transferred to a -80°C freezer.

Sample Preparation and Metatranscriptome Sequencing. Sample tissue was processed individually by flash freezing with liquid nitrogen and homogenizing with a mortar and pestle. Total RNA was extracted using the QIAGEN RNeasy Plus Mini kit. The RNA of some samples was pooled for downstream processing. The libraries in which the divergent flavi-like virus was found comprised single samples. Sequencing libraries were prepared using the Ribo-Zero Plus library preparation kit and sequenced on the Illumina NovaSeq 6000 platform. In total, 44 libraries were sequenced, including a “blank” negative control comprising a sterile water plus reagent mix.

Virus Identification. Raw reads were processed by removing sequencing adapters with Cutadapt v1.8.3 (49) and the parameters: $-u\ 5 -U\ 5 -q\ 25 -m\ 25$. The quality of trimming was assessed using Fastqc v0.11.8 (50). Reads mapping to rRNA were removed using SortMeRNA v4.3.3 (51) and the SILVA rRNA database (as of Dec. 2023). Contigs were assembled with MEGAHIT v1.2.9 (23) setting the minimum contig length to 200 nt. All contigs were screened against the RdRp-scan database (52) and a custom RNA virus databases using DIAMOND BLASTx v2.0.9 (53) with the setting ultrasensitive and an e-value cutoff of $1e-5$. The custom database included divergent viruses identified in ongoing research projects. Virus candidates (i.e., viruses with a detectable sequence similarity to viruses in one or both of the databases) that were at least 1 kb in length were collated for verification. To verify that putative virus sequences were not misassigned host genes, all candidates were screened against the NCBI nonredundant protein database (as of Sept. 2023) using DIAMOND BLAST with an e-value cutoff of $1e-5$ and the setting very-sensitive. Virus candidates sharing sequence similarity to host genes were excluded from further analysis. No virus candidates were detected in the negative control.

The sequence of the divergent flavi-like virus was translated using ExPasy (<https://web.expasy.org/translate/>) to verify the presence of a complete and uninterrupted open reading frame. UTRs were manually annotated in Geneious Prime v2023.2.1. To further verify that this sequence did not contain known host gene sequences, the polyprotein was screened for sequence and structure homologies using BLASTp and Phyre2 (54). The assembly of the contig was compared against assemblies generated using SPAdes v3.15.5 (24) and Trinity v2.8.6 (25).

The abundance of the remaining reads (i.e., non-rRNA reads) was estimated using RSEM v1.3.0 (55) implemented in Trinity v2.5.1 (25). Raw reads were aligned with bowtie2 v2.3.31 (56). The abundance of the contig encoding the flavi-like virus described in this study was calculated by dividing the expected count for that contig by the sum of the expected count for all rRNA reads.

Genome Sequencing Coverage. Sequencing coverage was assessed using BMAP v37.98 (26) with a mapping threshold of 95%. Calculation of the Q30 score and mapping visualization was performed by Geneious Prime (www.geneious.com).

Host Associations. To characterize the host composition of the library, non-rRNA reads and contigs were first screened using KMA v1.3.9a (57) and CCMetagen v1.1.3 (58). Reads were mapped to the prebuilt KMA database. Next, rRNA reads identified with SortMeRNA v4.3.3 (51) were assembled using MEGAHIT v1.2.9 (23) and screened against the NCBI Blast nt database (as of December 2023).

Strand-Specific Molecular Analysis.

cDNA Synthesis by Reverse Transcription. The first (+) and (−) strands were generated by reverse transcription using the SuperScript™ IV First-Strand Synthesis System (ThermoFisher). Reactions were performed using 100 ng of RNA as the template, 1 μL of the appropriate strand-specific primer flanked with a nonviral sequence tag (2 μM, Table 1), and 1 μL dNTPs (10 mM). The mixture was heated at 65 °C for 5 min and incubated on ice for 5 min. In a separate tube, 4 μL of SSIV Buffer (5X) was mixed with 1 μL of DTT (100 mM), 1 μL of Ribonuclease Inhibitor and 200 units of Superscript IV Reverse Transcriptase (Invitrogen). The RNA-primer mix and the RT reaction mix were mixed together and incubated at 55 °C for 30 min and subsequently inactivated by heating at 80 °C for 10 min. cDNAs were then used for the end-point and qPCR.

End-Point PCR Amplification. cDNAs produced by RT were used as templates for amplification in PCR carried out in a final volume of 50 μL that included 25 μL of Platinum™ SuperFi™ PCR Master Mix (2X, ThermoFisher), 2.5 μL of each respective forward and reverse primers (10 μM), 10 μL of cDNA, and 10 μL of nuclease-free water. PCRs were performed on a SimpliAmp™ (ThermoFisher) thermocycler with the following conditions: 98 °C for 30 s followed by 45 cycles of 98 °C for 10 s, 61 °C for 10 s, 72 °C for 10 s, and 72 °C for 5 min. The PCR products were analyzed on SYBR Safe (ThermoFisher) stained agarose gels.

Strand-Specific RNA Transcripts. Synthetic positive and negative RNA transcripts standards were synthesized by in vitro transcription using T7 RNA polymerase. First, the positive and negative cDNA products were produced by RT-PCR from the extracted RNA, using their respective primer pairs containing a T7 promoter at the 5' end (Table 1). Second, the in vitro transcription was performed using the HiScribe™ T7 High Yield RNA Synthesis Kit (NEB). Briefly, 10 μL of the reaction buffer (10X) was mixed with 2 μL of ATP, UTP, GTP, CTP (100 mM each), 1 μL of DTT (0.1 M), 2 μL of T7 RNA polymerase mix, 2 μL of T7-DNA products (1 μg) and 5 μL of nuclease-free water. The reaction was incubated at 37 °C for 2 h and RNA transcripts were then purified using the RNeasy MinElute Cleanup Kit (Qiagen). Concentrations of purified RNA were determined using a NanoDrop spectrophotometer (ThermoFisher).

Strand-Specific qPCR Assay. The QuantiNova SYBR Green PCR Kit (Qiagen) was used for the strand-specific quantification by qPCR. The mixture contained 10 μL of SYBR Green PCR Master Mix (2×), 1.25 μL of forward primer and reverse primers (0.7 μM), 6 μL of cDNA, and 2.5 μL of nuclease-free water. Assays were performed using the LightCycler® 480 Instrument II (Roche) with the following conditions: 95 °C for 2 min, followed by 40 cycles of 95 °C for 5 s, 60 °C for 10 s, 72 °C for 30 s, and data collection occurred during the 60 °C step. Synthetic RNA representing each strand, serially diluted from 10¹¹ to 10² was used to calculate the amount of viral RNA copies.

Annotation of Functional Domains. Preliminary annotation of the divergent flavi-like virus was performed using InterProScan v2.1 against the CDD, SuperFamily, and NCBIfam databases as implemented in Geneious Prime v2023.2.1. For manual annotation of protein boundaries, transmembrane (TM) domains were predicted using DeepTMMHMM v1.0.24 (59). These TM domains were then analyzed to determine whether they were signal peptidases by querying a 40 amino acid (aa) region encompassing each TM domain using a 15aa sliding window with SignalP v6.0 (60). The potential N-linked glycosylation residues on the polyprotein were identified using NetNGlyc v1.0 (61), with likely N-linked glycan residues considered above the threshold of 0.5. For functional protein domain prediction, the polyprotein sequence was queried against the Pfam v34.0 (62) and CDD v3.20 (63) databases utilizing the NCBI CD-search program (64) with an

e-value threshold of 0.1. To examine the presence of a protease, we aligned with NS3Pro of Classical swine fever virus (CSFV), Pangolin pestivirus, and the divergent flavi-like virus with MAFFT v7.511 (65) L-INS-I method. CSFV was chosen because it is a pestivirus and the catalytic triad in its protease is well documented. According to BLASTp, the highest sequence similarity to the putative NS3 region of the divergent flavi-like virus was found in Bovine viral diarrhea virus 3 (AFL65618.1). The putative envelope protein was identified using HHpred (66), which predicted its presence with 60.1% probability when the first 6,000 amino acids of the polyprotein were tested against the PDB_mmCIF70_24_Oct database.

Protein Structure Prediction and Homology Search. We exploited machine-learning approaches that allow high-confidence ab initio protein structure prediction from sequence data alone. The polyprotein sequence of the divergent flavi-like virus (12,694 amino acids) was split into 300 residue blocks each overlapping by 100 residues (resulting in a total of 125 sequence blocks; *SI Appendix, Fig. S6A*). The structure of each sequence block was predicted using both ColabFold-AlphaFold2 (29, 30) and ESMFold (31) (*SI Appendix, Fig. S6 B and C*). While the majority of the polyprotein did not yield confident predictions, peaks in confidence corresponding to well-folded domains were produced by both methods. ESMFold performed somewhat better than ColabFold (*SI Appendix, Fig. S6 D–F*), likely due to the latter's requirement for multiple sequence alignments, which cannot be readily assembled for highly divergent sequences, such as those found in this virus. Structure-guided homology searches were performed using the Foldseek server (32) and we focused on hits against experimentally solved structures from the protein database (PDB). Additionally, we used a local installation of Foldseek to query a protein structure prediction database covering the entire *Flaviviridae*, as described by Mifsud et al. (27). We evaluated the distribution of the identity of Foldseek hits to ensure that the reported top hit for each inferred structure reflected the majority of hits obtained through this screen, and we assessed concordance between Foldseek hits returned when either ESMFold or ColabFold inferred structures were used as input (*SI Appendix, Fig. S6D*). Tables of Foldseek hits for each structure block shown in the main figures are provided in *Datasets S2* and *S3*.

Phylogenetic Analysis. For the family-level virus analysis, all *Flaviviridae* (taxid 11050 and taxid 38144) available on the NCBI Virus were downloaded on Dec. 15, 2022. These data were supplemented with sequences from Mifsud et al. (67), Mifsud et al. (27), and the NCBI nucleotide database (searched using the phrase “flavi[All Fields] OR pesti[All Fields] OR hepac[All Fields] OR pegi[All Fields] AND viruses[filter]” on Dec. 15, 2022). We screened the BLASTx results for the divergent flavi-like virus for sequences published in 2023 and added the two that met these criteria (UHR49738.1 and WPA70758.1) to the dataset. We also screened the NCBI Virus database for pesti- and pesti-like viruses and large-genome flavi-like viruses published in 2023 that were longer than 20 kb, and consequently added *Macrosiphum euphorbiae* virus 1 isolate K01 (YP_009175071, length = 22.8 kb) to the dataset. Redundant nucleotide sequences were removed using CD-HIT (68) at a cutoff of 95%. Remaining sequences were translated using the Geneious Prime Find ORFs tool (v.2022.0) (www.geneious.com) and trimmed to the NS2/3 and NS5 regions, individually. Putative NS5 sequences that lacked the conserved GDD palm motif were excluded. Incomplete genomes were removed manually. The corresponding segments of the two closest protein BLASTp hits to the divergent flavi-like virus (Bovine viral diarrhea virus 3 [accession ANW09737] and *Pestivirus brazilense* [accession WEC89329]) were added to each dataset. Members of the *Potyviridae* were used as the outgroup for both trees. The final NS2/3 and NS5 datasets contained 427 and 464 sequences, respectively. All datasets were aligned with MAFFT v7.490 (65) using default parameters and the BLOSUM62 scoring matrix (NS2/3 length = 2,279aa; NS5 length = 2,537aa). Ambiguously aligned regions were removed with a conservation score of 0.15 using trimAl v1.4.1 (69).

For the genus-level analysis, pestivirus and pesti-like virus sequences were realigned using members of the genus *Orthoflavivirus* as the outgroup. Both MAFFT v7.490 (65) and MUSCLE v5.1 (70) alignment algorithms were used. Ambiguously aligned regions were removed at a range of thresholds (Table 2) using trimAl v1.4.1 (69).

All phylogenetic trees were inferred using the maximum likelihood method available in IQ-TREE v1.6.12 (71) with 1,000 ultrafast bootstraps. ModelFinder Plus was enabled for inference of the family-level phylogenies (NS3: LG+R, R8,

NS5: LG+F+R10), while substitution model selection was restricted to LG models for the *Pestivirus*-LFG clade because of the selection for the family-level analysis.

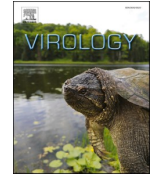
To infer the NPDk phylogenetic tree, the top 500 BLASTp results to the flavivirus-associated NPDk domain were downloaded. Redundant sequences were removed with CD-HIT (68) v4.8.1 at a threshold of 90%, leaving 139 sequences for subsequent analysis. Sequences were aligned with MAFFT v7.490 (65) and MUSCLE v5.1 (70), and the ends were manually trimmed to preserve the most conserved region (length = 269 amino acids, gaps inclusive). Ambiguously aligned regions were removed using trimAl v1.4.1 (69) with the setting "gappypout." This resulted in four alignments: i) MAFFT without ambiguities removed, ii) MAFFT with ambiguities removed, iii) MUSCLE without ambiguities removed, and iv) MUSCLE with ambiguities removed. Maximum likelihood trees were inferred using IQ-TREE v1.6.12 (71) with 1,000 SH-aLRT bootstraps and 1,000 ultrafast bootstraps. The ModelFinder was unrestricted. Selected models are shown in *SI Appendix, Fig. S7*.

Data, Materials, and Software Availability. Alignments and trees referred to in this manuscript as well as the nucleotide and amino acid sequences of the flavivirus-like virus are available on GitHub (https://github.com/maty-petrone/large_flavi) (72). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. This work was funded by a National Health and Medical Research Council (NHMRC) Investigator award and by AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong Special Administrative Region, China. E.C.H. is supported by a NHMRC Investigator Award (GNT2017197). J.G. is supported by the Wellcome Trust and Royal Society through a Sir Henry Dale Fellowship (107653/Z/15/Z) and the MRC-University of Glasgow Centre for Virus Research core support from the Medical Research Council/UKRI (MC_UU_00034/1). We thank Maxime Cailliau for inspiring the name *Maximus pesti*-like virus, and we thank Chris Cooney for helping with sample collection.

1. E. C. Holmes, *The Evolution and Emergence of RNA Viruses* (OUP, Oxford, UK, 2009).
2. X. Hou *et al.*, Artificial intelligence redefines RNA virus discovery. *bioRxiv* [Preprint] (2023). <https://doi.org/10.1101/2023.04.18.537342> (Accessed 14 June 2024).
3. N. Philippe *et al.*, Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281–286 (2013), 10.1126/science.1239181.
4. M. Shi *et al.*, The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018), 10.1038/s41586-018-0012-7.
5. M. Shi *et al.*, Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016), 10.1038/nature20167.
6. A. Zayed *et al.*, Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022), 10.1126/science.abm5847.
7. Y. I. Wolf *et al.*, Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat. Microbiol.* **5**, 1262–1270 (2020), 10.1038/s41564-020-0755-4.
8. Y. Zhou *et al.*, Genomic and biological characterization of a new cytopoxvirus isolated from *Dendrolimus punctatus*. *PLoS ONE* **9**, e113201 (2014), 10.1371/journal.pone.0113201.
9. E. C. Holmes, Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.* **11**, 543–546 (2003), 10.1016/j.tim.2003.10.006.
10. A. Acevedo, L. Brodsky, R. Andino, Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2014), 10.1038/nature12861.
11. R. Sanjuan, J. M. Cuevas, V. Furio, E. C. Holmes, A. Moya, Selection for robustness in mutagenized RNA viruses. *PLoS Genet.* **3**, e93 (2007), 10.1371/journal.pgen.0030093.
12. T. Sanderson *et al.*, A molnupiravir-associated mutational signature in global SARS-CoV-2 genomes. *Nature* **623**, 594–600 (2023), 10.1038/s41586-023-06649-6.
13. J. J. Bull, R. Sanjuan, C. O. Wilke, Theory of lethal mutagenesis for viruses. *J. Virol.* **81**, 2930–2939 (2007), 10.1128/JVI.01624-06.
14. R. Swanstrom, R. F. Schinazi, Lethal mutagenesis as an antiviral strategy. *Science* **375**, 497–498 (2022), 10.1126/science.abn0048.
15. E. Minskaia *et al.*, Discovery of an RNA virus 3'→5' exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5108–5113 (2006), 10.1073/pnas.0508200103.
16. L. D. Eckerle, X. Lu, S. M. Sperry, L. Choi, M. R. Denison, High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *J. Virol.* **81**, 12135–12144 (2007), 10.1128/jvi.01296-07.
17. E. C. Smith, H. Blanc, M. C. Surdel, M. R. Denison, Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: Evidence for proofreading and potential therapeutics. *PLoS Pathog.* **9**, e1003565 (2013), 10.1371/journal.ppat.1003565.
18. R. F. de Souza, L. Aravind, Identification of novel components of NAD-utilizing metabolic pathways and prediction of their biochemical functions. *Mol. Biosyst.* **8**, 1661–1677 (2012), 10.1039/C2MB05487F.
19. P. T. Nga *et al.*, Discovery of the first insect *Nidovirus*, a missing evolutionary link in the emergence of the largest RNA virus genomes. *PLoS Pathog.* **7**, e1002215 (2011), 10.1371/journal.ppat.1002215.
20. E. E. Matsumura, L. Nerva, J. C. Nigg, B. W. Falk, S. Nouri, Complete genome sequence of the largest known flavivirus-like virus, *Diaphorina citri* flavivirus, a novel virus of the Asian citrus psyllid, *Diaphorina citri*. *Genome Announc.* **4**, e00946-16 (2016), 10.1128/genomeA.00946-16.
21. H. Debat, N. Bejerano, Two novel flavivirus-like viruses shed light on the plant-infecting kosoviruses. *Arch. Virol.* **168**, 184 (2023), 10.1007/s00705-023-05813-7.
22. F. Ferron, B. Sama, E. Decroly, B. Canard, The enzymes for genome size increase and maintenance of large (+)RNA viruses. *Trends Biochem. Sci.* **46**, 866–877 (2021), 10.1016/j.tibs.2021.05.006.
23. D. Li, C. M. Liu, R. Luo, K. Sadakane, T. W. Lam, MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015), 10.1093/bioinformatics/btv033.
24. A. Bankevich *et al.*, SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012), 10.1089/cmb.2012.0021.
25. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011), 10.1038/nbt.1883.
26. B. Bushnell, J. Rood, E. Singer, BBMerge—Accurate paired shotgun read merging via overlap. *PLoS ONE* **12**, e0185056 (2017), 10.1371/journal.pone.0185056.
27. J. C. O. Mifsud *et al.*, Mapping glycoprotein structure reveals defining events in the evolution of the *Flaviviridae*. *bioRxiv* [Preprint] (2024), <https://doi.org/10.1101/2024.02.06.579159> (Accessed 18 June 2024).
28. F. Sievers *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011), 10.1038/msb.2011.75.
29. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021), 10.1038/s41586-021-03819-2.
30. M. Mirdita *et al.*, ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022), 10.1038/s41592-022-01488-1.
31. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023), 10.1126/science.ade2574.
32. M. van Kampen *et al.*, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2023), 10.1038/s41587-023-01773-0.
33. M. R. Oliver *et al.*, Structures of the Hepaci-, Pegi-, and Pestiviruses envelope proteins suggest a novel membrane fusion mechanism. *PLoS Biol.* **21**, e3002174 (2023), 10.1371/journal.pbio.3002174.
34. J. E. Blair, *The Timetree of Life* (Oxford University Press, 2009).
35. M. Rastogi, N. Sharma, S. K. Singh, Flavivirus NS1: A multifaceted enigmatic viral protein. *Virol. J.* **13**, 131 (2016), 10.1186/s12985-016-0590-7.
36. K. L. Harvey, V. M. Jarocki, I. G. Charles, S. P. Djordjevic, The diverse functional roles of elongation factor Tu (EF-Tu) in microbial pathogenesis. *Front. Microbiol.* **10**, 2351 (2019), 10.3389/fmicb.2019.02351.
37. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011), 10.1093/nar/gkr367.
38. F. Georgescauld, Y. Song, A. Dautant, Structure, folding and stability of nucleoside diphosphate kinases. *Int. J. Mol. Sci.* **21**, 6779 (2020), 10.3390/ijms21186779.
39. S. Jeudy, A. Lartigue, J.-M. Claverie, C. Abergel, Dissecting the unique nucleotide specificity of mimivirus nucleoside diphosphate kinase. *J. Virol.* **83**, 7142–7150 (2009), 10.1128/jvi.00511-09.
40. M. Shi *et al.*, Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the *Flaviviridae* and related viruses. *J. Virol.* **90**, 659–669 (2016), 10.1128/JVI.02036-15.
41. P. Guardado-Calvo, F. A. Rey, The viral class II membrane fusion machinery: Divergent evolution from an ancestral heterodimer. *Viruses* **13**, 2368 (2021), 10.3390/v13122368.
42. K. Brettle, M. Byrdin, Reaction mechanisms of DNA photolysis. *Curr. Opin. Struct. Biol.* **20**, 693–701 (2010), 10.1016/j.sbi.2010.07.003.
43. P. P. Dzeja, R. J. Zeleznikar, N. D. Goldberg, Adenylate kinase: Kinetic behavior in intact cells indicates it is integral to multiple cellular processes. *Mol. Cell Biochem.* **184**, 169–182 (1998).
44. E. C. Smith, M. R. Denison, Coronaviruses as DNA wannabes: A new model for the regulation of RNA virus replication fidelity. *PLoS Pathog.* **9**, e1003760 (2013), 10.1371/journal.ppat.1003760.
45. E. Harvey, E. C. Holmes, Diversity and evolution of the animal virome. *Nat. Rev. Microbiol.* **20**, 321–334 (2022), 10.1038/s41579-021-00665-x.
46. L. M. Iyer, A. M. Burroughs, V. Anantharaman, L. Aravind, Apprehending the NAD(+) ADPr-dependent systems in the virus world. *Viruses* **14**, 1977 (2022), 10.3390/v14091977.
47. F. Amaro, A. Martin-González, Microbial warfare in the wild—the impact of protists on the evolution and virulence of bacterial pathogens. *Int. Microbiol.* **24**, 559–571 (2021), 10.1007/s10123-021-00192-y.
48. C. Conaco *et al.*, Detection of prokaryotic genes in the *Amphimedon queenslandica* genome. *PLoS ONE* **11**, e0151092 (2016), 10.1371/journal.pone.0151092.
49. M. Martin, CUDAPT removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
50. S. Andrews, FastQC (2023). <https://github.com/s-andrews/FastQC>. Accessed 14 June 2024.
51. E. Kopylova, L. Noé, H. Touzet, SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012), 10.1093/bioinformatics/bts611.
52. J. Charon, J. P. Buchmann, S. Sadiq, E. C. Holmes, RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *Virus Evol.* **8**, veac082 (2022), 10.1093/ve/veac082.
53. B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021), 10.1038/s41592-021-01101-x.
54. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. Sternberg, The PySnpSnp portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015), 10.1038/nprot.2015.053.
55. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.* **12**, 323 (2011), 10.1186/1471-2105-12-323.
56. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012), 10.1038/nmeth.1923.
57. P. T. L. Clausen, F. M. Aarestrup, O. Lund, Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinf.* **19**, 307 (2018), 10.1186/s12859-018-2336-6.
58. V. R. Marcelino *et al.*, CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biol.* **21**, 103 (2020), 10.1186/s13059-020-02014-2.
59. J. Hallgren *et al.*, DeepTMMH predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv* [Preprint] (2022). <https://doi.org/10.1101/2022.04.08.487609> (Accessed 2 June 2024).
60. F. Teufel *et al.*, SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022), 10.1038/s41587-021-01156-3.
61. R. Gupta, S. Brunak, Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput.* **2002**, 310–322 (2002).
62. R. D. Finn *et al.*, The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010), 10.1093/nar/gkp985.

63. J. Wang *et al.*, The conserved domain database in 2023. *Nucleic Acids Res.* **51**, D384–D388 (2023), 10.1093/nar/gkac1096.
64. A. Marchler-Bauer *et al.*, CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222–D226 (2015), 10.1093/nar/gku1221.
65. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013), 10.1093/molbev/mst010.
66. L. Zimmermann *et al.*, A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018), 10.1016/j.jmb.2017.12.007.
67. J. C. O. Mifsud *et al.*, Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians. *Virus. Evol.* **9**, veac124 (2023), 10.1093/ve/veac124.
68. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012), 10.1093/bioinformatics/bts565.
69. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009), 10.1093/bioinformatics/btp348.
70. R. C. Edgar, Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022), 10.1038/s41467-022-34630-w.
71. B. Q. Minh *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020), 10.1093/molbev/msaa015.
72. M. E. Petrone *et al.*, Large flavi. GitHub. https://github.com/mary-petrone/large_flavi. Accessed 4 July 4, 2024.



Australian terrestrial environments harbour extensive RNA virus diversity

Sabrina Sadiq^a, Erin Harvey^a, Jonathon C.O. Mifsud^a, Budiman Minasny^b, Alex B. McBratney^b, Liana E. Pozza^b, Jackie E. Mahar^a, Edward C. Holmes^{a,*}

^a Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia

^b School of Life and Environmental Sciences & Sydney Institute of Agriculture, Faculty of Science, The University of Sydney, Sydney, NSW 2006, Australia

ARTICLE INFO

Handling Editor: Dr. Jasmine Tomar

Keywords:

RNA virus
Environmental
Meta-transcriptomics
Soil
Sediment
Australia
Ecology
Phylogenetics

ABSTRACT

Australia is home to a diverse range of unique native fauna and flora. To address whether Australian ecosystems also harbour unique viruses, we performed meta-transcriptomic sequencing of 16 farmland and sediment samples taken from the east and west coasts of Australia. We identified 2460 putatively novel RNA viruses across 18 orders, the vast majority of which belonged to the microbe-associated phylum *Lenarviricota*. In many orders, such as the *Nodamuvirales* and *Ghabrivirales*, the novel viruses identified here comprised entirely new clades. Novel viruses also fell between established genera or families, such as in the *Cystoviridae* and *Picornavirales*, while highly divergent lineages were identified in the *Sobelivirales* and *Ghabrivirales*. Viral read abundance and alpha diversity were influenced by sampling site, soil type and land use, but not by depth from the surface. In sum, Australian soils and sediments are home to remarkable viral diversity, reflecting the biodiversity of local fauna and flora.

1. Introduction

RNA viruses are ubiquitous, diverse, and can play important roles in multiple ecological processes. Yet a strong focus on viruses of clinical or agricultural significance has limited our understanding of RNA viruses as potentially key components of global ecosystems (French and Holmes, 2020). Similarly, studies of environmental viruses have largely considered marine systems and typically characterised DNA viromes (Trubl et al., 2020), leaving the terrestrial RNA virome understudied (Starr et al., 2019; Chen et al., 2022; Hillary et al., 2022).

Soil environments are complex and diverse, host to an intricate network of macro- and micro-organisms that form unique ecosystems with crucial global functions. An estimated 10^{10} microbes are present in 1 g of soil, with species diversity up to the tens of thousands (Raynaud and Nunan, 2014). Microbial community compositions both affect and depend on such factors as the physicochemical properties of soil (Singh et al., 2009; Dequiedt et al., 2011), land use (Tian et al., 2017), and depth (Xue et al., 2022). The abundance and diversity of organisms in soil implies that there must also be abundant and diverse viruses infecting these hosts, which may play their own roles in soil cycles and maintenance. Indeed, DNA viruses in soil have been estimated from below detection limits in hot deserts to over 10^{10} /g soil in wetlands (Williamson et al., 2017). DNA viruses have documented roles in carbon

metabolism (Trubl et al., 2018; Jin et al., 2019) and phosphorus metabolism (Han et al., 2022), as well as gene transfer to their bacterial hosts, aiding in host extremotolerance and adaptation to environmental stressors (Hwang et al., 2021; Huang et al., 2021). In contrast, the role of RNA viruses in soil ecosystems is less defined (Wu et al., 2021; Hillary et al., 2022).

Recent meta-transcriptomic (i.e., total RNA-sequencing) studies have led to the discovery and characterisation of novel RNA viruses from diverse ecosystems, including aquatic environments (Wolf et al., 2020) and soil (Starr et al., 2019; Chen et al., 2022; Hillary et al., 2022). In particular, meta-transcriptomic analyses of diverse soils and freshwater sediments have shown that non-marine environments are a rich source of viral diversity, with thousands of novel RNA viruses identified in every major lineage of RNA viruses (Chen et al., 2022) and revealing potential functions in soil carbon cycling (Starr et al., 2019) or interactions between hosts spanning different kingdoms (Hillary et al., 2022). Hence, soils provide a valuable means to characterise more of the terrestrial RNA virome. These studies have also led to a deeper understanding of RNA viruses in a broader ecological context, revealing the impacts of human land use, physicochemical properties, and geographical features of the sampling environment on viral abundance and diversity (French et al., 2022; Chen et al., 2022; Hillary et al., 2022).

Australia has been isolated from other continents for tens of millions

* Corresponding author.

E-mail address: edward.holmes@sydney.edu.au (E.C. Holmes).

<https://doi.org/10.1016/j.virol.2024.110007>

Received 11 October 2023; Received in revised form 20 January 2024; Accepted 22 January 2024

Available online 8 February 2024

0042-6822/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

of years and, as such, has developed many diverse and unique biomes. Flora and fauna have adapted to the continent's flat, dry, fire-prone, and nutrient-poor landmass, resulting in a remarkable level of biodiversity that is unique to Australia (Steffen, 2009). For example, many native Australian plants have hardened foliage (sclerophylly) and evergreen characteristics, causing herbivores to have slower metabolisms and reptiles to have predominantly invertebrate-based diets (Steffen, 2009). Insects play an active role in the dispersal of seeds and are key consumers of leaves in the absence of any native ruminant species (Steffen, 2009). The majority of Australia's mammals (87%), reptiles (93%), frogs (94%), and vascular plants (92%) found across the country are endemic (Chapman, 2009). Australia is also a "megadiverse" country, one of 17 that together comprise over 70% of the world's total biodiversity (Williams et al., 2001). With so much diversity in its animals, plants, and environments, it can be assumed that Australian microbes - including viruses - harbour similar levels of diversity.

Little is known about the RNA viromes of Australian soil and sediment systems. Herein, we aimed to provide an initial snapshot of the diversity, abundance, and composition of RNA viruses in Australian environmental samples. In particular, we asked whether the unique flora and fauna of Australia is mirrored by the presence of unique soil viruses. Accordingly, we performed meta-transcriptomic sequencing of 16 geographically and ecologically distinct farmland soil and riverbank sediment samples taken from New South Wales (NSW) and Western Australia (WA), respectively, two Australian states separated by approximately 3000 km.

2. Materials and methods

2.1. Sample collection

Soil samples from NSW were collected in December 2021 from vertosol, sodosol, and chromosol soil types (Isbell, 2016) in cropping/pasture fields and native vegetation at depths of 0–5 cm, 5–15 cm, and 15–30 cm at each site. NSW samples were collected from the University of Sydney-owned 'Nowley Farm' on the Liverpool plains because of the presence of multiple soil types and land uses. In the case of WA, sediment samples were collected in triplicate during April and May 2022 from four points along the riverbanks of the Swan River, Canning River, and Denmark River. Maps of the NSW and WA sampling sites are shown in Fig. 1, with more detail on the properties of the sampling sites

provided in Table S1. All samples were packed into sterile 50 ml conical tubes and either stored on ice (NSW samples) or between -30°C and -15°C (WA samples) until transported to a -80°C freezer for long-term storage.

2.2. RNA extraction, library construction, and sequencing

Total RNA was extracted from the samples using the RNeasy PowerSoil Total RNA Kit (Qiagen) as per the manufacturer's instructions. Extracted RNA was quantified using the Qubit RNA high sensitivity (HS) Assay Kit on the Qubit Fluorometer v3.0 (Thermo Fisher Scientific) and stored at -80°C prior to library construction and sequencing. Libraries were constructed using the Illumina Stranded Total RNA library preparation protocol and rRNA was removed using the Ribo-Zero Plus rRNA depletion kit (Illumina). Libraries were sequenced on the Illumina NovaSeq 6000 platform (paired-end, 150 bp). Library preparation and sequencing was performed by the Australian Genome Research Facility (AGRF).

2.3. Data processing and abundance measurements

Data processing and subsequent phylogenetic analyses utilised the BatchArtemisSRAMiner workflow (Mifsud, 2023). Sequence reads were adaptor- and quality-trimmed using Trimmomatic (v0.38) (Bolger et al., 2014), then assembled into contigs using MEGAHIT (v1.2.9) (Li et al., 2015) employing default assembly parameters. No eukaryotic or bacterial reads were filtered prior to assembly. Sequence quality was checked for both raw reads and trimmed reads using FastQC (v0.11.8) (Andrews, 2010). Assembled contigs were compared to an in-house curated database of viral sequences using DIAMOND BLASTX (v2.0.9) (Buchfink et al., 2015) with an e-value of 1×10^{-5} for sensitivity. This database was curated in October 2023 by querying NCBI's protein sequence database for entries with 'Viruses' in the organism field and built using the 'makedb' function in DIAMOND BLAST. Contigs returning positive hits were compared to the non-redundant (nr) protein database as of October 2023, using DIAMOND BLASTX to identify false positives. Contigs with hits to viral sequences were retained and sorted by the taxonomy of the closest relative to the level of virus family (or to the most specific available taxonomic level assigned to the closest relative).

The relative abundance of viral transcripts within each library was

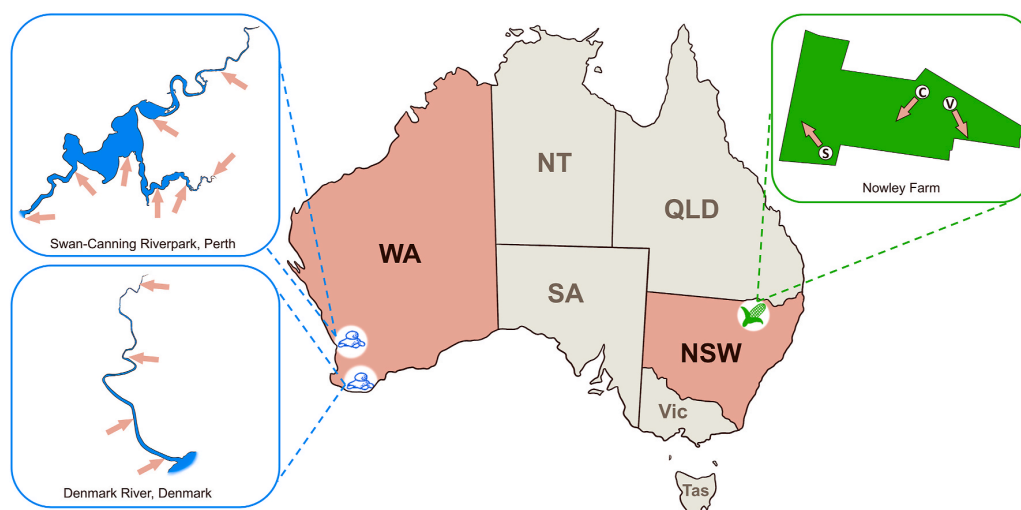


Fig. 1. Map of Australia showing riverbank sediment sampling sites along the Swan-Canning Riverpark (Perth, WA) and Denmark River (Denmark, WA), as well as farmland soil sampling sites in Nowley Farm, NSW. In the inset boxes, each sampling site is indicated by an arrow. In the case of the Nowley Farm samples, soil type is indicated by a letter: V - vertosol, C - chromosol, S - sodosol.

estimated using RSEM (v1.3.1) (Li and Dewey, 2011), calculated as the expected count of viral contigs divided by the total raw read count x 100. Virome compositions for each library were calculated as the expected count of contigs aligning to each viral family/order/phylum as a proportion of the total expected count of contigs aligning to sequences from the *Riboviria* (RNA viruses). The community compositions of potential host organisms in each library were estimated using CCMetagen (Clau-sen et al., 2018; Marcelino et al., 2020) by mapping the assembled contigs to NCBI's nucleotide database (nt) as of June 2019. Host organisms were largely grouped at the phylum level, or at higher taxonomic levels in the case of less frequently appearing organisms, such as those in the Metazoa or Viridiplantae. Virome and host organism community composition figures were generated in R using ggplot2 (v3.3.3) (Wickham, 2016).

Viral alpha diversity was described by: (i) richness, or the number of viral taxa in each library, (ii) Shannon diversity, which accounts for both richness and the evenness of their distribution, and (iii) "true" diversity or effective Shannon diversity, calculated as the exponential of each respective Shannon diversity index. These indices were calculated in RStudio (v4.1.1717) (RStudio Team, 2019), R (v4.1.0) (R Core Team, 2021) using an adaptation of the Rhea alpha diversity script (Lagkou-vardos et al., 2017; Wille et al., 2018). RNA virus read abundance and alpha diversity figures were generated in R using ggplot2 (v3.3.3) (Wickham, 2016). Soil characteristics were tested for their influence on viral read abundance and alpha diversity using generalised linear models. The significance of these models was evaluated using χ^2 tests and significant differences between pairs of groups within each ecological property were determined using post-hoc Tukey tests. To determine the factors impacting soil virus read abundance and alpha diversity, we conducted a best-subsets regression analysis. Models were evaluated using the Akaike information criterion (AIC) and the model with the lowest AIC was selected as the best-fit model. Soil characteristics included soil type (chromosol versus sodosol), land use (native vegetation, cropping, and pasture), environment (a combination of soil type and land use) and depth from the surface, and the amount of nucleic acid extracted. As sequence data was obtained from only one depth measurement for chromosol and sediment environments, only sodosol libraries were included in the comparison of viral read abundance and alpha diversity across different depths. Viral read abundance and alpha diversity were compared between the different sampling sites for sediment libraries. Due to the numerous confounding factors that may influence virus read abundance and alpha diversity, including the surrounding environment, geographical location, climate, and storage conditions prior to extraction, detailed comparisons between the NSW farmland soil and WA sediment samples were not conducted.

2.4. Identification of novel viruses and phylogenetic analysis

Contigs with DIAMOND BLASTX hits to the viral RNA-dependent RNA polymerase (RdRp) that were greater than 600 nucleotides (nt) in length (arbitrarily set so as to minimise inaccurate classification) and with less than 99% amino acid identity to their closest previously published relative were translated to amino acid sequences. The standard genetic code (i.e., code Table 1) was used in most cases, with the exception of 133 sequences from the family *Mitoviridae* (phylum *Lenarviricota*) for which the mitochondrial genetic code (i.e., code table 4) is more likely to be biologically accurate, and indeed provided ORFs of expected lengths where the standard code led to truncation. Translated sequences were checked for the presence of the conserved A, B, and C motifs that characterise viral RdRp. Contigs fulfilling these conditions were included for phylogenetic analysis as these likely represent RNA virus sequences.

Potentially novel virus sequences were aligned with members of the family, order, or multi-family clade of their respective closest DIAMOND BLASTX hit using MAFFT (v7.402) (Katoh and Standley, 2013). Sequence alignments were trimmed using trimAl (v1.4.1)

Table 1

Sequence data sets used for phylogenetic analysis (Figs. 6–8; Supplementary Figs. 2–22).

Virus group	Number of sequences			Alignment length	
	Total	Novel	Reference	Untrimmed	Trimmed
<i>Hepelivirales</i> + <i>Tymovirales</i> (Fig. 6A, Fig. S2)	224	16	208	8750	711
<i>Martellivirales</i> (Fig. 6B, Fig. S3)	249	22	227	9945	706
<i>Nodamuvirales</i> (Fig. S4)	259	101	158	6153	726
<i>Nodaviridae</i> (Nodamuvirales sub-tree) (Fig. 6C)	118	15	103	2839	704
<i>Tolivirales</i> (Fig. S5)	361	127	234	8263	702
<i>Leviviricetes</i> (Fig. S6)	2324	894	1430	19,225	711
<i>Mitoviridae</i> (Fig. S7)	1067	441	626	8555	710
<i>Narnaviridae</i> , <i>Botourmiaviridae</i> , <i>Narliviridae</i> (Fig. S8)	1157	445	712	12,089	725
<i>Marnaviridae</i> (Fig. S9)	396	70	326	8741	708
<i>Picornaviridae</i> , <i>Polycipiviridae</i> , <i>Solinviridae</i> (Fig. 7A, Fig. S10)	221	13	208	12,632	707
<i>Dicistroviridae</i> , <i>Secoviridae</i> , <i>Iflaviridae</i> (Fig. 7B, Fig. S11)	286	24	262	10,439	710
<i>Sobelivirales</i> (Fig. S12)	156	49	107	6283	714
<i>Astroviridae</i> (Fig. 8A, Fig. S13)	113	47	66	5966	704
<i>Hypoviridae</i> (Fig. S14)	59	3	56	8280	704
<i>Partiti-Picobirna</i> (Fig. S15)	1397	101	1296	3749	727
<i>Fusariviridae</i> (Fig. S16)	40	5	35	2688	702
<i>Monjiviricetes</i> (Fig. S17)	115	2	113	4569	704
<i>Bunyavirales</i> (Fig. S18)	42	21	21	5027	704
<i>Cystoviridae</i> (Fig. S19)	52	6	46	2308	702
<i>Reovirales</i> (Fig. S20)	76	5	71	2712	716
<i>Ghabrivirales</i> (Fig. S21)	289	64	225	7183	704
<i>Giardiavirus-like clade</i> (<i>Ghabrivirales</i> sub-tree) (Fig. 8B)	101	48	53	4911	702
<i>Birnaviridae</i> (Fig. S22)	52	4	48	2643	704

(Capella-Gutierrez et al., 2009) to retain only the most conserved amino acid positions (between 702 and 726 residues in length) and to remove ambiguously aligned regions (see Table 1). Alignments were also visually assessed to identify and remove poorly aligned sequences. Maximum likelihood phylogenetic trees were then estimated on each of these alignments using IQ-TREE (v1.6.12) (Nguyen et al., 2015), employing the Le-Gascuel (LG) model of amino acid substitution – determined using ModelFinder within IQ-TREE – with 1000 S H-aLRT replicates (Anisimova and Gascuel, 2006) to assess node support. All trees were visualised in R using packages 'ape' (v5.5) (Paradis and Schliep, 2019) and 'ggtree' (v3.0.2) (Yu et al., 2017). Probable host organisms for novel viruses were predicted based on the hosts of the established viruses with which they clustered most closely.

3. Results

3.1. Data generation

We generated 26 sequencing libraries, 16 of which were from NSW farmland soil samples, with the remaining 10 from sediment taken from Denmark River in WA. Eighteen soil samples, representing three distinct soil types (vertosol, chromosol, and sodosol), two categories of land use (native vegetation or agricultural), and three depths (0–5 cm, 5–15 cm, and 15–30 cm), were taken from NSW farmland environments (Supplementary Table 1). Vertosol, which is high in smectitic clay and has

high agricultural potential, and chromosol, which has a loamy texture and moderate agricultural potential, were both collected from sites containing native vegetation or crops. Sodosol has a sandy surface texture with a high concentration of sodium and is nutrient poor. Due to the generally low capacity for crop growth, sodosol samples were taken from pasture and native vegetation sites. We were able to successfully extract RNA from six of these 18 samples, five of which we extracted in technical triplicates, totalling 16 RNA libraries for meta-transcriptomic sequencing. No RNA was able to be extracted from vertosol or the 15–30 cm depths, and success in chromosol samples was restricted to the 0–5 cm depth despite multiple attempts on each sample.

The success of RNA extraction was similarly limited in the WA riverbank sediment samples. The 24 samples taken from a total of eight sites in the Swan-Canning Riverpark system (Perth, WA) were generally coarse and sandy in texture, while the 12 samples taken from four sites along Denmark River (Denmark, WA) were finer and muddier in comparison. No extractions from the Swan-Canning Riverpark system were successful, and RNA was only extracted from all three biological replicates in two of the four sites along Denmark River, for a total of 10 libraries able to be sequenced. Samples from which RNA was successfully extracted and sequenced are described in [Supplementary Table 1](#).

We generated approximately 5.26 billion paired-end reads from the 26 libraries successfully sequenced in this study, ranging from 133.5 million reads (native chromosol, NSW; library CN1A) to 396 million

reads (native sodosol, NSW; library SN2A) per library. From these data, 6.2 million contigs were assembled, with contig numbers from individual libraries ranging from 85,100 (native sodosol, NSW; library SN2A) to 701,601 (riverbank sediment, WA; library KCAP1), with a median value of approximately 182,600.

3.2. RNA virome composition

Across the data set as a whole there were 10,561 contigs greater than 600 nucleotides (nt) in length that had DIAMOND BLASTX hits to sequences from the *Riboviria* (i.e., RNA viruses). However, a proportion of these contigs did not robustly align to reference sequences and could not be reliably assigned to any RNA virus taxa. Thus, a subset of 9126 viral contigs was retained for analyses. The number of viral contigs per library ranged from 53 (native sodosol, NSW; SN2A) to 1336 (riverbank sediment, WA; KCAP3) and the majority of contigs (5823 out of 9126) had less than 50% amino acid identity to reference sequences ([Table S2](#)). The sample with the lowest RNA viral read abundance was NSW sodosol pasture (library SP2C) at 0.018%, while the highest abundance of 1.279% was found in WA riverbank sediment (library KCAP3) ([Fig. 2A](#)). Interestingly, some of the lowest and highest diversity indices were observed in samples taken from the same WA riverbank sites, namely, KCAP2 and MP2 (lowest) and MP1 and KCAP1 (highest) ([Fig. 2B](#)). While viral read abundance differed significantly between sampling locations

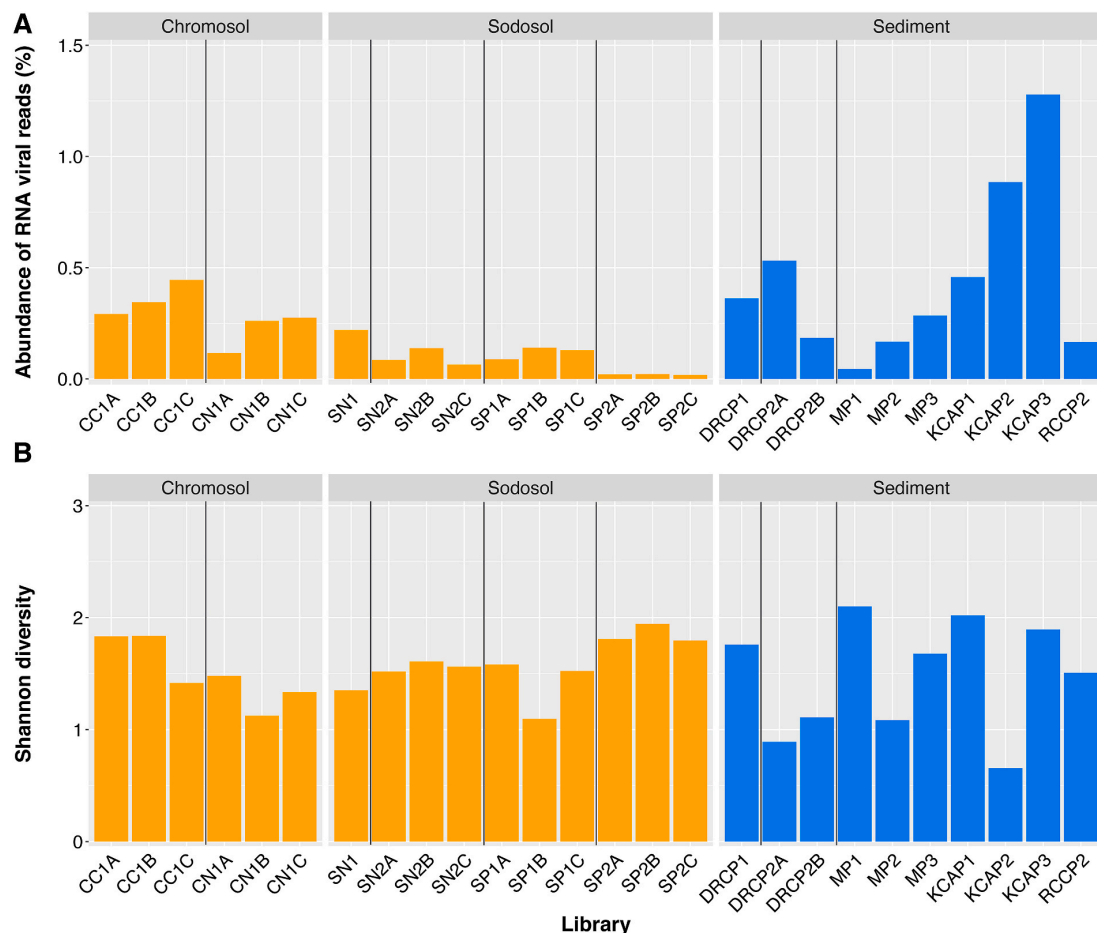


Fig. 2. (A) Abundance of RNA virus reads as a proportion of the number of total reads and (B) RNA virus Shannon diversity indices of the meta-transcriptomic sequencing libraries generated from samples taken from NSW farmland soil (left and middle, in orange) and WA riverbank sediment (right, in blue). Letters (A, B, C) at the end of each library name indicate technical replicate extractions of the same sample and are grouped within vertical lines.

($p = 0.008$), Shannon diversity did not ($p = 0.585$).

Sampling environment refers to the combination of soil type (chromosol, sodosol) and land use (native, cropping, pasture), which we found to be associated with soil virus read abundance and richness ($p < 0.001$ in both cases) (Fig. 3A). To explore which factors were most likely to contribute to this effect, best-fit models were estimated considering soil type and land use separately. Although soil type alone did not have a significant effect on richness ($p = 0.104$), the best fit model describing richness included both soil type and land use ($p < 0.001$), with agricultural soils (cropping and pasture) harbouring greater species richness than native soils in both chromosol and sodosol soil types (Fig. 3B). Soil

type did not have a significant effect on richness ($p = 0.58$). Land use was also associated with viral read abundance ($p < 0.001$), in which cropping soil had the highest read abundance, pasture soil had the lowest, and native soils fell in the middle (Fig. 3B). However, it is difficult to determine if the agricultural purpose of the soil (i.e., cropping or pasture) was truly impacting viral read abundance as soil type also had a significant influence on viral read abundance ($p < 0.001$), which was higher in chromosol than in sodosol (Fig. 3C). As the cropping soils were only chromosol and the pasture soils were only sodosol, the pattern of read abundance observed in Fig. 3C may have been due to the influence of soil type rather than agricultural purpose. In addition,

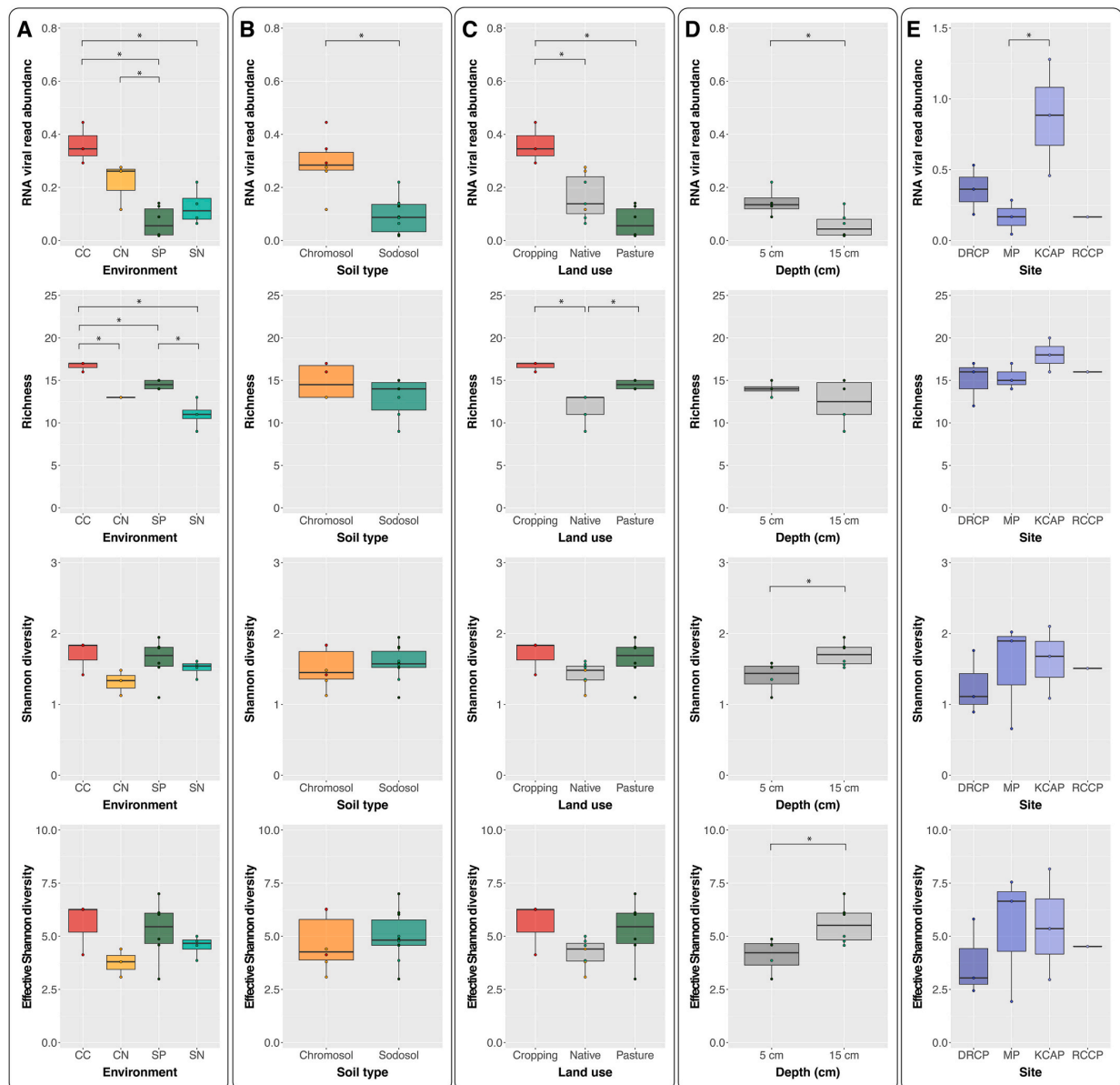


Fig. 3. Viral read abundance, richness, Shannon diversity, and effective Shannon diversity indices of NSW farmland soils plotted against (A) environment as a combination of soil type and land use, (B) soil type, (C) land use, (D) depth, and the same indices of WA riverbank sediments plotted against (E) sampling site. Asterisks indicate significant differences ($p < 0.05$) between pairs of ecological properties as determined by post hoc Tukey tests. Shorthand labels for environment indicate CC = cropping chromosol, CN = native chromosol, SP = pasture sodosol, and SN = native sodosol. Circles representing each library in columns B-D are coloured by sampling environment as per column A.

while native chromosol had lower viral read abundance than cropping chromosol, native sodosol had higher viral read abundance than its counterpart used for pasture (Fig. 3A). This suggests an influence of both soil type and the purpose of agricultural land use, supported by the inclusion of both variables in the best-fit model for viral read abundance. No ecological factors were associated with the Shannon and effective Shannon diversity indices of soil when including both soil types ($p = 0.165$ – 0.669 and $p = 0.109$ – 0.671 , respectively) (Fig. 3A–C), although sampling depth did significantly impact read abundance, Shannon diversity, and effective Shannon diversity indices of sodosol viruses (0.009 – 0.029) (Fig. 3D). Thus, the best fit models for these three indices included soil type, land use, and depth.

In the case of the sediment libraries, we also measured the effect of sampling site on read abundance, richness, Shannon diversity, and effective Shannon diversity (Fig. 3E). Only sampling site influenced viral read abundance ($p = 0.006$), with libraries from the densely vegetated KCAP site having higher viral read abundance than the other sediment sites, and higher than any soil libraries (Figs. 2 and 3E).

Also of note was that the amount of extracted RNA appeared to be partially dependent on sampling environment (Supplementary Figs. 1A and 1B). Native chromosol libraries had significantly higher RNA yields than other libraries, ranging between 459.8 and 564.5 ng/g, whereas the next highest yield was 219.7 ng/g from library KCAP3. Despite this association, RNA yields at the time of extraction did not correlate with read abundance or alpha diversity calculated from the subsequently generated sequencing data (Supplementary Figs. 1C–F).

Relative virus read abundances (i.e., the total expected counts of each viral group as a proportion of all *Riboviria* reads in each library) are shown in Fig. 4A. The phylum *Lenarviricota* and orders *Tolivirales* and *Nodamuvirales* were present in all libraries, with the *Lenarviricota* generally comprising a large proportion, if not the majority, of reads. A notable exception was library DRCP2B (riverbank sediment, WA) where unclassified *Picornavirales* sequences comprised the majority of reads. As they were the most abundant groups, we classified the virome compositions of the phylum *Lenarviricota* and order *Picornavirales* to the class and family level, respectively (Fig. 4). In general, *Picornavirales* sequences comprised a greater proportion of reads in sediment than soil libraries. Furthermore, these sequences were mostly unclassified *Picornavirales* in sediment libraries, whereas in the soil libraries over half of the reads comprising *Picornavirales* sequences were from the family *Dicistroviridae*. *Marnaviruses* also appeared more frequently in the sediment than the soil libraries, likely due to their typically aquatic hosts, although cropping chromosol library CC1A had a high proportion of *marnaviruses* relative to the other *Picornavirales* groups and other soil libraries.

The *Lenarviricota* were abundant in all libraries, as were bacteria (Fig. 4). However, there was an overall decrease in the proportion of fungal-associated class *Amabiliviricetes* and family *Mitoviridae* in the sediment compared to the soil libraries. Sediment libraries were instead dominated by bacteriophage of the class *Leviviricetes*, suggesting a switch in microbial community composition towards bacteria in sediment environments (Fig. 4A). It is interesting to note that pasture sodosol samples taken 0–5 cm from the surface (libraries SP1A–C) were an exception to this trend, in which the *Leviviricetes* represented a much higher proportion of *Lenarviricota* sequences than in the other soil libraries (Fig. 4A). The composition of potential host organisms reflected this switch in microbial community composition: although all libraries were largely dominated by bacteria, soil libraries had higher relative proportions of fungi (2.5–25.8%) than sediment libraries (0.3–4.5%) (Fig. 4B). The most predominant bacterial phyla included *Pseudomonadota* (10.2–32.9%), *Plantomycetota* (8.3–28.9%), and *Actinobacteriota* (1.5–20.1%), although the latter of the three was less prevalent in sediment libraries (1.5–6.8%). *Acidobacteriota* also comprised a notable proportion of sodosol communities (6.9–14.8%). In contrast, sediment libraries had higher relative proportions of hits to the domain Archaea than soil (4.1–45.9% compared to 0.6–4.6%), most of

which belonged to the phylum *Methanobacteriota* (Fig. 4B). Interestingly, library MP2, which harboured the highest relative proportions of *Lenarviricota* and *Leviviricetes* reads (95.6% and 84.8%, respectively), was also the library with the largest proportion of archaea (31.9%) (Fig. 4).

3.3. Phylogenetic analysis of novel virus species

We identified 2460 novel virus RdRp sequences across all five RNA virus phyla, including those with positive-sense single-stranded RNA genomes (*Astroviridae*, *Picornavirales*, *Sobelivirales*, from the phylum *Pisuviricota*; *Alsuviricetes*, *Nodamuvirales*, and *Tolivirales* from the phylum *Kitrinoviricota*; and the phylum *Lenarviricota*), negative-sense single-stranded RNA genomes (*Bunyavirales* and *Monjiviricetes* from the phylum *Negarnaviricota*), and double-stranded RNA genomes (*Hypoviridae*, *Partitiviridae*, *Picobirnaviridae*, and *Fusariviridae* from the phylum *Pisuviricota*; *Cystoviridae*, *Reovirales*, and *Ghabrivirales* from the phylum *Duplornaviricota*; and *Birnaviridae*) (Table 1, Fig. 5, Supplementary Table 2). By predicting host organisms based on phylogenetic clustering, we suggest that viruses with bacterial, fungal, and protist hosts comprised 38.7%, 41.9%, and 4.9% of novel sequences, respectively, and those with plant and invertebrate hosts comprised 2.6% and 7.6% of novel sequences, respectively. We were unable to confidently assign a host organism to 4.3% of novel sequences as these did not cluster closely enough to any reference sequences with a known host. Notably, no likely vertebrate-associated viruses were observed. The majority of novel sequences (1,780) fell into the microbe-associated phylum *Lenarviricota*. We now describe each of these groups in turn.

3.3.1. Positive-stranded RNA viruses – phyla *Kitrinoviricota*, *Lenarviricota*, and *Pisuviricota*

3.3.1.1. Class *Alsuviricetes* (*Kitrinoviricota*) and *bastroviruses* (*Pisuviricota*). Eight novel sequences clustered within the order *Tymovirales* (Fig. 6A, Supplementary Fig. 2). Four appeared to be novel *marafiviruses* (*Tymoviridae*), one fell basal to the family *Tymoviridae*, and the other three clustered with the divergent mycoflexivirus *Botrytis virus F* (*Gammaflexiviridae*). Within the *Hepelivirales*, four viruses were identified from riverbank sediment samples, three of which clustered with *bastroviruses*, which, despite being classified as members of the *Astroviridae*, encode a *Hepeviridae*-like RdRp protein due to a recombination event (Oude Munnink et al., 2016). Given the broad host range of *bastroviruses*, the host of these novel *bastroviruses* cannot be confidently inferred. Several novel sequences were also identified in the order *Martellivirales* (Fig. 6B, Supplementary Fig. 3). These included an *ampelovirus* (*Closteroviridae*), an *ilarvirus* (*Bromoviridae*), and 19 *alphaendornaviruses* (*Endornaviridae*).

3.3.1.2. *Nodamuvirales* (*Kitrinoviricota*). Our phylogenetic analysis of the *Nodaviridae* revealed a conflict with the current taxonomy (Fig. 6C). A clade including the *alphanodavirus* species *Pariacoto virus* (NP_620,109) formed a sister group to the genus *Betanodavirus* rather than clustering with the remaining *alphanodaviruses*. This clade comprised several environmental *noda*-like viruses, including the novel *noda*-like viruses *Sopodan virus*, *Ripokam virus*, *Sopoban virus*, *Chrocolird virus*, and *Ripokret virus*, and may represent the existence of a third genus within the *Nodaviridae*.

A total of 15 novel viruses fell within the *Nodaviridae*, four of which fell within a group of *noda*-like viruses forming a sister clade to the *Nodaviridae* (Fig. 6C, Supplementary Fig. 4). This sister clade – provisionally named *Ripobluf noda*-like group – may represent a fourth genus within the *Nodaviridae*. However, the largest expansion of virus diversity was in a sister clade to the *Nodaviridae* that likely comprises a new family. This putative family included many *noda*-like viruses previously identified from soil (Chen et al., 2022) and invertebrates (Shi et al.,

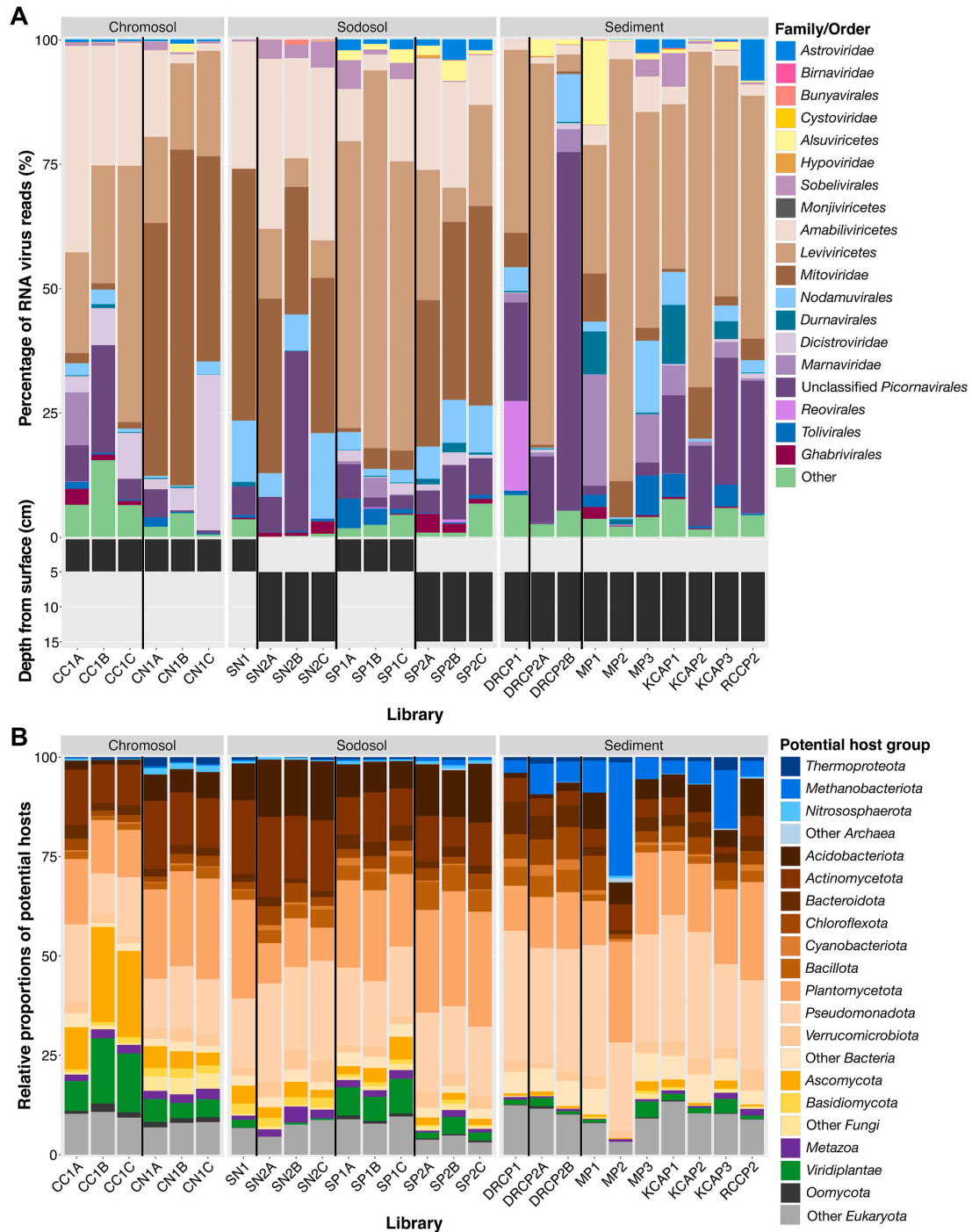


Fig. 4. (A) Virome compositions (family/order/phylum) and (B) potential host organism community compositions (phylum/kingdom/superkingdom) for each sampling library. Relative proportions for viruses were determined by the total number of reads corresponding to contigs with DIAMOND BLASTX hits in each virus clade and host community compositions were estimated using CCMetagen. Black bars underneath virome compositions represent the sampling depth of each sample (cm). Letters (A, B, C) at the end of each library name indicate technical replicate extractions of the same sample and are grouped within vertical lines. The more abundant phylum/order – *Lenarviricota* and *Picornavirales* – are further broken down into families or classes and are represented in shades of the same colour.

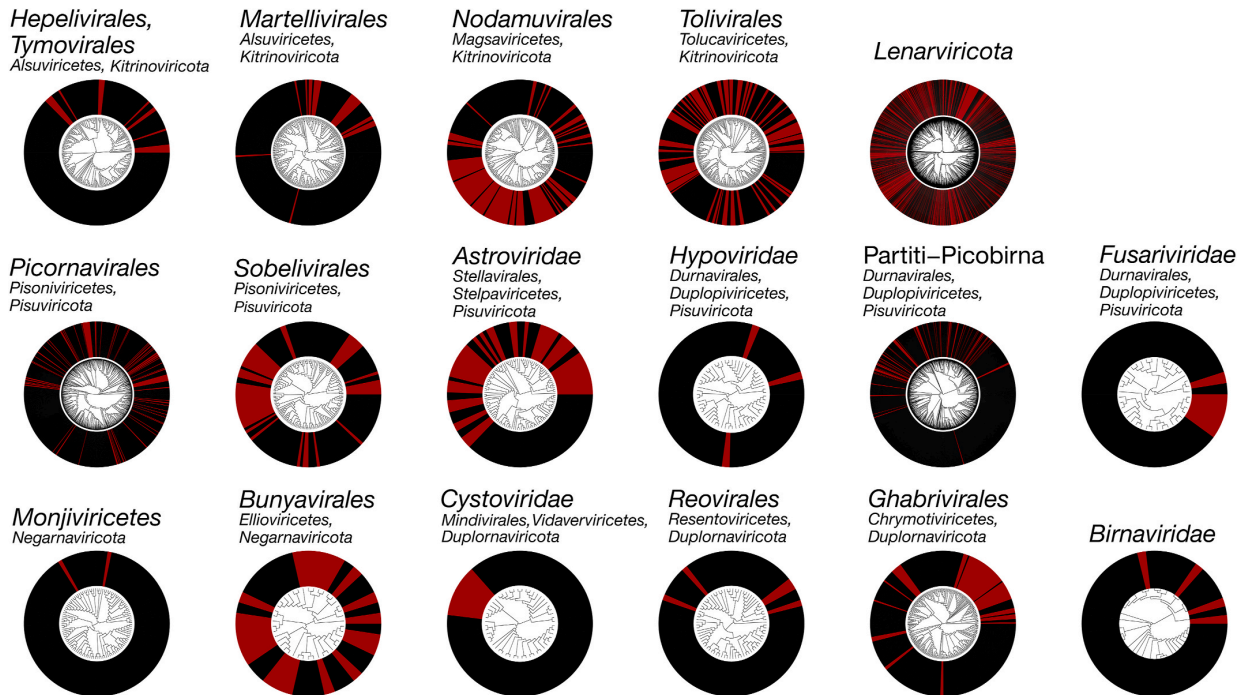


Fig. 5. Phylogenetic trees of the RdRp amino acid sequences of viruses identified in this study, with reference sequences. Known viruses are shown in black, while putative novel viruses identified here are shown in red. All trees are midpoint rooted for clarity only. Individual phylogenetic trees for each clade are shown in [Supplementary Figs. 2–22](#). Details on the alignments used to estimate these phylogenies are shown in [Table 1](#). Phylogenies of the phylum *Lenarviricota* and order *Picornavirales* were estimated for the purposes of visualising novel viruses but contain too much sequence divergence for robust alignments such that the phylogenies presented in the supplement are divided into multiple, smaller groups containing 1–3 related families each.

2016) in China, and a total of 86 novel viruses from all environments sampled ([Supplementary Fig. 4](#)).

3.3.1.3. Tolivirales (Kitrinoviricota). Novel *Tolivirales* sequences were also identified in all environments sampled, although 23 of the 28 sequences clustering within the plant-associated *Tombusviridae* were identified in riverbank sediment libraries ([Supplementary Fig. 5](#)). A total of 27 sequences (24 from riverbank sediments) formed sister lineages to the subfamilies *Procedovirinae* and *Calvusvirinae*.

All sampling environments were represented in the 20 novel sequences clustering within a clade of diverse tombus-like viruses ([Supplementary Fig. 5](#)). This clade also included the sole species within the *Carmotetraviridae* - Providence virus (AMQ67162) - a unique virus isolated from arthropod (lepidopteran) tissue that is also capable of replicating in plant and mammalian cell lines ([Jiwaji et al., 2019](#)). Finally, 28 novel sequences fell into the most divergent clade in this phylogeny, which largely comprised previously identified environmental viruses, as well as some sourced from invertebrates and animal faecal samples ([Supplementary Fig. 5](#)).

3.3.1.4. Lenarviricota. A remarkable 1780 novel sequences were identified within the phylum *Lenarviricota*, comprising the majority of novel viruses found in this study. Novel sequences were sourced from sediment and both farmland soil types and land uses. Due to the high level of sequence divergence across the *Lenarviricota*, phylogenetic trees were estimated on sub-alignments of (i) the class *Leviviricetes* (i.e., bacteriophage; [Supplementary Fig. 6](#)), (ii) the family *Mitoviridae* ([Supplementary Fig. 7](#)), and (iii) the families *Narnaviridae*, *Botourmiaviridae* and the newly proposed *Narliviridae* ([Supplementary Fig. 8](#)).

In total, 894 novel leviviruses were identified and from every sampling environment. In many large clades, the majority of viruses were

novel sequences identified in this study ([Supplementary Fig. 6](#)). Similarly, several clades largely comprising novel mitoviruses were identified in soil environments. Of particular interest was the Sopaluz mitovirus-like group, comprising seven native chromosol, five pasture sodosol, and one sediment viruses that clustered with a mitovirus - Kinsystemes virus (QQM16243) - identified in a termite, although their use of the mitochondrial genetic code suggests they likely infect the fungal hosts that are typical of mitoviruses ([Supplementary Fig. 7](#)). Similarly, the basal lineages of the Ripanov mito-like group contained viruses likely associated with the microbiomes of arthropods and vertebrates, although as these viruses were obtained from metagenomic studies their true host association is unclear. The Ripanov mito-like group also contained 61% of the total number of sediment-associated mitoviruses, with 129 of the 135 novel sequences in this clade being sourced from sediment samples ([Supplementary Fig. 7](#)).

In the *Narnaviridae*-*Botourmiaviridae*-*Narliviridae* phylogeny multiple new, distinct clades comprising viruses from different sampling sources were observed, typically biased towards sediment or soil environments. These novel viruses typically shared <50% amino acid identity with any reference sequence. The largest clade of entirely new viruses was identified within the *Narliviridae*, with this group of 117 novel viruses potentially representing a genus within this family ([Supplementary Fig. 8](#)). The viruses from this group included 24 from sediment, 22 and 18 from pasture and native sodosol, respectively, and 46 and 8 from cropping and native chromosol, respectively ([Supplementary Fig. 8](#)). Most of the novel *Botourmiaviridae* identified here were associated with soil samples. This contrasted with most of the viruses detected from other families within the *Lenarviricota* that were predominantly associated with sediments ([Supplementary Figs. 6–8](#)). Indeed, only 14 of the 81 novel botourmiaviruses were identified in sediment, whereas the percentage of sediment-associated sequences in the other families

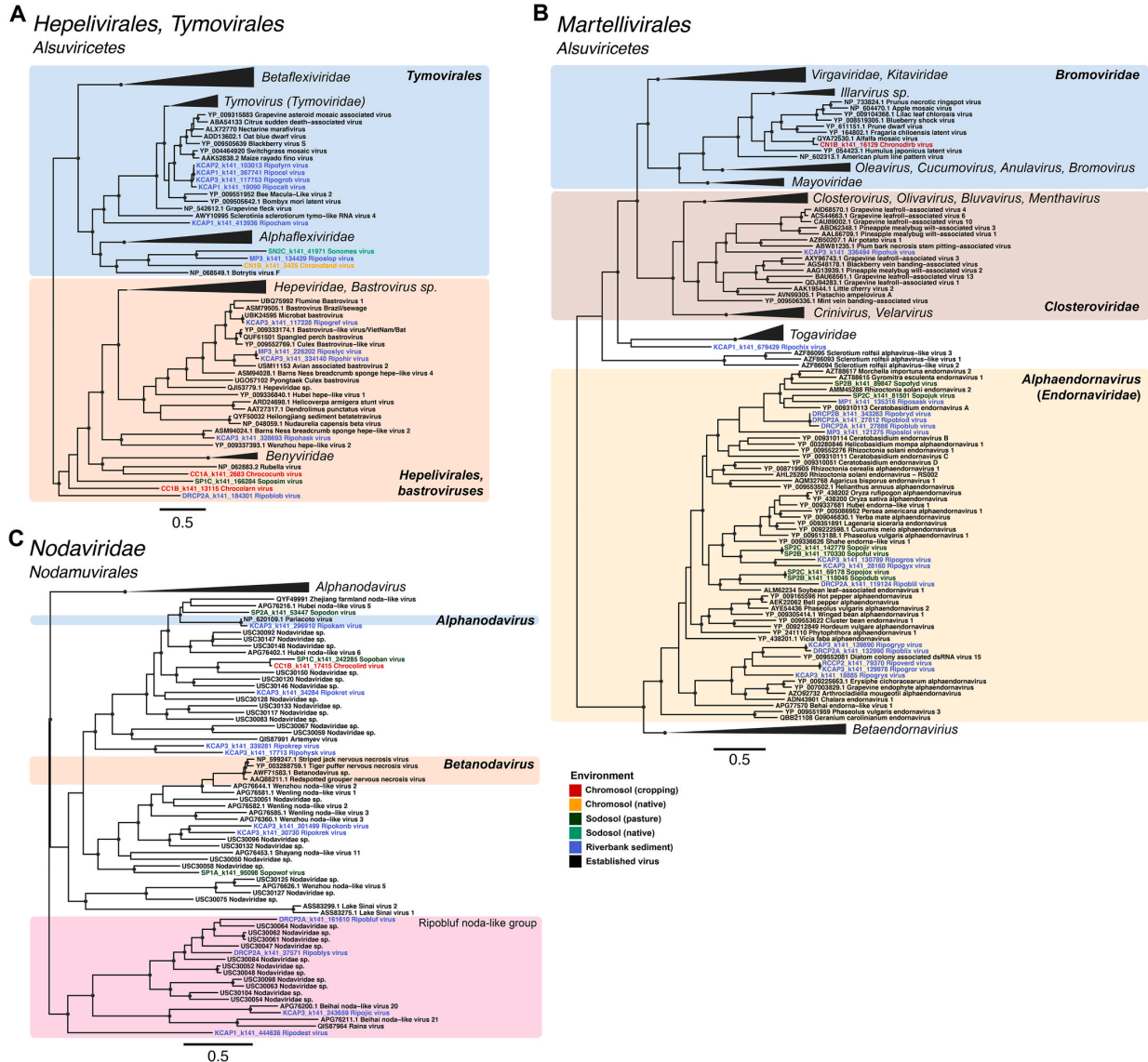


Fig. 6. Phylogenetic trees of RdRp amino acid sequences from (A) the orders *Hepelivirales* and *Tymovirales*, (B) the order *Martellivirales*, and (C) the family *Nodaviridae*. Known viruses are shown in black, while putative novel viruses identified here are coloured by sampling environment. Trees are midpoint rooted for clarity only and branch lengths are scaled according to the number of amino acid substitutions per site. Black circles represent node support $\geq 80\%$ using 1000 S-H-LRT replicates. Phylogenies with collapsed branches expanded are shown in [Supplementary Figs. 2–4](#).

ranged from 42 to 60%. A tendency for soil over sediment has previously been observed in the *Botourmiaviridae* (Chen et al., 2022). We also identified 19 novel *Narnaviridae* species in sediment, 8 and 7 in pasture and native sodosol, respectively, and 13 and 5 in cropping and native chromosol, respectively. Notably, the majority of the novel viruses identified in the *Narnaviricota* occurred in lineages that were divergent from known viral families, including viruses from all five sampling environments ([Supplementary Figs. 6–8](#)).

3.3.1.5. Picornavirales (*Pisuviricota*). The main expansion of novel sequences in this order was in the *Marnaviridae* and *Dicistroviridae*, as well as several unclassified *Picornavirales* species that fell outside of defined families. Again, the scale of novel diversity in this order warranted splitting it into smaller groups of 1–3 families for robust phylogenetic

analysis: (i) the *Marnaviridae* ([Supplementary Fig. 9](#)), (ii) the *Picornaviridae*, *Polycipiviridae*, *Soliniviridae* ([Fig. 7A](#), [Supplementary Fig. 10](#)), and (iii) the *Dicistroviridae*, *Secoviridae*, *Iflaviridae* ([Fig. 7B](#), [Supplementary Fig. 11](#)). Each phylogeny also included several clades of unclassified *Picornavirales*.

A total of 70 novel viruses with sequence similarity (22–94%) to the *Marnaviridae* were identified. The majority were sourced from riverbank sediment, which is unsurprising given that marnaviruses are typically associated with marine organisms and environments. At the genus level, 12 viruses clustered with the locarnaviruses, five with the salisharnaviruses, three with the sogarnaviruses, five with the kusarnaviruses, and 15 with the labyrinthaviruses ([Supplementary Fig. 9](#)). Finally, several unclassified virus lineages fell between certain genera: four clades between *Locarnavirus* and *Salisharnavirus*, five between *Bacillarnavirus* and

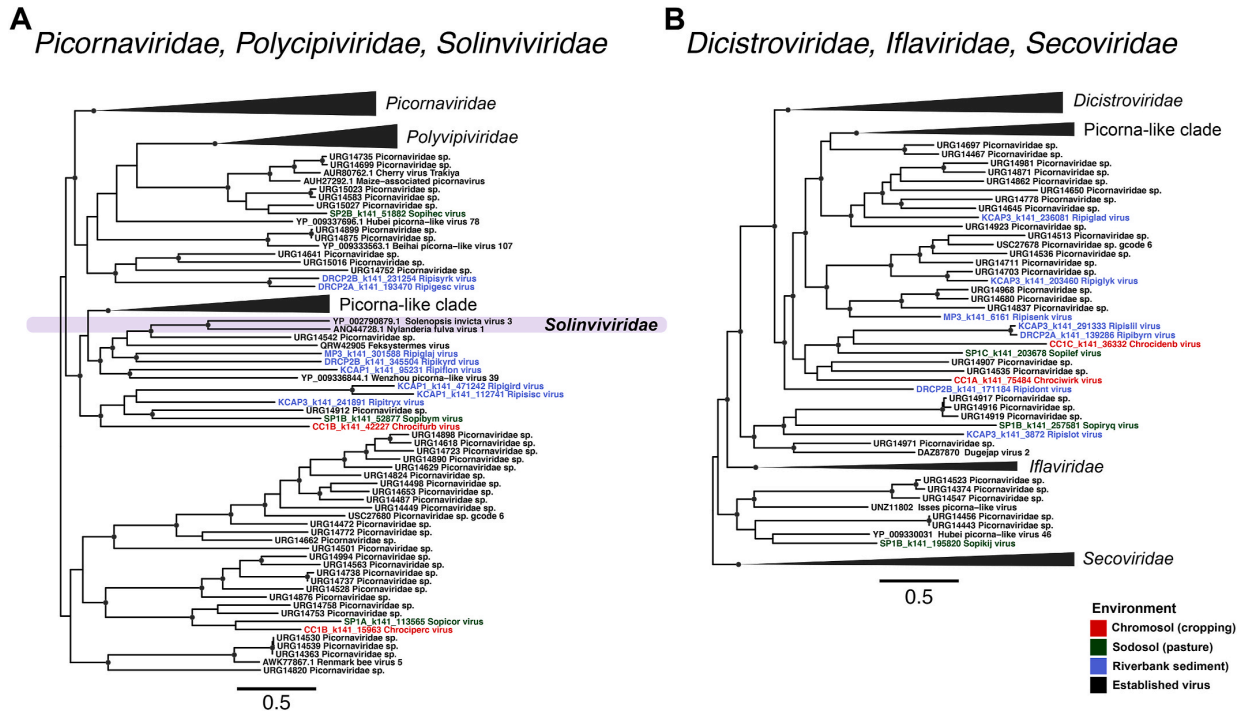


Fig. 7. Phylogenetic trees of RdRp amino acid sequences from the order *Picornvirales*, displaying the expansion of novel, unclassified picorna-like clades surrounding (A) the *Picornaviridae*, *Polycipiviridae*, *Solinviridae* and a clade of previously identified picorna-like sequences, and (B) the *Dicistroviridae*, *Iflavidae*, and *Secoviridae*. Known viruses are shown in black, while putative novel viruses identified here are coloured by the sampling environment. Trees are midpoint rooted with branch lengths scaled according to the number of amino acid substitutions per site. Black circles represent node support $\geq 80\%$ using 1000 S H-aLRT replicates. Phylogenetic trees with collapsed branches expanded are shown in [Supplementary Figs. 10 and 11](#).

Kusarnavirus, and one between *Kusarnavirus* and *Labymavirus* ([Supplementary Fig. 9](#)).

The 13 novel picorna-like sequences shown in [Fig. 7A](#) fell in divergent clades outside of the established families *Picornaviridae*, *Polycipiviridae*, and *Solinviridae*. Over 90% of the published sequences in these clades were metagenomically sourced from environmental or invertebrate samples such that their true hosts are difficult to determine ([Fig. 7A](#), [Supplementary Fig. 10](#)). Ten novel viruses were identified in the *Dicistroviridae*, and a further 14 novel picorna-like viruses fell in clades of diverse, unclassified *Picornavirales* that formed sister lineages to the *Dicistroviridae* ([Supplementary Fig. 11](#)). Another clade of interest, containing Sopikij virus, fell between the plant-infecting family *Secoviridae* and insect-associated families *Dicistroviridae* and *Iflavidae* ([Fig. 7B](#), [Supplementary Fig. 11](#)). The level of diversity in these unclassified clades, comprised mostly of metagenomically sourced, environmental picorna-like viruses, suggests the creation of several new families across the *Picornavirales* may be warranted.

3.3.1.6. Sobelivirales (*Pisuviricota*). Two viruses fell within the fungi-associated family *Barnaviridae*, while another 14 sequences formed a clade with the sole member of the dinoflagellate-associated *Alvernnaviridae* and three divergent sobemo-like viruses ([Supplementary Fig. 12](#)). Several other sobeli-like viruses fell in clades of unclassified *Sobeliviridae*. Of note were the six novel viruses comprising the Ripichaj sobeli-like group, which formed a cluster with solemoviruses predominantly associated with insects including termites, rabbit fleas, and flies ([Supplementary Fig. 12](#)).

3.3.1.7. Astroviridae (*Pisuviricota*). Two major, well-supported clades were identified in the *Astroviridae* ([Fig. 8A](#)). The first group contained vertebrate-associated genera *Mamastrovirus* and *Avastrovirus*, with a

sister group comprised of 19 novel sequences from soil and sediment clustering with several Flumine astrovirus species identified in river samples from New Zealand ([French et al., 2022](#)). The second group was similarly comprised of Flumine astrovirus species and 24 novel soil and sediment virus species ([Fig. 8A](#), [Supplementary Fig. 13](#)).

3.3.1.8. Durnavirales (*Pisuviricota*). A total of 109 novel viruses were identified in four families from the order *Durnavirales*; *Hypoviridae* ([Supplementary Fig. 14](#)), *Partitiviridae*, *Picobirnaviridae* ([Supplementary Fig. 15](#)) and *Fusariviridae* ([Supplementary Fig. 16](#)). Two novel *Hypoviridae* sequences were identified in sediment and one in cropping chromosol. Ripidost virus clustered with Jiangsu sediment hypovirus 3 (QYF49607), Chrocisint virus and Ripitrec virus were highly divergent and had low node support, such that their true phylogenetic placement could not be robustly determined ([Supplementary Fig. 14](#)). Novel sequences were identified in both the *Partitiviridae* and *Picobirnaviridae*, the majority of which were found in divergent, unclassified clades. All five sampling environments were represented in the *Partitiviridae*, while novel *Picobirnaviridae* were identified in sediment, cropping chromosol, and sodasol pasture, but not in native soil types ([Supplementary Fig. 15](#)). Of the five novel fusariviruses, four fell within the larger group comprising the majority of species in this family ([Supplementary Fig. 16](#)). The remaining sequence, Ripijis virus, clustered with the highly divergent species *Fusarium virguliforme* dsRNA mycovirus 1 and 2 (AEZ54148 and 6), although there was low node support (32%).

3.3.2. Negative-stranded RNA virus families (*Negarnaviricota*)

It is striking that only 23 putative novel species of negative-sense RNA viruses were identified from the entire phylum *Negarnaviricota*, two of which fell within the class *Monjiviricetes*. These viruses – Ripivex virus and Ripivax virus – were identified from a single riverbank

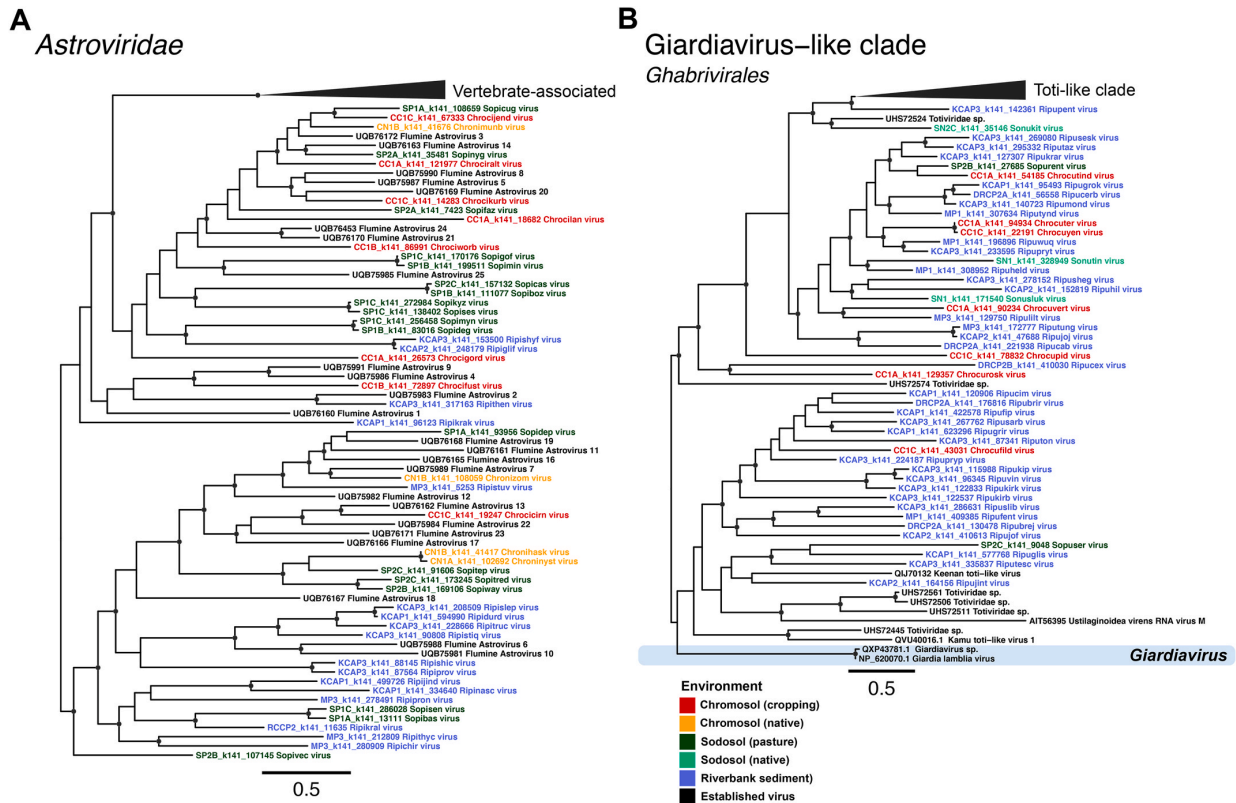


Fig. 8. Phylogenetic trees of RdRp amino acid sequences from (A) the *Astroviridae*, and (B) a clade of *Giardiavirus*-like sequences from the *Totiviridae*. Known viruses are shown in black, while putative novel viruses identified here are coloured by the sampling environment. Trees are midpoint rooted with branch lengths scaled according to the number of amino acid substitutions per site. Black circles represent node support $\geq 80\%$ using 1000 S-HA-RT replicates. Phylogenetic trees with collapsed branches expanded are shown in [Supplementary Figs. 13 and 21](#).

sediment library (MP3). Ripivex virus clustered within the family *Myonnaviridae* (although with low bootstrap support), while Ripivax fell basal to the family ([Supplementary Fig. 17](#)).

The remaining 21 negative-stranded RNA viruses fell in two clades of plant pathogenic fungi-associated *Bunyavirales*. The first clade, denoted the Ripeford bunya-like group, included eight sequences related to bunyaviruses found in grapevine downy mildew lesions caused by *Plasmopara viticola* ([Supplementary Fig. 18](#)). In the second clade, the Ripegrog bunya-like group, 13 sequences clustered with bunyaviruses associated with *Phytophthora cactorum* and *Halophytophthora* species ([Supplementary Fig. 18](#)). It is interesting that nine of the viruses in the Ripegrog bunya-like group were identified in soil, given that several of their relatives are associated with predominantly marine-inhabiting host organisms (*Halophytophthora* species).

3.3.3. Double-stranded RNA virus families (Duplomaviricota)

3.3.3.1. Cystoviridae. All six novel cysto-like viruses were identified from one WA riverbank sediment site (libraries KCAPI1 and KCAPI3). These sequences formed an intermediate lineage between the two major clades in this family ([Supplementary Fig. 19](#)).

3.3.3.2. Reoviridae. One divergent reo-like virus in cropping chromosol (denoted Chroculint virus) grouped with Hubei reo-like virus 10 and 11 (APG79149 and APG79051), falling as a sister lineage to the genus *Rotavirus* (a vertebrate-associated genus), although with low node support ([Supplementary Fig. 20](#)). Riputrev virus fell within the genus *Cypovirus*, while the remaining novel reo-like sequences (also identified

in riverbank sediment) fell in a small sister clade to the genus *Fijivirus* ([Supplementary Fig. 20](#)).

3.3.3.3. Ghabrivirales. Novel species in this order were identified predominantly in sediment, with a small number from both sodosol environments and cropping chromosol ([Supplementary Fig. 21](#)). Of note, 18 novel toti-like sequences sourced from sediment greatly expanded a clade that otherwise includes *Giardia lamblia* virus (NP.620,070), *Giardiavirus* sp. (QXP43781), and seven toti-like viruses ([Fig. 8B](#)). This may represent an expansion of the genus *Giardiavirus* within the *Totiviridae* that currently contains only one accepted species – *Giardia lamblia* virus. Other novel species in the *Totiviridae* included a trichomonas-like virus and a totivirus from riverbank sediment, three sequences identified from sediment samples that clustered with one of several lineages of unclassified *Totiviridae*, and 29 novel sequences clustering in the toti-like sister clade to the *Giardiavirus*-like group ([Supplementary Fig. 21](#)). A novel alphachryso-like virus (Riputesc virus) and a divergent megabirna-like virus (Ripusarb virus) were also identified in sediment. Lastly, three riverbank sediment and three sodosol pasture sequences formed a small sister clade to the *Chrysoviridae* and *Megabirnaviridae* ([Supplementary Fig. 21](#)).

3.3.3.4. Birnaviridae. All four novel in this group were identified from riverbank sediment and clustered with a group of birna-like viruses exclusively from Chinese environmental samples taken predominantly from sediment, along with farmland, forest, and wetlands ([Supplementary Fig. 22](#)).

4. Discussion

Numerous novel RNA viruses have been identified in Australia's unique ecosystems (Geoghegan et al., 2021; Harvey et al., 2018; Le Lay et al., 2020; Mahar et al., 2020; Mu et al., 2018; Pyke et al., 2021; Chang et al., 2021; Van Brussel et al., 2022; Wille et al., 2018). While efforts have been made to characterise the overall microbial communities of Australian soils (Bowd et al., 2022; Xue et al., 2022; Pino et al., 2023), little work has been done on the Australian soil virome. We generated meta-transcriptomic data on 26 libraries from 16 soil and sediment samples taken from eastern (NSW) and western (WA) Australia. From these, we identified a remarkable 2460 novel viral RdRp sequences across all five RNA virus phyla, of which 1780 belonged to the phylum *Lenarviricota*, classically associated with microbial species. The discovery of 2460 putative viruses from such a small number of meta-transcriptomic sequence libraries showcases the extensive, untapped diversity of RNA viruses in Australian soil and sediment environments. Viruses were detected across 18 viral orders and, in many cases, were so diverse that they would constitute new viral genera and even families comprised mostly, if not entirely, of novel Australian RNA viruses.

A relationship between virome composition and land use has been previously observed for both DNA viruses (Narr et al., 2017; Liao et al., 2022) and RNA viruses (Hillary et al., 2022) in soil on other continents including Europe and Asia. Likewise, local sampling environment has been shown to be associated with RNA and DNA viral abundance and diversity in Chinese and North American soils, respectively (Chen et al., 2022; Durham et al., 2022). We found land use and sampling environment to be significantly associated with the read abundance and richness of RNA viruses in Australian soils. However, more specific soil factors, such as pH and soil nutrient levels, have been identified as determinants of both DNA and RNA viral abundance and diversity (Narr et al., 2017; Chen et al., 2022; Liao et al., 2022). Of note, these factors also determine the community compositions of soil-dwelling microbial hosts (Wang et al., 2019; He et al., 2023), such that they likely play a role in the RNA viral read abundance and diversity trends observed in this study. Revealing the precise relationships between Australian soil ecosystems and the viruses within them will not only rely on expansive sampling of diverse environments, but also the generation of thorough ecological metadata.

Despite a small sample size, we detected viruses that spanned the entire diversity of the *Lenarviricota*. Such remarkable genetic diversity, as well as the presence of bacterial- and eukaryote-associated families (Hillman and Cai, 2013) and an often very simple genome structure, suggests that this phylum may comprise the oldest extant RNA viruses. Meta-transcriptomic studies of diverse ecosystems have consistently detected members of the *Lenarviricota*, such that the diversity of this phylum has greatly increased in recent years. Despite their microbial association, these viruses have been detected in studies of both vertebrates (Mahar et al., 2020; Wille et al., 2020) and invertebrates (Shi et al., 2016; Kondo et al., 2020; Le Lay et al., 2020; Thongsripong et al., 2021). Although they are unlikely to be infecting these animals directly, it is clear that these highly diverse viruses are present in virtually all environments (Starr et al., 2019; Wolf et al., 2020; Chen et al., 2022; Neri et al., 2022). Hence, it is no surprise that 1780 novel *Lenarviricota* sequences were identified from both soil types and land uses sampled here, as well as a considerable number from riverbank sediment.

The presence of RNA viruses in archaea is yet to be proven (Krupovic et al., 2018). This may in part reflect the difficulty in extracting sufficient RNA and generating high-quality transcriptomes from archaea-rich sites, as well as the likelihood that viruses infecting archaea are so highly divergent that current virus discovery tools are unable to accurately identify their sequences (Le Lay et al., 2023a). Due to the expansive microbial host range and basal phylogenetic positioning of the *Lenarviricota*, it is proposed that archaea-infecting RNA viruses are most likely to be members of or related to this phylum (Wolf et al., 2018;

Le Lay et al., 2023b). Although we found no evidence for archaea-infecting RNA viruses, the relatively high prevalence of archaea in the potential host organism communities from riverbank sediment libraries, combined with the remarkable expansion of novel viruses across the phylum *Lenarviricota*, suggests that aquatic sites remain a rich reservoir of viral diversity and should be screened for potentially archaea-associated RNA viruses.

Within the *Lenarviricota*, the *Narnaviridae* and *Botourmiaviridae* are currently placed in different virus classes, but have been proposed to comprise a single taxonomic class along with the newly proposed *Narliviridae* (Sadiq et al., 2022). The phylogeny generated here, with an additional 1780 viral sequences expanding known *Lenarviricota* diversity, supports the idea that the *Narliviridae* and *Botourmiaviridae* form sister clades to the exclusion of the *Narnaviridae* (Sadiq et al., 2022). Furthermore, the genus *Ourmiavirus* did not fall within the *Botourmiaviridae*, where it is currently classified, but instead clustered with the *Narliviridae*. This corroborates previous findings that these three ourmiaviruses are more closely related to the narliviruses than the botourmiaviruses (Sadiq et al., 2022).

Based on the nature of the samples (soil and sediment) and the known host ranges of families with which these novel viruses share sequence similarity, the majority of host taxa evaluated in this study were most likely associated with plant or microbial rather than animal hosts. The limited proportion of viruses that were predicted to have animal hosts (<10%) were also most likely associated with invertebrates, such as insects, rather than vertebrates. The novel mitoviruses identified here appeared to utilise the mitochondrial genetic code, which changes the UGA codon from a stop codon to Tryptophan and is typical of fungi-infecting mitovirus genomes, suggesting that these mitoviruses also infect fungi (Cole et al., 2000; Nibert, 2017). As the *Leviviricetes* and *Cystoviridae* are a class and family of bacteriophage (King et al., 2012; Poranon et al., 2017), the 900 novel viruses within them are also likely bacteriophage. Similarly, the detection of large numbers of picobirnaviruses in environmental samples supports the hypothesis that these viruses do not infect vertebrate hosts (as is routinely assumed in the case of vertebrate faecal samples – Malik et al., 2014; Delmas et al., 2019), but are instead associated with microflora present in animal gastrointestinal systems or components of their diet (Ghosh and Malik, 2021). Novel species in the *Tombusviridae* were likely to be associated with plants due to their phylogenetic proximity to established plant viruses. Similarly, the ten novel viruses that clustered within the established dicistroviruses are likely to be insect-associated. Therefore, while the metagenomic nature of this study makes it inherently difficult to confidently determine the hosts of the novel viruses identified, assuming similar host ranges to their closest phylogenetic relatives reveals a remarkable variety of potential host organisms ranging from microbes to plants to invertebrates. Hence, RNA viruses are infecting the majority, if not all, key players in the ecological processes of terrestrial and aquatic soil systems.

A limitation of this study was the initial extraction of genetic material. Of 90 potential sequencing libraries (including technical triplicates for the NSW farmland soil samples), RNA was only successfully extracted from 26 samples at low but detectable concentrations of 11.9–564.4 ng per gram of soil or sediment. This prevented us from conducting robust statistical analyses on the various ecological factors that might shape RNA viral abundance and diversity. The inability to extract RNA from any samples collected 15–30 cm from the soil surface may result from the reduction in host organisms at this depth, as deeper layers of soil have been shown to harbour limited microbial diversity compared to surface layers (Hao et al., 2021; Zhao et al., 2021). We were also unable to extract RNA from the vertosol, even from surface level soil which was successful in both chromosol and sodosol. This is likely due to the high clay content of vertosol, which has a negative effect on RNA yield under a variety of extraction protocols (Novinscak and Filion, 2011). Yields may have also been impacted due to limitations in sample preservation, indicating a clear need for a more refined sampling and a

storage cold chain to effectively extract RNA from remote soil and sediment environments.

5. Conclusions

The discovery of 2460 novel viruses spanning all five RNA viral phyla and a potential host range of bacteria, protists, fungi, plants, and invertebrates shows that Australian terrestrial environments are evidently an untapped resource for RNA virus diversity. These environments may harbour entire families of ecological and evolutionary importance, likely reflecting the vast array of flora and fauna that is unique to the continent. Our work provides an initial view of the Australian terrestrial RNA virosphere, as well as the broad environmental properties such as land use and soil type that may be driving viral composition.

Data availability

The raw sequence read data for this study are available in the NCBI Sequence Read Archive (SRA) database under the BioProject PRJNA981585 and SRA accession numbers SRR26298330 – SRR2629835. All genome sequences used in the phylogenetic analysis are available in NCBI GenBank under the accession numbers PP171752 – PP174210.

Funding

This work was funded by an Australian Research Council Australian Laureate Fellowship to E.C.H. (FL170100022).

CRedit authorship contribution statement

Sabrina Sadiq: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing, Visualization.

Erin Harvey: Investigation, Writing – review & editing.

Jonathon C. O. Mifsud: Investigation, Writing – review & editing.

Budiman Minasny: Data curation, Writing – review & editing.

Alex. B. McBratney: Data curation, Writing – review & editing.

Liana E. Pozza: Data curation, Investigation, Writing – review & editing.

Jackie E. Mahar: Investigation, Writing – original draft, Writing – review & editing.

Edward C. Holmes: Conceptualization, Resources, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Dr Michelle Wille for previously providing the modified Rhea alpha diversity script sets. We also thank Dr Edward Jones for assisting with sample collection at the Nowley Farm sites.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.virol.2024.110007>.

References

- Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.* 55, 539–552.
- Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.
- Bowd, E.L., Banks, S.C., Andrew, A., May, T.W., 2022. Disturbance alters the forest soil microbiome. *Mol. Ecol.* 31, 419–447.
- Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60.
- Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973.
- Chang, W.-S., Rose, K., Holmes, E.C., 2021. Meta-transcriptomic analysis of the virome and microbiome of the invasive Indian myna (*Acridotheres tristis*) in Australia. *One Health* 13, 100360.
- Chapman, A.D., 2009. Numbers of Living Species in Australia and the World, second ed. Australian Govt., Dept. of the Environment, Water, Heritage, and the Arts, Parkes, ACT.
- Chen, Y.-M., Sadiq, S., Tian, J.-H., Chen, X., Lin, X.-D., Shen, J.-J., et al., 2022. RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nature Microbiology* 7, 1312–1323.
- Clausen, P.T.L.C., Aarestrup, F.M., Lund, O., 2018. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinf.* 19.
- Cole, T.E., Hong, Y., Brasier, C.M., Buck, K.W., 2000. Detection of an RNA-dependent RNA polymerase in mitochondria from a mitovirus-infected isolate of the Dutch Elm Disease fungus, *Ophiostoma novo-ulmi*. *Virology* 268, 239–243.
- Delmas, B., Attoui, H., Ghosh, S., Malik, Y.S., Mundt, E., Vakharia, V.N., et al., 2019. ICTV virus taxonomy profile: *Picobirnaviridae*. *J. Gen. Virol.* 100, 133–134.
- Dequiedt, S., Saby, N.P.A., Lelievre, M., Jolivet, C., Thioulouse, J., Toutain, B., et al., 2011. Biogeographical patterns of soil molecular microbial biomass as influenced by soil characteristics and management: biogeography of soil microbial biomass. *Global Ecol. Biogeogr.* 20, 641–652.
- Durham, D.M., Sieradzki, E.T., Ter Horst, A.M., Santos-Medellin, C., Bess, C.W.A., Geonczy, S.E., et al., 2022. Substantial differences in soil viral community composition within and among four northern California habitats. *ISME Communications* 2, 100.
- French, R., Charon, J., Lay, C.L., Muller, C., Holmes, E.C., 2022. Human land-use impacts viral diversity and abundance in a New Zealand river. *Virus Evolution* 8, veac032.
- French, R.K., Holmes, E.C., 2020. An ecosystems perspective on virus evolution and emergence. *Trends Microbiol.* 28, 165–175.
- Geoghegan, J.L., Di Giallonardo, F., Wille, M., Ortiz-Baez, A.S., Costa, V.A., Ghaly, T., et al., 2021. Virome composition in marine fish revealed by meta-transcriptomics. *Virus Evolution* 7, veab005.
- Ghosh, S., Malik, Y.S., 2021. The true host/s of picobirnaviruses. *Front. Vet. Sci.* 7, 615293.
- He, Z., Yuan, C., Chen, P., Rong, Z., Peng, T., Farooq, T.H., et al., 2023. Soil microbial community composition and diversity analysis under different land use patterns in Taojia River Basin. *Forests* 14, 1004.
- Hillary, L.S., Adriaenssens, E.M., Jones, D.L., McDonald, J.E., 2022. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME Communications* 2, 34.
- Hillman, B.I., Cai, G., 2013. The family Narnaviridae. *Adv. Virus Res.* 86, 149–176.
- Huang, D., Yu, P., Ye, M., Schwarz, C., Jiang, X., Alvarez, P.J.J., 2021. Enhanced mutualistic symbiosis between soil phages and bacteria with elevated chromium-induced environmental stress. *Microbiome* 9, 150.
- Hwang, Y., Rahlff, J., Schulze-Makuch, D., Schlöter, M., Probst, A.J., 2021. Diverse viruses carrying genes for microbial extremotolerance in the Atacama desert hyperarid soil. *mSystems* 6, e00385–21.
- Isbell, R.F., 2016. The Australian Soil Classification. CSIRO Publishing, Victoria, Australia.
- Jin, M., Guo, X., Zhang, R., Qu, W., Gao, B., Zeng, R., 2019. Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* 7, 58.
- Jiwaji, M., Matcher, G.F., Bruyn, M.-M. de, Awando, J.A., Moodley, H., Waterworth, D., et al., 2019. Providence virus: an animal virus that replicates in plants or a plant virus that infects and replicates in animal cells? *PLoS One* 14, e0217494.
- Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
- King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., 2012. *Leviviridae*. In: *Virus Taxonomy*, first ed. Elsevier.
- Kondo, H., Fujita, M., Hisano, H., Hyodo, K., Andika, I.B., Suzuki, N., 2020. Virome analysis of aphid populations that infest the barley field: the discovery of two novel groups of nege/kita-like viruses and other novel RNA viruses. *Front. Microbiol.* 11, 509.
- Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D., Koonin, E.V., 2018. Viruses of archaea: structural, functional, environmental and evolutionary genomics. *Virus Res.* 244, 181–193.
- Lagkouravos, I., Fischer, S., Kumar, N., Clavel, T., 2017. Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. *PeerJ* 5, e2836.
- Le Lay, C., Shi, M., Buček, A., Bourguignon, T., Lo, N., Holmes, E., 2020. Unmapped RNA virus diversity in termites and their symbionts. *Viruses* 12, 1145.

- Le Lay, C., Hamm, J.N., Williams, T.J., Shi, M., Cavicchioli, R., Holmes, E.C., 2023a. Viral community composition of hypersaline lakes. *Virus Evolution* 9, vead057.
- Le Lay, C., Stott, M.B., Shi, M., Sadiq, S., Holmes, E.C., 2023b. A metatranscriptomic analysis of geothermal hot springs reveals diverse RNA viruses including the phylum *Lenarviricota*. *Virology* 587, 109873.
- Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.* 12, 323.
- Li, D., Liu, C.-M., Luo, R., Sadakane, K., Lam, T.-W., 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676.
- Liao, H., Li, H., Duan, C.-S., Zhou, X.-Y., Luo, Q.-P., An, X.-L., et al., 2022. Response of soil viral communities to land use changes. *Nat. Commun.* 13, 6027.
- Mahar, J.E., Shi, M., Hall, R.N., Strive, T., Holmes, E.C., 2020. Comparative analysis of RNA virome composition in rabbits and associated ectoparasites. *J. Virol.* 94, e02119-19.
- Malik, Y.S., Kumar, N., Sharma, K., Dhama, K., Shabbir, M.Z., Ganesh, B., et al., 2014. Epidemiology, phylogeny, and evolution of emerging enteric picobirnaviruses of animal origin and their relationship to human strains. *BioMed Res. Int.* 2014, 780752.
- Marcelino, V.R., Clausen, P.T.L.C., Buchmann, J.P., Wille, M., Iredell, J.R., Meyer, W., et al., 2020. CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biol.* 21.
- Mifsud, J.C.O., 2023. BatchArtemisSRAMiner v1.0.3. Zenodo. Available at: <https://github.com/JonathonMifsud/BatchArtemisSRAMiner/>.
- Mu, F., Xie, J., Cheng, S., You, M.P., Barbetti, M.J., Jia, J., et al., 2018. Virome characterization of a collection of *S. sclerotiorum* from Australia. *Front. Microbiol.* 8, 2540.
- Narr, A., Nawaz, A., Wick, L.Y., Harms, H., Chatzinotas, A., 2017. Soil viral communities vary temporally and along a land use transect as revealed by virus-like particle counting and a modified community fingerprinting approach (fRAPD). *Front. Microbiol.* 8, 1975.
- Neri, U., Wolf, Y.I., Roux, S., Camargo, A.P., Lee, B., Kazlauskas, D., et al., 2022. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* 185, 4023–4037.
- Nguyen, L.-T., Schmidt, H.A., Haeseler, A. von, Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.
- Nibert, M.L., 2017. Mitovirus UGA(Trp) codon usage parallels that of host mitochondria. *Virology* 507, 96–100.
- Novinscak, A., Filion, M., 2011. Effect of soil clay content on RNA isolation and on detection and quantification of bacterial gene transcripts in soil by quantitative reverse transcription-PCR. *Appl. Environ. Microbiol.* 77, 6249–6252.
- Oude Munnink, B.B., Cotton, M., Canuti, M., Dejin, M., Jebbink, M., van Hemert, F.J., et al., 2016. A novel astrovirus-like RNA virus detected in human stool. *Virus Evolution* 2, vew005.
- Paradis, E., Schliep, K., 2019. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528.
- Pino, V., Fajardo, M., McBratney, A., Minasny, B., Wilson, N., Baldock, C., 2023. Australian soil microbiome: a first sightseeing regional prediction driven by cycles of soil temperature and pedogenic variations. *Mol. Ecol.* 32, 6243–6259.
- Poranon, M.M., Mäntynen, S., ICTV Report Consortium, 2017. ICTV: virus taxonomy profile: *Cystoviridae*. *J. Gen. Virol.* 98, 2423–2424.
- Pyke, A.T., Shivas, M.A., Darbro, J.M., Onn, M.B., Johnson, P.H., Crunkhorn, A., et al., 2021. Uncovering the genetic diversity within the *Aedes notoscriptus* virome and isolation of new viruses from this highly urbanised and invasive mosquito. *Virus Evolution* 7, veab082.
- Raynaud, X., Nunan, N., 2014. Spatial ecology of bacteria at the microscale in soil. *PLoS One* 9, e87217.
- RStudio Team, 2019. RStudio. Integrated Development for R. RStudio, Inc., Boston, MA.
- R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Sadiq, S., Chen, Y.-M., Zhang, Y.-Z., Holmes, E.C., 2022. Resolving deep evolutionary relationships within the RNA virus phylum *Lenarviricota*. *Virus Evolution* 8, veac055.
- Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., et al., 2016. Redefining the invertebrate RNA virosphere. *Nature* 540, 539–543.
- Singh, B.K., Dawson, L.A., Macdonald, C.A., Buckland, S.M., 2009. Impact of biotic and abiotic interaction on soil microbial communities and functions: a field study. *Appl. Soil Ecol.* 41, 239–248.
- Starr, E.P., Nuccio, E.E., Pett-Ridge, J., Banfield, J.F., Firestone, M.K., 2019. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl. Acad. Sci. USA* 116, 25900–25908.
- Steffen, W., 2009. Australia's Biodiversity and Climate Change. CSIRO Publishing. <https://ebooks.publish.csiro.au/content/ISBN/9780643098190>.
- Thongsripong, P., Chandler, J.A., Kittayapong, P., Wilcox, B.A., Kapan, D.D., Bennett, S. N., 2021. Metagenomic shotgun sequencing reveals host species as an important driver of virome composition in mosquitoes. *Sci. Rep.* 11, 8448.
- Tian, Q., Taniguchi, T., Shi, W.-Y., Li, G., Yamanaka, N., Du, S., 2017. Land-use types and soil chemical properties influence soil microbial communities in the semiarid Loess Plateau region in China. *Sci. Rep.* 7 (45289), 9.
- Trubl, G., Hyman, P., Roux, S., Abedon, S.T., 2020. Coming-of-age characterization of soil viruses: a user's guide to virus isolation, detection within metagenomes, and viromics. *Soil Systems* 4, 23.
- Trubl, G., Jang, H.B., Roux, S., Emerson, J.B., Solonenko, N., Vik, D.R., et al., 2018. Soil viruses are underexplored players in ecosystem carbon processing. *mSystems* 3. <https://doi.org/10.1128/mSystems.00076-18>.
- Van Brussel, K., Mahar, J.E., Ortiz-Baez, A.S., Carrai, M., Spielman, D., Boardman, W.S. J., et al., 2022. Faecal virome of the Australian grey-headed flying fox from urban/suburban environments contains novel coronaviruses, retroviruses and sapoviruses. *Virology* 576, 42–51.
- Wang, C., Zhou, X., Guo, D., Zhao, J., Yan, L., Feng, G., et al., 2019. Soil pH is the primary factor driving the distribution and function of microorganisms in farmland soils in northeastern China. *Ann. Microbiol.* 69, 1461–1473.
- Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag. <https://ggplot2.tidyverse.org>.
- Wille, M., Eden, J., Shi, M., Klaassen, M., Hurt, A.C., Holmes, E.C., 2018. Virus–virus interactions and host ecology are associated with RNA virome structure in wild birds. *Mol. Ecol.* 27, 5263–5278.
- Wille, M., Harvey, E., Shi, M., Gonzalez-Acuña, D., Holmes, E.C., Hurt, A.C., 2020. Sustained RNA virome diversity in Antarctic penguins and their ticks. *ISME J.* 14, 1768–1782.
- Williams, J., Read, C., Norton, A., Dovers, S., Burgman, M., Proctor, W., Anderson, H., 2001. Biodiversity, Australia State of the Environment Report 2001 (Theme Report). CSIRO Publishing on behalf of the Department of the Environment and Heritage, Canberra, Australia.
- Williamson, K.E., Fuhrmann, J.J., Wommack, K.E., Radosevich, M., 2017. Viruses in soil ecosystems: an unknown quantity within an unexplored territory. *Annual Review of Virology* 4, 201–219.
- Wolf, Y.I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J.H., Krupovic, M., et al., 2018. Origins and evolution of the global RNA virome. *mBio* 9, e02329-18.
- Wolf, Y.I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., et al., 2020. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nature Microbiology* 5, 1262–1270.
- Xue, P., McBratney, A.B., Minasny, B., O'Donnell, T., Pino, V., Fajardo, M., et al., 2022. Soil bacterial depth distribution controlled by soil orders and soil forms. *Soil Ecology Letters* 4, 57–68.
- Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T., 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36.

Current Biology

Host adaptive radiation is associated with rapid virus diversification and cross-species transmission in African cichlid fishes

Highlights

- African cichlid fishes contain a rich pool of virus diversity
- Viruses are frequently shared between diverse cichlid fauna in Lake Tanganyika
- Virome composition is not impacted by aspects of host relatedness or ecology
- Hepacivirus evolution accelerated alongside the diversification of cichlids

Authors

Vincenzo A. Costa, Fabrizia Ronco, Jonathon C.O. Mifsud, Erin Harvey, Walter Salzburger, Edward C. Holmes

Correspondence

edward.holmes@sydney.edu.au

In brief

Costa et al. show that the African cichlid radiation in Lake Tanganyika has had a profound impact on the diversity and evolution of the viruses they carry. Viruses frequently jumped between cichlid species, irrespective of their ecological niche or phylogenetic relatedness, particularly during a period of rapid host diversification.



Costa et al., 2024, Current Biology 34, 1247–1257
March 25, 2024 © 2024 The Author(s). Published by Elsevier Inc.
<https://doi.org/10.1016/j.cub.2024.02.008>



Article

Host adaptive radiation is associated with rapid virus diversification and cross-species transmission in African cichlid fishes

Vincenzo A. Costa,¹ Fabrizia Ronco,² Jonathon C.O. Mifsud,¹ Erin Harvey,¹ Walter Salzburger,³ and Edward C. Holmes^{1,4,*}¹Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia²Natural History Museum, University of Oslo, 0562 Oslo, Norway³Zoological Institute, Department of Environmental Sciences, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland⁴Lead contact*Correspondence: edward.holmes@sydney.edu.au<https://doi.org/10.1016/j.cub.2024.02.008>

SUMMARY

Adaptive radiations are generated through a complex interplay of biotic and abiotic factors. Although adaptive radiations have been widely studied in the context of animal and plant evolution, little is known about how they impact the evolution of the viruses that infect these hosts, which in turn may provide insights into the drivers of cross-species transmission and hence disease emergence. We examined how the rapid adaptive radiation of the cichlid fishes of African Lake Tanganyika over the last 10 million years has shaped the diversity and evolution of the viruses they carry. Through metatranscriptomic analysis of 2,242 RNA sequencing libraries, we identified 121 vertebrate-associated viruses among various tissue types that fell into 13 RNA and 4 DNA virus groups. Host-switching was commonplace, particularly within the *Astroviridae*, *Metahepadnavirus*, *Nackednavirus*, *Picornaviridae*, and *Hepacivirus* groups, occurring more frequently than in other fish communities. A time-calibrated phylogeny revealed that hepacivirus diversification was not constant throughout the cichlid radiation but accelerated 2–3 million years ago, coinciding with a period of rapid cichlid diversification and niche packing in Lake Tanganyika, thereby providing more closely related hosts for viral infection. These data depict a dynamic virus ecosystem within the cichlids of Lake Tanganyika, characterized by rapid virus diversification and frequent host jumping, and likely reflecting their close phylogenetic relationships that lower the barriers to cross-species virus transmission.

INTRODUCTION

Since the origin of vertebrates some 540 million years ago (mya), the evolution of the viruses that infect these animals has been shaped by a combination of host-virus co-divergence over many millions of years, coupled with cross-species virus transmission (i.e., host-switching) events that occasionally result in disease outbreaks.^{1–7} Although viruses regularly emerge in new hosts, there is increasing evidence that viruses are natural and abundant components of ecosystems and may be transmitted among species with no apparent disease.^{8–10} Teleost fishes, which are the largest clade of living vertebrates and display extensive ecological diversity, are among those groups that naturally exhibit highly diverse viromes, making them an informative model system for determining the host and virological factors that impact cross-species virus transmission.^{10,11} Indeed, relatives of coronaviruses, hepatitis C virus, influenza viruses, filoviruses, and paramyxoviruses have all been identified in teleosts.^{2,10–15}

Adaptive radiations—the rapid formation of a species complex through adaptations to new or existing ecological niches—have generated substantial biodiversity in teleost fishes, with iconic examples including the three-spined sticklebacks, Antarctic

notothenioids, and most notably the cichlid fishes inhabiting the East African Great Lakes—Tanganyika, Malawi, and Victoria.^{16–18} The African cichlid radiations rank among the most rapidly evolving species assemblages in vertebrates, comprising nearly 2,000 species, all of which have evolved in less than 10 million years.^{19–21} The cichlid radiation in Lake Tanganyika is the oldest of these species flocks, with approximately 240 endemic species arranged into 12 phylogenetic subclades (i.e., “tribes”).^{19,20} The cichlids in Lake Tanganyika dominate the lake’s ichthyofauna in most habitats, covering multiple trophic levels and showing an extraordinary diversity in habitat use and morphological features.^{19,20} Despite their phenotypic diversity, African cichlids exhibit lower levels of genetic diversity than most non-radiating vertebrate populations, and many alleles are shared among species from different lakes as well as between those with strikingly different morphologies, body sizes, and ecologies.²² From a virological perspective, the limited genetic diversity among Tanganyikan cichlids should also mean that these animals will exhibit broadly similar cell receptors, in turn lowering the adaptive barriers for viruses to jump species boundaries. Conversely, it is possible that viruses might co-diverge with the rapidly evolving cichlids into different ecological niches. To date, these ideas have not been tested.



Determining the viromes of African cichlids can also provide insights into how key host transitions have shaped the macro- and microevolution of viruses. The cichlid radiation in Lake Tanganyika, for which a robust phylogenetic framework based on genome-wide data is available,²⁰ began approximately 9–10 mya,^{20,23} and the divergence times of these host species can be used to estimate the rate of virus diversification and determine whether and how it has been shaped by major events in host evolution.²⁴ For instance, as host populations become denser and more diverse, it is expected that rates of virus evolution, including those of virus speciation, will similarly increase. This question has yet to be addressed for viruses. Here, we examine the relationship between host adaptive radiation and virus evolution, using the exceptionally diverse cichlid fauna of Lake Tanganyika as a model system. In particular, we sought to determine whether viruses have co-diverged with the radiating cichlids or whether there has been more frequent cross-species transmission within a restricted geographical space. In addition, we aimed to reveal the rate of virus diversification in cichlids and assess whether it was constant throughout cichlid evolutionary history or whether it increased during periods of rapid host diversification.

RESULTS

The Tanganyikan cichlid virome

We screened 23.7 billion RNA sequencing reads (median 62,909 assembled contigs per library) from 74 Lake Tanganyikan cichlid species ($n = 6$ replicates per species) for vertebrate-associated viral transcripts. The sequencing libraries screened were generated from a previous study of host gene expression dynamics²⁵ and are available on the NCBI Sequence Read Archive (SRA) under BioProject: PRJNA552202. The taxonomic sample covers the entire spectrum of the ecological, morphological, and taxonomic (phylogenetic) diversity^{19,20} of the cichlid adaptive radiation in Lake Tanganyika and includes all 12 tribes: Bathybatini, Benthochromini, Boulengerochromini, Cyphotilapiini, Cyprichromini, Ectodini, Eretmodini, Lamprologini, Limnochromini, Perissodini, Trematocarini, and Tropheini/Haplochromini.

From this analysis, we identified 121 likely vertebrate-associated viruses that fell within the following taxonomic groups (virus genus or family): *Hepacivirus* ($n = 33$ viruses), *Flavivirus* ($n = 7$), *Astroviridae* ($n = 11$), *Matonaviridae* ($n = 5$), *Caliciviridae* ($n = 1$), *Picornaviridae* ($n = 5$), *Coronaviridae* ($n = 3$), *Hepeviridae* ($n = 2$), *Nanghoshaviridae* ($n = 2$), *Piscichuvirus* ($n = 4$), *Arenaviridae* ($n = 1$), *Paramyxoviridae* ($n = 2$), *Birnaviridae* ($n = 2$), *Ichthamaparvovirus* ($n = 2$), *Adomaviridae* ($n = 7$), *Metahepadnavirus* ($n = 19$), and *Nackednavirus* ($n = 15$) (Figure 1B; Table S1). Among the tribes, we found a positive relationship between the number of cichlid species examined and the number of viruses detected ($R^2 = 0.74$, $p < 0.001$), with the Lamprologini harboring 40% of the individual viruses identified, followed by the Ectodini (30.4%), Tropheini/Haplochromini (9.6%), and Cyprichromini (8%), reflecting the species richness of these tribes in Lake Tanganyika.¹⁹

To explore the potential tissue tropisms of the cichlid viruses identified here, we calculated the transcript abundances of viruses identified in the brain, gill, liver, lower pharyngeal jaw (LPJ; an important component of the cichlid's feeding apparatus), ovary, and testis. This revealed that 94% of the viral

groups were present in the gills and the LPJ, implying that these organs, which are in direct contact with the aquatic environment, are a common transmission route among cichlid viruses (Figure 1C). Although the cichlids examined displayed no overt signs of disease, we detected 41 viruses in multiple organs, suggestive of systemic infection (Table S1). Notably, the piscichuviruses were predominantly found in brain tissue, while the hepaciviruses showed a higher prevalence in the liver.

Genome organization of cichlid viruses

The organization of cichlid virus genomes resembled those of other fish viruses and, in many instances, those of their mammalian counterparts (Figure S2). For example, all positive-sense single-stranded RNA viruses had the same number of open reading frames (ORFs) and conserved domains as related mammalian viruses, with the picornaviruses containing a polyprotein with rhinovirus (rhv)-like capsid, helicase, proteinase, and RdRp domains, astroviruses comprising three ORFs, and the hepeviruses and matonaviruses containing a non-structural polyprotein and structural capsid protein, with the latter similar to that of rubella virus (Figure S2A).²⁶

Similarly, the genome of *Neolamprologus multifasciatus* paramyxovirus had the same number and order of proteins as the orthoparamyxoviruses²⁷ (Figure S2C). The N protein encoded an RNA-binding nucleocapsid domain and the three membrane-associated proteins contained a viral matrix protein, fusion glycoprotein domain, and conserved receptor binding protein encoding a haemagglutinin-neuraminidase (HN) domain with a predicted β -propeller fold consisting of six anti-parallel β sheets (i.e., "propeller blades").²⁸ The second propeller blade had a starting sequence of "HRKSCA," resembling that of the "NRKSCS" hexapeptide motif (i.e., sialic-acid binding site) found in the respiroviruses, jeilongviruses, ferlaviruses, orthorubulaviruses, and avulaviruses that use sialic acids for cell entry, suggesting that this mode of entry originated in the aquatic environment.^{27,28}

Among DNA viruses, all hepadnaviruses had circular and intact genomes, with the expected length of approximately 3.2 kb strongly implying that these are exogenous viruses (Figure S2E). The metahepadnaviruses exhibited their typical genome organization, with the viral P protein containing the terminal protein (TP), reverse transcriptase and RNase H domains (RT/RH), as well as the viral core (C) protein and envelope glycoproteins (PreS/S).^{29,30} All nakednaviruses contained viral C and P proteins, with the absence of an envelope³⁰ (Figure S2F).

Ecology and evolutionary history of Tanganyikan cichlid viruses

To determine whether host-switching has shaped the evolutionary history of cichlid viruses, or whether it reflects the co-divergence of viruses and cichlids into different ecological niches during the Tanganyikan cichlid radiation, we performed a phylogenetic reconciliation analysis. Accordingly, virus and host phylogenetic trees were mapped against each other, from which we calculated the proportion of the following biological events: co-divergence, duplication, extinction, and host-switching.^{1,31} Co-divergence occurs when the virus phylogeny mirrors the host phylogeny, indicative of linked speciation events. Duplication is a process where a virus lineage evolves independently of the host lineage, and host-switching indicates that viruses are

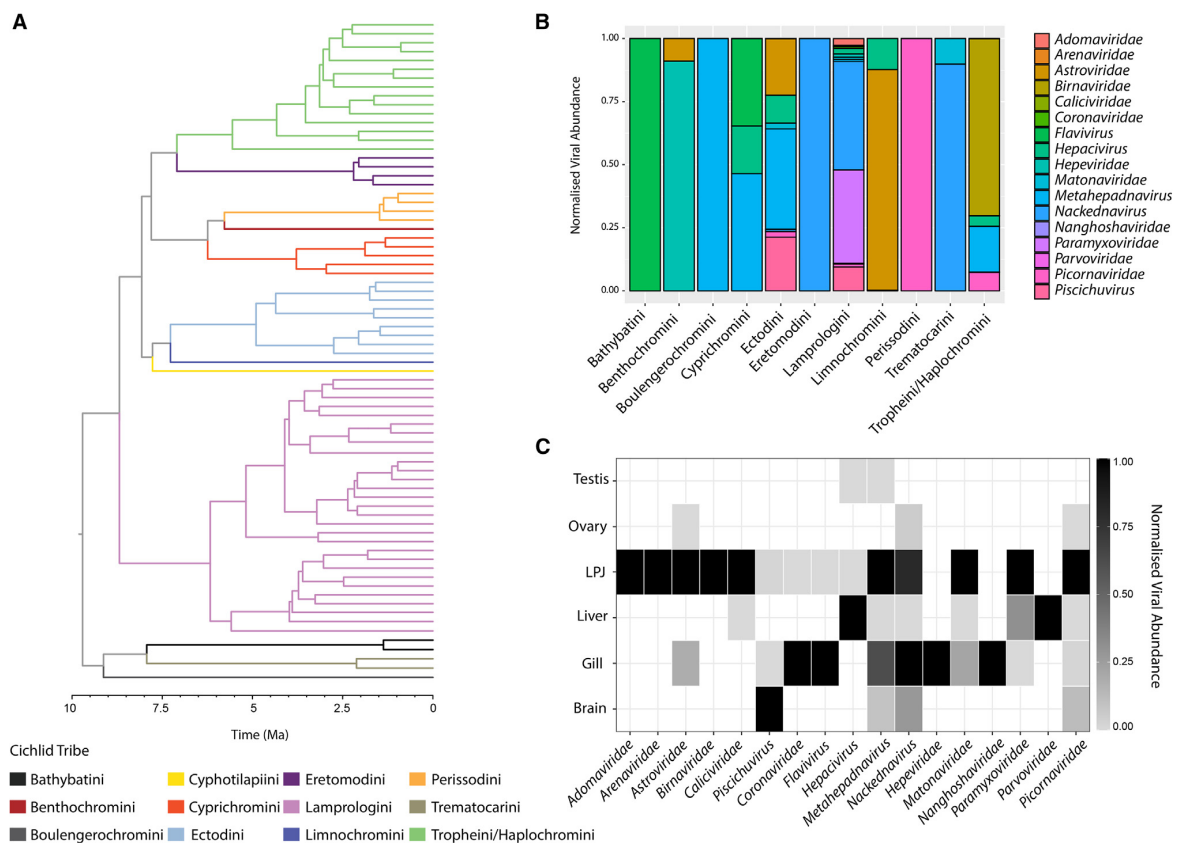


Figure 1. Abundance and diversity of cichlid viromes

(A) Time-calibrated phylogeny of the cichlid (host) adaptive radiation in Lake Tanganyika. The phylogeny was obtained from the literature²⁰ and pruned to the species examined in this study. Tree branches are colored according to cichlid tribe.

(B) Normalized viral abundance of cichlid tribes.

(C) Normalized viral abundance for each organ.

See also [Figures S1](#) and [S2](#) and [Table S1](#).

horizontally transmitted between different species.^{1,31,32} We focused on the *Hepacivirus*, *Astroviridae*, *Matonaviridae*, *Adomaviridae*, *Picornaviridae*, *Piscichuvirus*, *Metahepadnavirus*, and *Nakednavirus* groups as these contained viruses from multiple cichlid species. This analysis revealed that host-switching was the most common event among virus-host associations ($\chi^2 = 141.52$, $p < 0.00001$), representing 60% of all events, followed by co-divergence (23%), duplication (15%), and extinction (4%). Cross-species virus transmission was particularly common in the astroviruses (72% host-switching events), hepaciviruses (70%), nakednaviruses (65%), metahepadnaviruses (63%), and picornaviruses (61%). Hence, cross-species transmission has played a major role in shaping cichlid virus evolution, occurring in both RNA and DNA viruses (Figure 2).

To further investigate the impact of host relatedness and ecology on virus evolution, particularly whether cichlid viruses are associated with specific host phylogenetic lineages or ecological niches, we evaluated the phylogenetic and ecological diversity of the infected and uninfected host species based on the host

phylogeny and available carbon ($\delta^{13}\text{C}$)- and nitrogen ($\delta^{15}\text{N}$)-stable-isotope measurements. Stable-isotope composition is a reliable measure of benthic-pelagic positioning ($\delta^{13}\text{C}$) and trophic level ($\delta^{15}\text{N}$) among fish communities and therefore a strong indicator of host ecological diversity.²⁰ For instance, fishes with high $\delta^{13}\text{C}$ and low $\delta^{15}\text{N}$ occur mainly in the benthic and/or littoral zone and are lower in the food chain (e.g., littoral herbivores), whereas those with low $\delta^{13}\text{C}$ and high $\delta^{15}\text{N}$ occur in the pelagic or deep zone and have a higher trophic level (e.g., pelagic predators).²⁰ Overall, we found no clear ecological or phylogenetic partitioning for all or individual viruses (Figures 3A and 3B). For example, the hepaciviruses, hepadnaviruses, picornaviruses, and astroviruses were found in both pelagic and benthic species, with no structuring according to host ecology. We further compared the ecological and phylogenetic diversity (measured as mean pairwise phylogenetic distances) of both infected and uninfected species in our dataset to 1,000 subsets of species that we randomly sampled across the cichlid radiation. Differences between observed diversity and the random samples were

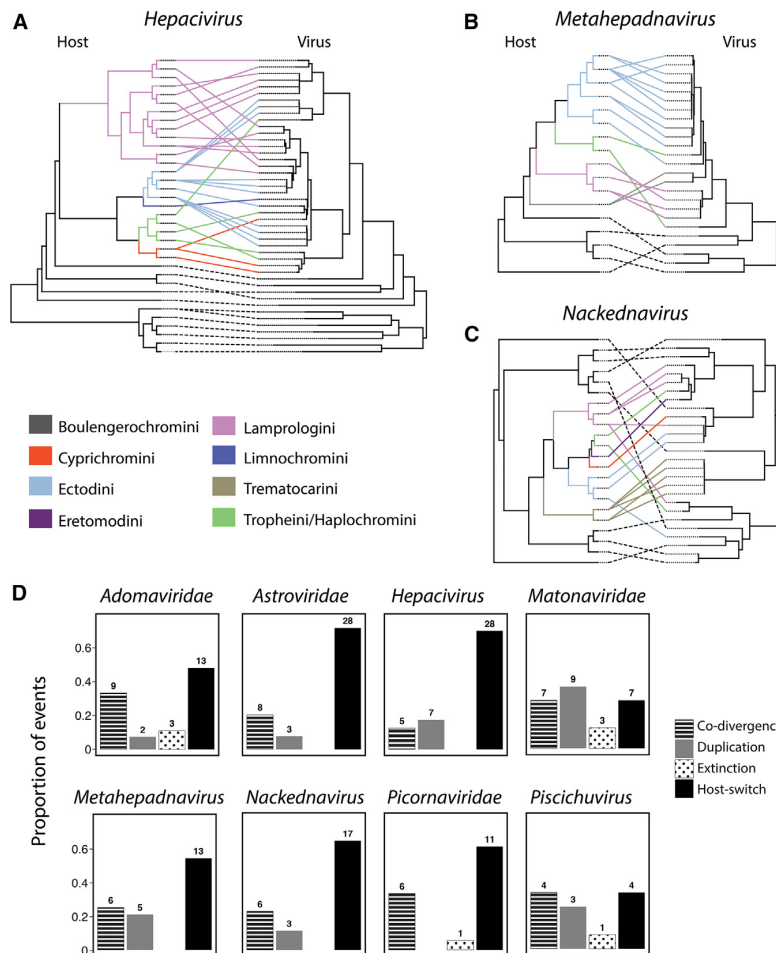


Figure 2. Co-phylogenetic analysis reveals high levels of cross-species virus transmission among African cichlids

(A–C) Co-phylogenetic analysis of the genera *Hepacivirus*, *Metahepadnavirus*, and *Nakednavirus*, with the host phylogeny on the left and virus phylogeny on the right. Lines connecting taxa are colored according to cichlid tribe. Dashed black lines represent background sequences of non-cichlid fish viruses.

(D) Phylogenetic reconciliation analysis of cichlid viral groups. Plots show the proportion of events.

assessed using a one-tailed test, which, under the null hypothesis, assumes that virus infectivity and transmission is random. This analysis showed that both groups represented random subsets of the radiating cichlids that are not restricted to certain areas in the host phylogeny (infected, $p = 0.097$; uninfected, $p = 0.796$) or to particular ecological niches (infected, $p = 0.063$; uninfected, $p = 0.275$) (Figure 3C). However, we found slightly lower phylogenetic diversity in the infected hepacivirus group compared with random subsets of equal sample size ($p = 0.049$), which may be driven by the presence (and transmission) of these viruses among multiple lamprologine cichlid species (see below). Collectively, these data strongly suggest that cichlid viruses are generalists in Lake Tanganyika and may be readily transmitted among species, irrespective of their ecological niche and, in most cases, also independent of the host's phylogenetic position.

Close phylogenetic relationships among cichlid RNA viruses

We identified 77 RNA viruses likely associated with vertebrate hosts. Of these, 28% were associated with recent cross-species

transmission, as the viruses in question were very closely related (>99.5% amino acid similarity). A case in point were the hepaciviruses, which fell into three clades (Figure 4). Although there was a relatively high level of genetic divergence between these clades (e.g., 60%–66% amino acid similarity between clades I and II), there were multiple instances of cross-species virus transmission within clades. Within clade I, we identified four viruses that had 99%–100% pairwise identity in the NS5 gene in *Julidochromis ornatus*, *Neolamprologus savoryi*, *Neolamprologus furcifer*, and *Chalinochromis brichardi*. We also detected African cichlid picornavirus in five species from the tribes Lamprologini ($n = 2$ individuals), Perissodini ($n = 1$), Ectodini ($n = 1$), and Tropheini ($n = 1$), as well as African cichlid piscichuvirus in the brain tissue of three species (Figure 4). Similarly, all three piscichuviruses shared 99.2% L protein (RdRp), 98.4% M protein (glycoprotein), and 98.9% S protein (nucleoprotein) similarity. We also identified African cichlid hepevirus in *Altolamprologus compressiceps* and *Benthochromis horii*, and African cichlid flavivirus in two Cyprichromini and *Bathybates fasciatus* (Bathybatini) (Figure 4). Among other RNA viruses, 63% fell within the same genus or species as another cichlid virus, such that they represent a mixture of common ancestry and past host-switching throughout evolutionary history (Figure 4). For example, we identified a novel group of five rubella-like viruses (*Matonaviridae*) that had ~80% similarity and two birnaviruses that exhibited 93% similarity. Although the remaining RNA viruses were divergent (or identified in a single host species), they exhibited phylogenetic relationships to other fish viruses (e.g., *Paramyxoviridae*, *Arenaviridae*) (Figure S3).

Host-switching and phylogenetic relationships of cichlid DNA viruses

As well as RNA viruses, our analysis identified 43 vertebrate DNA viruses, the majority of which fell within the *Hepadnaviridae*: 19 metahepadnaviruses and 15 nakednaviruses (Figure 5). With the exception of *Lamprologus ocellatus* nakednavirus, all other

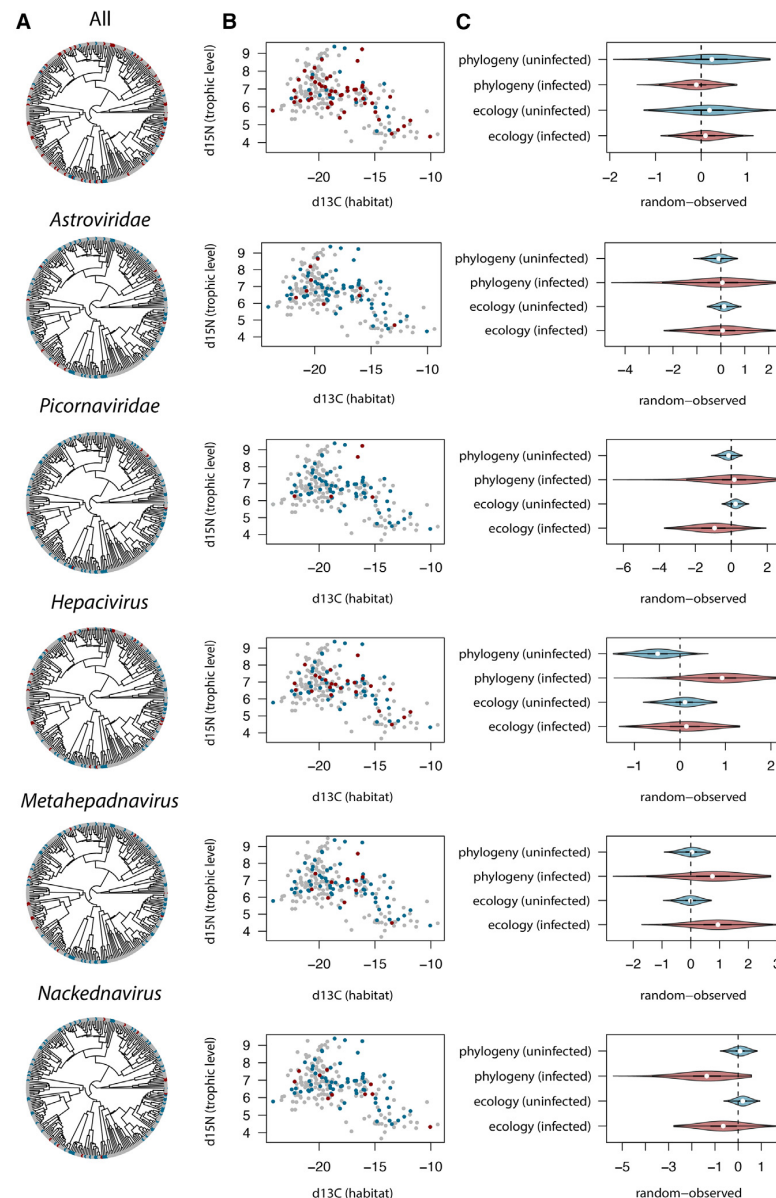


Figure 3. Random transmission of cichlid viruses

(A and B) Phylogenetic (A) and ecological diversity (B) of infected (red) and uninfected (blue) species compared with the overall (gray) diversity of cichlids in Lake Tanganyika.

(C) Violin plots of the difference between the diversity (mean pairwise phylogenetic and ecological distances) of random sets of species (1,000 subsamples of the radiation) minus the diversity of the observed set of infected and uninfected species, respectively.

levels of sequence similarity within the adomaviruses (93%) strongly suggests that these viruses were associated with cross-species transmission in Lake Tanganyika (Figure 5).

Temporal dynamics of Hepacivirus diversification in Lake Tanganyika

To infer the temporal dynamics of cichlid virus evolution, we performed a molecular clock analysis using the Lake Tanganyika cichlid hepaciviruses as a model system. We based our analysis on this group because they contained the largest number of viruses ($n = 33$) and are monophyletic, strongly suggesting that their diversification occurred within Lake Tanganyika. Accordingly, the formation of Lake Tanganyika at 9–12 mya,²³ which also corresponds to the age of the cichlid radiation recently estimated at 9.7 mya,^{19,20} can be used as a biogeographic calibration point to infer the timescale of cichlid hepacivirus evolution. This molecular clock analysis revealed that the mean time to the most recent common ancestor (tMRCA) of all the cichlid hepaciviruses and *Chaetodon aureofasciatus* hepacivirus (i.e., the Percomorpha clade¹⁰) was 12.3 mya (95% higher posterior density [95% HPD]: 10.1–14 mya), while (as expected given the fixed calibration time) the mean tMRCA of the cichlid hepaciviruses was 10.3 mya (95% HPD: 8.9–11.6 mya) (Figure 6). Similarly, clade I has a mean tMRCA of 5 mya (95% HPD: 4.3–6.2 mya), clade II at 9.2 mya (95% HPD: 7.7–10.8 mya), and clade III the most recently emerged at 3.5 mya (95% HPD: 2.8–4.2 mya).

cichlid hepadnaviruses clustered together with other cichlid viruses in our phylogenetic trees, reflecting cross-species transmission or recent common ancestry (Figure 5). This was particularly apparent in the case of metahepadnaviruses, with examples of cross-species virus transmission occurring within clade II (subclades II.1, II.2, and II.3) and clade I (between *Telmatochromis temporalis* and *Lamprologus lemairi*). This pattern was also observed within clades I and II of the nakednaviruses (Figure 5).

We also identified a novel group of adomaviruses that had ~70% amino acid similarity to tilapia adomavirus,³³ identified in *Oreochromis niloticus*, which shares a common ancestor with Tanganyikan cichlids about 16 mya³⁴ (Figure 5). The high

diversification rate of cichlid hepaciviruses was 10.3 mya (95% HPD: 8.9–11.6 mya) (Figure 6). Similarly, clade I has a mean tMRCA of 5 mya (95% HPD: 4.3–6.2 mya), clade II at 9.2 mya (95% HPD: 7.7–10.8 mya), and clade III the most recently emerged at 3.5 mya (95% HPD: 2.8–4.2 mya).

To test whether the rate of cichlid hepacivirus diversification changed over time, we computed the gamma statistic (γ), the null model of which assumes that rates of clade diversification have remained constant over evolutionary history.³⁵ Our time-calibrated phylogeny revealed that the rate of hepacivirus diversification was not constant throughout the cichlid radiation ($\gamma = 3.864$, $p = 0.001$), but seemingly began to accelerate around

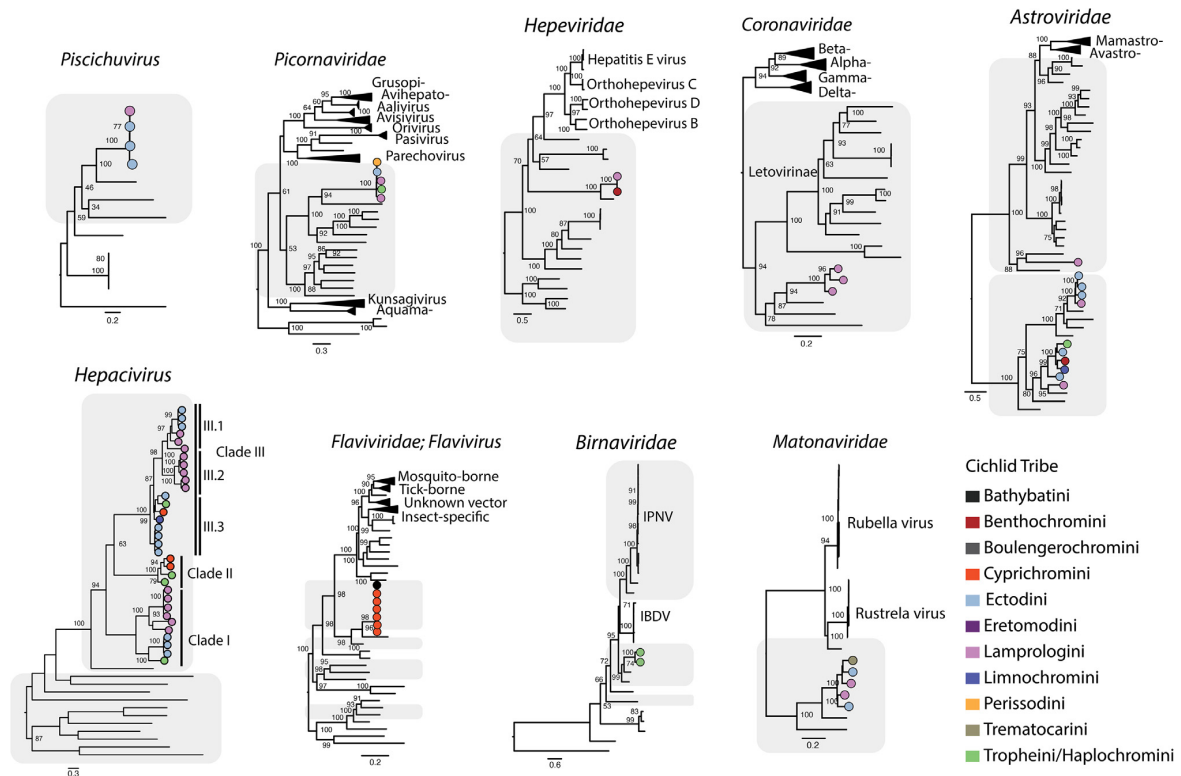


Figure 4. Close evolutionary relationships among cichlid RNA viruses

Phylogenetic analysis reveals close evolutionary relationships among cichlid RNA viruses—in most cases forming monophyletic groups—reflecting recent common ancestry and cross-species virus transmission. Phylogenies were estimated using amino acid sequences of the RdRp gene. Trees were midpoint rooted for clarity only. Circles on branch tips represent individual viruses identified in this study and colored according to tribe. *Birnaviridae*: IPNV, infectious pancreatic necrosis virus; IBDV, infectious bursal disease virus. Shaded branches show fish viruses. The scale bar represents amino acid substitutions per site.

See also Figure S3 and Table S2.

2–3 mya, with multiple viral species forming during the last 1–2 million years (Figure 6). Overall, this coincides with the most recent of three pulses of cichlid diversification in Lake Tanganyika, a period of rapid host diversification.¹⁹

DISCUSSION

Evolutionary radiations are responsible for much of the earth's biodiversity and often involve complex interactions between both biotic and abiotic factors.³⁶ The interplay between these factors is also central to the generation of virus diversity: viruses routinely adapt to genetic and immunological barriers when they invade new hosts, and there must be adequate exposure between host species for intra- and inter-host transmission.³⁷ By analyzing the viromes of 74 closely related cichlid species endemic to Lake Tanganyika, we show that host adaptive radiation is associated with rapid virus diversification and cross-species transmission. From a total of 121 vertebrate-associated DNA and RNA viruses identified, 94% were characterized by both recent common ancestry and host-switching.

The evolutionary patterns observed here support the idea that viruses more readily jump between closely related host species, perhaps because of an underlying similarity in their cell receptors, which in turn lowers the adaptive barriers to viruses when they encounter a new host species.^{37,38} The maximum pairwise genetic distance between cichlids in Lake Tanganyika is roughly equivalent to that of human and chimpanzee, with some species having pairwise distances of less than 0.03%, which are among the lowest values observed in all vertebrate groups.²² It is noteworthy that the genetic distance between humans and chimpanzees is often sufficient for successful cross-species virus transmission,³ and in this context we assume that the remarkably close genetic relationships among African cichlids, in line with their rapid speciation, should therefore provide viruses access to a constant supply of diverse cichlid hosts within Lake Tanganyika. A case in point is the presence of African cichlid picornavirus in five species that are members of four tribes: the Lamprologini, Ectodini, Perissodini, and Tropheini (Figure 4). Moreover, that our analysis detected no evidence for assortative transmission with respect to host phylogeny and ecology (Figure 3) further implies that cichlid viruses have evolved to maximize their

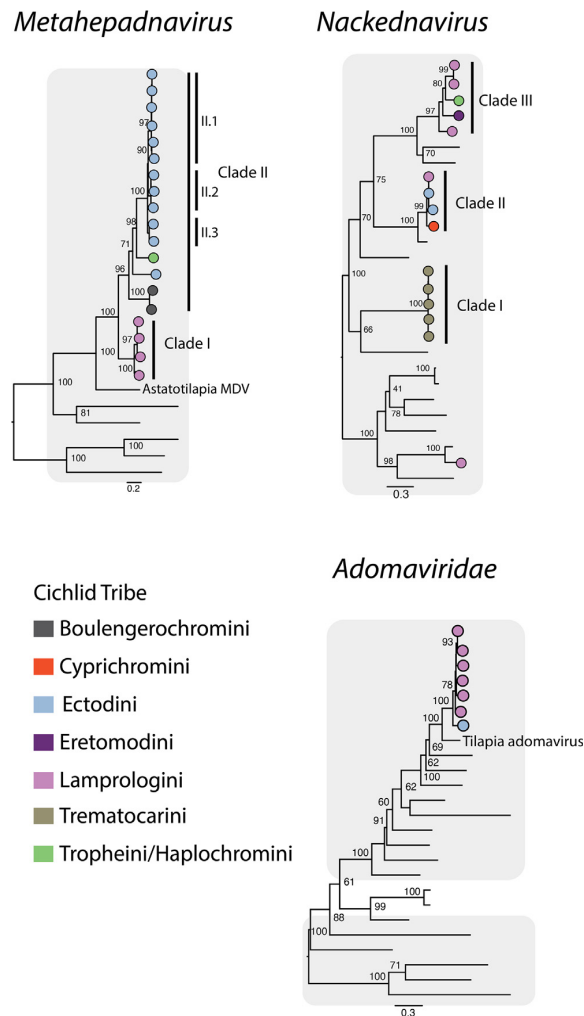


Figure 5. Close evolutionary relationships among cichlid DNA viruses

Phylogenies were estimated using amino acid sequences of P protein for the hepadnaviruses and the LO7 protein for adomaviruses. Trees were midpoint rooted for clarity only. Circles on branch tips represent individual viruses identified in this study and colored according to tribe. The scale bar represents amino acid substitutions per site.

See also [Figure S3](#) and [Table S2](#).

fitness by infecting multiple host species in Lake Tanganyika, and co-evolving with cichlids into ecological niches might therefore be a poor evolutionary strategy.

It is also likely that African cichlids will have similar immune properties, increasing virus susceptibility, and are frequently exposed to viruses. For example, a 125-m² area in Lake Tanganyika can harbor up to 39 different cichlid species.³⁹ In contrast, reef fishes—that are characterized by multiple families with greater genetic diversity—show minimal virus cross-species transmission, despite ample exposure among species.¹⁰ Indeed, a recent study found that only four of the 38 viruses

identified in a spatially restricted—100 m²—coral reef fish community, comprising 60 fish species (16 families), were associated with host-switching, and only on relatively deep evolutionary timescales.¹⁰ Moreover, in this reef system fish are consistently found in groups of up to 25 species in a 3.5-m² area, such that they represent the largest species densities among all extant vertebrate species.⁴⁰ Despite such substantially high rates of contact and exposure, the lack of host-switching in the reef ecosystem implies that there are strong host barriers to infection, particularly given the large degree of genetic divergence among reef fish families. Indeed, the cross-species transmission events observed were only identified in members of the same fish family, the Gobiidae. This further supports the notion that close genetic relatedness is a key determinant of rapid virus host-switching in African cichlids, occurring more frequently than in other fish communities.

We assembled a time-calibrated phylogeny of hepaciviruses from the Lake Tanganyika cichlid radiation. It was notable that the rate of hepacivirus evolution was not constant throughout the cichlid radiation, but increased rapidly during the last 2–3 million years, coinciding with a period of rapid diversification, packing of niche space, and micro-niche partitioning in Tanganyikan cichlids ([Figure 6](#)).²⁰ This strongly suggests that, as cichlid communities became denser and more diverse, viruses exploited a larger number of closely related host species for infection through cross-species transmission. For the future, it will be important to test this theory using a larger virus sample size and more detailed analyses, such as through a piece-wise constant birth-death model.

As well as hepaciviruses, our analysis identified 34 hepadnaviruses of which 97% were closely related, indicative of recent cross-species transmission ([Figure 5](#)). In particular, a novel group of metahepadnaviruses fell sister to *Astatotilapia* metahepadnavirus, which was recently discovered in *Astatotilapia* species from Lake Masoko. *Astatotilapia* belongs to the Haplochromini, which is the most species-rich clade of cichlids and widespread in Africa, where it diversified in many rivers and lakes, including Lake Malawi.²¹ This suggests that cichlid metahepadnaviruses are widely distributed in East African freshwater systems and have likely diversified alongside cichlids during multiple radiation events. A recent molecular clock analysis of the *Hepadnaviridae* estimated that the genus *Metahepadnavirus*, characterized exclusively by ray-finned fish hosts, split from orthohepadnaviruses, a group of mammalian viruses (e.g., hepatitis B virus), approximately 240 mya, with both *Astatotilapia* and Mexican tetra (*Astyanax mexicanus*) metahepadnaviruses diverging ~80–100 mya.³⁰ However, the Haplochromini formed in Lake Tanganyika ~5.5 mya, and *A. mexicanus*, both surface-dwelling and cave populations, originated very recently during the late Pleistocene, approximately 20 thousand years ago.⁴¹ This strongly suggests that metahepadnaviruses have likely diverged much more recently than previously estimated.³⁰

We identified piscichuviruses with high abundances in the brain tissue of three cichlid species. Currently, the genus *Piscichuvirus* (*Chuviridae*) is described by viruses associated with ectothermic vertebrates such as snakes, turtles, and ray-finned fish, as well as decapods.⁴² Similarly, these viruses were recently detected in the brain and spinal cord of boa constrictors and three aquatic turtle species, causing lymphocytic meningoencephalomyelitis.^{43,44}

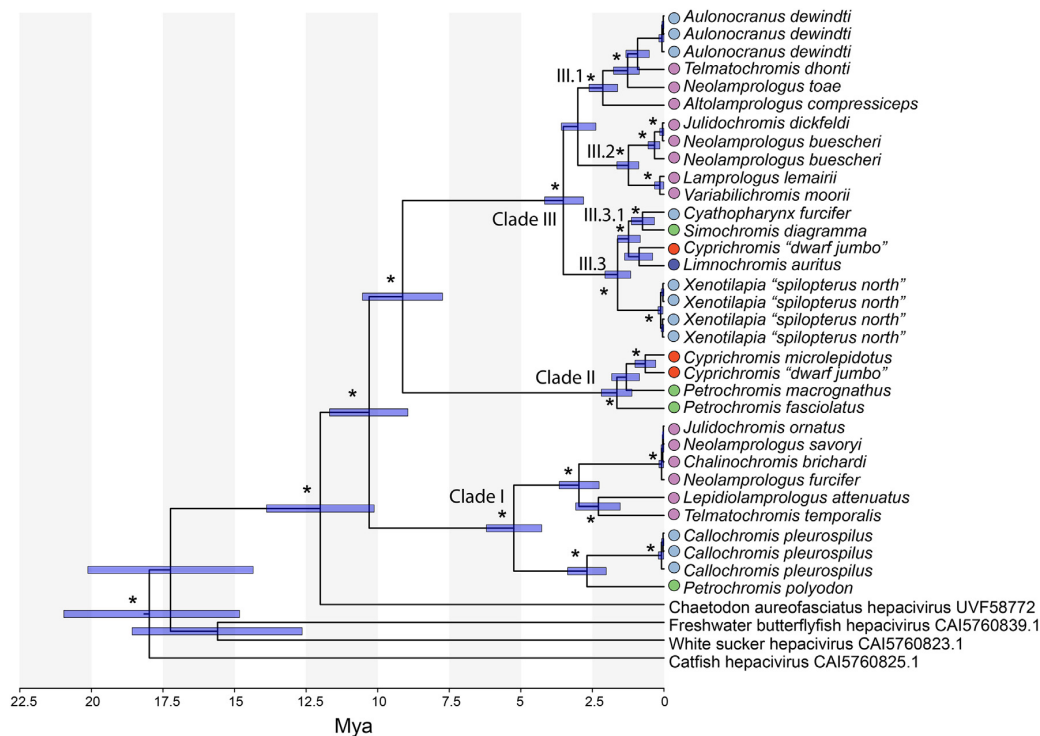


Figure 6. Temporal dynamics of hepacivirus evolution in Tanganyikan cichlids

Time-calibrated maximum clade credibility tree of all currently described ray-finned fish (Actinopterygii) hepaciviruses. The time axis represents millions of years ago (mya). Node bars represent the 95% HPD interval for age. Stars denote a posterior probability of over 0.85. Tip labels are colored according to cichlid tribe.

Despite the divergence between fish and reptile piscichuviruses (~48% similarity), our data show that this emerging group is localized to the central nervous system and have likely maintained this tropism over millions of years.

Given their exposure to the external environment, it was not surprising that, across all virus groups, the highest number of viral transcripts were identified in the gills and the LPJ, which together share a developmental origin and exhibit similar patterns of gene expression.²⁵ Conversely, we detected a very low number of viral transcripts in the gonads. This suggests that the viruses in this ecosystem might utilize cell receptors that are not naturally expressed in cichlid reproductive organs. It is also possible that the gonads might contain a rich source of antimicrobial peptides, as observed in the seminal fluid proteome of sparrows.⁴⁵ Indeed, it was recently shown that the rate of transcriptome evolution and the number of both protein-coding genes and long noncoding (lnc) RNAs are higher in the gonads compared with all other cichlid organs examined in this study.²⁵ Transcriptome evolution in the gonads should therefore experience strong selective pressures given their function in sperm competition and mate choice, and perhaps might select for antiviral mechanisms that protect the gametes from infection.⁴⁶

Although our analysis predicted multiple host-switching events and potential tissue tropisms, these are necessarily based on sequence analysis alone, such that the precise mechanisms of infection and the mode of transmission are not

covered here and require further investigation. This is of particular importance given the geographical proximity and evolutionary relatedness to one of the most farmed and consumed fish species globally, the Nile tilapia, which often experiences devastating impacts from emerging viral infections.⁴⁷ Indeed, because the viruses identified are novel, their receptors are unknown. Moreover, the transcriptome data screened were gene expression profiles generated for host evolutionary analyses only²⁵ and therefore not optimized for virus discovery. Nevertheless, we discovered a large diversity of viruses in 72% of the cichlid species examined. It is important to note that, given the nature of transcriptomic sequencing, we were limited to the detection of RNA viruses and DNA viruses that were actively expressing genes at the time of sampling.

Taken together, we show that the African cichlid radiation has had a profound impact on the extent, pattern, and evolution of virus diversity in Lake Tanganyika. Although virus-host co-divergence is a common macroevolutionary pattern that forms the “backbone” of many family-level virus phylogenies, in the context of a host adaptive radiation virus, diversification is accelerated by frequent cross-species transmission, especially during a time of rapid host diversification and increased niche packing in Lake Tanganyika.^{19,20} Hence, the rich viral diversity shared by the cichlids is likely the product of close genetic relatedness in combination with frequent interactions in Lake Tanganyika.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
 - Lead contact
 - Materials availability
 - Data and code availability
- **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**
- **METHOD DETAILS**
 - Assembly of cichlid viromes
 - Identification and removal of endogenous viral elements
 - Taxonomic assignment of cichlid viruses
 - Annotation of cichlid viral genomes
 - Phylogenetic analysis
 - Co-phylogenetic analysis of cross-species virus transmission and co-divergence
 - Time calibration and diversification rates of cichlid hepaciviruses
 - Phylogenetic and ecological diversity analysis
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.cub.2024.02.008>.

ACKNOWLEDGMENTS

This work was funded by an Australian Research Council Discovery Project (DP200102351) and a National Health and Medical Research Council Investigator grant (GNT2017197) to E.C.H., and the Swiss National Science Foundation to F.R. (206869) and W.S. (208002). We acknowledge the University of Sydney for providing the high-performance computer cluster “Artemis,” which was used for this study.

AUTHOR CONTRIBUTIONS

V.A.C., F.R., W.S., and E.C.H. conceptualized the study. V.A.C., F.R., and J.C.O.M. performed the analyses. V.A.C. wrote and prepared the original draft. V.A.C., J.C.O.M., E.C.H., E.H., F.R., and W.S. edited and revised the manuscript. E.C.H. funded the project. E.H. and E.C.H. supervised the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 19, 2023
Revised: December 18, 2023
Accepted: February 6, 2024
Published: February 29, 2024

REFERENCES

1. Geoghegan, J.L., Duchêne, S., and Holmes, E.C. (2017). Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog.* **13**, e1006215.
2. Shi, M., Lin, X.D., Chen, X., Tian, J.H., Chen, L.J., Li, K., Wang, W., Eden, J.S., Shen, J.J., Liu, L., et al. (2018). The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202.
3. Sharp, P.M., and Hahn, B.H. (2010). The evolution of HIV-1 and the origin of AIDS. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 2487–2494.
4. Webby, R.J., and Webster, R.G. (2001). Emergence of influenza A viruses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356**, 1817–1828.
5. Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452.
6. Pekar, J.E., Magee, A., Parker, E., Moshiri, N., Izhikevich, K., Havens, J.L., Gangavarapu, K., Malpica Serrano, L.M., Crits-Christoph, A., Matteson, N.L., et al. (2022). The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science* **377**, 960–966.
7. Worobey, M., Levy, J.I., Malpica Serrano, L.M., Crits-Christoph, A., Pekar, J.E., Goldstein, S.A., Rasmussen, A.L., Kraemer, M.U.G., Newman, C., Koopmans, M.P.G., et al. (2022). The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. *Science* **377**, 951–959.
8. Holmes, E.C. (2022). The ecology of viral emergence. *Annu. Rev. Virol.* **9**, 173–192.
9. French, R.K., and Holmes, E.C. (2020). An ecosystems perspective on virus evolution and emergence. *Trends Microbiol.* **28**, 165–175.
10. Costa, V.A., Bellwood, D.R., Mifsud, J.C.O., Van Brussel, K., Geoghegan, J.L., Holmes, E.C., and Harvey, E. (2023). Limited cross-species virus transmission in a spatially restricted coral reef fish community. *Virus Evol.* **9**, veab011.
11. Geoghegan, J.L., Di Giallonardo, F., Wille, M., Ortiz-Baez, A.S., Costa, V.A., Ghaly, T., Mifsud, J.C.O., Turnbull, O.M.H., Bellwood, D.R., Williamson, J.E., et al. (2021). Virome composition in marine fish revealed by meta-transcriptomics. *Virus Evol.* **7**, veab005.
12. Miller, A.K., Mifsud, J.C.O., Costa, V.A., Grimwood, R.M., Kitson, J., Baker, C., Brosnahan, C.L., Pande, A., Holmes, E.C., Gemmell, N.J., et al. (2021). Slippery when wet: cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the *Coronaviridae*. *Virus Evol.* **7**, veab050.
13. Mifsud, J.C.O., Costa, V.A., Petrone, M.E., Marzinelli, E.M., Holmes, E.C., and Harvey, E. (2023). Transcriptome mining extends the host range of the *Flaviviridae* to non-bilaterians. *Virus Evol.* **9**, veac124.
14. Costa, V.A., Mifsud, J.C.O., Gilligan, D., Williamson, J.E., Holmes, E.C., and Geoghegan, J.L. (2021). Metagenomic sequencing reveals a lack of virus exchange between native and invasive freshwater fish across the Murray–Darling Basin, Australia. *Virus Evol.* **7**, veab034.
15. Petrone, M.E., Parry, R., Mifsud, J.C.O., Van Brussel, K., Vorhees, I.E.H., Richards, Z.T., and Holmes, E.C. (2023). Evidence for an ancient aquatic origin of the RNA viral order *Articulavirales*. *Proc. Natl. Acad. Sci. USA* **120**, e2310529120.
16. Taylor, E.B., and McPhail, J.D. (1999). Evolutionary history of an adaptive radiation in species pairs of threespine sticklebacks (*Gasterosteus*): insights from mitochondrial DNA. *Biol. J. Linn. Soc.* **66**, 271–291.
17. Clarke, A., and Johnston, I.A. (1996). Evolution and adaptive radiation of Antarctic fishes. *Trends Ecol. Evol.* **11**, 212–218.
18. Fryer, G., and Iles, T.D. (1972). The Cichlid Fishes of the Great Lakes of Africa: Their Biology and Evolution (Oliver and Boyd).
19. Ronco, F., Büscher, H.H., Indermaur, A., and Salzburger, W. (2020). The taxonomic diversity of the cichlid fish fauna of ancient Lake Tanganyika, East Africa. *J. Great Lakes Res.* **46**, 1067–1078.
20. Ronco, F., Matschiner, M., Böhne, A., Boila, A., Büscher, H.H., El Taher, A., Indermaur, A., Malinsky, M., Ricci, V., Kahmen, A., et al. (2021). Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature* **589**, 76–81.
21. Salzburger, W. (2018). Understanding explosive diversification through cichlid fish genomics. *Nat. Rev. Genet.* **19**, 705–717.
22. Svardal, H., Salzburger, W., and Malinsky, M. (2021). Genetic variation and hybridization in evolutionary radiations of cichlid fishes. *Annu. Rev. Anim. Biosci.* **9**, 55–79.

23. Cohen, A.S., Soreghan, M.J., and Scholz, C.A. (1993). Estimating the age of formation of lakes: An example from Lake Tanganyika, East African Rift system. *Geology* 21, 511–514.
24. Harvey, E., and Holmes, E.C. (2022). Diversity and evolution of the animal virome. *Nat. Rev. Microbiol.* 20, 321–334.
25. El Taher, A., Böhne, A., Boileau, N., Ronco, F., Indermaur, A., Widmer, L., and Salzburger, W. (2021). Gene expression dynamics during rapid organismal diversification in African cichlid fishes. *Nat. Ecol. Evol.* 5, 243–250.
26. Dominguez, G., Wang, C.Y., and Frey, T.K. (1990). Sequence of the genome RNA of rubella virus: evidence for genetic rearrangement during togavirus evolution. *Virology* 177, 225–238.
27. Rima, B., Balkema-Buschmann, A., Dundon, W.G., Duprex, P., Easton, A., Fouchier, R., Kurath, G., Lamb, R., Lee, B., Rota, P., et al. (2019). ICTV Virus Taxonomy Profile: *Paramyxoviridae*. *J. Gen. Virol.* 100, 1593–1594.
28. Azarm, K.D., and Lee, B. (2020). Differential features of fusion activation within the *Paramyxoviridae*. *Viruses* 12, 161.
29. Dill, J.A., Camus, A.C., Leary, J.H., Di Giallonardo, F., Holmes, E.C., and Ng, T.F.F. (2016). Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses. *J. Virol.* 90, 7920–7933.
30. Lauber, C., Seitz, S., Mattei, S., Suh, A., Beck, J., Herstein, J., Börold, J., Salzburger, W., Kaderali, L., Briggs, J.A.G., et al. (2017). Deciphering the origin and evolution of hepatitis B viruses by means of a family of non-enveloped fish viruses. *Cell Host Microbe* 22, 387–399.e6.
31. Santichaivekin, S., Yang, Q., Liu, J., Mawhorter, R., Jiang, J., Wesley, T., Wu, Y.C., and Libeskind-Hadas, R. (2021). eMPress: a systematic copy-homology reconciliation tool. *Bioinformatics* 37, 2481–2482.
32. Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. (2010). Jane: a new tool for the copy-homology reconstruction problem. *Algorithms Mol. Biol.* 5, 16.
33. Welch, N.L., Tisza, M.J., Starrett, G.J., Belford, A.K., Pastrana, D.V., Pang, Y.-Y.S., Schiller, J.T., An, P., Cantalupo, P.G., Pipas, J.M., et al. (2020). Identification of adomavirus virion proteins. Preprint at biorxiv. <https://doi.org/10.1101/341131v1>.
34. Matschiner, M., Böhne, A., Ronco, F., and Salzburger, W. (2020). The genomic timeline of cichlid fish diversification across continents. *Nat. Commun.* 11, 5895.
35. Pybus, O.G., and Harvey, P.H. (2000). Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc. Biol. Sci.* 267, 2267–2272.
36. Naciri, Y., and Linder, H.P. (2020). The genetics of evolutionary radiations. *Biol. Rev. Camb. Philos. Soc.* 95, 1055–1072.
37. Parrish, C.R., Holmes, E.C., Morens, D.M., Park, E.C., Burke, D.S., Calisher, C.H., Laughlin, C.A., Saif, L.J., and Daszak, P. (2008). Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol. Mol. Biol. Rev.* 72, 457–470.
38. Longdon, B., Brockhurst, M.A., Russell, C.A., Welch, J.J., and Jiggins, F.M. (2014). The evolution and genetics of virus host shifts. *PLoS Pathog.* 10, e1004395.
39. Widmer, L., Heule, E., Colombo, M., Rueegg, A., Indermaur, A., Ronco, F., and Salzburger, W. (2019). Point-Combination Transect (PCT): incorporation of small underwater cameras to study fish communities. *Methods Ecol. Evol.* 10, 891–901.
40. Bellwood, D.R., Hoey, A.S., Ackerman, J.L., and Depczynski, M. (2006). Coral bleaching, reef fish community phase shifts and the resilience of coral reefs. *Glob. Change Biol.* 12, 1587–1594.
41. Fumey, J., Hinaux, H., Noiret, C., Thermes, C., Rétaux, S., and Casane, D. (2018). Evidence for late Pleistocene origin of *Astyanax mexicanus* cave-fish. *BMC Evol. Biol.* 18, 43.
42. Di Paola, N., Dheilly, N.M., Junglen, S., Paraskevopoulou, S., Postler, T.S., Shi, M., and Kuhn, J.H. (2022). *Jingchuvirales*: a new taxonomical framework for a rapidly expanding order of unusual Monjiviricete viruses broadly distributed among arthropod subphyla. *Appl. Environ. Microbiol.* 88, e0195421.
43. Argenta, F.F., Hepojoki, J., Smura, T., Szilovicza, L., Hammerschmitt, M.E., Driemeier, D., Kipar, A., and Hetzel, U. (2020). Identification of reptarenaviruses, hartmaniviruses, and a novel chuvirus in captive native Brazilian boa constrictors with boid inclusion body disease. *J. Virol.* 94, e00001-00020.
44. Laovechprasit, W., Young, K.T., Stacy, B.A., Tillis, S.B., Ossiboff, R.J., Vann, J.A., Subramaniam, K., Agnew, D., Zhang, J., Whitaker, S., et al. (2023). Piscichuvirus-Associated Severe Meningoencephalomyelitis in Aquatic Turtles, United States, 2009–2021. *Emerg. Infect. Dis.* 30, 280–288. Preprint at biorxiv. <https://doi.org/10.1101/2023.02.24.528524v1>.
45. Rowe, M., Whittington, E., Borziak, K., Ravinet, M., Eroukhanoff, F., Sætre, G.P., and Dorus, S. (2020). Molecular diversification of the seminal fluid proteome in a recently diverged passerine species pair. *Mol. Biol. Evol.* 37, 488–506.
46. Necsulea, A., and Kaessmann, H. (2014). Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* 15, 734–748.
47. Machimbirike, V.I., Jansen, M.D., Senapin, S., Khunrae, P., Rattanarajpong, T., and Dong, H.T. (2019). Viral infections in tilapines: more than just tilapia lake virus. *Aquaculture* 503, 508–518.
48. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.
49. Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676.
50. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60.
51. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649.
52. Nayfach, S., Camargo, A.P., Schulz, F., Elie-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* 39, 578–585.
53. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12, 323.
54. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.
55. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
56. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240.
57. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858.
58. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973.
59. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.
60. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10, e1003537.
61. Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29, 1969–1973.



62. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904.
63. McMurdie, P.J., and Holmes, S. (2013). phyloseq: An R Package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8, e61217.
64. Wickham, H. (2011). ggplot2. *WIREs Comp. Stats.* 3, 180–185.
65. Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528.
66. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223.
67. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., and Khan, A.S. (2018). A Reference Viral Database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *mSphere* 3, e00069–00018.
68. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermini, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw RNA sequencing reads	El Taher et al. ²⁵	NCBI SRA PRJNA552202
Viral consensus sequences	This study	NCBI/GenBank OR350242-OR350390
Software and algorithms		
Trimmomatic (v.0.38)	Bolger et al. ⁴⁸	https://github.com/usadellab/Trimmomatic
MEGAHIT (v.1.2.9)	Li et al. ⁴⁹	https://github.com/voutcn/megahit
Diamond (v.2.0.9)	Buchfink et al. ⁵⁰	https://github.com/bbuchfink/diamond
Geneious Prime (v.2022.0)	Kearse et al. ⁵¹	https://www.geneious.com/
CheckV (v.0.8.1)	Nayfach et al. ⁵²	https://bitbucket.org/berkeleylab/checkv/src/master/
RSEM (v.1.3.0)	Li and Dewey ⁵³	https://github.com/deweylab/RSEM
Bowtie2 (v.2.3.3.1)	Langmead and Salzberg ⁵⁴	https://github.com/BenLangmead/bowtie2
MAFFT (v.7.450)	Katoh and Standley ⁵⁵	https://github.com/GSLBiotech/mafft
InterProScan (v.5.56)	Jones et al. ⁵⁶	https://github.com/ebi-pf-team/interproscan
Phyre2 (v.2.0)	Kelley et al. ⁵⁷	http://www.sbg.bio.ic.ac.uk/phyre2
TrimAl (v.1.2)	Capella-Gutierrez et al. ⁵⁸	https://vicfero.github.io/trimal/
IQ-TREE (v1.6.12)	Nguyen et al. ⁵⁹	https://github.com/iqtree/iqtree2
FigTree (v.1.4.4)	N/A	https://github.com/rambaut/figtree/
eMPress (v.1.2)	Santichaivekin et al. ³¹	https://github.com/ssantichaivekin/empress
BEAST2 (v.2.7.3)	Bouckaert et al. ⁶⁰	https://github.com/CompEvol/beast2
BEAUti (v.2.7.3)	Drummond et al. ⁶¹	https://www.beast2.org/beauti/
Tracer (v.1.7.2)	Rambaut et al. ⁶²	https://github.com/beast-dev/tracer
TreeAnnotator (v.2.7.3)	N/A	https://github.com/beast-dev/beast-mcmc/blob/master/src/dr/app/tools/TreeAnnotator.java
R Package 'phyloseq' (v.1.42)	McMurdie and Holmes ⁶³	https://github.com/joey711/phyloseq
R Package 'ggplot2' (v.3.3.6)	Wickham ⁶⁴	https://github.com/tidyverse/ggplot2/releases
R Package 'ape' (v.5.6-2)	Paradis and Schliep ⁶⁵	https://github.com/cran/ape
R Package 'phytools' (v.1.0-3)	Revell ⁶⁶	https://github.com/liamrevell/phytools

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Prof. Edward C. Holmes (edward.holmes@sydney.edu.au).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This study analyses publicly available data from the NCBI/SRA database under Bioproject: PRJNA552202. All viral sequences discovered have been deposited in NCBI/GenBank under the accessions GenBank: OR350242-OR350390. All sequence alignments, phylogenetic trees, and tables are available on GitHub under the repository <https://github.com/vcosta16/cichlidvirome>.
- This study does not report original code.
- Any additional information required to reanalyse the data reported in this paper is available from the [lead contact](#) upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study did not use any experimental model or study participants.

METHOD DETAILS

Assembly of cichlid viromes

Brain, liver, gill, lower pharyngeal jaw, ovary and testis transcriptomes from 74 Tanganyikan cichlid species ($n = 445$ individuals, 2242 RNA sequencing libraries) were obtained from a previous study²⁵ (available at the NCBI Sequence Read Archive (SRA) under BioProject: PRJNA552202; Cichlid taxonomy followed ref. ¹⁹ and for detailed information on sampling localities refer to ref. ²⁵). Raw reads were quality trimmed using Trimmomatic (v.0.38)⁴⁸ with the parameters SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5, and MINLEN:25, and then *de novo* assembled into contigs using MEGAHIT (v.1.2.9)⁴⁹ with default parameter settings. Assembled contigs were compared against the protein version of the Reference Viral Database (RVDB) (August 2022), NCBI non-redundant protein (nr) and nucleotide (nt) databases (August 2022) using DIAMOND (BLASTX) (v.2.0.9) and BLASTn.^{50,67} To enable the identification of divergent viral sequences, we used an e-value search threshold of 1×10^{-5} as in previous studies.^{2,10–13} Contigs with top matches to the kingdom “Viruses” (NCBI taxid: 10239) were predicted as open reading frames (ORFs) using Geneious Prime (v.2022.0) (www.geneious.com).⁵¹ These represented 0.00002% of all contigs generated. All putative viral ORFs were translated into amino acid sequences and used as a query to perform a second search (DIAMOND BLASTP) against the NCBI nr database to remove false positives. ORFs with top matches to fish or bacterial genes were deemed as probable false positives and removed from further analysis. Viral contig contamination and completion was determined using CheckV.⁵² Contig abundances were calculated using RNA-Seq by Expectation Maximization (RSEM) (v.1.3.0) and coverage was assessed by mapping using Bowtie2 (v.2.3.3.1).^{53,54} Virus abundance plots were generated using the phyloseq (v.1.42) and ggplot2 (v.3.3.6) packages in R (v.4.2.2).^{63,64}

Identification and removal of endogenous viral elements

To determine whether our putative viral contigs were expressed endogenous viral elements (EVEs) rather than exogenous viruses, we screened for: (i) disrupted ORFs and (ii) flanking host regions using CheckV and BLASTn.^{10,52} Contigs that contained intact ORFs with no flanking host genes were compared to all available cichlid genomes using TBLASTN with an e-value search threshold of 1×10^{-20} . Sequences identified as EVEs (14% of all viral contigs) were removed from further analysis.

Taxonomic assignment of cichlid viruses

To taxonomically assign our newly discovered viruses, we aligned their predicted amino acid sequences (complete or partial genomes) with the complete genomes of related viruses available on NCBI/GenBank using MAFFT (v.7.450) (E-INS-i algorithm).⁵⁵ To infer whether our viruses represented novel species, we used levels of sequence similarity (e.g., amino acid p-distances) and phylogenetic grouping (see below) as specified by the International Committee of Viral Taxonomy (ICTV) (<https://talk.ictvonline.org>) for each viral genus/family. We also used these criteria to determine whether a virus was likely vertebrate-specific (i.e., infecting cichlid fishes) or associated with fish diet, microbiome or environment (i.e., “non-vertebrate”), particularly as viruses from these different major host groups are usually phylogenetically distinct.^{10,11,14} Viruses identified as likely of non-vertebrate origin were excluded from further analysis.

Annotation of cichlid viral genomes

Viral genomes were annotated with the “Live Annotate and Predict” tool implemented in Geneious using related sequences from NCBI/GenBank, with a similarity threshold of 20%. To annotate specific protein domains, we used InterProScan with the TIGRFAMs (v.15.0), SFLD (v.4.0), PANTHER (v.15.0), SuperFamily (v.1.75), PROSITE (v.2022_01), CDD (v.3.18), Pfam (v.34.0), SMART (v.7.1), PRINTS (v.42.0), and CATH-Gene3D databases (v.4.3.0).⁵⁶ To annotate divergent regions, we employed Phyre2 (confidence level of $\geq 95\%$) which uses homology detection methods for protein structure prediction.⁵⁷

Phylogenetic analysis

To infer the evolutionary histories of cichlid viruses, we first used MAFFT (E-INS-i algorithm) to align the sequences determined here with representative background sequences for each viral family/subfamily/genus. In instances where viruses identified in this study were detected on more than one occasion, a single sequence from each individual fish was selected for the alignment. In the case of RNA viruses, we utilised amino acid sequences of the most conserved RNA-dependent RNA polymerase (RdRp), while the DNA polymerase and capsid proteins were used for DNA viruses. Background sequences were selected from the ICTV and downloaded from NCBI/GenBank. Amino acid sequence alignments were trimmed using TrimAl (v.1.2) to remove ambiguously aligned regions with a gap threshold of 0.9 and a variable conserve value.⁵⁸ For each data set the best-fit model of amino acid substitution was estimated with the “ModelFinder Plus” (–m MFP) flag in IQ-TREE (Version 1.6.12).^{59,68} We used a maximum likelihood approach to estimate phylogenetic trees using IQ-TREE, with 1000 bootstrap replicates used to assess nodal support. All trees were annotated using FigTree (v.1.4.4) (<http://tree.bio.ed.ac.uk/software/figtree/>).

Co-phylogenetic analysis of cross-species virus transmission and co-divergence

To visualise the relative frequency of cross-species virus transmission and virus-host co-sdivergence within each viral group, we determined the co-phylogenetic relationship between viruses and their hosts. A complete phylogeny based on genome-wide data of Tanganyikan cichlids was downloaded²⁰ and pruned to those taxa in which a virus was identified using the R packages phytools (v.1.0-3)⁶⁶ and ape (v.5.6-2).⁶⁵ A cladogram of the non-cichlid hosts was generated using the phyloT software, a phylogenetic tree

generator based on NCBI taxonomy (<http://phylot.biobyte.de/>). The pruned cichlid and non-cichlid cladograms were then concatenated to form a host cladogram for each viral group. Tanglegrams that illustrate the link between host and virus trees were created using phytools and ape. The virus phylogenies used in the co-phylogenies were constructed as described above. The relative frequencies of cross-species virus transmission versus virus-host co-divergence were quantified using eMPress,³¹ which uses a maximum parsimony approach to determine the optimal “map” of the virus phylogeny onto the host phylogeny. The event costs followed those in the literature,^{1,2,12,13} specifically, duplication, host-jumping, and extinction event types were set to 1.0, whereas the cost of virus-host co-divergence was set to zero since it is regarded as the null event. The number of generations and the population size for both virus and host were set to 100.

Time calibration and diversification rates of cichlid hepaciviruses

We used BEAST2 (v.2.7.3)⁶⁰ to estimate a time-calibrated phylogeny of representative ray-finned fish (Actinopterygii) hepaciviruses, utilising the NS5 amino acid sequence alignment (see above). To achieve this, we employed a Yule speciation model with a single, normally distributed calibration point at the root node of the cichlid hepacivirus clade based on prior age estimates of the formation of Lake Tanganyika (9–12 Ma) and the tMRCA of the cichlid radiation that was recently estimated to have diverged around 9.7 Ma.^{20,23} Our calibration was therefore set with a mean of 10.5 and standard deviation of 0.6 (interval 9–12 Ma). The best-fit evolutionary model was selected using BEAUti (v.2.7.3).⁶¹ We tested a combination of JTT, WAG, and LG amino acid substitution models with strict and relaxed clocks (lognormal or gamma) with an Effective Sample Size (ESS) cut-off of 200. Based on tree topology, ESS and marginal likelihoods, a strict clock with the LG substitution model, a gamma distribution of among-site rate variation and a proportion invariant of 0.077 were identified as the most appropriate model. We ran the model using a Markov chain Monte Carlo (MCMC) chain with 50 million iterations, sampling every 10000, with convergence and ESS assessed using Tracer (v.1.7.2).⁶² The maximum clade credibility (MCC) tree was constructed using TreeAnnotator (v.2.7.3) after a 10% burn-in protocol (<https://www.beast2.org/treeannotator/>). Finally, we computed the gamma statistic to test the null hypothesis that the rate of hepacivirus diversification was constant over the cichlid radiation using the ape package in R.^{35,65}

Phylogenetic and ecological diversity analysis

To determine whether virus transmission across Lake Tanganyika cichlids was random with respect to phylogenetic position and the ecology of the species, we evaluated the ecological and phylogenetic diversity of the infected and uninfected host species compared to random subsets. Phylogenetic diversity was calculated as the mean pairwise phylogenetic distances, which were extracted from the host tree using the cophenetic function in the R package ape (v.5.6-2).⁶⁵ Ecological diversity was calculated as the mean pairwise Euclidean distance of available carbon ($\delta^{13}\text{C}$) and nitrogen ($\delta^{15}\text{N}$) stable-isotope measurements of the Tanganyika cichlid species.²⁰ We then assessed whether the diversity of the infected and uninfected set of species (observed) is smaller than the diversity of a random subsample of the radiation (random) by randomly subsampling the total number of infected and uninfected species from the entire radiation 1000 times. If the observed diversity is lower (random-observed > 0) in more than 95% of the random samples (one-tailed test), we considered the difference significant (transmission and infectivity is assortative). This resampling approach was applied once across all viruses together, and once separately for each virus group that was present in at least five species in our data set.

QUANTIFICATION AND STATISTICAL ANALYSIS

Viral contig abundances were estimated using RSEM as described under [method details](#). We used linear regression to test for a correlation between observed viral richness (i.e., number of viruses) and the number of cichlid species examined for each tribe. Subsampling analysis was performed to determine whether virus transmission was random with respect to phylogenetic position and host species ecology. Phylogenetic diversity was calculated as the mean pairwise phylogenetic distances, while ecological diversity was calculated as the mean pairwise Euclidean distance of available carbon ($\delta^{13}\text{C}$) and nitrogen ($\delta^{15}\text{N}$) stable-isotope measurements. A one-tailed test was employed to test for any significant differences. All statistical analyses were performed using R. The R packages can be found under [method details](#) and [key resources table](#). Support for nodes in phylogenetic trees were calculated using bootstrap methods as described under [method details](#).



OPEN ACCESS

EDITED BY

Jie Cui,
Fudan University, China

REVIEWED BY

Akio Adachi,
Tokushima University, Japan

*CORRESPONDENCE

Uri Neri
✉ uri.neri@gmail.com

RECEIVED 17 January 2024

ACCEPTED 15 February 2024

PUBLISHED 04 April 2024

CITATION

Charon J, Olendraite I, Forgia M, Chong LC, Hillary LS, Roux S, Kupczok A, Debat H, Sakaguchi S, Tahzima R, Nakagawa S, Babaian A, Abroi A, Beijerman N, Ben Mansour K, Brown K, Butkovic A, Cervera A, Charriat F, Chen G, Chiba Y, De Coninck L, Demina T, Dominguez-Huerta G, Dubrulle J, Gutierrez S, Harvey E, Jayaraj Mallika FR, Karapliafis D, Lim SJ, Kasibhatla SM, Mifsud JCO, Nishimura Y, Ortiz-Baez AS, Raco M, Rivero R, Sadiq S, Saghaei S, San JE, Shaikh HM, Sieradzki ET, Sullivan MB, Sun Y, Wille M, Wolf YI, Zrelavs N and Neri U (2024) Consensus statement from the first RdRp Summit: advancing RNA virus discovery at scale across communities. *Front. Virol.* 4:1371958. doi: 10.3389/fviro.2024.1371958

COPYRIGHT

© 2024 Charon, Olendraite, Forgia, Chong, Hillary, Roux, Kupczok, Debat, Sakaguchi, Tahzima, Nakagawa, Babaian, Abroi, Beijerman, Ben Mansour, Brown, Butkovic, Cervera, Charriat, Chen, Chiba, De Coninck, Demina, Dominguez-Huerta, Dubrulle, Gutierrez, Harvey, Jayaraj Mallika, Karapliafis, Lim, Kasibhatla, Mifsud, Nishimura, Ortiz-Baez, Raco, Rivero, Sadiq, Saghaei, San, Shaikh, Sieradzki, Sullivan, Sun, Wille, Wolf, Zrelavs and Neri. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Consensus statement from the first RdRp Summit: advancing RNA virus discovery at scale across communities

Justine Charon¹, Ingrida Olendraite², Marco Forgia³, Li Chuin Chong⁴, Luke S. Hillary⁵, Simon Roux⁶, Anne Kupczok⁷, Humberto Debat^{8,9}, Shoichi Sakaguchi¹⁰, Rachid Tahzima^{11,12}, So Nakagawa¹³, Artem Babaian^{14,15}, Aare Abroi¹⁶, Nicolas Beijerman^{8,9}, Karima Ben Mansour^{17,18}, Katherine Brown², Anamarija Butkovic¹⁹, Amelia Cervera²⁰, Florian Charriat²¹, Guowei Chen²², Yuto Chiba^{23,24}, Lander De Coninck²⁵, Tatiana Demina²⁶, Guillermo Dominguez-Huerta²⁷, Jeremy Dubrulle²⁸, Serafin Gutierrez^{21,29}, Erin Harvey³⁰, Philmar Raj Jayaraj Mallika³¹, Dimitris Karapliafis⁷, Shen Jean Lim³², Sunitha Manjari Kasibhatla^{33,34}, Jonathon C. O. Mifsud³⁰, Yosuke Nishimura²⁴, Ayda Susana Ortiz-Baez³⁰, Milica Raco³⁵, Ricardo Rivero³⁶, Sabrina Sadiq³⁰, Shahram Saghaei³⁷, James Emmanuel San^{38,39}, Hisham Mohammed Shaikh^{40,41}, Ella Tali Sieradzki⁴², Matthew B. Sullivan⁴³, Yanni Sun²², Michelle Wille⁴⁴, Yuri I. Wolf⁴⁵, Nikita Zrelavs⁴⁶ and Uri Neri^{47*}

¹Fruit Biology and Pathology Unit, University of Bordeaux, INRAE, Bordeaux, France, ²Division of Virology, Department of Pathology, Addenbrookes Hospital, University of Cambridge, Cambridge, United Kingdom, ³Istituto per la Protezione Sostenibile Delle Pianta, CNR, Torino, Italy, ⁴Institute for Experimental Virology, TWINCORE Centre for Experimental and Clinical Infection Research, a Joint Venture Between the Hannover Medical School (MHH) and the Helmholtz Centre for Infection Research (HZI), Hannover, Germany, ⁵Department of Plant Pathology, University of California, Davis, Davis, CA, United States, ⁶DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, United States, ⁷Bioinformatics Group, Wageningen University, Wageningen, Netherlands, ⁸Instituto de Patología Vegetal, Centro de Investigaciones Agropecuarias, Instituto Nacional de Tecnología Agropecuaria (IPAVE, CIAP, INTA), Córdoba, Argentina, ⁹Unidad de Fitopatología y Modelización Agrícola, Consejo Nacional de Investigaciones Científicas y Técnicas, Córdoba, Argentina, ¹⁰Department of Microbiology and Infection Control, Faculty of Medicine, Osaka Medical and Pharmaceutical University, Osaka, Japan, ¹¹University of Liège, Gembloux AgroBioTech - TERRA, Gembloux, Belgium, ¹²Flanders Research Institute for Agriculture, Fisheries and Food, Plant Sciences Unit - Virology, Ghent, Belgium, ¹³Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Japan, ¹⁴Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada, ¹⁵The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada, ¹⁶Institute of Technology, University of Tartu, Tartu, Estonia, ¹⁷Ecology, Diagnostics and Genetic Resources of Agriculturally Important Viruses, Fungi and Phytoplasmas, Crop Research Institute, Prague, Czechia, ¹⁸Department of Plant Protection, Faculty of Agrobiology, Food and Natural Resources Czech University of Life Sciences, Prague, Czechia, ¹⁹Archaeal Virology Unit, Institut Pasteur, Université Paris Cité, CNRS, Paris, France, ²⁰Instituto de Biología Molecular y Celular de Plantas (IBMCP), Consejo Superior de Investigaciones Científicas (CSIC) - Universitat Politècnica de València (UPV), València, Spain, ²¹ASTRE, CIRAD, INRAE, University of Montpellier, Montpellier, France, ²²Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China, ²³School of Agriculture, Meiji University, Kawasaki, Japan, ²⁴Research Centre for Bioscience and Nanoscience, Japan Agency for Marine-Earth Science and

Technology (JAMSTEC), Yokosuka, Japan, ²⁵KU Leuven, Department of Microbiology, Immunology, and Transplantation, Rega Institute, Division of Clinical and Epidemiological Virology, Laboratory of Viral Metagenomics, Leuven, Belgium, ²⁶Department of Microbiology, Faculty of Agriculture and Forestry, University of Helsinki, Helsinki, Finland, ²⁷Department of Oceanography and Global Change, Centro Oceanográfico de Málaga (IEO-CSIC), Fuengirola, Spain, ²⁸Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand, ²⁹Department of Virology, Montpellier University Hospital, Montpellier, France, ³⁰Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW, Australia, ³¹Molecular Virology Laboratory, ICAR-National Research Centre for Banana, Tiruchirappalli, Tamil Nadu, India, ³²College of Marine Science, University of South Florida, St Petersburg, FL, United States, ³³Bioinformatics Centre, SP Pune University, Pune, India, ³⁴High Performance Computing (HPC)-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing, Pune, India, ³⁵Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, United States, ³⁶Paul G. Allen School for Global Health, Washington State University, Pullman, WA, United States, ³⁷Bioinformatics/High-Throughput Analysis, Faculty of Mathematics and Computer Science, Friedrich Schiller University Jena, Jena, Germany, ³⁸KwaZulu-Natal Research and Innovation Sequencing Platform (KRISP), UKZN, Durban, South Africa, ³⁹School of Laboratory Medicine and Medical Sciences, University of KwaZulu-Natal, Durban, South Africa, ⁴⁰Research Department, Flanders Marine Institute (VLIZ), InnovOcean Site, Ostend, Belgium, ⁴¹Department of Microbiology and Biochemistry, Faculty of Sciences, Ghent University, Ghent, Belgium, ⁴²Laboratoire Ampère, École Centrale de Lyon, Écully, France, ⁴³Departments of Microbiology and Civil, Environmental, and Geodetic Engineering, and Center of Microbiome Science, Ohio State University, Columbus, OH, United States, ⁴⁴Centre for Pathogen Genomics, Department of Microbiology and Immunology, University of Melbourne, at the Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia, ⁴⁵National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, United States, ⁴⁶Latvian Biomedical Research and Study Centre, Riga, Latvia, ⁴⁷The Shmunis School of Biomedicine and Cancer Research, Tel Aviv University, Tel Aviv, Israel

Improved RNA virus understanding is critical to studying animal and plant health, and environmental processes. However, the continuous and rapid RNA virus evolution makes their identification and characterization challenging. While recent sequence-based advances have led to extensive RNA virus discovery, there is growing variation in how RNA viruses are identified, analyzed, characterized, and reported. To this end, an RdRp Summit was organized and a hybrid meeting took place in Valencia, Spain in May 2023 to convene leading experts with emphasis on early career researchers (ECRs) across diverse scientific communities. Here we synthesize key insights and recommendations and offer these as a first effort to establish a consensus framework for advancing RNA virus discovery. First, we need interoperability through standardized methodologies, data-sharing protocols, metadata provision and interdisciplinary collaborations and offer specific examples as starting points. Second, as an emergent field, we recognize the need to incorporate cutting-edge technologies and knowledge early and often to improve omic-based viral detection and annotation as novel capabilities reveal new biology. Third, we underscore the significance of ECRs in fostering international partnerships to promote inclusivity and equity in virus discovery efforts. The proposed consensus framework serves as a roadmap for the scientific community to collectively contribute to the tremendous challenge of unveiling the RNA virosphere.

KEYWORDS

RNA virus discovery, viral metagenomics, RNA-dependent RNA polymerase, viral genome annotation, metagenomic metadata standards, virus evolution and diversity

1 Introduction

RNA viruses (*Orthornavirae*) are genetic elements with RNA-based genomes that replicate using their encoded RNA-dependent RNA polymerase (RdRp) and by hijacking their host's cellular machinery. Progeny viruses are then transmitted to new hosts either vertically, or horizontally most often in protein-based viral particles that can sometimes be surrounded by a lipid envelope. Viruses are widely diverse, infect all life forms (1), and include many human pathogens of medical and epidemiological importance (2), as well as various species with strong deleterious impact in agriculture (3). Additionally, by infecting unicellular eukaryotic and prokaryotic life forms, RNA viruses play a role in shaping microbial ecosystems, from the oceans to the human gut (4–8).

Historically, RNA virus discovery and characterization relied on direct cultivation or isolation of the infective agents via experiments that are often laborious and inherently restricted to viruses infecting hosts amenable to laboratory cultivation or propagation. The procedures typically involve the concentration of infectious particles from symptomatic or diseased host cells or tissues, followed by various identification techniques like microscopy (imaging), neutralization (antibody), hemadsorption, hemagglutination and plaque assays, and animal, plant, tissue or cell culture inoculation. In most cases, isolated concentrated particles would then undergo (viral) RNA extraction and purification, followed by reverse transcription into cDNA and subsequent sequencing enabling further genomic investigations (phylogeny, genotyping, etc).

1.1 Recent developments in omic-based RNA virus discovery: more, bigger, faster is the new pace

The advent of the genomic era has gradually expanded RNA virus discovery beyond experimental cultivation and isolation methodologies. The substantial decrease in costs associated with high-throughput nucleic acid sequencing, coupled with advances in computational capacities for big data storage and processing greatly facilitates the development of RNA virus discovery projects through large-scale sequencing (omic-based). Importantly, this genomic-data-driven exploration of the RNA virosphere using computational tools offers a unique opportunity to bypass many biases and limitations of traditional approaches, and goes hand in hand with the growing recognition of global viral diversity in ecological systems as a whole, including public health and one health [“pandemic preparedness”, surveillance - (9)].

With an expanded diversity of environments sampled as well as a growing re-assessment of publicly available sequencing data and the continuous development of tools and resources, this field has experienced massive growth in recent years, with no signs of deceleration in sight (Figure 1) (8, 10–29).

1.2 Interdisciplinarity nature of the field limits data uniformity and standardization

RNA virus discovery is at the interface of various disciplines (*i.e.* virology, molecular, and structural biology, evolution, genomics, ecology, and computational sciences) and spans various fields within virology itself, each with its specific virus groups of interest, priorities, approaches, concepts, resources, etc. leading to apparent discrepancies throughout the scientific process. This heterogeneity manifests in both the experimental design, in the interpretation of the data and in the eventual conclusions and data sharing (novelty estimation, host inference, risk assessment, choice of data deposition location).

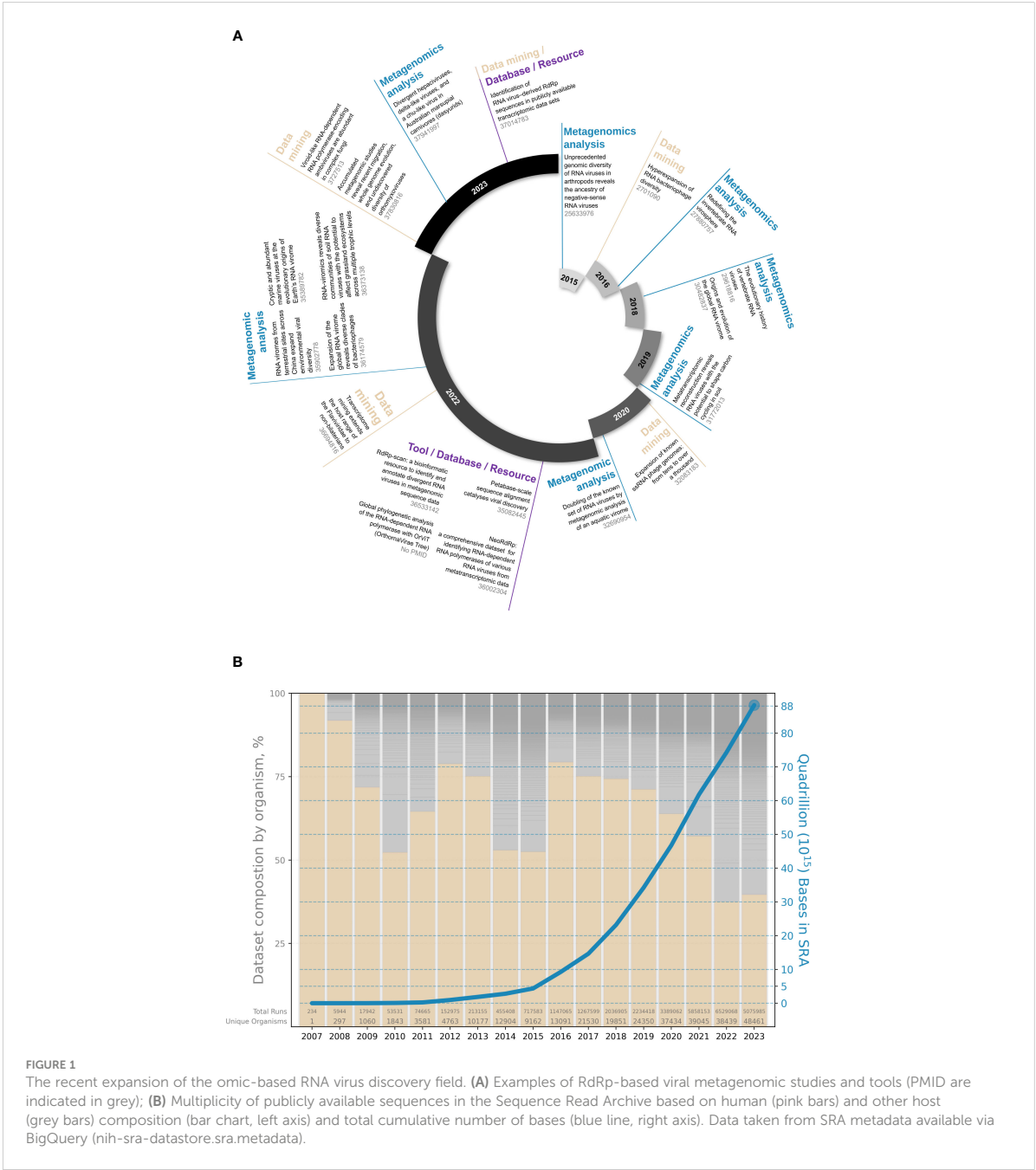
This growing lack of standardization directly and severely hampers interoperability - *i.e.* the ability to review, compare, reproduce, share, and build on each other's efforts. For instance, what one study may consider as a new RNA virus group based on coat or movement protein sequence similarity, another study may consider part of an existing group using the RdRp comparison.

In recognition of these issues as detrimental to the advancement of the field, we recently held the first “RdRp summit” (<https://rdrp.io/>) - a discussion-centric event with the goals of fostering reproducibility, collaboration and interoperability in omics-derived RNA virus discovery. The event was attended by over 70 participants (60% in-person and 40% remotely), from 50 research institutions across the world. Most attendees were ECRs, with half of the participants listed as PhD students. To promote inclusion and exchanges between all participants, the meeting featured both open discussion sessions and traditional lectures, given by key bioinformaticians and experimentalists. Herein, we summarize the major insights and consensus that emerged from the workshop.

2 Current challenges in RNA virus discovery

2.1 Multiplicity of experimental and computational practices in RNA virus discovery workflows

The initial source and type of environment, the preservation, and preparation of RNA input have profound implications on the whole analysis and final RNA virus discovery. The input for RNA-virus metagenomic studies is often the total extractable RNA from an environmental, vector or host-associated sample (12, 17, 30). Alternatively, studies can focus on a size-selected fraction where host cells are excluded and virus-like-particles (VLPs) are enriched using filtration and/or centrifugation, otherwise known as viromics or virion-associated nucleic acids (VANA)-based sequencing (22, 31). Double-stranded RNA (dsRNA) purification can also be applied to total RNA samples to specifically target dsRNA virus genomes and replicative intermediate of single-stranded RNA (ssRNA) viruses instead of single-stranded transcripts and



ribosomal RNAs (32). Beyond dsRNA enrichment, targeted approaches such as the Fragmented and primer-Ligated dsRNA Sequencing (FLDS) method also feature the ability to sequence both ends of the genome (33, 34), from which pairs of segmented or multipartite viral genomes can be searched (33). Deep sequencing of small RNAs (sRNAs) can also be advantageous to plant and mycovirus discovery, by using various sRNA size profiles depending on the organism for RNA genome assembly (35–40). On the other

hand, untargeted total RNA extraction followed by RNA-seq better reflects the global sample complexity, including host and viral diversity, and can assist with host association, further answering ecological questions (41). However, the choice of the kit used for RNA extraction can substantially influence the downstream analysis (42). Prior to sequencing, classical treatments include genomic DNA digestion and either targeted ribosomal RNA (rRNA) depletion, or poly(A) enrichment steps prior to reverse

transcription of RNA into cDNA. Along with the choice of sequencing platforms/technologies, those methods will directly impact the subsequent RNA virus findability and identification.

Current computational identification of RNA viruses from “-omic” data (mainly transcriptomic and metatranscriptomic sequencing, *i.e.* bulk RNA-seq of either a single organism or a community of organisms, respectively) is typically conducted *via* direct comparative approaches following quality control and filtering of raw reads and *de novo* assembly. The discovery of viruses through omics data primarily depends on identifying sequence similarities with existing RNA virus genomes, protein sequences, or protein sequence profiles. This is often accomplished using methods such as Hidden Markov Models (HMMs) or Position-Specific Scoring Matrices (PSSMs). Similarity is defined as a set of minimal statistical thresholds; typically established arbitrarily in each study and thus further stresses the need for standardized protocols.

Predominantly, the RdRp, which is the only protein shared by all known RNA viruses, is used as the marker gene for RNA virus identification (14). Virus RdRps share a right-hand-shape structure, typical of DNA/RNA polymerases, with a palm-based active site comprising several catalytic and structural motifs, which may require additional host factors to constitute a mature, complete replicase domain (43). Assignment of a query sequence as a potential viral RdRp usually requires the identification of at least the three “core” motifs, commonly referred to as A, B, and C, with the presence of any additional motifs increasing the reliability of the assignment and the presumed completeness of the analyzed sequence (44). The presence of these motifs (or roughly the region they occupy) is most often identified via sequence search engines (BLAST, DIAMOND, MMseqs, etc) (45–47) or profile-based approaches (HMM via HH-Suite or HMMER, or PSSM/PWM using PSI-BLAST or MEME) (48–50), often used in conjunction with public databases and repositories of RNA virus-derived RdRp *e.g.* subsetting NCBI nr, or custom databases like TSA-database derived RdRps (29), NeoRdRp (23), Palmscan (51), or RdRp-scan (18).

Furthermore, the enhancement of RNA extraction methods, sequencing technologies, and the rapid advancement in the development of new AI-based techniques, among other factors, is playing a crucial role in advancing RNA virus discovery. These advances facilitate the improved identification of potentially divergent and low-concentration viruses within overlooked environmental or host taxa. Nevertheless, these developments also reinforce the methodology gaps and heterogeneity between studies and severely limit interoperability.

2.2 Consequences of procedural inconsistency for comparative analyses across studies

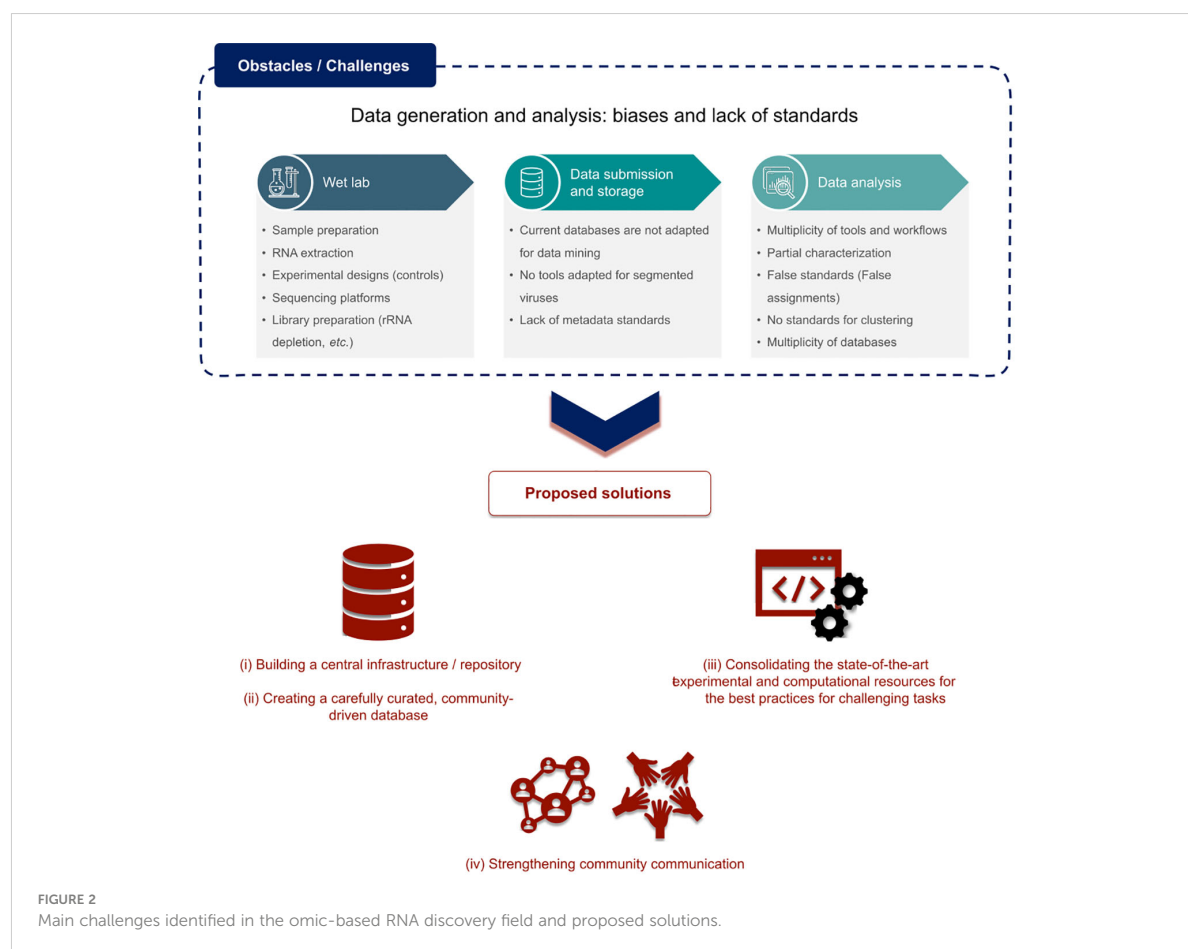
Choices in both the experimental procedures and subsequent *in silico* analyses play a crucial role in how different studies handle, share, report, and reach conclusions regarding the suspected viral sequences in the corresponding data. Discussions held during this

first RdRp summit pinpointed the global lack of procedure and good practice standards at every level (sampling, extraction, sequencing, read and contig processing, as well as data analysis, storage, submission, mining, survey, etc) (Figure 2). Ultimately, those differences in the computational and experimental aspects can constitute strong obstacles for the ability to adequately compare the results of different studies and it seemed important to first identify them.

Regarding RNA virus detection itself, there is a crucial lack of standard minimal alignment statistics (*e-value*, %ID, %coverage, etc). Also, the inconsistency of what is considered a genuine viral genome/viral hit/viral sequence versus a potential “false-positive” or contaminant poses a major risk of misinterpretation of results. The ability to identify and discriminate true replicative RNA virus signals from active or integrated viruses replicating via reverse-transcriptases (RT) (divergent other palm-like polymerases) (kingdom *Pararnavirae*), endogenous viral elements (EVEs), and non-viral hits or contaminants, is absolutely crucial in our field. However, there is currently no widely agreed-upon consensus about defining quality standards for viral sequences and how to ensure their identification as such. Compounding this issue is the lack of definition for real RNA-virus derived sequences that are either chimeric or misassembled, and thus are not likely to represent a functional infectious entity. Plus, expanding our knowledge of the RNA virosphere revealed an ever-increasing plurality of genome architectures and RdRp properties, which make it even harder to define one single rule for all of them. The recently-described divided RdRps confirmed and validated *in silico* (52–54), which are encoded by two distinct ORFs from separate genomic segments, constitute the best example of such unexpected plurality. Such challenges require continuous adaptation of standard practices and motivate to establish community-driven, up-to-date guidelines for RNA virus discovery.

Standardizing RNA virus detection would strongly require a community-built consensus about performance evaluation pipelines (sensitivity, recall, F1, precision, algorithm resilience, etc.), similarly to the ongoing efforts in microbial and DNA virus metagenomics (55–58). Directly linked to this, unequivocal agreements on the plurality of operational taxonomic unit (OTU) definitions, clustering thresholds, and minimal procedure for genome completeness estimation of novel and divergent viruses will help set gold standards for the scientific community (Figure 2).

The aforementioned considerations would also dramatically decrease common inconsistencies regarding the multiplicity of repositories that host the data as well as metadata associated with viral metagenomics projects deposited in standard databases. Indeed, one could note the major confusion between host and sample source, arbitrary taxonomy assignment and gene, protein, and genome annotations, the lack of information relating to sample preparation, sequencing, and computational analysis, the inadequacy of current tools for uploading viruses with divided RdRps or segmented genomes, or sequences with alternative START codons, and the inability for external users to revise/re-assess/edit/annotate the deposited metadata. All of which leads to an absurd rate of unclassified/unannotated sequences when dealing with remote homology searches.



Considering the rising scale of viral metagenomic studies and the pace of virus taxonomy expansion, such misannotations or misassignments in reference databases can have dramatic consequences when propagated to new studies and data submission, and drastically limit the scope and efficiency of data mining projects, yet are increasingly essential in our field.

3 Solutions and future perspectives

While incredibly valuable, the interdisciplinarity in the RNA virus discovery field also requires concerted efforts from researchers to build connections between those communities, share, and adapt our respective practices, tools, vocabulary, terminologies, and standards to fit in with everyone's domain language.

To tackle the first challenge consisting of establishing minimum standards for RNA virus genome (or viral RdRp) annotations, cut-off for parameters (alignment scores, e-values, query and reference lengths, etc.) need to be agreed upon when comparing candidate sequences to known valid RdRp sequence database *vs.* decoy - (RT)-like - databases and sets of unclassified/unannotated sequences. Annotation could then be automatically assigned based on these comparison scores (true complete RdRp *vs.* RT-like hits *vs.*

unclassified/unknown). Discussions and lectures at the summit also highlighted the importance of integrating additional procedures into classical workflows such as placement within phylogenies, genomic context scan (untranslated regions, RNA structure, nucleotide and kmer composition among others), and structural homology assessment using cutting-edge AI-based prediction tools (59, 60), essential for distant homology detection or validation. While some quality criteria and cut-offs can easily be built in, some others may be very challenging. Defining boundaries for the RdRp gene (minimal length to describe a RdRp, structural attributes, minimal presence of the catalytic motifs, presence/absence of additional domains, such as Nidovirus RdRp associated nucleotidyl transferase domain - NiRAN) remains a complex task and requires expertise and extended knowledge of the viral strain (61–63).

In addition, the standardization effort should also promote the integration into the standard discovery pipelines of the most recent and state-of-the-art concepts in RNA virology such as the search for potential additional segments (64), the screen for divided RdRp (52–54), etc.

Another challenge consists of homogenization of manual and automatic clusterization procedures and taxonomic assignment of viral-like sequences. Formal virus classification by the International

Committee on Taxonomy of Viruses (ICTV) plays a vital role in providing a reference language for scientists to communicate, collaborate, and share knowledge about viruses. By coupling the robust, updated, and standardized ICTV classification framework with the power of omic-based RNA virus discovery, we can collectively improve our description and global understanding of RNA virus diversity (65).

Concretely, all these proposals will be pursued by the RdRp summit community with different initiatives that are intended to be maintained and updated over time by the community members. In particular, the community will focus on:

- I) Building a central infrastructure for the RNA virus discovery community, which would work as a central repository for data and knowledge.
- II) Creating a curated database with automated quality scores for data deposition based on the information provided. The database should be community-driven and open to feedback from end users to aid further curation.
- III) Consolidating state-of-the-art experimental and computational resources and knowledge to ensure the best up-to-date practices among the community (pipelines, scripts, protocols, glossary, guides, international journal club). Among others, this could consist of recommended workflows for the challenging identification of novel/remote RdRp, the Do's and Don'ts of annotating new, uncultivated RNA virus genomes, and metadata recommendations for RNA viruses.
- IV) In the same manner as the European Virus Bioinformatics Center (<https://evbc.uni-jena.de/>), particular effort will be put into enhancing the communication and uniting our expanding community into one single spot through a potential membership system, forums, round tables, workshops, online chat channels, etc.

To conclude, the major consensus that emerged from the rich discussions led during this first RdRp summit lies in the current lack of interoperability and reproducibility in our field but also the possible concrete solutions to tackle these obstacles (summarized in Figure 2).

The omic-based RNA virus discovery community, as an open science-to-society-oriented community, should be aware of its roles and responsibilities to make its scope as transparent and accessible as possible. Through the collective development of a user-friendly open platform, we aim to build a solid foundation for communicating, sharing, and performing comparable analyses using optimal and state-of-the-art tools across a wide array of biological contexts by reaching the broadest audience possible. With similar issues faced in microbial metagenomics and omic-based DNA virus discovery, we also intend to inspire from the emerging solutions and infrastructures being developed in these related fields and learn from their experiences in tackling these challenges (e.g. 66–68).

We believe these efforts will lay the groundwork to promote ECRs insertion into the community, best practices and repeatability, and ultimately ensure the best future for our exploration of the RNA virosphere.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

JC: Writing – original draft, Writing – review & editing, Conceptualization, Project administration. IO: Writing – original draft, Writing – review & editing, Visualization. MF: Writing – original draft, Writing – review & editing. LC: Writing – original draft, Writing – review & editing, Visualization. LH: Writing – original draft, Writing – review & editing. SR: Writing – original draft, Writing – review & editing. AK: Writing – original draft, Writing – review & editing. HD: Writing – original draft, Writing – review & editing. SS: Writing – original draft, Writing – review & editing. RT: Writing – original draft, Writing – review & editing. SN: Writing – original draft, Writing – review & editing. ABa: Writing – original draft, Writing – review & editing. AA: Writing – review & editing. NB: Writing – review & editing. KB: Writing – review & editing. KBM: Writing – review & editing. ABu: Writing – review & editing. AC: Writing – review & editing. FC: Writing – review & editing. GC: Writing – review & editing. YC: Writing – review & editing. LDC: Writing – review & editing. TD: Writing – review & editing. GD: Writing – review & editing. JD: Writing – review & editing. SG: Writing – review & editing. EH: Writing – review & editing. FJ: Writing – review & editing. DK: Writing – review & editing. SL: Writing – review & editing. SMK: Writing – review & editing. JM: Writing – review & editing. YN: Writing – review & editing. AO: Writing – review & editing. MR: Writing – review & editing. RR: Writing – review & editing. SS: Writing – review & editing. SS: Writing – review & editing. JS: Writing – review & editing. HS: Writing – review & editing. ES: Writing – review & editing. MS: Writing – review & editing. YS: Writing – review & editing. MW: Writing – review & editing. YIW: Writing – review & editing. NZ: Writing – review & editing. UN: Writing – original draft, Writing – review & editing, Conceptualization, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. JC was supported by the French National Research Agency (ANR-21-CE35-0009). UN was supported by the European Research Council (ERC-AdG 787514) and by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University, and by fellowship from the Shmunis School of Biomedicine and Cancer Research. The work conducted by the U.S. Department of Energy Joint Genome Institute (S.R.) (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. AB was supported by the postdoctoral fellowship from Fondation Recherche Médicale (FRM) and

Canadian Institutes for Health Research (PJT-496709). ABu was supported by a postdoctoral fellowship from Fondation Recherche Médicale (FRM). KBM was supported by the Ministry of Agriculture of the Czech Republic, institutional support (MZE-RO0423). RT was funded by the Belgian Federal Public Service Health, Food Chain Safety and Environment through the contract (GenoPREDICT - RF 22/635). ES was supported by Marie Skłodowska-Curie fellowship “DIVOBIS” awarded by Horizon Europe 2022. SN was supported by KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas 16H06429, 16K21723, 19H04843 and by JST CREST JPMJCR20H6. YIW was supported by intramural funds of the US Department of Health and Human Services (National Institutes of Health, National Library of Medicine). EH was supported by AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong Special Administrative Region, China. AO-B was supported by a National Health and Medical Research Council (NHMRC) Investigator Grant (GNT2017197) awarded to Prof. Edward C. Holmes. YS was supported by Hong Kong Research Grants Council (RGC) General Research Fund (GRF) (11206819, 11217521) and Hong Kong Innovation and Technology Fund (ITF) MRP/071/20X. RR was supported by funding to Verena (viralemergence.org) from the U.S. National Science Foundation, including NSF BII 2021909 and NSF BII 2213854 and by a Chan Zuckerberg Initiative (CZI) Global Voice Travel Award from CZI, Redwood City, California, USA. YC and YN were supported by JSPS KAKENHI Grant Number JP22H05714. YC was supported by JSPS KAKENHI Grant Number JP23KJ1995. TD was supported by the Research Council of Finland grant 330977 and the Kone Foundation grant. FM was supported by a Chan Zuckerberg Initiative (CZI) Global Voice Travel Award from CZI, Redwood City, California, USA. DK and AK were supported by the Graduate School Experimental Plant Sciences (project V-GENE), LDC was supported by the Research Foundation Flanders (11L1323N). IO and KB were supported by Wellcome Trust 220814/Z/20/Z awarded to Prof Andrew Firth. HS was supported by Flanders Marine Institute (VLIZ). AC was supported by Ministerio de Ciencia e Innovación of Spain and FEDER grant PID2020-116008GB-I00 awarded to Marcos de la Peña. JS was supported by the South African Medical Research Council (SAMRC) with funds received from the National Department of Health. MS was supported by the US National Science Foundation awards #OCE1829831 and ABI#2149505. FC and SG were supported by the French Direction Générale de l’Alimentation (grant C-2023-016). LH was supported by the U.S.

Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research (BER), Genomic Science Program, award number DE-SC0023127 grant to Joanne B. Emerson, grant PI Sydney Glassman. SS was supported by KAKENHI Grant-in-Aid for Scientific Research (22K14999).

Acknowledgments

We would like to thank Amy Heather Fitzpatrick, Anastasia Gulyaeva, Chien-Fu Wu, Heli Mõnttinen, Liubov Chuprikova, Fabiana Neves, Nikolay Simankov, Noriko Cassman, Nuria Fontdevila Pareta, Rebecca Grimwood, Stephanie Waller, Zivile Buivydaite, Alexander Allman, Artem Baidaliuk, Benjamin Lee, Cadhla Firth, Chuan Cao, Clarence Le, Luca Nishimura, Dehan Cai, Manja Marz, Elizabeth Fahsbender, Gabriel Lencioni Lovate, Hassan Z. A. Ishag, Isa de Vries, Janelle Wierenga, Javier Rodriguez-Grille, Jordan Taylor, Katarina Bačnik, Katrina Kalantar, Lauren Lim, Mark Paul S. Rivarez, Nina A. H. Madsen, Nolwenn Dheilly, Olve Peersen, Robert Edgar, Sandra Triebel, Satyabrata Satapathy, Sung won Lim, Xubo Tang and Yuanyuan Zhang for their precious contribution to the RdRp summit and the fruitful exchanges. We are also thankful to the sponsors of the first RdRp summit, the Chan-Zuckerberg Initiative (CZI) and Frontiers in Virology, for their support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Dominguez-Huerta G, Wainaina JM, Zayed AA, Culley AI, Kuhn JH, Sullivan MB. The RNA virosphere: How big and diverse is it? *Environ Microbiol.* (2023) 25:209–15. doi: 10.1111/1462-2920.16312
2. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. *Nature.* (2008) 451:990–3. doi: 10.1038/nature06536
3. Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. The global burden of pathogens and pests on major food crops. *Nat Ecol Evol.* (2019) 3:430–9. doi: 10.1038/s41559-018-0793-y
4. Lang AS, Rise ML, Culley AI, Steward GF. RNA viruses in the sea. *FEMS Microbiol Rev.* (2009) 33:295–323. doi: 10.1111/j.1574-6976.2008.00132.x
5. Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G. Are we missing half of the viruses in the ocean? *ISME J.* (2013) 7:672–9. doi: 10.1038/ismej.2012.121
6. Liang G, Bushman FD. The human virome: assembly, composition and host interactions. *Nat Rev Microbiol.* (2021) 19:514–27. doi: 10.1038/s41579-021-00536-5

7. Dominguez-Huerta G, Wainaina JM, Zayed AA, Culley AI, Kuhn JH, Sullivan MB, et al. The RNA virosphere: How big and diverse is it? *Environ Microbiol.* (2022) 25:209–15. doi: 10.1111/1462-2920.16312
8. Neri U, Wolf YI, Roux S, Camargo AP, Lee B, Kazlauskas D, et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell.* (2022) 185:4023–37.e18. doi: 10.1016/j.cell.2022.08.023
9. Keusch GT, Amuasi JH, Anderson DE, Daszak P, Eckerle I, Field H, et al. Pandemic origins and a One Health approach to preparedness and prevention: Solutions based on SARS-CoV-2 and other RNA viruses. *Proc Natl Acad Sci USA.* (2022) 119:e2202871119. doi: 10.1073/pnas.2202871119
10. Li C, Shi M, Tian J, Lin X, Kang Y, Chen L, et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife.* (2015) e05378. doi: 10.7554/eLife.05378
11. Krishnamurthy SR, Janowski AB, Zhao G, Barouch D, Wang D. Hyperexpansion of RNA bacteriophage diversity. *PLoS Biol.* (2016) 14:1–17. doi: 10.1371/journal.pbio.1002409
12. Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, et al. Redefining the invertebrate RNA virosphere. *Nature.* (2016) 540:539–43. doi: 10.1038/nature20167
13. Shi M, Lin X-D, Chen X, Tian J-H, Chen L-J, Li K, et al. The evolutionary history of vertebrate RNA viruses. *Nature.* (2018) 556:197–202. doi: 10.1038/s41586-018-0012-7
14. Wolf YI, Kazlauskas D, Iranzo J, Lucia-Sanz A, Kuhn JH, Krupovic M, et al. Origins and evolution of the global RNA virome. *MBio.* (2018) 9:e02329–18. doi: 10.1128/mBio.02329-18
15. Starr EP, Nuccio EE, Pett-Ridge J, Banfield JF, Firestone MK. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc Natl Acad Sci U S A.* (2019) 116:25900–8. doi: 10.1073/pnas.1908291116
16. Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C. Expansion of known ssRNA phage genomes: From tens to over a thousand. *Sci Adv.* (2020) 6:eay5981. doi: 10.1126/sciadv.aay5981
17. Wolf YI, Silas S, Wang Y, Wu S, Bocsek M, Kazlauskas D, et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol.* (2020) 5:1262–70. doi: 10.1038/s41564-020-0755-4
18. Charon J, Buchmann JP, Sadiq S, Holmes EC. RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *Virus Evol.* (2022) 8:1–15. doi: 10.1093/ve/veac082
19. Chen YM, Sadiq S, Tian JH, Chen X, Lin XD, Shen JJ, et al. RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nat Microbiol.* (2022) 7:1312–23. doi: 10.1038/s41564-022-01180-2
20. Cheng D-Q, Kolundžija S, Lauro FM. Global phylogenetic analysis of the RNA-dependent RNA polymerase with OrVIT (OrthornaVirae Tree). *Front Virol.* (2022) 2:981177. doi: 10.3389/fviro.2022.981177
21. Edgar RC, Taylor J, Lin V, Altman T, Barbera P, Meleshko D, et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature.* (2022) 602:142–7. doi: 10.1038/s41586-021-04332-2
22. Hillary LS, Adriaenssens EM, Jones DL, McDonald JE. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME Commun.* (2022) 2:1–10. doi: 10.1038/s43705-022-00110-x
23. Sakaguchi S, Urayama SI, Takaki Y, Hirotsuna K, Wu H, Suzuki Y, et al. NeoRdRp: A comprehensive dataset for identifying RNA-dependent RNA polymerases of various RNA viruses from metatranscriptomic data. *Microbes Environ.* (2022) 37:1–7. doi: 10.1264/JSME2.ME22001
24. Zayed AA, Wainaina JM, Dominguez-Huerta G, Pelletier E, Guo J, Mohssen M, et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science.* (2022) 376:156–62. doi: 10.1126/SCIENCE.ABM5847
25. Chong LC, Lauber C. Viroid-like RNA-dependent RNA polymerase-encoding ambiviruses are abundant in complex fungi. *Front Microbiol.* (2023) 14:1144003. doi: 10.3389/fmicb.2023.1144003
26. Dudas G, Batson J. Accumulated metagenomic studies reveal recent migration, whole genome evolution, and undiscovered diversity of orthomyxoviruses. *J Virol.* (2023) 97:1–15. doi: 10.1128/jvi.01056-23
27. Harvey E, Mifsud JCO, Holmes EC, Mahar JE. Divergent hepaciviruses, delta-like viruses, and a chu-like virus in Australian marsupial carnivores (dasyurids). *Virus Evol.* (2023) 9:1–14. doi: 10.1093/ve/vead061
28. Mifsud JCO, Costa VA, Petrone ME, Marzinelli EM, Holmes EC, Harvey E. Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians. *Virus Evol.* (2023) 9:1–17. doi: 10.1093/ve/veac124
29. Olendraitte I, Brown K, Firth AE. Identification of RNA virus-derived RdRp sequences in publicly available transcriptomic data sets. *Mol Biol Evol.* (2023) 40:1–24. doi: 10.1093/molbev/msad060
30. Ni XB, Cui XM, Liu JY, Ye RZ, Wu YQ, Jiang JF, et al. Metavirome of 31 tick species provides a compendium of 1,801 RNA virus genomes. *Nat Microbiol.* (2023) 8:162–73. doi: 10.1038/s41564-022-01275-w
31. Ma Y, Marais A, Lefebvre M, Theil S, Svanella-Dumas L, Faure C, et al. Phytovirus analysis of wild plant populations: comparison of double-stranded RNA and virion-associated nucleic acid metagenomic approaches. *J Virol.* (2019) 94:e01462–19. doi: 10.1128/jvi.01462-19
32. Decker CJ, Steiner HR, Hoon-Hanks LL, Morrison JH, Haist KC, Stabell AC, et al. DsRNA-seq: Identification of viral infection by purifying and sequencing dsRNA. *Viruses.* (2019) 11:1–21. doi: 10.3390/v11100943
33. Urayama S, Takaki Y, Nunoura T. FLDS: A comprehensive dsRNA sequencing method for intracellular RNA virus surveillance. *Microbes Environ.* (2016) 31:33–40. doi: 10.1264/jsme2.me15171
34. Hirai M, Takaki Y, Kondo F, Horie M, Urayama SI, Nunoura T. RNA viral metagenome analysis of subnanogram dsRNA using fragmented and primer ligated dsRNA sequencing (FLDS). *Microbes Environ.* (2021) 36:1–6. doi: 10.1264/jsme2.ME20152
35. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, et al. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology.* (2009) 388:1–7. doi: 10.1016/j.virol.2009.03.024
36. Vainio EJ, Jurvansuu J, Streng J, Rajamäki ML, Hantula J, Valkonen JPT. Diagnosis and discovery of fungal viruses using deep sequencing of small RNAs. *J Gen Virol.* (2015) 96:714–25. doi: 10.1099/jgv.0.000003
37. Nerva L, Ciuffo M, Vallino M, Margaria P, Varese GC, Gnani G, et al. Multiple approaches for the detection and characterization of viral and plasmid symbionts from a collection of marine fungi. *Virus Res.* (2016) 219:22–38. doi: 10.1016/j.virusres.2015.10.028
38. Zheng Y, Gao S, Padmanabhan C, Li R, Galvez M, Gutierrez D, et al. VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology.* (2017) 500:130–8. doi: 10.1016/j.virol.2016.10.017
39. Adalia EJM, Diez JJ, Fernández MM, Vainio JHE. Characterization of small RNAs originating from mitoviruses infecting the conifer pathogen *Fusarium circinatum*. *Arch Virol.* (2018) 163:1009–18. doi: 10.1007/s00705-018-3712-2
40. Raco M, Vainio EJ, Sutela S, Eichmeier A, Hakalová E, Jung T, et al. High diversity of novel viruses in the tree pathogen *Phytophthora castaneae* revealed by high-throughput sequencing of total and small RNA. *Front Microbiol.* (2022) 13:911474. doi: 10.3389/fmicb.2022.911474
41. Coclet C, Sorensen PO, Karaoz U, Wang S, Brodie EL, Eloe-Fadrosch EA, et al. Virus diversity and activity is driven by snowmelt and host dynamics in a high-altitude watershed soil ecosystem. *Microbiome.* (2023) 11:1–19. doi: 10.1186/s40168-023-01666-z
42. Hjelmsø MH, Hellmér M, Fernandez-Cassi X, Timoneda N, Lukjancenko O, Seidel M, et al. Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing. *PLoS One.* (2017) 12:1–17. doi: 10.1371/journal.pone.0170199
43. Te Velthuis AJW. Common and unique features of viral RNA-dependent polymerases. *Cell Mol Life Sci.* (2014) 71:4403–20. doi: 10.1007/s00018-014-1695-z
44. Poch O, Sauvaget I, Delarue M, Tordo N. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J.* (1989) 8:3867. doi: 10.1002/emboj.1989.8.issue-12
45. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinf.* (2009) 10:1–9. doi: 10.1186/1471-2105-10-421
46. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* (2015) 12:59–60. doi: 10.1038/nmeth.3176
47. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol.* (2017) 35:1026–8. doi: 10.1038/nbt.3988
48. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* (1997) 25:3389–402. doi: 10.1093/genetics/156.4.1997
49. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol.* (2011) 7:e1002195. doi: 10.1371/journal.pcbi.1002195
50. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinf.* (2019) 20:1–15. doi: 10.1186/s12859-019-3019-7
51. Babaian A, Edgar R. Ribovirus classification by a polymerase barcode sequence. *PeerJ.* (2022) 10:e14055. doi: 10.7717/PEERJ.14055
52. Sutela S, Forgia M, Vainio EJ, Chiappello M, Daghighi S, Vallino M, et al. The virome from a collection of endomycorrhizal fungi reveals new viral taxa with unprecedented genome organization. *Virus Evol.* (2020) 6:veaa076. doi: 10.1093/VE/VEAA076
53. Chiba Y, Oiki S, Yaguchi T, Urayama SI, Hagiwara D. Discovery of divided RdRp sequences and a hitherto unknown genomic complexity in fungal viruses. *Virus Evol.* (2021) 7:1–11. doi: 10.1093/ve/veaa101
54. Chiba Y, Oiki S, Zhao Y, Nagano Y, Urayama SI, Hagiwara D. Splitting of RNA-dependent RNA polymerase is common in Narnaviridae: Identification of a type II divided RdRp from deep-sea fungal isolates. *Virus Evol.* (2021) 7:1–9. doi: 10.1093/ve/veab095
55. Szczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat Methods.* (2017) 14:1063–71. doi: 10.1038/nmeth.4458

56. Meyer F, Hofmann P, Belmann P, Garrido-Oter R, Fritz A, Szczyrba A, et al. AMBER: assessment of metagenome BinnERS. *Gigascience*. (2018) 7:1–8. doi: 10.1093/gigascience/giy069
57. Pratama AA, Bolduc B, Zayed AA, Zhong Z, Guo J, Vik DR, et al. Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. *PeerJ*. (2021) 9:e11447. doi: 10.7717/peerj.11447
58. Ho SFS, Wheeler NE, Millard AD, van Schaik W. Gauge your phage: benchmarking of bacteriophage identification tools in metagenomic sequencing data. *Microbiome*. (2023) 11:1–15. doi: 10.1186/s40168-023-01533-x
59. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. (2021) 596:583–9. doi: 10.1038/s41586-021-03819-2
60. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. (2023) 379:1123–30. doi: 10.1126/science.ade2574
61. Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, Janssen GMC, Ruben M, Overkleef HS, et al. Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. *Nucleic Acids Res*. (2015) 43:8416–34. doi: 10.1093/nar/gkv838
62. Slanina H, Madhugiri R, Bylapudi G, Schultheiß K, Karl N, Gulyaeva A, et al. Coronavirus replication-transcription complex: Vital and selective NMPylation of a conserved site in nsp9 by the NiRAN-RdRp subunit. *Proc Natl Acad Sci U S A*. (2021) 118:2–13. doi: 10.1073/pnas.2022310118
63. Malone B, Urakova N, Snijder EJ, Campbell EA. Structures and functions of coronavirus replication-transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat Rev Mol Cell Biol*. (2022) 23:21–39. doi: 10.1038/s41580-021-00432-z
64. Candresse T, Svanella-Dumas L, Marais A, Depasse F, Faure C, Lefebvre M. Identification of seven additional genome segments of grapevine-associated Jivivirus 1. *Viruses*. (2023) 15:39. doi: 10.3390/v15010039
65. Adriaenssens EM, Roux S, Brister JR, Karsch-Mizrachi I, Kuhn JH, Varsani A, et al. Guidelines for public database submission of uncultivated virus genome sequences for taxonomic classification. *Nat Biotechnol*. (2023) 41:898–902. doi: 10.1038/s41587-023-01844-2
66. Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. *Nucleic Acids Res*. (2021) 49:D121–4. doi: 10.1093/nar/gkaa967
67. Eloe-Fadrosh EA, Ahmed F, Anubhav, Babinski M, Baumes J, Borkum M, et al. The National Microbiome Data Collaborative Data Portal: An integrated multi-omics microbiome data resource. *Nucleic Acids Res*. (2022) 50:D828–36. doi: 10.1093/nar/gkab990
68. Parks DH, ChuvoChina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res*. (2022) 50:D785–94. doi: 10.1093/nar/gkab776

Detection of exotic biosecurity threat ribgrass mosaic virus and novel tobamoviruses through metatranscriptomic sequencing of animal gut content

Jackie Mahar^{a*}, Jonathon C. O. Mifsud^a, Kate Van Brussel^a, Anna E. Lachenauer^a, Erin Harvey^a, Olivia M. H. Turnbull^{a*}, Stefanie Bonat^b, Thomas M. Newsome^b, Annabelle Olsson^c, Antje Chiu-Werner^d, Menna E. Jones^d, Edward C. Holmes^a, Solomon Maina^e

^aSchool of Medical Sciences, University of Sydney, Sydney, NSW 2006, Australia.

^bSchool of Life and Environmental Sciences, University of Sydney, Sydney, NSW 2006, Australia.

^cWildlife Health and Conservation Hospital, Sydney School of Veterinary Science, University of Sydney, Camden, NSW 2570, Australia.

^dSchool of Natural Sciences, University of Tasmania, Hobart, TAS 7000, Australia.

^ePlant Biosecurity Research and Diagnostics, NSW Department of Primary Industries and Regional Development, Menangle, NSW 2563, Australia.

#Address correspondence to :

Prof. Edward C. Holmes,

School of Medical Sciences, University of Sydney, Sydney, NSW 2006, Australia.

Tel: +61 2 9351 5591

Email: edward.holmes@sydney.edu.au

*Current address: Jackie Mahar: CSIRO Australian Centre for Disease Control, Geelong, VIC 3220, Australia; Olivia Turnbull: Centre for Molecular Oncology, Faculty of Medicine, University of New South Wales, Kensington, NSW 2052, Australia

Running title: Tobamovirus detection in animal gut metatranscriptomes

Word count: 6,395; Abstract word count: 207

Abstract

Ribgrass mosaic virus (RMV) and related viruses of the genus *Tobamovirus* (*Virgaviridae*) are cruciferous plant pathogens that represent a threat to global horticultural systems. In Australia, they are considered exotic biosecurity threats, and an incursion of these viruses would require rapid and strict control efforts. However, current surveillance methods for these viruses are limited. We examined whether RMV and related tobamoviruses could be detected by deep sequencing of gut metatranscriptomes of vertebrate animals and ticks. Using this method, we discovered that RMV, as well as a novel relative of RMV, and two highly diverse novel tobamoviruses are present in Australia. RMV was detected in multiple sites in both the Australian Capital Territory (ACT) and Tasmania, two regions separated by approximately 700km of land and 200km of water. The novel relative of RMV was detected in the ACT and New South Wales (NSW), while the highly divergent novel tobamoviruses were each detected in a single state, NSW and Queensland (QLD). In addition, Tobacco mild green mosaic virus, which is already known to be present in Australia, was detected in QLD using this method. This work highlights the potential utility of metatranscriptomic sequencing of wild animal gut for the surveillance of biosecurity threats to native and agricultural plant species.

Importance

Plant viruses can have devastating impacts on global horticulture. Tobamoviruses (family *Virgaviridae*, genus *Tobamovirus*) are among the most damaging seed-borne viruses in horticultural crops, and Australia is free of many of the tobamoviruses that cause major crop losses in other countries. These viruses are extremely difficult to eradicate. Consequently, early detection of incursions is key to the control of these viruses in Australia, alongside rapid deployment of eradication and management plans. Current biosecurity surveillance methods in Australia rely on visual inspection, immunological assays, and molecular methods such as screening of imported seed lots. This study introduces a complementary approach that utilises unbiased metatranscriptomic sequencing of animal gut material to detect cryptic plant viruses circulating in nature. Using this approach, we detected five different tobamovirus circulating in Australia, including a virus thought to be exotic and three novel viruses. This unique approach highlights alternative options for surveillance/detection of exotic crop viruses.

Keywords: Metatranscriptomics, *Tobamovirus*, surveillance, plant viruses, biosecurity, incursion.

Introduction

Worldwide, virus disease outbreaks, epidemics and pandemics pose a threat to agricultural and horticultural systems (1). Viral diseases have been associated with impaired crop growth and vigour, leading to diminished or total loss of gross yields (2). This has been accelerated by agricultural globalization, climate change and other factors, such as pesticide resistance in virus vectors, that together are complicating disease management (3). Tobamoviruses (family *Virgaviridae*, genus *Tobamovirus*) are among the most damaging seed-borne viruses in horticultural crops (4). These viruses have a single stranded, positive-sense genome of about 6.3-6.5 kb that contain six open reading frames (ORFs): an RNA replicase read-through derivative (ORF1 and 2), an undefined protein (ORF3), movement protein (MP; ORF4), coat protein (CP; ORF5) (5), and occasionally an ORF6, that overlaps ORF4 and ORF5 (6). Tobamoviruses have an extensive host range and on this basis are divided into three subgroups: subgroup 1 infect solanaceous species, subgroup 2 infect Cactaceae, Cucurbitaceae, Fabaceae, Malvaceae and Passifloraceae species, and subgroup 3 infect a range of hosts, including Cruciferae/Brassicaceae and *Plantago* (Plantaginaceae) species (7, 8). Tobamoviruses have been associated with substantial economic losses linked to reduced crop yield as well as fruit damage which triggers a significant reduction in marketability (9, 10). For example, tomato brown rugose fruit virus (ToBRFV), is a damaging new tobamovirus that infects tomatoes and peppers that was first discovered in Jordan (11) and Israel (12), later spreading to many countries worldwide (13). Similarly, cucumber green mottle mosaic virus (CGMMV) is a recently introduced tobamovirus in Australia (14) that causes serious disease and damage in the *Cucurbitaceae*, such that it is one of the most economically important cucurbit viruses (15).

Agriculture of non-native crops in Australia commenced after colonization by Europeans in 1788 and in some regions, as recently as 60 years ago (16). Such a recent timescale, as well as the isolation of this large island nation, means that many viruses and virulent virus strains that co-evolved with these plants and cause damaging crop diseases in other parts of the world are absent or have only recently arrived in Australia. Human activities within the global trade in plants and plant products, mostly involving unknowingly infected seed or vegetative propagules, may spread viruses from other countries to Australia (3, 17, 18). Evidence that this may have already occurred in the Australian horticultural industry include recent detections of CGMMV and pepper vein yellows virus (14, 19). Alternatively, viruses can spill over from indigenous wild plant populations to managed cultivated populations (20–22). As agriculture and horticulture continues to expand in Australia, so too does the

103 overlap of wild plant and crop habitat, increasing chances of virus spillover (23). As such,
 104 understanding the range of viruses that exist on wild vegetation and their interactions with
 105 local animal fauna is an important complementary route for managing crop viral diseases.
 106 Indeed, plant viruses may colonize new environments through grazing animal teeth,
 107 ingestion of plants, and subsequent passing of the viruses through the gut (24). However,
 108 very few studies have comprehensively explored persistent viruses in wild plants (23, 25).
 109
 110 Recently, we inadvertently detected ribgrass mosaic virus (RMV) in Australia in wild animal
 111 gut content while exploring animal viromes using metatranscriptomic (i.e. total RNA)
 112 sequencing (26). RMV is considered a biosecurity risk and was previously considered to be
 113 exotic to Australia, but is endemic in neighbouring New Zealand (27). Its natural hosts
 114 include members of the Cruciferae and Plantaginaceae, and it is known to have a broad
 115 host range compared to other tobamoviruses (28). RMV belongs to *Tobamovirus* subgroup
 116 3, and is most closely related to youcai mosaic virus (YoMV), turnip vein clearing virus
 117 (TVCV) and wasabi mottle virus (WMOV) (29), all of which are exotic plant pathogens and
 118 Australian biosecurity risks ([https://www.agriculture.gov.au/biosecurity-trade/pests-](https://www.agriculture.gov.au/biosecurity-trade/pests-diseases-weeds/plant/national-priority-plant-pests-2019)
 119 [diseases-weeds/plant/national-priority-plant-pests-2019](https://www.agriculture.gov.au/biosecurity-trade/pests-diseases-weeds/plant/national-priority-plant-pests-2019), accessed 21/10/24).
 120
 121 Tobamoviruses are highly contagious, easily spread through mechanical transmission
 122 between plants and fomites (30), and in some cases are seed transmitted. Additionally, they
 123 can remain viable in soil for many months (even up to a year) and are very difficult to
 124 eradicate once present (31). As there are no available treatments, plant destruction and soil
 125 decontamination are the only means of control (32). Consequently, early detection and
 126 prevention of incursion is key to the control of these viruses in Australia, alongside rapid
 127 deployment of eradication and management plans in the event of incursions. Current
 128 biosecurity surveillance methods in Australia rely on visual inspection, immunological
 129 assays, and molecular methods such as screening of imported seed lots (32). Efforts to
 130 include targeted plant viral genomics are in development (33).
 131
 132 Herein, we explored the use of an alternative surveillance strategy, based on the
 133 metatranscriptomic sequencing of animal gut content, to unravel unknown circulating
 134 biosecurity threats within wild or cultivated vegetation. This approach was used (i) to
 135 determine the presence of biosecurity threat tobamoviruses in Australia, at the same time
 136 assessing the utility of this method, and (ii) to reveal the phylogenetic relationships among

the tobamoviruses detected and their implications for the biosecurity of the Australian horticultural industry.

Materials and Methods

Metatranscriptomic sequence data

Metatranscriptomic data sets previously generated from studies of Australian animal viromes were opportunistically screened for the presence of tobamoviruses. These data sets were generated by sequencing total RNA extracted from whole ticks, and mammalian gut content in the form of faecal samples, caecal content samples, or anal/rectal swabs (Table 1). Total RNA was extracted using the Maxwell 16 LEV simply RNA tissue kit in combination with a Maxwell nucleic acid extraction robot (Promega, WI, USA) for rabbit caecal content; the Qiagen RNeasy Plus Universal Mini Kit (Hilden, Germany) for deer and kangaroo carcass swabs; and the Qiagen RNeasy Plus Mini Kit (sometimes in conjunction with Qiagen QiaShredder columns) for dog swabs, ticks and marsupial carnivore faecal samples. The rabbit RNA libraries were prepared using the TruSeq Total RNA library preparation protocol (Illumina, CA, USA), with rRNA removal using Illumina Ribo-Zero gold epidemiology rRNA removal kit. The Illumina Stranded Total RNA Prep with Ribo-Zero Plus kit was used for all the other libraries. Paired-end sequencing was conducted on an Illumina HiSeq 2500 (200 cycles) for the rabbit libraries, and Illumina NovaSeq 6000 S4 lane (300 cycles) was used for all the other libraries (Illumina, CA, USA). All library preparation and sequencing was conducted at the Australian Genomic Research Facility (AGRF). The raw reads obtained were trimmed using Trimmomatic v0.38 (34) and *de novo* assembled using MEGAHIT v1.2.9 (35) or Trinity v2.5.1 (36). Further details on data generation and ethics for data sets collected from rabbits can be found in Mahar et al. 2020 (26). Manuscripts are in preparation for the remaining studies.

Ethics

For deer and kangaroo carcasses, scientific licenses and collection permits were obtained to relocate carcasses (SL102334) and research was approved by the University of Sydney Animal Ethics Committee, project number: 2019/1640. Dog and tick sampling was approved by the Taronga Conservation Society Australia's Animal Ethics Committee, approval number 4c/10/21. Ethics for Tasmanian carnivores was obtained from the University of Tasmania under Animal Ethic Committee permit number A0018012. Ethics for rabbit sampling is as described in Mahar et al., 2020 (26).

172 *Tobamovirus detection*

173 The assembled contigs were annotated using DIAMOND BLASTx (37), and blastn (38)
174 against NCBI's non-redundant protein and nucleotide databases, respectively. Contigs with
175 top hits to tobamovirus were selected for further analysis. Open reading frames were
176 identified using the Find ORFs tool within Geneious Prime ® 2022.2.2 using the standard
177 genetic code and UAG read-through. ORFs were adjusted and annotated based
178 annotations copied from related tobamoviruses using the Annotate from Database tool
179 available within Geneious Prime. ORF gene assignments were confirmed by identification of
180 conserved domains using RSP-TBLASTN v2.6.0 (39) and the NCBI CDD database.

Table 1. Metadata of metatranscriptome libraries containing tobamoviruses.

Collection location	Library name	Animal species (common name)	Sample type	No. samples in pool	D.O.C	Tobamovirus(es) detected
Crackenback/ Moonbah, NSW	12-Kang2B	<i>Macropus giganteus</i> ^b (Eastern grey kangaroo)	Mouth and anal swab	4	Feb 2021	SG3-like tobamovirus
Crackenback/ Moonbah, NSW	13-Deer2A	<i>Dama dama</i> ^b (Fallow deer)	Mouth and anal swab	4	Dec 2020 and Jan 2021	SG3-like tobamovirus
Crackenback/ Moonbah, NSW	15-Deer2B	<i>Dama dama</i> ^b (Fallow deer)	Mouth and anal swab	4	Jan 2021	Bambi tobamovirus
Crackenback/ Moonbah, NSW	16-Deer2B	<i>Dama dama</i> ^b (Fallow deer)	Mouth and anal swab	5	Feb and Mar 2021	SG3-like tobamovirus, Bambi tobamovirus
Hope Vale, QLD	HPV3-R	<i>Canis lupus familiaris</i> (Domestic dog)	Rectal swab	2	Dec 2021	Bluey tobamovirus
Hope Vale, QLD	HPV4-R	<i>Canis lupus familiaris</i> (Domestic dog)	Rectal swab	3	Dec 2021	Bluey tobamovirus
Hope Vale, QLD	HPV6-R	<i>Canis lupus familiaris</i> (Domestic dog)	Rectal swab	1	Dec 2021	Bluey tobamovirus, TMGMV
Hope Vale, QLD	HPV15-2R	<i>Canis lupus familiaris</i> (Domestic dog)	Rectal swab	2	Feb 2022	Bluey tobamovirus
Hope Vale, QLD	HPV16-2R	<i>Canis lupus familiaris</i> (Domestic dog)	Rectal swab	3	Feb 2022	Bluey tobamovirus
Hope Vale, QLD	HPV35-2E	<i>Rhipicephalus sanguineus</i> (tick), attached to domestic dog	Whole animal	1	Feb 2022	Bluey tobamovirus
Ben Lomond, TAS	P04	<i>Dasyurus maculatus</i> (Spotted-tailed quoll)	Faecal	2	Apr-Aug 2021	RMV
Northern Slopes, TAS	P19	<i>Felis catus</i> (feral cat)	Faecal	1	Apr-Aug 2021	RMV

Northern Slopes, TAS	P21	<i>Dasyurus maculatus</i> (Spotted-tailed quoll)	Faecal	1	Apr-Aug 2021	RMV
Northern Slopes, TAS	P22	<i>Sarcophilus harrisii</i> (Tasmanian devil)	Faecal	5	Apr-Aug 2021	RMV
Crace, ACT	GUJ-CC ^a	<i>Oryctolagus cuniculus</i> ^b (European rabbit)	Caecal content	20	Dec 2016-Jan 2017	RMV, SG3-like tobamovirus
Booth, ACT	Gudg-CC ^a	<i>Oryctolagus cuniculus</i> ^b (European rabbit)	Caecal content	18	Feb 2017	RMV

a - published in Mahar et al., 2020

b – samples taken from carcasses, not live animals.

144 *Phylogenetic classification of tobamoviruses*

145 Contigs that had a top blast hit to a tobamovirus were confirmed through sequence
146 alignment and phylogenetic analysis with other published tobamovirus sequences.
147 Accordingly, the tobamovirus NCBI nucleotide RefSeqs, the top blast hits to the
148 tobamoviruses detected here, as well as all published nucleotide sequences of RMV, TVCV,
149 WMoV, YoMV, and TMGMV were downloaded from the NCBI nucleotide database
150 (<https://www.ncbi.nlm.nih.gov/>) and aligned with the tobamovirus nucleotide sequences
151 detected in this study using MAFFT v7.490 (40). Ambiguously aligned regions were
152 removed using trimAl v1.4.1 (41). Nucleotide substitution model selection was performed
153 (42) and maximum likelihood phylogenies inferred using IQ-TREE v2.1.3 (43). Branch
154 supports were estimated using bootstrapping (1,000 replicates). Phylogenetic trees were
155 inferred for both the complete genome (n=131, 8,341 nt), as well as the coat gene, the most
156 commonly sequenced gene among tobamoviruses (n=167, 528 nt). A distance matrix for
157 subgroup 3 tobamoviruses was inferred by aligning all available subgroup 3 nucleotide
158 sequences from NCBI with the new sequences from this study, using MAFFT. Identical
159 sequences were removed and genetic distances were calculated (as percentage nucleotide
160 identity) and visualized in Geneious Prime ® 2022.2.2.

161

162 *Recombination detection*

163 All new tobamovirus species were screened for the presence of recombination using the
164 recombination detection program v4.96 (RDP4) (44), using the default parameters. A full
165 genome alignment including the new tobamoviruses detected here, NCBI RefSeqs for
166 subgroup 1 and 3 viruses, and the top blast hits for each tobamovirus contig was used as
167 input (n=75 sequences). The highest acceptable p-value was set to 0.05 with Bonferroni
168 correction, and recombination events were only considered if detected by at least three
169 different methods.

170

171 *Meta-transcriptomic identification of potential plant host species*

172 CCMetagen (45) was used to classify the non-viral transcripts in the libraries where
173 tobamoviruses were detected, using the default method of abundance calculation.
174 Classification was based on screening against the curated indexed database compiled by
175 the CCMetagen creators, which contains the NCBI nucleotide collection, with the exception
176 of most artificial and environmental sequences that lack taxids.

177

178 *PCR confirmation*

179 The presence of the novel or exotic tobamoviruses detected within each
180 metatranscriptomic library were confirmed by amplifying a small fragment of the relevant
181 virus (RMV, 565 bp; SG3-like tobamovirus, 571 bp; bluey tobamovirus, 226 bp; bambi
182 tobamovirus, 540 bp). For RMV, the SuperScript™ One-Step RT-PCR System with
183 Platinum™ Taq DNA polymerase (Invitrogen, MA, USA) was used according to the
184 manufacturer's instructions. RMV isolate PV-051 infected lyophilized plant material
185 imported from DSMZ Germany was used as a positive control. Optimum cycling conditions
186 were as follows: 50°C for 30 min for reverse transcription, 95°C for 15 min followed by 35
187 cycles at 95°C for 30 s, 55°C for 40 s, and 72°C for 45 s, with a final extension at 72°C for
188 10 min. For all other viruses, the RT-PCRs were conducted in a different lab using the
189 SuperScript™ IV One-Step RT-PCR system (Invitrogen) according to the manufacturer's
190 protocol. The cycling conditions were as follows: 50°C for 10 mins and 98°C for 2 mins,
191 followed by 35 cycles of 98°C for 10 s, 55°C for 10 s and 72°C for 30 s, with a final
192 extension of 72°C for 5 mins. The amplified specific PCR products generated were
193 confirmed by gel electrophoresis followed by SYBR safe staining. At least one set of PCR
194 products for each virus were purified and confirmed by Sanger sequencing at the Australian
195 Genome Research Facility (AGRF) at Westmead, NSW.

196

197 *Amplification and Sanger sequencing confirmation of SG3-like tobamovirus virus genome*
198 To confirm the consensus genome sequence of SG3-like tobamovirus virus, RT-PCR and
199 Sanger sequencing was performed on sample 12.9 from library 16-Deer2B. The
200 tobamovirus sequence was amplified in four segments (~1040 bp to ~2100 bp) using the
201 SuperScript™ IV One-Step RT-PCR system (Invitrogen) following the manufacturer's
202 protocol and using four primer sets (Table S1). The cycling conditions were as follows: 50°C
203 for 10 mins and 98°C for 2 mins, 35 cycles of 98°C for 10 s, 55°C for 10 s and 72°C for 1
204 min, and a final extension of 72°C for 5 mins. DNA amplicons were purified using the
205 GenElute PCR clean-up kit (Sigma-aldrich, MO, USA) and sent for single direction sanger
206 sequencing at AGRF at Westmead, NSW, Australia using the amplification primers plus
207 additional sequencing primers (Table S1). The resulting chromatograms were trimmed and
208 mapped to the original metatranscriptomic-assembled genome sequence using Geneious
209 Prime version 2022.1.1, and a consensus generated.

210

211 **Data Availability**

212 The consensus sequences for tobamoviruses assembled in this study have been submitted
213 to NCBI/GenBank and assigned accession numbers XXXXX-XXXXX.

214

215 Results

216 Detection and classification of tobamoviruses in animal gut metatranscriptomes

217 Viral contigs related to tobamovirus species were detected by Blast analyses of the gut
218 metatranscriptomes of Australian wild animals collected from four states/territories in
219 eastern Australia; NSW, TAS, ACT and QLD. Phylogenetic classification based on complete
220 genome sequences (Figure 1A) revealed that five different tobamovirus species were
221 present in the animal metatranscriptomes sampled here: Tobacco mild green mosaic virus
222 (TMGMV), two subgroup 3 viruses, and two distinct novel tobamoviruses. Tobamovirus
223 contigs varied in length, but the complete coding region of each of the five tobamoviruses
224 detected here was assembled from at least one library.

225

226 Classification of the subgroup 3 viruses was hindered by the fact that many published virus
227 sequences (particularly CP sequences) in the subgroup 3 clade have been misclassified.
228 Despite this, it was possible to identify defined clades that represent each viral species
229 *Tobamovirus plantagonis* (RMV), *Tobamovirus rapae* (TVCV), *Tobamovirus wasabi* (WMOV),
230 and *Tobamovirus youcai* (YoMV) (Figure 1), each with 100% bootstrap support in the
231 genome sequences tree. A distance matrix based on complete genome nucleotide
232 sequences also supported that the species should be defined by these clades, in
233 accordance with the ICTV species demarcation criteria for these viruses
234 (<https://ictv.global/report/chapter/virgaviridae/virgaviridae/tobamovirus>), with sequences in
235 the same clade sharing >90% similarity and those in separate clades sharing less than 90%
236 nucleotide similarity (with a single outlier in YoMV) (Table 2).

237

238 **Table 2. Nucleotide identity within and between tobamovirus subgroup 3 species.**

	YoMV	WMOV	TVCV	RMV	SG3-like tobamovirus
Maximum % identity (identical sequences have been removed)					
YoMV	99.9 ^a	85.0	82.4	82.5	81.7
WMOV	85.0	99.7	82.7	83.3	82.6
TVCV	82.4	82.7	98.4	89.0	88.0
RMV	82.5	83.3	89.0	99.9	88.2
SG3-like tobamovirus	81.7	82.6	88.0	88.2	98.8
Minimum % identity					
YoMV	89.1 ^b	83.1	80.6	79.5	80.5
WMOV	83.1	98.4	81.9	80.3	82.2
TVCV	80.6	81.9	92.6	85.4	87.2
RMV	79.5	80.3	85.4	90.3	86.0

SG3-like tobamovirus	80.5	82.2	87.2	86.0	98.8
Mean % identity					
YoMV	94.2	84.5	81.6	81.0	81.3
WMoV	84.5	99.1	82.3	81.8	82.5
TVCV	81.6	82.3	94.0	87.0	87.6
RMV	81.0	81.8	87.0	93.6	86.7
SG3-like tobamovirus	81.3	82.5	87.6	86.7	98.8

239 a - Cells with >90% nucleotide identity are shaded grey
240 b - Minimum identity within YoMV sequences drops below 90% due to one sequence only

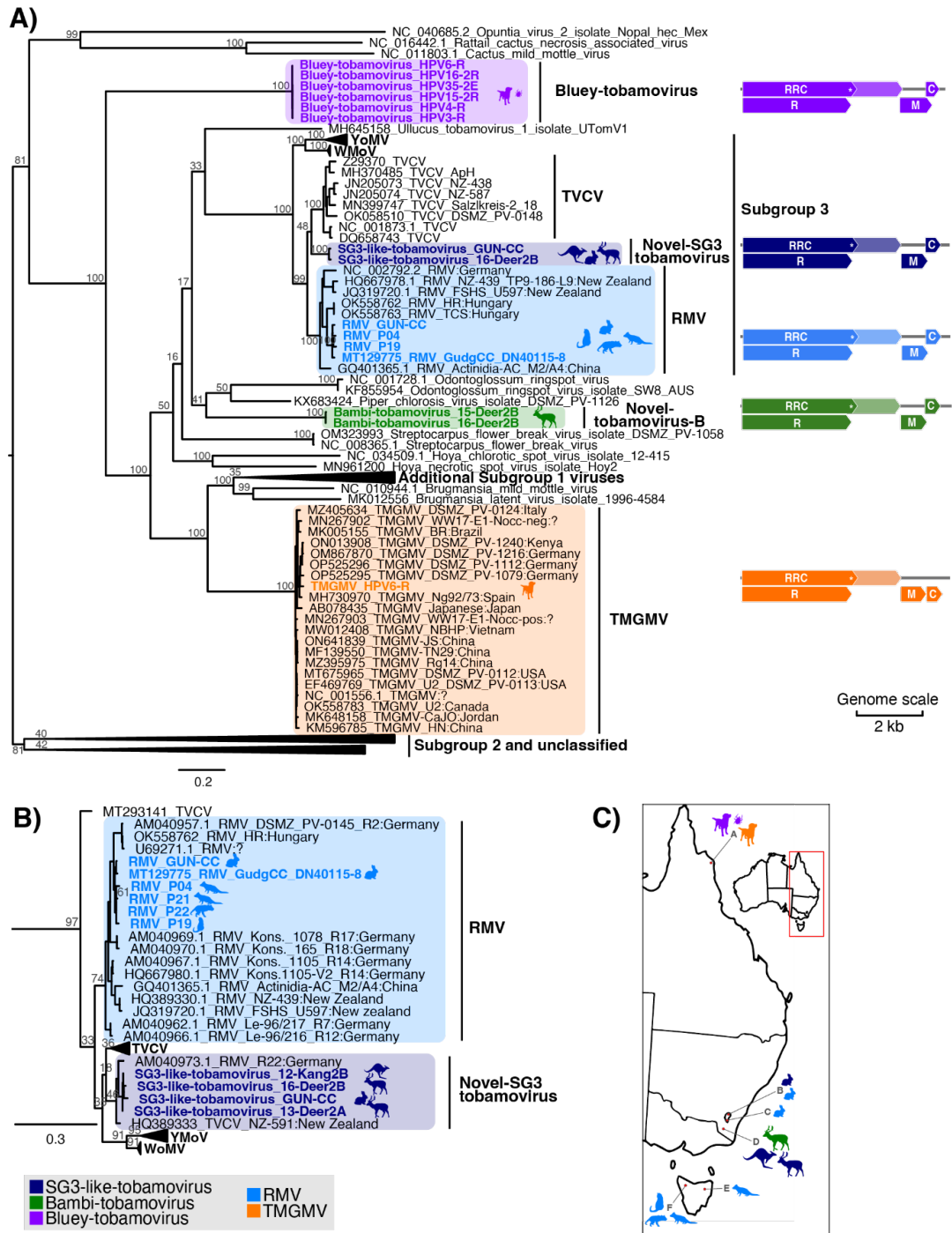


Figure 1. Diversity and geographic distribution of tobamoviruses found in animal meta-transcriptomes in Australia. (A) ML phylogenetic tree of the genome sequences of the tobamoviruses detected in Australian animal metatranscriptomes together with published tobamovirus genomes. **(B)** ML phylogenetic tree of the coat gene sequences of tobamoviruses detected in Australian animal metatranscriptomes together with published

coat sequences, showing only the subgroup 3 clade (CP tree is shown in Supplementary Figure 1). In the ML trees, sequences obtained from animal metatranscriptomes in Australia are indicated in bold and coloured by viral species, and the clades that they cluster within are highlighted in the same colour: purple=bluey tobamovirus; dark blue=SG3-like tobamovirus; blue=RMV; green=bambi tobamovirus; orange=TMGMV. Animal silhouettes beside the clades indicate the animal metatranscriptomes from which these viruses were obtained and are coloured by viral species. The GenBank accession number for published sequences is indicated at the start of the taxon name. Numbers at the nodes indicate the percentage support from 1,000 bootstrap replicates and the trees are midpoint rooted. The location (country) of collection is indicated in the taxa name in relevant clades. Genome schematics for viruses detected in this study are shown to the right of the relevant viral species, with coloured arrow-tipped boxes representing forward open reading frames for the following gene products: RNA replicase read-through component (RRC), RNA replicase (R), movement protein (M) and capsid protein (C). The read-through stop codon in the RRC is indicated by an asterisk and a break in the box, while the RNA-dependent RNA polymerase region translated via the read-through mechanism is indicated by lighter shading. **(C)** Map of eastern Australia indicating where tobamoviruses were detected. Animal silhouettes indicate the animal metatranscriptomes where these viruses were obtained and are coloured by viral species. Locations on the map are indicated by a letter and are as follows: A=Hope Vale, QLD; B=Crace, ACT; C=Booth, ACT; D=Crackenback/Moonbah, NSW; E=Ben Lomond, TAS; F=Northern Slopes, TAS.

In clearly defining the subgroup 3 species clades, it was possible to determine that the two distinct SG3 viruses detected here belonged to RMV and a distinct novel SG3 group, tentatively named SG3-like tobamovirus (Figure 1A and B). The SG3-like tobamovirus clade contained sequences previously published as RMV or TVCV, including the complete genome of a virus detected by Mahar et al., 2020 (26) in rabbits (accession MT129780) and CP sequences from New Zealand (HQ389333) and Germany (AM040973) (Figure 1B). While phylogenetic analyses grouped these sequences with RMV and TVCV with strong bootstrap support, these viruses are distinct from both RMV and TVCV (Figure 1). Support for the SG3-like tobamovirus clade of viruses is not robust in the CP tree (bootstrap support of 46%; Figure 1B) due to limited informative sites, but the complete genome phylogeny strongly supports the existence of this small clade (Figure 1A). Furthermore, according to the *Tobamovirus* genus demarcation criteria (<90% nucleotide similarity between species) the SG3-like tobamovirus clade is diverse enough to constitute a new species, with all

sequences in the SG3-like tobamovirus clade sharing no more than 88.2% identity with RMV and other viruses (Table 2), although this will need to be confirmed with additional data. It should be noted that because this virus was novel but closely related to RMV and TVCV, the metatranscriptome assembled genome sequence was confirmed by amplification and Sanger sequencing. The entire genome sequence was confirmed with the exception of a 395 nt region within ORF1 which could not be completed by Sanger sequencing due to sample insufficiency.

The other two novel viruses detected in this study, tentatively named Bluey Tobamovirus and Bambi Tobamovirus, were highly divergent from each other and all previously classified Tobamoviruses (Figure 1A). Bambi tobamovirus was most genetically similar to Piper chlorosis virus (PChV) isolate DSMZ PV-1126 (accession KX683424) based on nucleotide sequence identity but shared only 60.3% nucleotide similarity across the genome and these sequences did not cluster together (Figure 1). Similarly, bluey tobamovirus had the highest nucleotide identity to TVCV (accession OK058510), but shared only 57.5% nucleotide identity with this virus and did not form a monophyletic group with TVCV (Figure 1). These two highly novel viruses did not group into a known subgroup but were most closely related to subgroup 3 viruses. Finally, no viruses discovered here were determined to be recombinants.

Distribution and diversity of detected tobamoviruses

TMGMV was detected in domestic dog rectal swabs from one location in QLD, while RMV was found in rabbit gut in two locations in the ACT, and in marsupial carnivore faeces in two locations in Tasmania (Figure 1C). Genetic diversity within the detected RMV viruses ranged from 97.4 – 99.3% nucleotide identity. SG3-like tobamovirus was detected in both the ACT in rabbit gut and in deer and kangaroo carcass swabs in NSW, sharing 98.1 – 99.9% nucleotide identity between them. Bluey tobamovirus was detected in a single location in QLD in five libraries sequenced from domestic dog rectal swabs as well as a library from a tick that was attached to a domestic dog. The bluey tobamovirus viruses discovered here shared 98.1 – 100% nucleotide identity, including the tick virus that exhibited 100% identity with a virus from a dog rectal swab. Bambi tobamovirus was found in two libraries from fallow deer anal and mouth swabs, both libraries sampled from the same deer 2-4 weeks apart in a single NSW location and these sequences exhibited 99.8% nucleotide identity.

317 *Potential source and spread of tobamoviruses in Australia*

318 As there are limited published sequences for RMV, TMGMV, and related viruses, it is
 319 challenging to determine when these viruses arrived in Australia and where they came from.
 320 The closest relatives to the Australian RMVs were sampled from Germany and Hungary
 321 (according to the CP phylogeny which has the most available sequence data, Figure 1B),
 322 while the closest relatives of the SG3-like tobamovirus viruses were from Hungary,
 323 Germany and New Zealand. However, these countries are not necessarily the source of the
 324 introduction but may have simply contributed a large proportion of the published
 325 tobamovirus sequences (particularly in the case of multiple RMV sequence from Germany).
 326 The Australian TMGMV clustered with viruses from Italy and Japan, based on the CP gene
 327 tree (Supplemental Figure 1), but based on the full genome phylogeny was most closely
 328 related to viruses from Germany. Accession OP525296, from Germany shared the highest
 329 identity with the virus detected here at 96.7% nucleotide identity across the complete
 330 genome, although there are a range of viruses from various countries that share more than
 331 96% identity, including one from China (MZ395975), and Kenya (ON013908), that share
 332 96.6% and 96.5% identity, respectively. Indeed, the most divergent TMGMV (MH730970)
 333 still shares 94.5% identity with the new Australian isolate. Considering the high levels of
 334 identity between isolates from different countries, it is not possible to get an indication of
 335 the source country. However, for each viral species, the Australian sequences clustered
 336 together (Figure 1A), suggesting that there was a single introduction of each of the five
 337 tobamoviruses into Australia, and subsequent spread to other states, although this is again
 338 subject to sampling biases.

339

340 *Abundance of plant transcripts in gut-metatranscriptomes*

341 To assess potential hosts of the tobamoviruses detected in these libraries, we examined
 342 the metatranscriptomic data to identify transcripts from plants (Figure 2). The Poaceae
 343 family (grasses) were detected almost universally in the dog libraries isolated in QLD. This
 344 was the only plant family detected in three out of the five libraries sequenced and were only
 345 absent in one library, which was made up entirely of Fabaceae (legumes). Thus, the
 346 Poaceae and Fabaceae are strong candidates as hosts for bluey tobamovirus. Notably,
 347 grasses were present in most metatranscriptomes sampled in this study (Figure 2), albeit in
 348 lower proportions than in the dog libraries. The libraries from TAS, ACT and NSW contained
 349 a greater variety of plant transcripts making it harder to identify potential hosts (Figure 2).
 350 Both libraries containing bambi tobamovirus had transcripts from the Fabaceae,
 351 Plantaginaceae, and Poaceae, but among a range of many other potential hosts. Most

RMV- and SG3-like tobamovirus-containing libraries prepared from samples sourced from herbivores (with the exception of Gudg-CC and 12-Kang2B) contained Plantaginaceae and/or Brassicaceae transcripts (among a range of other plant transcripts), which are known hosts for SG3 tobamoviruses (27, 29). However, the Plantaginaceae and Brassicaceae families were absent in the RMV-containing carnivore libraries, suggesting potential alternate hosts (Figure 2). There were no detectable plant transcripts in the tick library.

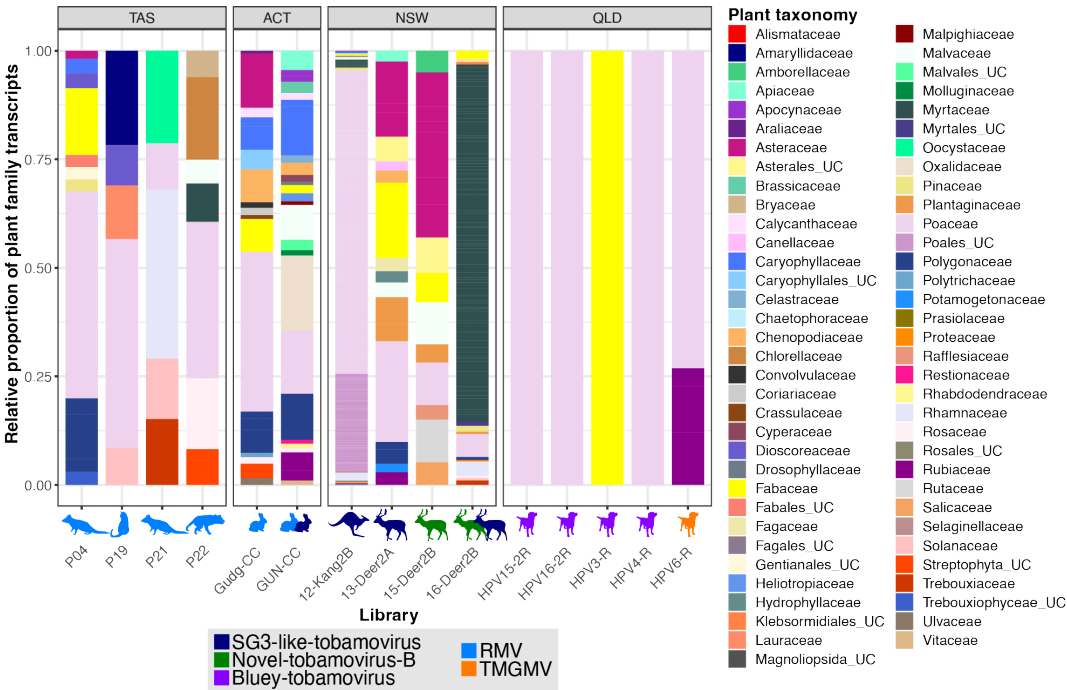


Figure 2. Abundance of plant transcripts in animal gut metatranscriptomes. Abundance of plant families is represented as a proportion of total plant transcripts present in each library. Animal silhouettes underneath each bar indicate the animal metatranscriptomes from which the plant transcripts were obtained and are coloured by viral species (purple=bluey tobamovirus; dark blue=SG3-like tobamovirus; blue=RMV; green=bambi tobamovirus; orange=TMGMV). Multiple animal silhouettes represent multiple viruses in that library. Note that libraries not pictured in the plot did not contain any plant transcripts.

PCR confirmation

We employed RT-PCR of original RNA samples to verify the true presence of novel and exotic viruses in each individual library. The following viruses were detected in multiple

373 libraries sequenced on the same lane: bluey tobamovirus, bambi tobamovirus, SG3-like
374 tobamovirus, and RMV. RNA was not available for the ACT libraries, but for the libraries
375 from NSW, QLD and TAS, the presence of the relevant viruses in individual RNA samples
376 was determined by RT-PCR (except for libraries P19 and P21 for which there was
377 insufficient RNA for all samples) (Table 3). Importantly, where RNA was available for testing,
378 we confirmed the presence of all viruses detected in metatranscriptomes in one or more
379 samples pooled in the library, including confirming the presence of the bluey tobamovirus in
380 tick samples. While insufficient RNA precluded the use of RT-PCR to confirm that the RMV
381 contigs in P19 and P21 are not the result of index hopping from other libraries, we
382 determined that the contigs from these libraries are not identical to those from other
383 libraries sequenced on the same lane, suggesting that RMV is indeed present in P19 and
384 P21. Specifically, the P21 library had multiple RMV contigs which shared 96.4-99.4%
385 nucleotide identity with contigs from P22, and 95.4-98.4% nucleotide identity with the
386 single RMV contig from P04. P19 RMV contigs shared 97.6-98.4% and 97.8-98.1%
387 nucleotide identity with contigs from P22 and P04, respectively. Equally, while RMV was
388 detected in both of the libraries from the ACT that were sequenced on the same lane, the
389 RMV contigs in these two libraries (Gudg-CC and GUN-CC) were not identical, sharing
390 98.4% nucleotide identity, and therefore RMV is likely present in both libraries.

Table 3. RT-PCR confirmation of detected viruses.

Library and RT-PCR testing details					No. of samples testing positive by RT-PCR per library				
State	Library	Animal(s) sampled	No. samples in library	No. samples tested ^a	Viruses tested ^b	RMV	SG3-like tobamovirus	Bambi tobamovirus	Bluey tobamovirus
NSW	12-Kang2B	Eastern grey kangaroo	4	4	SG3-like tobamovirus	NA ^c	1	NA	NA
NSW	13-Deer2A	Fallow deer	4	4	SG3-like tobamovirus	NA	1	NA	NA
NSW	15-Deer2B	Fallow deer	4	4	Bambi tobamovirus	NA	NA	1	NA
NSW	16-Deer2B	Fallow deer	5	5	Bambi tobamovirus, SG3-like tobamovirus	NA	4	2	NA
QLD	HPV3-R	Domestic dog	2	2	Bluey tobamovirus	NA	NA	NA	2
QLD	HPV4-R	Domestic dog	3	3	Bluey tobamovirus	NA	NA	NA	3
QLD	HPV6-R	Domestic dog	1	1	Bluey tobamovirus	NA	NA	NA	1
QLD	HPV15-2R	Domestic dog	2	2	Bluey tobamovirus	NA	NA	NA	2
QLD	HPV16-2R	Domestic dog	3	3	Bluey tobamovirus	NA	NA	NA	3
QLD	HPV35-2E	tick	1	1	Bluey tobamovirus	NA	NA	NA	1
TAS	P04	Spotted-tailed quoll	2	1	RMV	1	NA	NA	NA
TAS	P19	feral cat	1	0	RMV	Insuff ^d	NA	NA	NA

TAS	P21	Spotted-tailed quoll	1	0	RMV	Insuff	NA	NA	NA
TAS	P22	Tasmanian devil	5	5	RMV	3	NA	NA	NA

a - No. samples tested refers to the actual number of samples from this library tested by RT-PCR (as some samples had insufficient RNA)

b - Viruses tested - refers to the viruses tested by RT-PCR for this library; determined by those found in the metatranscriptome of this library as well as other libraries on the same lane (TMGMV was not included in testing as it was not a novel or exotic finding and it was only found in a single library).

c - NA=testing not required, indicated in grey

d - Insuff=Insufficient RNA for all samples in library (in both cases each library only had one RNA sample), indicated in grey.

390 Discussion

391 Tobamoviruses are known to have devastating impacts on agricultural and horticultural
392 crops. Consequently, they are listed as biosecurity threats in many countries, including
393 Australia, where many tobamoviruses are not known to be present. Australia imports seeds
394 for the horticultural industry under stringent phytosanitary inspections. However, with
395 increasing climate change and globalisation, Australia remains vulnerable to the
396 introduction of seed or vector transmitted viruses (46, 47). Surveillance has been used in
397 Australia to monitor virulent viruses and their vectors and has been critical for the early
398 detection of exotic viruses, as well as demonstration of freedom from exotic or endemic
399 pathogens (47). The surveillance methods vary by region, but mainly involve the use of
400 targeted molecular approaches and immunoassays applied to survey particular field crops
401 and samples at the border (32). These approaches are limited in that they are unlikely to
402 detect unknown viruses and potential endemic variants, and may fail to detect incursions in
403 a timely manner (15, 48). In addition, viruses may invade or persist in wild or native host
404 vegetation habitat before spilling over to cultivated land. As such, combining surveillance in
405 field crops and border samples with metatranscriptomic surveillance of wild and domestic
406 animal gut material, offers an alternative virus surveillance strategy for early detection of
407 exotic virus incursions, and demonstration of eradication or containment measures. In this
408 study, metatranscriptomic sequencing of opportunistically sampled wild and domestic
409 animal gut material in Australia revealed the presence of five different tobamovirus species,
410 including one that was thought to be exotic (RMV) and three novel viruses (a *tentative* novel
411 relative of subgroup 3 tobamoviruses – SG3-like tobamovirus – and two highly diverse
412 novel tobamoviruses, bluey tobamovirus and bambi tobamovirus). This is the first study to
413 explore the use of metatranscriptomic sequencing of animals as a strategy to detect exotic
414 crop viruses.

415
416 The significance and potential impact of the presence of each of the five tobamoviruses
417 detected here varies. The detection of RMV in Australia in native and feral animal gut
418 metatranscriptomes is critical as this virus was previously considered to be exotic in
419 Australia. Likewise, the detection of the SG3-like tobamovirus is also important, as its
420 closest relatives, RMV and TVCV are known plant pathogens (27). Furthermore, this virus
421 clusters within the larger subgroup 3 clade of viruses containing RMV, TVCV, WMoV, and
422 YoMV, which are all important pathogens of Brassicaceae and Plantaginaceae (27, 49–52).
423 Therefore, it is highly possible that this virus could be pathogenic in Brassicaceae and
424 Plantaginaceae plants like its subgroup 3 relatives (27, 49–52) as transcripts from these

host families were found in three of four libraries containing SG3-like tobamovirus. In addition, the New Zealand virus that clusters with the SG3-like tobamovirus based on CP sequence (HQ389333; originally characterised as TVCV), was found in a *Plantago* species presenting with mild chlorosis or mottles (27). Nevertheless, a broader survey of the SG3-like tobamovirus species in various hosts should be conducted in the future, along with virus host studies.

The detection of TMGMV is not alarming as this virus is among a number of tobamoviruses that have previously been detected in Australia (53), including CGMMV (14), yellow tailflower mild mottle virus (YTMMV) (54), clitoria yellow mottle virus (55), YoMV (48), pepper mild mottle virus (48), hibiscus virus (AY664875), tomato mosaic virus (NC_002692), odontoglossum ringspot virus (KF855954.1), and tomato mottle mosaic virus (56). Indeed, TMGMV has been present in Australia since the late 1800s (53). TMGMV seems well adapted to members of *Nicotiana* (57) and is particularly common in wild *Nicotiana glauca* wherever it is present around the world, including Australia (53). However, TMGMV is known to infect a wide range of hosts, including members of the Solanaceae, Umbelliferae, Gesneriaceae, Rubiaceae and Poaceae families (28, 58), and transcripts of the latter two were present in the TMGMV containing library, suggesting these as potential hosts in this case.

It is difficult to speculate on the pathogenic potential or likely hosts of the two highly novel viruses detected in this study – bluey tobamovirus and bambi tobamovirus – because they are both distinct from known viruses. However, analysis of the eukaryotic transcripts in the libraries containing bluey tobamovirus and bambi tobamovirus viruses suggests grasses (Poaceae) and potentially legumes (Fabaceae) as strong candidate hosts for bluey tobamovirus and bambi tobamovirus, along with Plantaginaceae for bambi tobamovirus. It is also possible that the true natural host for these viruses were not detectable in the metatranscriptome due to low abundance or RNA degradation, as eukaryotic RNA transcripts would not be protected like encapsulated viral RNA. Indeed, there were no detectable plant transcripts in the tick library (which contained bluey tobamovirus).

As these viruses have not been detected elsewhere, is it possible that the natural hosts of bluey tobamovirus and bambi tobamovirus are wild Australian native plants, as previously shown for yellow tailflower mild mottle virus (54). As a tobamovirus originating from native Australian plants has been shown to naturally spread to exotic plant species (22), a broader

understanding of the distribution and the natural hosts of these viruses could be beneficial for both native conservation, and their management to deter potential spillover or spread to cultivated crops. Notably, there was no evidence that any of the tobamoviruses detected in animal metatranscriptomes are present in horticultural farms in Australia. Thus, increased field surveillance efforts remain vital to protect the horticultural industry.

We detected these viruses in a variety of animal guts, including native and feral wild-living species and domestic animals, and assume that the viral hosts are plants within the animal diet. Interestingly, we also found some of these viruses in carnivores and hypothesize that the viruses were in plants that were incidentally consumed or in the gut of their ingested prey. This is possible due to the high stability of tobamoviruses and their ability to remain viable in the environment for a long time (30, 31). As tobamoviruses are spread by mechanical transmission (30), it is likely that vertebrate species may play a role in the spread of these viruses. It has already been shown that plant viruses may colonize new environments through grazing animal teeth, ingestion of plants, and subsequent passing of the viruses through the gut (24). As such, these data emphasize the potential ecological importance of wild animals in the spread of tobamoviruses and other contact transmitted viruses.

Interestingly, bluey tobamovirus was found in high abundance in a tick metatranscriptome with a near identical viral genome sequence (99.8-100% nucleotide identity) to those found in dog faeces. As this virus is highly divergent from known tobamoviruses and no plant transcripts were found in the tick library, we cannot rule out an animal tropism. However, since no tobamoviruses are known to be arboviruses, it is unlikely that this virus was replicating in ticks or dogs. Further, it is not uncommon for plant viruses to be detected in (and transmitted by) invertebrates (24, 26, 48). Indeed, 80% of plant viruses rely on insect vectors for mechanical transmission (59), and invertebrate pollinators (such as the honeybee and bumblebee) have been shown to transmit tobamoviruses (60, 61). As ticks are not herbivores, it is unlikely that they would play a major role in plant virus transmission. However, ticks are commonly found in plants, as they climb into vegetation to be in a position to fall on passing animals. It is therefore possible that the bluey tobamovirus mechanically adhered to tick body parts following tick contact with infected sap, through physical abrasion of an infected plant, as known to occur with plant viruses transmitted by bees and other pollinator insects (60–63). As such, ticks could play a role in wider dispersal of tobamoviruses. Indeed, a number of tobamoviruses were found in tick viromes in China,

including Tobacco mosaic virus, YoMV, and CGMMV (64). As the dog and tick sampling was done at an outdoor pop-up vet clinic in regional Queensland, it is also possible that the dog and tick samples were contaminated at the point of sampling. Regardless of the host origins or transmission pathways of these viruses, ectoparasites should also be considered during the management and control of these viruses in the field.

These findings demonstrate that total RNA-sequencing of wild and domestic animal gut swabs and faecal material is a potentially useful alternative strategy for the surveillance of economically important plant viruses in Australia, in addition to targeting specific field crops or natural plant hosts. This work also demonstrates that tobamoviruses are detectable in a range of different animals (including domestic, native and introduced wild animals, herbivores, carnivores, and even invertebrates) and that any method of sampling the gut content (i.e. anal/rectal swab, faecal sample, gut content sample) can be applied. While current surveillance methods involve direct sampling of individual plants or seed lots, gut metatranscriptomics allows a larger scale sampling approach whereby a pool of animal faeces, representing an entire range of ingested plants can be sequenced. This may reduce the labour involved in the field sampling effort. In addition, where RNA extraction from plant material can be challenging, the plant material in animal gut has already been broken down, and therefore extraction from animal material could be a simpler process requiring less complex tissue disruption methods. Conversely, a limitation of the gut metatranscriptomic approach is that viruses that are not highly stable in the environment may be less likely to be detected (24) due to degradation within the animal gut. It may also often fail to detect DNA viruses as only RNA transcripts from these viruses are expected to be sequenced and these may be degraded in the animal gut since they would not be protected by capsids. However, considering RNA viruses are the most abundant plant pathogens, causing nearly one-third of global crop losses (65), this approach could be a highly beneficial addition to current surveillance techniques, enabling detection of a broad range of potential viral threats. In particular, animal gut metatranscriptome-based surveillance would be a valuable tool for detection of potential viral threats circulating in the wild, facilitating mitigation techniques to avoid spill over into cultivated crops. Interestingly, in Australia, metatranscriptomics has also been employed to detect plant viruses on honeybees (48), including the devastating CGMMV known to have caused major losses in the horticultural industry (66). Although CGMMV was first detected in Australia from a survey of cucurbit crops in the Northern Territory conducted in July 2014 (14), metatranscriptomic surveillance of honeybees detected low levels of CGMMV in additional states, Queensland and Western

530 Australia, prior to its identification in field crops (48). This further demonstrates the utility of
531 metatranscriptomics of animals for plant virus surveillance.

532

533 Due to a lack of global sequence data for RMV and related viruses, it is difficult to
534 determine their source of introduction into Australia. The TMGMV detected in this study is
535 most closely related to viruses detected in other countries, including Germany, Japan and
536 Italy. However, this does not necessarily indicate that the detected virus was recently
537 introduced from overseas as TMGMV has been endemic in Australia since the early 1900s
538 (53) and is known to demonstrate little genetic variation over time (53). However, our work
539 indicates that there has been at least one recent incursion event of a known virus – RMV.
540 As the Australian RMV sequences formed a monophyletic group, the current evidence
541 suggests that there was a single incursion of this virus. The SG3-like tobamovirus also
542 seems likely to be from a single introduction and, based on CP sequence, is closely related
543 to New Zealand and German sequences (although support for the German and New
544 Zealand viruses' position in the CP tree was weak). Finally, the two highly diverse novel
545 tobamoviruses, bluey tobamovirus and bambi tobamovirus, are so distinct from known
546 viruses that it is difficult to determine if they represent exotic incursions or whether they are
547 endemic in Australian plants, like YTMMV (54).

548

549 Importantly, tobamoviruses were detected in animals in four Australian states and
550 territories, and RMV and SG3-like tobamovirus were each found in two states and
551 territories indicating spread of these viruses within Australia. RMV was detected in both the
552 ACT and Tasmania, even though these locations are more than 700 kms apart and
553 separated from the mainland of Australia by Bass strait, >200km of water. Viruses may
554 have spread over Bass strait through their ability to be contact transmitted via fomites (15) –
555 potentially on shoes or equipment travelling by air or on car tyres on ferries, or even by
556 migratory birds. Future studies should focus on understanding the distribution and the
557 impact of RMV and the other tobamoviruses detected here.

558

559 In sum, we determined that metatranscriptomic screening of animal gut content could serve
560 as a non-targeted surveillance tool, particularly in conjunction with existing methods of
561 plant biosecurity surveillance. Future work should focus on characterizing the host range,
562 geographic distribution and the impact potential of the new and known tobamoviruses
563 detected here. The significant distribution of these viruses in Australia demonstrates the
564 importance of early detection to prevent cryptic spread. This study also solidifies the

importance of continuous integration of innovative rapid diagnostic and surveillance
genomics tools to support viral disease management for sustainable food production.

Funding

ECH and JM were supported by an Australian Research Council Australian Laureate Fellowship (FL170100022). SM was funded by the Centre for Invasive Species Solutions through a grant titled Enhanced preparedness to diagnose 200+ priority pests and diseases of NSW and Australian Agriculture (A-023). MJ was funded by an Australian Research Council Discovery grant (DP170101653), ACW was supported by a Holsworth Wildlife Research Endowment – Equity Trustees Charitable Foundation from the Ecological Society of Australia to ACW. ACW and MJ were funded by the Dr. Eric Guiler Tasmanian Devil Research Grant. TMN and SB were funded by Hermon Slade Foundation and also by the Holsworth Wildlife Research Endowment. AL was supported by a Fulbright Future Scholarship, funded by the Kinghorn Foundation.

Acknowledgements

We thank Wendy Forbes assisting with samples submission to AGRF and ordering the positive controls. We acknowledge the University of Sydney's high-performance computing cluster Artemis for providing the computing resources used for this study.

References

1. Wamaitha MJ, Nigam D, Maina S, Stomeo F, Wangai A, Njuguna JN, Holton TA, Wanjala BW, Wamalwa M, Lucas T, Djikeng A, Garcia-Ruiz H. 2018. Metagenomic analysis of viruses associated with maize lethal necrosis in Kenya. *Virology* 15:90.
2. Jones RAC, Sharman M, Trębicki P, Maina S, Congdon BS. 2021. Virus diseases of cereal and oilseed crops in Australia: Current position and future challenges. *Viruses* 13:2051.
3. IPPC Secretariat. 2021. Scientific review of the impact of climate change on plant pests: A global challenge to prevent and mitigate plant-pest risks in agriculture, forestry and ecosystems. FAO on behalf of the IPPC Secretariat, Rome, Italy. <https://www.fao.org/documents/card/en/c/cb4769en>. Retrieved 13 June 2023.
4. Alishiri A, Rakhshandehroo F, Zamanizadeh H-R, Palukaitis P. 2013. Prevalence of *Tobacco mosaic virus* in Iran and evolutionary analyses of the coat protein gene. *Plant Pathology* 29:260–273.
5. Cooper B. 2014. Proof by synthesis of *Tobacco mosaic virus*. *Genome Biology* 15:R67.

- 600 6. Canto T, MacFarlane SA, Palukaitis P. 2004. ORF6 of *Tobacco mosaic virus* is a
601 determinant of viral pathogenicity in *Nicotiana benthamiana*. J Gen Virol 85:3123–
602 3133.
- 603 7. Chavan RR, Pearson MN. 2016. Molecular characterisation of a novel recombinant
604 *Ribgrass mosaic virus* strain FSHS. Virol J 13:29.
- 605 8. Adkins S, Kamenova I, Achor D, Lewandowski DJ. 2003. Biological and molecular
606 characterization of a novel tobamovirus with a unique host range. Plant Dis 87:1190–
607 1196.
- 608 9. Jones RAC, Naidu RA. 2019. Global dimensions of plant virus diseases: Current status
609 and future perspectives. Annu Rev Virol 6:387–409.
- 610 10. Hanssen IM, Lapidot M. 2012. Major tomato viruses in the Mediterranean basin. Adv
611 Virus Res 84:31–66.
- 612 11. Salem N, Mansour A, Ciuffo M, Falk BW, Turina M. 2016. A new tobamovirus infecting
613 tomato crops in Jordan. Arch Virol 161:503–506.
- 614 12. Luria N, Smith E, Reingold V, Bekelman I, Lapidot M, Levin I, Elad N, Tam Y, Sela N,
615 Abu-Ras A, Ezra N, Haberman A, Yitzhak L, Lachman O, Dombrovsky A. 2017. A new
616 Israeli tobamovirus isolate infects tomato plants harboring tm-22 resistance genes.
617 PLoS One 12:e0170429.
- 618 13. Ghorbani A. 2024. Genetic analysis of tomato brown rugose fruit virus reveals
619 evolutionary adaptation and codon usage bias patterns. Sci Rep 14:21281.
- 620 14. Tesoriero LA, Chambers G, Srivastava M, Smith S, Conde B, Tran-Nguyen LTT. 2015.
621 First report of cucumber green mottle mosaic virus in Australia. Australasian Plant Dis
622 Notes 11:1.
- 623 15. Dombrovsky A, Tran-Nguyen LTT, Jones RAC. 2017. *Cucumber green mottle mosaic*
624 *virus*: rapidly increasing global distribution, etiology, epidemiology, and management.
625 Annu Rev Phytopathol 55:231–256.
- 626 16. Maina S, Coutts BA, Edwards OR, de Almeida L, Kehoe MA, Ximenes A, Jones RAC.
627 2017. *Zucchini yellow mosaic virus* populations from East Timorese and Northern
628 Australian cucurbit crops: Molecular properties, genetic connectivity, and biosecurity
629 implications. Plant Dis 101:1236–1245.
- 630 17. Gibbs AJ, Mackenzie AM, Wei K-J, Gibbs MJ. 2008. The potyviruses of Australia. Arch
631 Virol 153:1411–1420.
- 632 18. Jones RAC. 2009. Plant virus emergence and evolution: Origins, new encounter
633 scenarios, factors driving emergence, effects of changing world conditions, and
634 prospects for control. Virus Res 141:113–130.

- 635 19. Maina S, Edwards OR, Jones RAC. 2016. First complete genome sequence of *Pepper*
636 *vein yellows virus* from Australia. *Genome Announc* 4:e00450-16.
- 637 20. Jones RAC. 2020. Disease pandemics and major epidemics arising from new
638 encounters between indigenous viruses and introduced crops. *Viruses* 12:1388.
- 639 21. Roossinck MJ, García-Arenal F. 2015. Ecosystem simplification, biodiversity loss and
640 plant virus emergence. *Curr Opin Virol* 10:56–62.
- 641 22. Xu W, Li H, Sivasithamparam K, Tran DT, Jones MGK, Chen X, Wylie SJ. 2022.
642 Spillover of a tobamovirus from the Australian indigenous flora to invasive weeds.
643 *Viruses* 14:1676.
- 644 23. Malmstrom CM, Alexander HM. 2016. Effects of crop viruses on wild plants. *Curr Opin*
645 *Virol* 19:30–36.
- 646 24. Roossinck MJ. 2013. Plant virus ecology. *PLoS Pathog* 9:e1003304.
- 647 25. Roossinck MJ. 2012. Plant virus metagenomics: Biodiversity and ecology. *Annu Rev*
648 *Genet* 46:359–369.
- 649 26. Mahar JE, Shi M, Hall RN, Strive T, Holmes EC. 2020. Comparative analysis of RNA
650 virome composition in rabbits and associated ectoparasites. *J Virol* 94:e02119-19.
- 651 27. Cohen D, Chavan RR, Blouin AG, Pearson MN. 2012. First report of *Turnip vein*
652 *clearing virus* and *Ribgrass mosaic virus* from New Zealand. *Australasian Plant Dis*
653 *Notes* 7:67–69.
- 654 28. Wetter C. 1986. *Tobacco mild green mosaic virus*, p. 205–219. *In* Van Regenmortel,
655 MHV, Fraenkel-Conrat, H (eds.), *The plant viruses: The rod-shaped plant viruses*.
656 Springer US, Boston, MA.
- 657 29. Heinze C, Lesemann D-E, Ilmberger N, Willingmann P, Adam G. 2006. The
658 phylogenetic structure of the cluster of tobamovirus species serologically related to
659 ribgrass mosaic virus (RMV) and the sequence of streptocarpus flower break virus
660 (SFBV). *Arch Virol* 151:763–774.
- 661 30. Smith E, Dombrovsky A. 2019. Aspects in tobamovirus management in intensive
662 agriculture, Ch. 3. *In* *Plant diseases-Current threats and management trends*.
663 IntechOpen.
- 664 31. Lovelock D, Mintoff S, Kurz N, Neilsen M, Patel S, Constable F, Tran-Nguyen L. 2022.
665 Investigating the longevity and infectivity of cucumber green mottle mosaic virus in
666 soils of the Northern Territory, Australia. *Plants (Basel)* 11:883.
- 667 32. Dall DJ, Lovelock DA, Penrose LDJ, Constable FE. 2023. Prevalences of tobamovirus
668 contamination in seed lots of tomato and capsicum. 4. *Viruses* 15:883.

- 669 33. Maina S, Zheng L, Rodoni BC. 2021. Targeted genome sequencing (TG-Seq)
670 approaches to detect plant viruses. *Viruses* 13:583.
- 671 34. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
672 sequence data. *Bioinformatics* 30:2114–2120.
- 673 35. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-
674 node solution for large and complex metagenomics assembly via succinct de Bruijn
675 graph. *Bioinformatics* 31:1674–1676.
- 676 36. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,
677 Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di
678 Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-
679 length transcriptome assembly from RNA-Seq data without a reference genome. *Nat*
680 *Biotechnol* 29:644–652.
- 681 37. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life
682 scale using DIAMOND. *Nat Methods* 18:366–368.
- 683 38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment
684 search tool. *J Mol Biol* 215:403–410.
- 685 39. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997.
686 Gapped BLAST and PSI-BLAST: a new generation of protein database search
687 programs. *Nucleic Acids Res* 25:3389–3402.
- 688 40. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
689 improvements in performance and usability. *Mol Biol Evol* 30:772–780.
- 690 41. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAL: a tool for
691 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*
692 25:1972–3.
- 693 42. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017.
694 modelfinder: fast model selection for accurate phylogenetic estimates. 6. *Nat Methods*
695 14:587–589.
- 696 43. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and
697 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol*
698 *Evol* 32:268–74.
- 699 44. Martin DP, Varsani A, Roumagnac P, Botha G, Maslamoney S, Schwab T, Kelz Z,
700 Kumar V, Murrell B. 2021. RDP5: a computer program for analyzing recombination in,
701 and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol*
702 7:veaa087.

- 703 45. Marcelino VR, Clausen PTLC, Buchmann JP, Wille M, Iredell JR, Meyer W, Lund O,
704 Sorrell TC, Holmes EC. 2020. CCMetagen: comprehensive and accurate identification
705 of eukaryotes and prokaryotes in metagenomic data. *Genome Biol* 21:103.
- 706 46. Davis RI, Jones LM, Pease B, Perkins SL, Vala HR, Kokoa P, Apa M, Dale CJ. 2021.
707 Plant virus and virus-like disease threats to Australia's north targeted by the northern
708 Australia quarantine strategy. *Plants (Basel)* 10:2175.
- 709 47. Maina S, Jones RAC. 2023. Enhancing biosecurity against virus disease threats to
710 Australian grain crops: current situation and future prospects. *Front Hortic* 2:1263604.
- 711 48. Roberts JMK, Ireland KB, Tay WT, Paine D. 2018. Honey bee-assisted surveillance for
712 early plant virus detection. *Ann Appl Biol* 173:285–293.
- 713 49. Melcher U. 2003. *Turnip vein-clearing virus*, from pathogen to host expression profile.
714 *Mol Plant Pathol* 4:133–140.
- 715 50. MacDonald JL, Punja ZK, Xiang Y, Bouthillier MJ, Reade R, DeYoung RM, Bhagwat B,
716 Betz EC, Li YQ, Chen X. 2021. First report of wasabi mottle virus causing ringspot and
717 vein-clearing symptoms on wasabi (*Wasabia japonica*) in North America. *Can J Plant*
718 *Pathol* 43:311–322.
- 719 51. Kim J-S, Cho J-D, Choi H-S, Lee S-H, Choi G-S, Lee S-Y, Kim H-J, Yoon M-K. 2010.
720 Ribgrass mosaic tobamovirus occurred on Chinese cabbage in Korea. *Plant Pathol J*
721 26:328–339.
- 722 52. Aguilar I, Sánchez F, Martín AM, Martínez-Herrera D, Ponz F. 1996. Nucleotide
723 sequence of *Chinese rape mosaic virus* (oilseed rape mosaic virus), a crucifer
724 tobamovirus infectious on *Arabidopsis thaliana*. *Plant Mol Biol* 30:191–197.
- 725 53. Fraile A, Escriu F, Aranda MA, Malpica JM, Gibbs AJ, García-Arenal F. 1997. A century
726 of tobamovirus evolution in an Australian population of *Nicotiana glauca*. *J Virol*
727 71:8316–8320.
- 728 54. Wylie SJ, Li H, Jones MGK. 2014. *Yellow tailflower mild mottle virus*: a new
729 tobamovirus described from *Anthocercis littorea* (Solanaceae) in Western Australia.
730 *Arch Virol* 159:791–795.
- 731 55. Wei K, Gibbs A, Mackenzie A. 2012. *Clitoria yellow mottle virus*: a tobamovirus from
732 Northern Australia. *Australasian Plant Dis Notes* 7:59–61.
- 733 56. Lovelock DA, Kinoti WM, Bottcher C, Wildman O, Dall D, Rodoni BC, Constable FE.
734 2020. *Tomato mottle mosaic virus* intercepted by Australian biosecurity in *Capsicum*
735 *annuum* seed. *Australasian Plant Dis Notes* 15:8.
- 736 57. de Andrés-Torán R, Guidoum L, Zamfir AD, Mora MÁ, Moreno-Vázquez S, García-
737 Arenal F. 2023. *Tobacco mild green mosaic virus* (TMGMV) isolates from different plant

- 738 families show no evidence of differential adaptation to their host of origin. *Viruses*
739 15:2384.
- 740 58. Zamfir AD, Babalola BM, Fraile A, McLeish MJ, García-Arenal F. 2023. Tobamoviruses
741 show broad host ranges and little genetic diversity among four habitat types of a
742 heterogeneous ecosystem. *Phytopathology* 113:1697–1707.
- 743 59. Hohn T. 2007. Plant virus transmission from the insect point of view. *PNAS*
744 104:17905–17906.
- 745 60. Darzi E, Smith E, Shargil D, Lachman O, Ganot L, Dombrovsky A. 2018. The honeybee
746 *Apis mellifera* contributes to *Cucumber green mottle mosaic virus* spread via
747 pollination. *Plant Pathol* 67:244–251.
- 748 61. Okada K, Kusakari S, Kawaratani M, Negoro J, Ohki ST, Osaki T. 2000. *Tobacco*
749 *mosaic virus* is transmissible from tomato to tomato by pollinating bumblebees. *J Gen*
750 *Plant Pathol* 66:71–74.
- 751 62. Aramburu J, Galipienso L, Aparicio F, Soler S, López C. 2010. Mode of transmission of
752 *parietaria mottle virus*. *J Plant Pathol* 92:679–684.
- 753 63. Sdoodee R, Teakle DS. 1993. Studies on the mechanism of transmission of pollen-
754 associated tobacco streak ilarvirus virus by *Thrips tabaci*. *Plant Pathol* 42:88–92.
- 755 64. Ni X-B, Cui X-M, Liu J-Y, Ye R-Z, Wu Y-Q, Jiang J-F, Sun Y, Wang Q, Shum MH-H,
756 Chang Q-C, Zhao L, Han X-H, Ma K, Shen S-J, Zhang M-Z, Guo W-B, Zhu J-G, Zhan
757 L, Li L-J, Ding S-J, Zhu D-Y, Zhang J, Xia L-Y, Oong X-Y, Ruan X-D, Shao H-Z, Que T-
758 C, Liu G-Y, Du C-H, Huang E-J, Wang X, Du L-F, Wang C-C, Shi W-Q, Pan Y-S, Zhou
759 Y-H, Qu J-L, Ma J, Gong C-W, Chen Q-Q, Qin Q, Lam TT-Y, Jia N, Cao W-C. 2023.
760 Metavirome of 31 tick species provides a compendium of 1,801 RNA virus genomes.
761 1. *Nat Microbiol* 8:162–173.
- 762 65. Pagán I, Firth C, Holmes EC. 2010. Phylogenetic analysis reveals rapid evolutionary
763 dynamics in the plant RNA virus genus *Tobamovirus*. *J Mol Evol* 71:298–307.
- 764 66. Mackie J, Campbell PR, Kehoe MA, Tran-Nguyen LTT, Rodoni BC, Constable FE.
765 2023. Genome characterisation of the CGMMV virus population in Australia—
766 informing plant biosecurity policy. *Viruses* 15:743.

Tunicate metatranscriptomes reveal ancient virus-host co-divergence and inter-order recombination in the evolutionary history of disease-causing viruses

Mary E. Petrone^{1,2}, Joe Grove³, Rhys H. Parry⁴, Kate Van Brussel¹, Jonathon C.O. Mifsud¹, Zuhairah Dindar⁵, Shi-qiang Mei⁶⁻⁸, Mang Shi⁶⁻⁹, Olivia M. H. Turnbull¹, Ezequiel M. Marzinelli¹⁰, Edward C. Holmes^{1,2*}

¹School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.

²Laboratory of Data Discovery for Health Limited, Hong Kong SAR, China.

³MRC-University of Glasgow Centre for Virus Research, Glasgow G61 1QH, UK.

⁴School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4067, Australia.

⁵Centre for Marine Science and Innovation, Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, UNSW, Sydney, NSW 2033, Australia.

⁶State Key Laboratory for Biocontrol, School of Medicine, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.

⁷National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, Sun Yat-sen University, Shenzhen, China.

⁸Shenzhen Key Laboratory for Systems Medicine in Inflammatory Diseases, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.

⁹Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China.

¹⁰School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales, Australia.

*Correspondence:

Prof. Edward C. Holmes

University of Sydney

edward.holmes@sydney.edu.au

32 **SUMMARY**

33 Tunicates are a key transitional taxon in animal evolution as the closest extant invertebrate
 34 relatives of the vertebrates. Their viruses may also reflect this transitional state. Yet, it is not
 35 known whether tunicate viruses are more closely related to vertebrate- or invertebrate-infecting
 36 viral lineages. We analysed primary and publicly available RNA libraries to extend the known
 37 diversity of tunicate-associated viruses and determine their relationship to viruses of other
 38 animals. We present evidence that influenza viruses, alphaviruses, and some mononegaviruses
 39 emerged prior to the evolution of vertebrates. We also show that the recombination of
 40 glycoproteins between different orders of RNA viruses, including between positive- and
 41 negative-sense viruses, may have shaped the evolution of multiple lineages. Our study reveals
 42 that some disease-causing RNA virus lineages were present in early chordates and highlights
 43 that the evolution of structural genes may be incongruent with that of the highly conserved
 44 RNA-dependent RNA polymerase.

45

46 **KEY WORDS**

47 RNA virus, evolutionary virology, influenza virus, recombination, protein structure prediction

INTRODUCTION

Tunicates, sometimes referred to as urochordates, are filter-feeding marine invertebrate members of the phylum Chordata. The Tunicata subphylum comprises three classes – the *Ascidiacea*, *Thaliacea*, and *Appendicularia* – that exhibit distinct life cycles and body plans. The *Ascidiacea* (“sea squirts”) are the canonical tunicate, as it was their thick, tunic-like encasement made of cellulose that inspired the name “Tunicata” (Lamarck, 1816). Ascidiarians begin life as free-swimming ‘tadpole’ larvae with a simple chordate body-plan, before attaching to manmade or natural surfaces where they undergo profound morphological changes and become sessile adults. Despite a simple outward appearance, adult ascidians possess a complement of organs, including a pharynx, reproductive organs, digestive organs, neural gland, and a heart that circulates hemolymph. The *Thaliacea* (salps and doliolids) do not attach to surfaces when they mature and remain free floating at a wide range of ocean depths for the duration of their lives¹. Like that of the ascidians, the doliolid life cycle includes a tadpole stage. In contrast to the *Ascidiacea* and *Thaliacea*, the *Appendicularia* do not undergo morphological changes as they mature, instead crafting intricate mucosal “houses” around their tadpole-like bodies to trap food².

The Tunicata are of evolutionary importance because all tunicate larvae possess a notochord. It was the discovery of this feature by Alexander Kovalevsky in 1866 that prompted the reclassification of the Tunicata from Mollusca to Chordata³. Since then, ascidians, in particular *Ciona intestinalis* (Phlebobranchia, *Ascidiacea*), have been used to elucidate the mechanisms driving gastrulation and the formation of the notochord during the early stages of chordate development^{4,5}. The precise phylogenetic relationship of the Tunicata to the Vertebrata has sparked considerable debate. While it has been argued that cephalochordates (lancelets) are the sister group of the vertebrates^{6,7}, molecular data support the conclusion that tunicates are indeed the closest relatives of the vertebrates⁸⁻¹⁰. Thus, tunicates represent a key transitional taxon in animal evolution as the “most vertebrate-like invertebrate”.

Despite their scientific importance, very little is known about the viruses that infect tunicates aside from studies of bacteriophage diversity in tunicate gut microbiomes¹¹⁻¹³ and incidental findings from metagenomic sequencing. Outbreak investigations into the cause of soft tunic syndrome, a disease that affects edible ascidians, identified a novel birnavirus^{14,15}. However, a pathogenic protist was later determined to be the aetiological agent¹⁶⁻¹⁸. Data mining of tunicate metatranscriptomes identified two members of the *Medioniviridae* (*Botrylloides leachii*

nidovirus, NCBI accession MK956105 and Tognidovirus botryllus22085, accession BK066634¹⁹) that utilise togavirus-like class II fusion proteins. Tunicate-associated Beihai orthomyxo-like virus 2 was originally classified within the *Orthomyxoviridae*²⁰, but was later found to be a member of the divergent “Cnidenomoviridae” (order *Articulavirales*)²¹. Lastly, Chowder bay tunicate associated flavi-like virus (OX394159) was identified as a member of the *Flaviviridae* related to other marine invertebrate viruses among the “tamanaviruses”²². The diversity of tunicate-associated viruses in such a small data set hints at unrealised RNA virus diversity that has yet to be fully explored.

As well as uncovering more of the global virosphere, a thorough characterisation of the tunicate virome could address a key evolutionary question: as tunicates are the sister-group to the vertebrates, do their RNA viruses share a closer phylogenetic relationship with viruses found in vertebrate or invertebrate hosts? If the former, these vertebrate-infecting viral lineages may have predated the evolution of the Vertebrata themselves. In addition, the emergence of the vertebrates was marked by major evolutionary innovations, including the development of an adaptive immune system, which may have in turn impacted virus diversity. Hence, the identification of vertebrate-infecting-like viruses in tunicates would provide a starting point for understanding how these innovations in vertebrate evolution shaped the long-term evolutionary history of disease-causing viruses. Herein, we explored this question by systematically screening primary and publicly available tunicate metatranscriptomes for RNA viruses. We combined phylogenetic analysis and protein structural prediction to re-assess the evolutionary origins of some vertebrate-infecting viral lineages and examine instances of recombination between viral orders.

RESULTS

Tunicate metatranscriptomes are a rich source of viral diversity

We assembled a data set of 3,499 tunicate metatranscriptomes from publicly available SRA libraries and primary samples. The SRA libraries comprised all three classes within the Tunicata (*Ascidacea*, *Appendicularia*, and *Thaliacea*), albeit with an uneven distribution: the majority of libraries were *Ascidacea* (93%), and orders within each class were not uniformly represented (**Fig. 1a**). The primary samples sequenced for this study comprised 31 samples of *Ascidacea* collected from Chowder Bay, Sydney Harbour, Australia at two time points approximately one year apart (November 2021 and October 2022) (**Fig. 1b**).

116 This data set yielded extensive viral diversity, particularly in the primary samples. In the SRA
 117 and primary libraries combined, we detected putative viral contigs in 242 of the 3,480 libraries
 118 analysed (7.0%) (**Fig. 1c**). However, we identified putative viral contigs in all the libraries
 119 generated from primary samples, which may be indicative of differences in viral prevalence
 120 between primary and laboratory reared organisms and library selection protocols (e.g., the use
 121 of poly-A selection in 440 libraries). Of the 1,370 putative viral contigs detected, 33 belonged
 122 to known viral taxa and an additional 187 shared sequence similarity with viruses that are
 123 unclassified by NCBI or the International Committee on Taxonomy of Viruses (ICTV). We
 124 further categorised these contigs as “vertebrate-like candidates” and “probable non-vertebrate
 125 infecting” according to the viral order to which they belonged (**Fig. 1d**). Contigs assigned to
 126 “vertebrate-like candidates” belonged to taxa with known vertebrate hosts: the *Amarillovirales*,
 127 *Articulavirales*, *Birnaviridae*, *Bunyavirales*, *Hepelivirales*, *Jingchuvirales*, *Martellivirales*,
 128 *Mononegavirales*, *Nidovirales*, *Nodamuvirales*, *Picornavirales*, *Reovirales*, and *Stellavirales*
 129 (**Fig. 1e**). Contigs sharing sequence similarity to viruses in taxa that have no known vertebrate
 130 hosts (e.g., the *Ourlivirales*) were classified as “probable non-vertebrate infecting” and
 131 excluded from further analysis. No vertebrate-like candidates were found in *Appendicularia*
 132 libraries and only three candidates were detected among the *Thaliacea* libraries (**Fig. 1d**).
 133
 134 Because each viral taxon with known vertebrate hosts also includes families of viruses that
 135 infect non-vertebrates (i.e., invertebrates and organisms outside of the animal kingdom), we
 136 performed preliminary phylogenetic analyses to identify and remove vertebrate-like candidates
 137 that fell within known invertebrate-infecting virus clades. Virus candidates belonging to the
 138 orders *Amarillovirales*, *Birnaviridae*, *Bunyavirales*, *Hepelivirales*, *Nodamuvirales*,
 139 *Picornavirales*, and *Reovirales* shared closer phylogenetic relationships to non-vertebrate-
 140 associated viruses in their respective taxa (**Figs. S1-9**). Notably, two divergent *Amarillovirales*
 141 NS5 sequences were positioned close to Chowder bay tunicate associated flavi-like virus²² near
 142 the tamanaviruses (**Fig. S1b**). Two tunicate-associated birna-like viruses formed a clade with
 143 Chicken proventricular necrosis virus and a bat faecal birnavirus (UVF62127.1) with low
 144 support when aligned with MAFFT (**Fig. S2a**), but because this relationship was not apparent
 145 when a different alignment method (MUSCLE²³) was used (**Fig. S2b**), we were unable to draw
 146 a confident conclusion on the phylogenetic relationships among these viruses. Similarly, two
 147 viruses found in Stolidobranchia libraries (*Polycarpa*, accession: SRR18265385; *Botrylloides*,
 148 primary sample) fell at the base of a fish astrovirus clade (**Fig. S9a**). However, this relationship
 149 was not recapitulated when sequences were aligned with MUSCLE²³. Instead, these viruses

150 fell at the base of a clade comprising bat-associated astroviruses and Duck hepatitis virus 3
151 (albeit with low support; sh-aLRT/UFboot = 50/59) (**Fig. S9b**). In the case of the *Nidovirales*,
152 tunicate-associated nidoviruses were not closely related to the vertebrate-infecting members of
153 this order and exhibited unique features that we discuss below.

154

155 As a consequence of these analyses, we removed viruses belonging to the *Amarillovirales*,
156 *Birnaviridae*, *Bunyavirales*, *Hepelivirales*, *Nodamuvirales*, *Picornavirales*, and *Reovirales*
157 from our vertebrate-like candidate data set. Virus candidates belonging to the *Articulavirales*,
158 *Jingchuvirales*, *Martellivirales*, *Mononegavirales*, and *Nidovirales* were retained for
159 subsequent analysis.

160

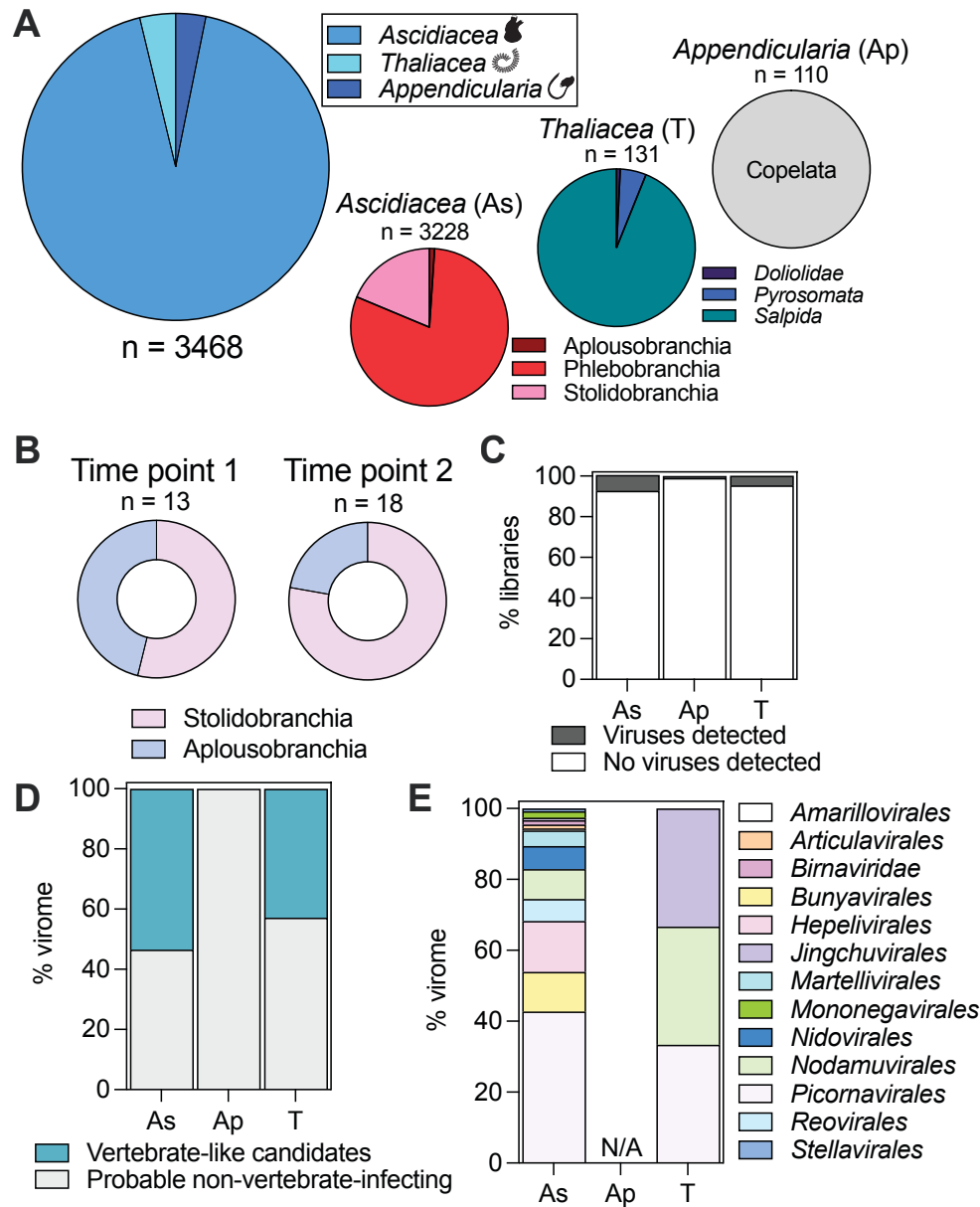


Figure 1. Diversity of RNA viruses identified in Tunicata metatranscriptomes. (a) Distribution of host organisms represented in SRA libraries as of February 2024. (b) Distribution of host organisms of primary samples collected in November 2021 (time point 1) and October 2022 (time point 2). (c) Percentage of libraries in which viral contigs were detected by host class. (d) Distribution of viruses belonging to orders with known vertebrate hosts (vertebrate-like candidates) and with no known vertebrate hosts (probable non-vertebrate infecting). (e) Distribution of vertebrate-like candidates in libraries by host class. As: *Ascidacea*, Ap: *Appendicularia*, T: *Thaliacea*.

Evolutionary history of the influenza virus lineage may have predated the Vertebrata

We detected articulavirus-like segments in eight *Ascidacea* libraries (**Fig. 2a**). Three libraries (accessions SRR7503161, SRR7503162, SRR7503170) were from the same BioProject (PRJNA480375) and their corresponding segments were identical. These and the library generated from *Clavelina lepadiformis* (accessions ERR11767553, ERR11767551) were derived from pharyngeal tissue. Only a small fragment (692nt) of a polymerase basic (PB) 1-like segment was found in a primary library that we collected at the first time point. Two fragments of a PB3-like segment were also identified in this library but could not be joined with PCR. We identified complete or near-complete additional segments in five of the eight libraries and fragments in two libraries (**Fig. 2a**). All segments were at low abundance in their respective libraries (<0.01% non-rRNA reads) (**Supp. Data 1**).

Notably, the PB1 segment in the *Clavelina* library encoded a complete PB1 protein with two overlapping ORFs that could make the complete PB1 through a programmed ribosomal frameshift (-1 PRS). This -1 PRF mechanism involves a "slippery" heptanucleotide sequence (X_XXY_YYZ, here A_AAA_AAA), followed by a downstream stem-loop RNA structure, stimulating ribosomal slippage. The placement of this frameshift (position 841) was not consistent with the position of the PB1-F2 ORF found in some influenza A PB1 segments²⁴. The sequencing coverage for this contig was low (median fold = 6, BMap, **Fig. S10**), but it was not indicative of an assembly error. We detected five additional segments in this library (PB2, PB3, NP, HA, M), which supported the conclusion that this virus was exogenous rather than an endogenous sequence.

Four of the articulavirus-like PB1 segments were influenza-like, falling basal to the hagfish-associated PB1-like sequence with long branch lengths (**Fig. 2b,c**). Three of these (*Ciona savignyi*, *Clavelina lepadiformis*, and *Ascidia mentula* influenza-like viruses) formed a clade (**Fig. 2b,c**). Although the relationship of these sequences to that associated with a hagfish (jawless vertebrate) and most other vertebrate-associated influenza viruses was consistent with virus-host co-divergence over timescales of hundreds of millions of years (**Fig. 2c**), their relationship to each other was not. In particular, *Ciona* and *Ascidia* are both members of the Phlebobranchia, while *Clavelina*, the sister group to the *Ciona*-associated sequence, belongs to the Aplousobranchia. Additionally, two fish-associated sequences (Grass carp influenza-like

204 virus and Sturgeon influenza-like virus) fell at the base of the clade, albeit with low support
205 (sh-aLRT/UFboot = 49.9/44), suggestive of possible ancient host jumps. This placement was
206 consistent in the other two polymerase segments (PB2 and PB3, **Fig. S11**), but not in the
207 structural proteins (HA and NP, **Fig. S12,13**) that may be too divergent to compare.

208

209 The other segments from the tunicate libraries similarly shared a close relationship with the
210 canonical influenza viruses (i.e., Influenza A, B, C, and D) (**Fig. S11-13**). In all cases, the PB2
211 and PB3 segments shared a closer relationship with influenza C and D viruses rather than
212 falling as distinct lineages (**Fig. S11**). The placement of the NP and HA segments was uncertain
213 and changed depending on the alignment method used (MAFFT: C/D-like, MUSCLE: A/B-
214 like) (**Fig. S12-13**). The PB1 segment from the *Aplidium* library (ERR6745738), which was
215 the only segment we could identify in this library, was not influenza-like and formed a sister
216 group with the divergent “Cnidenomoviridae”²¹ (**Fig. 2a,b**).

217

218 Given that influenza-like viruses have not been previously documented in invertebrates and
219 ascidians are filter-feeders that consume particulate organic matter, an alternative hypothesis
220 was that these viruses were derived from the diet of the host animal. However, this was not
221 supported through an analysis of the library composition. In three cases, other chordate reads
222 comprised a small percentage of each library (6% ERR1037086, 1% SRR7503170, 0.7%
223 Primary Sample #17), and no other chordate reads were detected in the remaining four libraries
224 (**Fig. S14**). The consistently close phylogenetic relationship of the tunicate-associated
225 segments further suggested that tunicates were the *bona fide* host because the libraries were
226 derived from organisms collected across multiple continents (e.g., ERR10378036, Europe;
227 SRR7503170, Asia; Primary sample #17, Oceania). We therefore conclude that these viruses
228 likely represent the first observed instances of invertebrate-associated influenza-like viruses,
229 and, as a consequence, that the influenza virus lineage evolved prior to the Vertebrata.

230

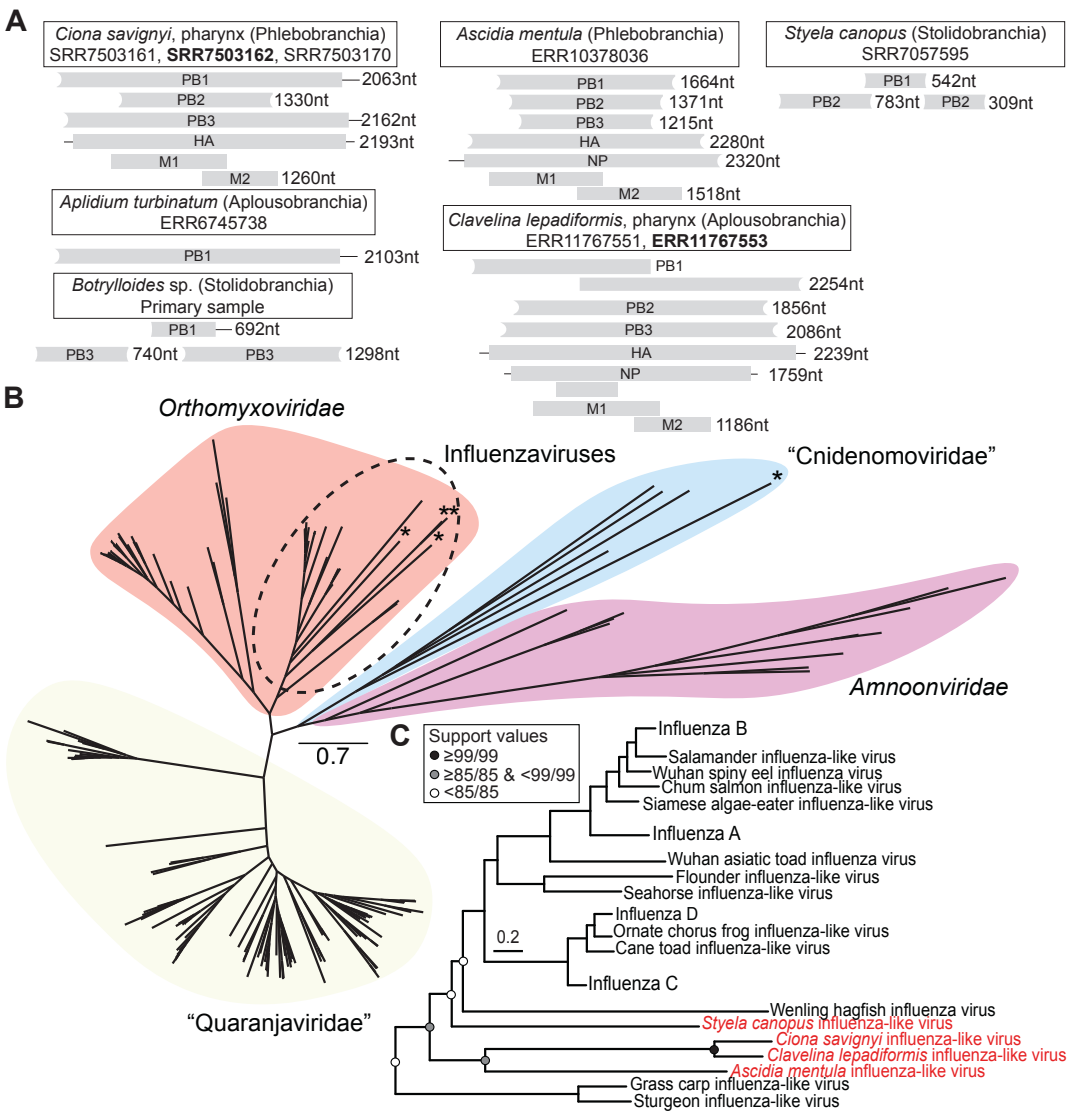


Figure 2. Influenza-like virus segments in seven ascidian metatranscriptomes. (a) Schematic of articlaviruses-like segments detected in SRA (n = 8) and primary (n = 1) metatranscriptome libraries. The lengths of each segment are denoted in nucleotides (nt). Curved edges indicate incomplete ORFs. Bold face indicates the representative library from which segments were visualised. (b) Unrooted phylogeny of the *Articulavirales* inferred using the PB1 gene. Viruses identified in this study are indicated with stars. (c) Extract of the influenza virus clade from the order-level phylogeny. Viruses identified in this study are denoted by red tip labels. Support values for select nodes are reported as sh-aLRT/UFboot. The branches of both phylogenies are scaled by the number of amino acid substitutions per site.

Evidence for aquatic, invertebrate origins of multiple vertebrate disease-causing viral lineages

Given the discovery of tunicate-associated influenza-like viruses we asked whether other disease-causing viral lineages might have originated early in chordate evolution. We identified one such example in the novirhabdoviruses, negative-sense RNA viruses from the order *Mononegavirales* that can cause overt disease in finfish²⁵. Specifically, five putative viral contigs in our data set formed a “stepwise” branching pattern at the base of the canonical *Novirhabdovirus* clade (**Fig. 3a**). The family-level tree topology recapitulated the position of the novirhabdoviruses as the only members of the *Gammarhabdovirinae*, which form a sister group to the other lineages within the *Rhabdoviridae*²⁶ (**Fig. 3b**). Two novirhabdo-like viruses were associated with Stolidobranchia (*Botryllus schlosseri*, accession SRR10352255; *Botrylloides leachii*, primary sample). Three identical and highly divergent Phlebobranchia-associated viruses from the same BioProject (PRJNA248170, *Ciona intestinalis*) fell between them (**Fig. 3b**). The *Ciona* libraries were derived from nervous tissue in animals undergoing regeneration (**Fig. S15**). Two novirhabdo-like contigs encoded seemingly complete genomes (13306nt) with six ORFs, while the third contig was partial (8764nt) (**Fig. 3b**).

The genomic architecture of the seemingly complete novirhabdo-like virus differed from that of canonical novirhabdoviruses. Specifically, the *Ciona*-associated virus (*Ciona intestinalis* novirhabdo-like virus 1) was ~200nt longer than other viruses in this genus, and the ORF encoding the glycoprotein appeared to be split with a frameshift (**Fig. 3c**). In addition, we were unable to identify the non-virion (NV) protein unique to the novirhabdoviruses using sequence-based similarity methods. It may be that the tunicate-associated NV is too divergent in sequence, and we hypothesised that it could be encoded between ORFs 3 and 4 (**Fig. 3c**), although it is also plausible that these viruses do not encode a NV protein at all. We therefore concluded that these ascidian-associated viruses were novirhabdo-like and represented divergent, invertebrate members of the genus.

In a similar fashion, we detected two fragments of a chu-like virus closely related to Tasmanian devil chu-like virus (TDCV)²⁷ in a library generated from *Ciona intestinalis* larvae undergoing tail regression (accession SRR18365523, BioProject PRJNA725676). We assumed that the ascidian was the host because no other plausible host was present in the library, which comprised >99% Phlebobranchia contigs. One fragment encoded the RNA-dependent RNA polymerase (RdRp) and was present at low abundance (0.003% of non-rRNA reads). The 5’

terminal of the ORF was incomplete, precluding us from determining whether this virus is segmented. Nevertheless, the RdRp sequence formed a sister clade with that of the TDCV (sh-aLRT/UFboot = 100/100), and this clade did not fall within any known *Jingchuvirales* lineages (**Fig. 3d, Fig. S16**) as previously reported²⁸. The other fragment encoded three ORFs (a metapneumo-like glycoprotein, a protein of unknown function, and a bornavirus-like nucleoprotein), which reflected the genome organisation of TDCV (**Fig. 3d, inset**) and was more abundant than the RdRp-encoding fragment (0.13% non-rRNA reads). The nucleoprotein and the glycoprotein sequences similarly formed clades with their TDCV counterparts (**Fig. S17**). Taken together, these evolutionary relationships, the shared genome organisation, and the phylogenetic distance from other vertebrate-infecting viruses in this order, suggested that there are multiple unrelated clades of chordate-infecting viruses in the *Jingchuvirales*.

Lastly, we identified a fragment (3481nt) of an alpha-like virus in a library generated from a hemolymph sample of an *Ascidia sydneiensis samea* (Phlebobranchia) at low abundance (0.002% of non-rRNA reads). Although the ORFs encoding non-structural and structural proteins were both partial, they exhibited well-defined alphavirus characteristics. The structural ORF encoded an alphavirus capsid protein (HHPred probability = 97.16%, e-value = 0.00027). The RdRp encoded by the non-structural ORF fell at the base of the canonical alphavirus clade with robust support (sh-aLRT/UFboot = 99.9/100) (**Fig. 3e, Fig. S18**). Of note, its relationship basal to the two known aquatic alphaviruses (southern elephant seal virus and salmon pancreas disease virus, **Fig. 3e, icons**) was consistent with virus-host co-divergence and an aquatic origin of the *Togaviridae*²⁹.

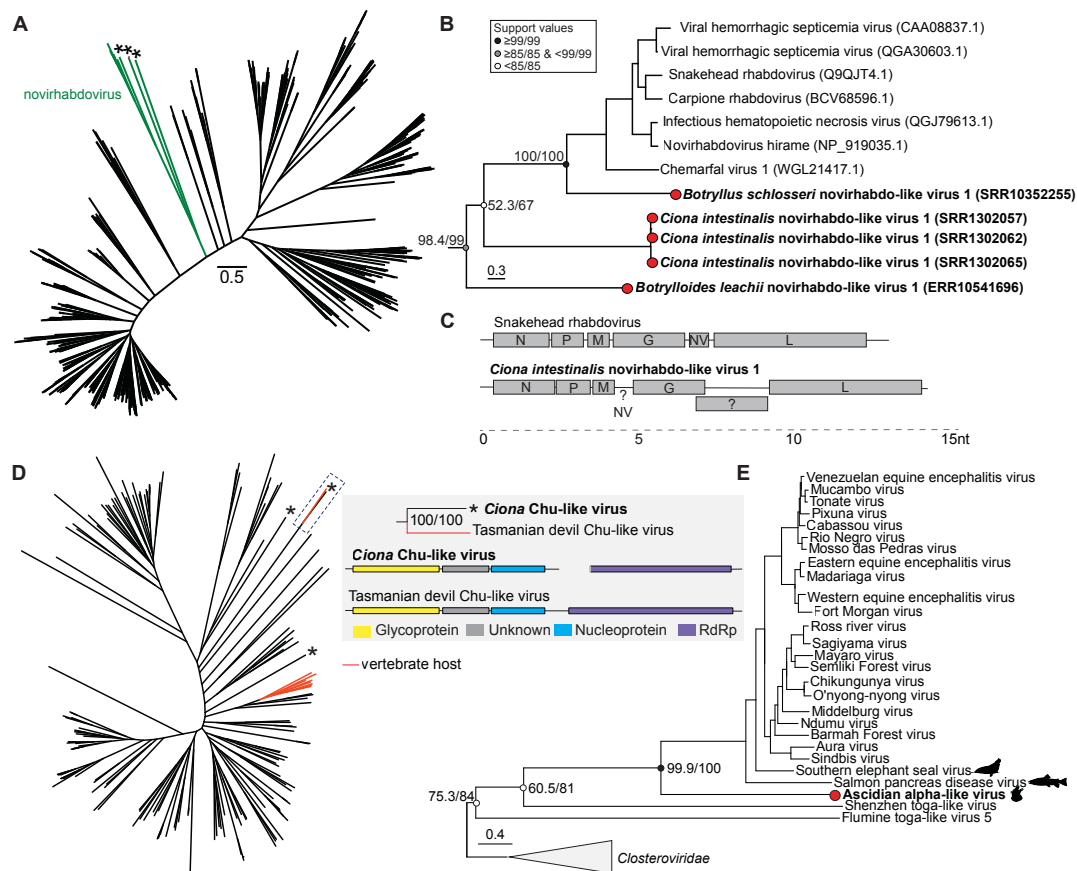


Figure 3. Evolutionary origin of the novirhabdoviruses, vertebrate-infecting *Jingchuvirales*, and the *Togaviridae* may have predated Vertebrata. For all phylogenies, branches are scaled according to the number of amino acid substitutions per site. (a) Phylogenetic diversity of the *Rhabdoviridae* with the novirhabdoviruses indicated by green branches. Stars denote viruses identified in this study. (b) The *Novirhabdovirus* clade from (a). Support values are shown as sh-aLRT/UFboot at select nodes. Red tip circles denote viruses identified in this study. (c) Genomic organisation of a canonical novirhabdovirus (Snakehead rhabdovirus) and a novirhabdo-like virus identified in this study (*Ciona intestinalis* novirhabdo-like virus 1). Genomes are scaled by nucleotides (nt). (d) Phylogenetic diversity of the order *Jingchuvirales* RdRp-encoding ORFs. Viruses that are associated with vertebrate hosts are indicated by red branches. Stars denote viruses identified in this study. Inset: Genomic organisation of *Ciona* Chu-like virus identified in this study and Tasmanian devil Chu-like virus. (e) Rooted phylogeny of the *Togaviridae* (order *Martellivirales*). Support values are shown as sh-aLRT/UFboot at select nodes. The red circle indicates the placement of the

ascidian-associated alpha-like virus identified in this study. The icons represent the putative host of select viruses (source: phylopic.org).

Recombinant history of ascidian-associated nidovirus glycoproteins

Thus far, we have examined the evolutionary history of RNA viral taxa using the RdRp. However, other viral proteins, such as those that facilitate virus entry into the host cell, can exhibit markedly different histories³⁰. Recombination involving glycoproteins has already been reported in tunicate-associated nidoviruses. Specifically, two of these viruses belong to the *Medioniviridae* (positive-sense RNA viruses; order *Nidovirales*) and encode Class II fusion proteins with apparent similarity to togavirus E1 glycoproteins (*Botrylloides leachii* nidovirus, MK956105 and Tognidovirus botryllus22085, BK066634¹⁹) – hence the name “Tognidoviruses”. This contrasts with the glycoproteins encoded by other chordate- and invertebrate-infecting families in the order: the *Coronaviridae*³¹, *Tobaniviridae*³², and *Mesoniviridae*³³ that utilise Class I fusion proteins, and the *Arteriviridae*, whose fusion protein has yet to be identified^{34,35}. Thus, recombination among diverse virus groups may have played a central role in shaping the evolution of the *Nidovirales* within chordates.

To investigate recombination in more detail, we first identified six additional tunicate-associated nido-like viruses. Two of these – an apparently complete 28kb putative viral contig in a *Botrylloides* ascidian and a partial RdRp fragment in a *Didemnum* ascidian – shared sequence similarity with members of the *Medioniviridae* (**Fig. 4a**). The remaining four were highly divergent and present in 53 libraries (51 *Phlebobranchia*, 2 *Aplousobranchia*) (**Supp. Data 2**). We recovered only fragments of two of the four viruses, but the other two appeared to be complete and were ~33kb and ~40kb in length. They did not encode a frameshift in their putative ORF1ab and expressed either one or two ORFs following the putative glycoprotein (**Fig. 4a**). We assigned names to these viruses according to their putative hosts (**Fig. 4a**) and will use them herein.

We next used protein structural prediction to identify and characterise the glycoproteins of these viruses and of members of nidovirus families for which the class of fusion protein was not known. Class II fusion proteins were a defining characteristic of the *Medioniviridae*, including the members we identified in this study (“*Botrylloides medionivirus 1*”, “*Didemnum medionivirus 1*”, **Fig 4b**). Of note, they shared nearly equal structural homology (defined by

FoldSeek probabilities) with the glycoproteins of flaviviruses (Spondweni virus and Bole Tick virus 4), togaviruses (chikungunya virus), and bunyaviruses (Severe Fever with Thrombocytopenia Syndrome Virus), suggesting that their glycoproteins share a deep ancestry with all Class II fusion proteins and are not necessarily “toga-like” (Fig 4b). The glycoproteins of the non-medioniviruses identified here (“*Ciona robusta* nido-like virus 1”, “*Ciona savignyi* nido-like virus 1 & 2”) were more difficult to classify, with sequence-based analyses returning ambiguous results (i.e., low probabilities and high e-values, Table 1). However, by structure-based analysis, *Ciona robusta* nido-like virus 1 and *Ciona savignyi* nido-like virus 2 glycoproteins were inferred to be Class II-like with medium to high probability (0.912 and 0.692, respectively, Fig 4b). The protein structure of the *Ciona savignyi* nido-like virus 1 putative glycoprotein could not be predicted confidently, suggesting it is highly divergent.

The glycoproteins of previously identified nidoviruses were also predominantly Class II-like. According to sequence-based analyses, the *Roniviridae* and *Euroniviridae* utilise fusion proteins with similarity to both alpha- and bunyavirus glycoproteins (Table S1). Unlike other members of the *Nidovirales*, the *Mononiviridae* encode a single polyprotein, and the location of the glycoprotein therein was previously unknown. We therefore searched for glycoprotein-like features in the Flatworm mononi-like virus polyprotein using protein structure prediction and found a fold (amino acid residues 9650-9800) that was consistent with the C-terminal portion of a Class II fusion protein (comprising Domain-III and the transmembrane region; Fig. S19). Similarly, divergent nidoviruses previously identified using artificial intelligence-based analyses³⁶ encode Class II-like fusion proteins, albeit with varying degrees of likelihood (Table S2, Fig. S20).

Table 1. HHPred results for glycoproteins encoded by divergent ascidian-associated nidoviruses

Virus name	HHPred reference name	E-value	Probability (%)
<i>Ciona savignyi</i> nido-like virus 1	Envelope gB; Varicella Zoster Virus	4.2	67.5
<i>Ciona savignyi</i> nido-like virus 2	Glycoprotein, zika virus	5.1	84.85
<i>Ciona robusta</i> nido-like virus 1	Protocadherin-15	3.3	81.37

Given the frequent occurrence of Class II fusion proteins among nidovirus families, we mapped fusion protein class onto the *Nidovirales* RdRp phylogeny to assess the frequency of inter-virus

377 recombination. Notably, fusion protein types fell on multiple branches on the RdRp
 378 phylogenies, indicative of recombinant evolutionary histories. The two tunicate medioniviruses
 379 identified here fell within the diversity of the Class II-encoding *Medioniviridae*, which forms
 380 a sister group with the Class I-encoding *Mesioniviridae* (**Fig. 4c**). The remaining four newly
 381 identified tunicate nido-like viruses did not fall within a known family, and the long branches
 382 characterising the latter combined with their range of genome size (between ~33kb and 40kb)
 383 suggest that each may represent novel virus families (**Fig. 4c**). Although there was an apparent
 384 division of vertebrate- and invertebrate-infecting lineages, this pattern again did not correspond
 385 to glycoprotein class. Thus, as with the flaviviruses³⁰, glycoprotein composition in the
 386 *Nidovirales* is not congruent with the evolutionary history of the RdRp, such that
 387 recombination likely played a substantial role in shaping this order.
 388

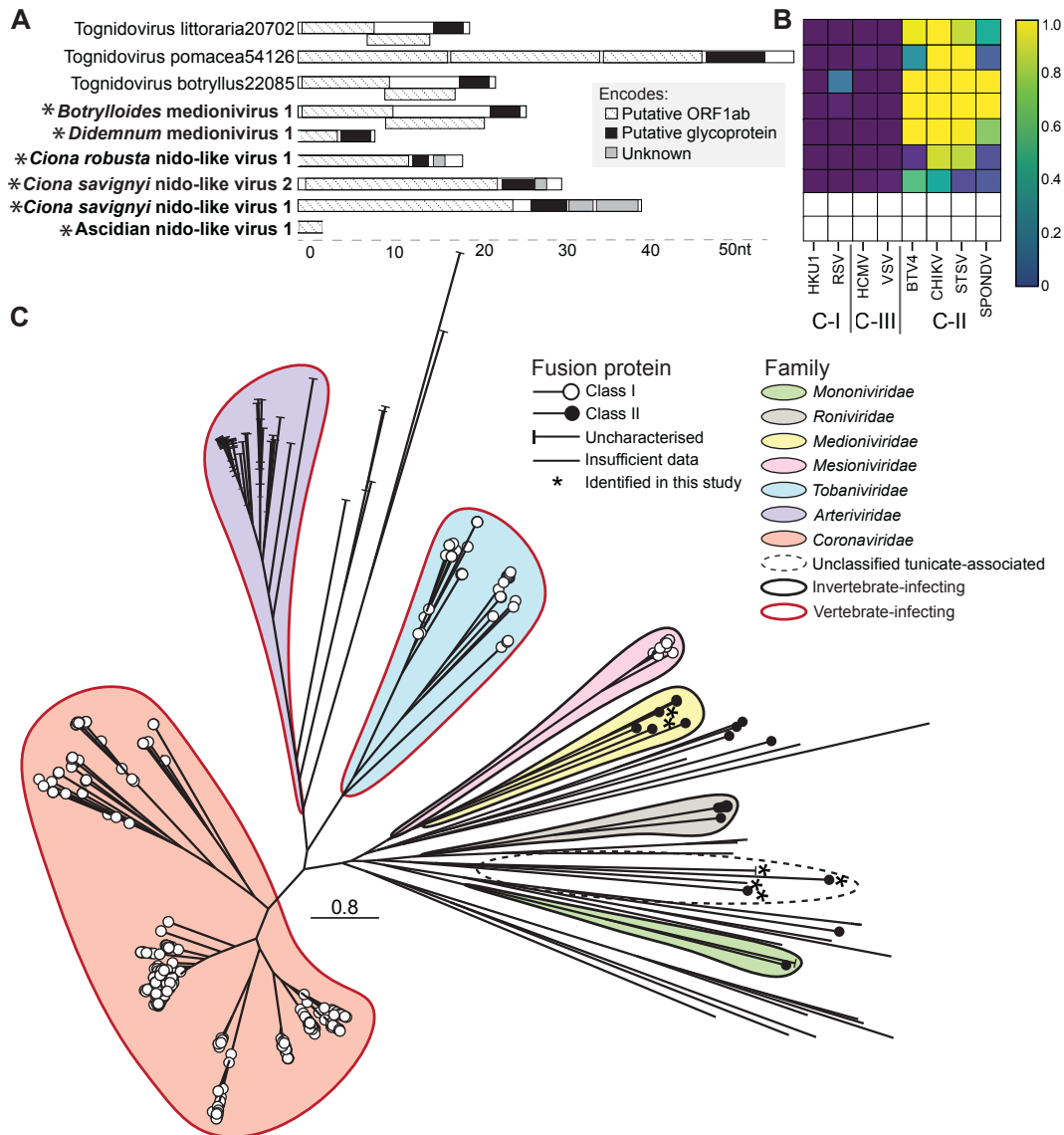


Figure 4. Utilisation of Class I and Class II fusion proteins across the *Nidovirales*. (a) Genomic architecture of tunicate-associated nido-like viruses and previously identified “Tognidoviruses”. Asterisks indicate viruses identified in this study. Genomes are scaled by genome size (nt). (b) Protein structure predictions for sequential overlapping sequence blocks, spanning entire viral polypeptides, were compared to class-I (C-I), III (C-III) and II (C-II) fusion protein reference structures using Foldseek³⁷. Heat map displays highest detected Foldseek probabilities for each nido- and nido-like virus against each fusion protein reference. White squares are shown where no probability value was returned by Foldseek. (c) Unrooted phylogeny of the *Nidovirales* RdRp. Branches are scaled by the number of amino acid substitutions per site.

Virus-host co-divergence and recombination in a divergent mononega-like virus

The examination of Class II fusion proteins in our tunicate virus data set revealed an additional example of inter-virus recombination of a glycoprotein in a highly divergent virus found in the *Styela rustica* hemolymph library (accession SRR16495746). The organisation of this putative virus resembled that of members of the *Rhabdoviridae*, a family of negative-sense RNA viruses (order *Mononegavirales*), encoding six complete ORFs, including a nucleoprotein, a glycoprotein, and an RdRp that utilises the GDD amino acid motif in the palm C catalytic triad (**Fig. 5a**). The first ORF was mononega-like, sharing similarity with *Borna-* and *Paramyxoviridae* nucleoproteins (**Table 3**). Although the length and position of ORFs 2 and 3 may correspond to rhabdo-like phosphoprotein (P) and matrix protein (M), respectively, they shared neither sequence nor structural homology with known proteins (**Fig. 5a**). We were also unable to determine the possible function of the sixth ORF (**Supp. Data 3**). The RdRp was highly divergent, sharing minimal sequence similarity (28.57%) with members of this order (**Table 4**). Phylogenetic analysis of the *Negarnaviricota* placed this virus sequence at the base of the vertebrate-infecting *Paramyxoviridae* and *Pneumoviridae*, suggesting that the ancestors of these families evolved among invertebrate, chordate hosts (**Fig. 5b**). We named this putative virus *Styela rustica* patchyvirus 1.

In a similar manner to the *Rhabdoviridae*, but unlike the *Paramyxoviridae* and *Pneumoviridae*, the genomic architecture of *Styela rustica* patchyvirus 1 suggested that it encoded a single glycoprotein ORF (**Fig. 5a**). Strikingly, this glycoprotein was not rhabdo-like (Class I), but instead shared low (25.93%) but detectable sequence similarity to the E1 glycoprotein of alphaviruses, a group of positive-sense RNA viruses (*Togaviridae*) (**Table 4**). When the phylogenetic relationship of all so-called “toga-like” glycoproteins was inferred relative to the canonical alphavirus glycoproteins, this sequence formed a sister group to the glycoprotein of Ascidian alpha-like virus described above, albeit with long branches (**Fig. 5c**). Due to the divergent nature of these sequences, we could not conclude that *Styela rustica* patchyvirus 1 acquired its Class II fusion protein from the *Togaviridae*; however, the shared sequence identity suggests that a recombination event between positive- and negative-sense RNA viruses may have occurred during the evolutionary history of this viral lineage. Thus, given its unique composition, *Styela rustica* patchyvirus 1 may be the first representative of a distinct family within the *Mononegavirales* that exemplifies a complex, recombinant evolutionary history.

434

435 **Table 3. HHPred results for the putative nucleoprotein encoded by *Styela rustica***
436 **patchyvirus 1**

HHPred reference name	E-value	Probability (%)
Borna disease virus nucleoprotein	6.3e-33	100
Newcastle disease virus nucleoprotein	2.7e-6	98.6
Parainfluenza virus 5 nucleocapsid	3.3e-6	98.58
Nipah virus nucleoprotein	4.4e-6	98.53

437

438 **Table 4. BLASTx nr results for *Styela rustica* patchyvirus 1 ORFs**

Reference name (accession)	Query cover (%)	E-value	Identity (%)
<i>RdRp (ORF4)</i>			
Beihai rhabdo-like virus 5 (YP_009333422.1)	20	4e-04	28.57
Hubei rhabdo-like virus 5 (YP_010084250.1)	26	0.001	22.26
<i>Artemisia capillaris</i> nucleorhabdovirus 1 (YP_010805763.1)	29	0.002	23.59
Jungle carpet python virus (YP_009508484.1)	20	0.002	26.92
<i>Glycoprotein (ORF3)</i>			
Western equine encephalitis virus (ABD57956.1)	49	3e-08	25.93
Highlands J virus (ANJ02316.1)	49	2e-06	25.83
Fort Morgan virus (YP_003324599.1)	59	3e-06	23.15
Eastern equine encephalitis virus (ADB08674.1)	55	6e-06	25.86

439

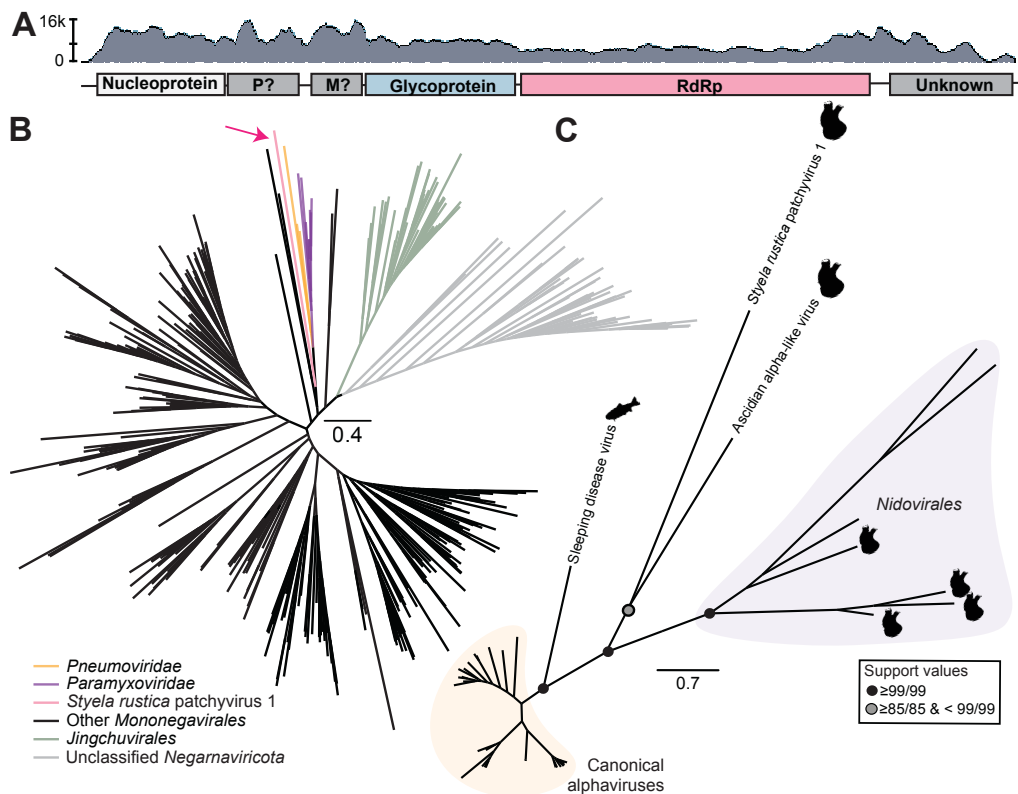


Figure 5. Recombination of Class II glycoproteins to a tunicate-associated mononega-like virus. (a) Schematic of divergent mononega-like viral genome sequence with sequencing coverage. Question marks indicate ORFs that shared no detectable sequence or structural similarity with known proteins but whose positions may be indicative of their function according to the genome organisation of the *Rhabdoviridae*. Abbreviations: P (phosphoprotein), M (matrix protein). (b) Unrooted maximum likelihood phylogeny of the *Negarnaviricota*. Branches are scaled by amino acid substitutions per site. The arrow indicates the placement of the *Styela rustica patchyvirus 1* identified in this study. Branches are scaled by amino acid substitutions per site. (c) Unrooted maximum likelihood phylogeny of alphavirus E1 glycoproteins and class II glycoproteins encoded by tunicate-associated RNA viruses. Branches are scaled by amino acid substitutions per site, and support values are shown at select nodes as sh-aLRT/UFboot. Animal icons indicate the known or associated host of each putative virus (source: phylopic.org).

DISCUSSION

Our analysis of tunicate metatranscriptomes sheds new light on the evolutionary history of animal RNA viruses. Specifically, we present evidence that five vertebrate-infecting viral lineages may have evolved prior to the origin of the Vertebrata. The phylogenetic placement of the RdRp of some tunicate-associated influenza-, novirhabdo-, alpha-, chu-, and mononega-like viruses was consistent with a process of virus-host co-divergence that commenced in basal chordates. Perhaps most notable among these were the influenza viruses, which we previously hypothesised first emerged in early vertebrates due to basal fish-associated lineages²¹. In contrast, other tunicate-associated viruses belonging to nine viral lineages (the *Amarillovirales*, *Birnaviridae*, *Bunyavirales*, *Hepelivirales*, *Nidovirales*, *Nodamuvirales*, *Picornavirales*, *Reovirales*, and *Stellavirales*) were invertebrate-like or fell with non-animal infecting viruses. Whether this pattern reflects the evolutionary histories of those orders or is the result of insufficient sampling requires further investigation.

In addition to enriching our understanding of the origins of some disease-causing viral lineages, vertebrate-like tunicate-associated viruses exhibited unique genetic features. *Ciona* Chu-like virus and its closest relative Tasmanian devil Chu-like virus encode an ORF of unknown function between their chu-like glyco- and nucleoproteins. Similarly, *Ciona intestinalis* novirhabdo-like virus 1 may not encode the auxiliary non-virion (NV) protein. It could be that this protein, which has been shown to interfere with the host immune response³⁸, was an innovation that evolved when the novirhabdoviruses emerged in vertebrates or that the tunicate-associated homolog is too divergent to detect with sequence similarity-based methods. This protein or its derivative could be encoded by the second glycoprotein-like domain immediately preceding the RdRp, congruent with the canonical novirhabdovirus organisation. Characterising these tunicate virus domains in cell culture would directly address the question of how the evolution of the vertebrate immune system may have influenced that of RNA viruses.

More striking was the example of *Styela rustica* patchyvirus 1, which comprises a negative-sense mononega-like nucleoprotein and RdRp, coupled with a Class II fusion protein that shares detectable sequence similarity to the E1 glycoprotein of a positive-sense alphavirus. The placement of the RdRp at the base of the *Paramyxoviridae* and *Pneumoviridae* is indicative of a viral lineage that extends back to the basal chordates. Both of those vertebrate-infecting families encode Class I fusion proteins, which suggests that the switch from Class II to Class I formed part of the transition from invertebrate chordate to vertebrate hosts. A similar pattern

of major changes of glycoprotein composition concordant with the use of vertebrate hosts has been observed in members of the *Flaviviridae*³⁰. However, the retention of sequence similarity to alphavirus E1 in the patchyvirus is puzzling, particularly because this degree of similarity was not retained by the putative glycoprotein encoded by Ascidian alpha-like virus. It is also intriguing because it suggests a recombination event between positive- and negative-sense viruses that may have occurred more recently than the transition to vertebrate hosts. Calibrating the timing and direction of this event will require additional sampling of the diversity of the patchyvirus lineage.

Despite these uncertainties, incongruence between the glycoprotein and RdRp phylogenies can be observed multiple times in tunicate-associated viruses, and we found evidence that this pattern is not specific to tunicate hosts. Tunicate-associated nidoviruses utilise an array of Class II-like fusion proteins that differ markedly from both each other and those used by other families in the order. Similarly, glycoproteins encoded by the non-chordate-infecting families the *Roniviridae*, *Euroniviridae*, and *Mononiviridae* share sequence and/or structural homology with Class II fusion proteins. The *Medioniviridae*, which comprise a combination of tunicate and non-chordate invertebrate hosts, also exhibit this feature. In contrast, Class I fusion proteins are only utilised by the *Coronaviridae*, *Tobaniviridae*, and *Mesioniviridae* within the *Nidovirales*, and the RdRps of these families are polyphyletic. Unlike in the *Flaviviridae*, fusion protein type does not map onto the host: not all vertebrate-infecting families utilise Class I (e.g., the *Arteriviridae*) and not all invertebrate-infecting families encode Class II (e.g., the *Mesioniviridae*). Thus, host type does not account for such flexibility, and additional sampling is required to elucidate the evolutionary drivers that underpinned these recombination events.

Genome size also appears to be fluid within the *Nidovirales*. The discovery of Nam Dinh virus (family *Mesioniviridae*) prompted the hypothesis that this family represented an “evolutionary intermediate” between the large nidoviruses (the *Coronaviridae* and *Roniviridae*) and the small nidoviruses (the *Arteriviridae*), bolstering the argument that the acquisition of a proof-reading mechanism enabled the size of the nidovirus genome to expand³³. The nido-like viruses discovered in this study and previously^{19,36} refute such a history. In particular, *Ciona savignyi* nido-like virus 1 (~40kb) and *Ciona savignyi* nido-like virus 2 (~30kb) formed a clade outside of any known family in the order. Not only is this relationship indicative of unrealised viral diversity, but it suggests that the genome size of nidoviruses has fluctuated, growing both larger and becoming smaller, at multiple points throughout the evolution of this order. Whether this

is due to the acquisition of proof-reading mechanisms, to patterns of horizontal gene transfer or gene loss, or to other key genetic features requires experimental validation. It also raises the question of whether other orders of RNA viruses are equally disposed to take up genetic material from cellular organisms in this way. A large flavi-like virus that we previously identified in a sea sponge³⁹, notable for both its length and apparent acquisition of bacterial genes, remains the sole example outside of the *Nidovirales*. Further exploration of the marine invertebrate virosphere is likely to reveal additional examples.

This study has some limitations. A primary challenge of metatranscriptomic analysis in the absence of experimental validation is the inability to directly determine the host associations of the newly identified viruses. We were instead reliant on indirect lines of evidence such as the library composition and the phylogenetic relationship of divergent viruses to one another to conclude that some of the viruses described in this study were likely infecting ascidians. Our data set was also affected by sampling biases, specifically the overrepresentation of *Ciona spp.* in the SRA. We were therefore unable to fully explore the viromes of the *Thaliacea* and *Appendicularia*, and a concerted effort to sample these organisms from their natural environments is needed to rectify these gaps. Additionally, although we could conclude that some viral lineages likely emerged before the origin of the vertebrates, we were unable to determine the timing of these events as we did not observe clustering by host order. For example, *Ciona intestinalis* novirhabdo-like virus 1 (host: *Phlebobranchia*) fell between two *Stolidobranchia*-associated viruses. A similar absence of clustering was observed for the influenza-like viruses, and the placement of the sturgeon- and carp-associated clade at the base of this group added another layer of virus-host incongruence. A further complication is a lack of consensus on the phylogenetic relationships within the Tunicata and even the *Ascidacea*^{40,41}, such that whether or not tunicate-associated viruses adhere to virus-host co-divergence within this subphylum cannot be assessed currently.

Despite this, our findings address and raise intriguing questions about the mechanisms and antiquity of animal RNA virus evolution. Virus-host co-divergence has, in multiple instances, been punctuated by the acquisition of genetic material through recombination across virus taxa that may have had a marked effect on virus evolution (e.g., *Styela rustica* patchyvirus 1). As this was the first systematic analysis of tunicate-associated RNA viruses, more systematic sampling of the Tunicata may find that the origins of additional disease-causing vertebrate viruses predated the Vertebrata. In particular, mapping the acquisition and loss of genetic

features to the transition from basal chordates to vertebrates in these lineages could reveal viral traits that are correlated with disease emergence in vertebrates. Not only are tunicates a rich source of RNA virus diversity, but their viruses stand to further reshape our understanding of animal RNA virus evolution.

METHODS

Sample collection

Primary samples were collected as previously described³⁹ from Chowder Bay, Sydney Harbour, Australia on November 24, 2021 (n = 12) and October 20, 2022 (n = 18). To minimise the risk of cross-contamination, divers wore latex gloves and cleaned forceps with 96% ethanol between sampling events. Samples were stored in RNA/DNA-free cryogenic tubes and placed immediately in liquid Nitrogen before being transferred to a -80C freezer for long-term storage.

Classification of primary samples

All ascidians were identified based on morphology. Further classification was performed in two ways. The composition of each library was first analysed using CCMetagen⁴², KMA⁴³, and the NCBI nt database (June 2019). In many cases, this proved inconclusive due to the complex nature of the libraries, which include the ascidian, its diet, and the invertebrate symbionts that can inhabit the organism (**Fig. S21**). Second, the rRNA in each library was assembled and screened against the NCBI nucleotide (nt) database using blast+ v.2.12.0⁴⁴ (**Supp. Data**).

RNA extractions and sequencing

To extract total RNA, samples were processed individually and homogenised by flash freezing with liquid nitrogen and grinding with a sterile mortar and pestle (one mortar/pestle pair per sample). Total RNA was extracted using the QIAGEN RNeasy Plus Mini kit, with the RNA from some libraries pooled for downstream processing. Samples from the first time point were sequenced individually, and samples from the second time point were combined into eleven pools. Library preparation was performed using the Ribo-Zero Plus library preparation kit, and all libraries were sequenced on the Illumina NovaSeq 6000 platform. One "blank" negative control that contained sterile water and reagent mix was included for each run.

Library processing and assembly

Nextera adapters were removed from raw reads using cutadapt v1.8.3⁴⁵ and the following parameters: minimum length = 25, quality-cutoff = 24, number of bases trimmed at the beginning and end of reads = 5. The quality of trimming was checked FastQC v0.11.8⁴⁶. In cases where removal was incomplete, the contigs were manually trimmed and reassembled if needed. Reads mapping to rRNA were removed using sortMeRNA v4.3.3⁴⁷. Contigs were assembled with MEGAHIT v1.2.9⁴⁸. A total of 3,468 libraries were successfully assembled (Supp. Data 4).

SRA screening

To capture all publicly available tunicate metatranscriptomic data, we queried the NCBI SRA database for all libraries sequenced from "Tunicata", "Urochordata", "Sea squirt", "Ascidiacea" as of March 2024, at least 0.5Gb in size (n = 3937). Ten libraries were removed because they were in fact associated with mollusc metatranscriptomes (1 Katharina tunicata, 1 Crobula tunicata, 8 Clione limacina).

Identifying putative viral contigs

To detect RNA virus-like sequences, we screened all assembled libraries against the protein version of the Reference Viral Database (RVDB, November 2023)^{49,50} and RdRp-Scan databases using DIAMOND BLASTx v2.0.9⁵¹. We specified an e-value of 1×10^{-5} and the parameter “-very-sensitive”. All contigs that returned hits to viruses were checked against the NCBI non-redundant (nr) protein database (as of September 2024) and contigs that shared detectable sequence similarity to known cellular genes were removed from the data set. Minimum length thresholds for contigs were assigned according to order (Table S2).

All contigs were checked for open reading frames (ORFs) using the standard genetic code (all virus contigs), invertebrate mitochondrial genetic code (*Cryptavirales*), or ciliate genetic code (*Zhaovirus*). Contigs that encoded truncated ORFs or did not encode ORFs were removed from the data set. The sequencing coverage of contigs with unique features such as frameshifts or instances of putative recombination was assessed using BMap v37.98⁵². Nucleotide sequences were translated using the Expasy web translate tool (<https://web.expasy.org/translate/>), EMBOSS getorf (<https://www.bioinformatics.nl/cgi-bin/emboss/getorf>), or InterProScan⁵³. Abundance was measured according to the expected count generated by RSEM v1.3.0⁵⁴.

PCR of ascidian-associated PB3 fragments

PCR was performed using the SuperScript IV One-Step RT-PCR kit following the manufacturer's instructions. An annealing temperature of 55C was used. Custom primers were designed (Table 5).

Table 5 Primers used for RT-PCR

Primer direction	Sequence
Forward	CGAACTGATTGAGGAAAGAATCGAG
Reverse	CACCTATGAAACAAAAACACTGTCT

Genome annotation and protein structure prediction

We used multiple sequence-based methods to identify functional domains in the genomes of putative viral contigs: NCBI Conserved Domains⁵⁵, InterProScan (CDD, NCBIfam, TMHMM)⁵³, HHPred⁵⁶, and Phyre2⁵⁷.

We used protein structure prediction to search for putative glycoproteins as described previously^{30,39}. Briefly, viral polyprotein sequences were split into sequential 300 residue blocks, overlapping by 100 residues. Structures were predicted using AlphaFold2-ColabFold and ESMFold⁵⁸⁻⁶⁰ and then compared to reference protein structures using Foldseek³⁷. The following reference structures were used: Class I – HKU1 Spike (PDB: 5I08), RSV F (6APB). Class III: HCMV Gb (7KDP), VSV G (5I2S); Class II – CHIKV E1 (6NK7), SPONDV E (6ZQI), SFTSV Gc (8ILQ) and Bole Tick Virus 4 E glycoprotein identified in a previous study^{30,61-67}. For each virus species, we identified structure blocks with the best Foldseek probabilities (highest values) and e-values (lowest values) against each of the reference structures.

Phylogenetic analysis

Preliminary invertebrate/vertebrate analyses

We first curated reference data sets for RNA virus orders with known vertebrate and invertebrate hosts using the NCBI Virus resource. For the *Bunyavirales*, *Martellivirales*, and *Picornavirales*, we selected representative members of each family within these orders due to the high levels of sequence divergence. For the other viral orders, we downloaded all publicly available RdRp amino acid sequences and removed redundant sequences at 85% identity with

CD-HIT v4.8.1⁶⁸. The remaining sequences were aligned with MAFFT v7.490⁶⁹, and sequences that did not align correctly were removed. Misaligned sequences were mostly misannotated genes (i.e., they were not RdRps). We aligned the *Bunyavirales*, *Martellivirales*, and *Nidovirales* using MAFFT and MUSCLE v5²³. MUSCLE performed better for the *Bunyavirales* (i.e., it did not incorrectly pull fragmented sequences to the centre of the alignment), and hence this alignment was used for further analysis. All other phylogenetic trees were inferred from sequence alignments generated with MAFFT v7.490. We removed ambiguities at specified thresholds for the *Martellivirales* (70%) and *Nidovirales* (75%) and with the parameter gappout in all other cases using trimAl v1.4.1⁷⁰. Phylogenies were inferred using the maximum likelihood (ML) method in IQ-TREE v1.6.12⁷¹, with the ModelFinder limited to LG models with “-mset”. Node support was measured with 1000 ultrafast bootstraps (UFboot) and sh-aLRT.

Family and lineage-specific analyses

For the phylogenetic analysis of influenza virus segments (i.e., PB2, PB3, NP, HA), representative sequences from the canonical influenza viruses A-D were included, along with amphibian and fish-associated influenza-like viruses. Representative thogotoviruses and quaranjaviruses were included as outgroups. Sequences were aligned with both MAFFT v7.490 and MUSCLE v5. No ambiguities were removed, but the ends of the alignments were trimmed for PB2 and NP.

In the case of the *Togaviridae*, we downloaded all polymerase sequences from this family that were at least 400 amino acid residues in length (n = 1,041) from NCBI. Redundant sequences were removed at 90% identity with CD-HIT, leaving 45 unique sequences in our data set. We then added the first 100 BLASTx hits for Ascidian alpha-like virus and removed redundancies at 90% identity with CD-HIT. For the *Jingchuvirales* nucleoprotein and glycoprotein phylogenies, we downloaded all corresponding sequences available on NCBI and removed redundant sequences at 99% using CD-HIT. Sequences from each data set were aligned with MAFFT v7.490 and the edges of the alignments were manually trimmed. Ambiguities were removed with trimAl and the parameter gappout, and ML phylogenies were then inferred with IQ-TREE v1.6.12 and the best-fitting model was chosen by ModelFinder. Similarly, we downloaded all alphavirus glycoproteins from NCBI (n = 2307) and removed redundancies at 85% with CD-HIT. The resulting phylogenetic trees were inferred as described for the *Jingchuvirales* glycoproteins. We compiled reference sequences from the *Negarnaviricota*

692 from NCBI, the LucaProt database³⁶, and studies identifying viruses in environmental
693 samples^{36,72-75}. Amino acid sequences were aligned with MAFFT v7.480 and ambiguities were
694 removed with trimAl v1.2rev59 (parameters: -gt 0.8 -st 0.005 -cons 30). The phylogeny was
695 inferred using the ML method available in PhyML v3.3.3⁷⁶, and the model was limited to LG.

696 **RESOURCE AVAILABILITY**

697 Nucleotide sequences, amino acid sequences, multiple sequence alignments, and tree files are
698 publicly available at https://github.com/mary-petrone/tunicate_viruses_2024.

699 **ACKNOWLEDGMENTS**

701 This work was funded by a National Health & Medical Research Council (NHMRC)
702 Investigator grant (GNT2017197) to E.C.H., an Australian Research Council (ARC)
703 Discovery Project grant (DP240101313) to E.C.H. and M.E.P., and by AIR@InnoHK
704 administered by the Innovation and Technology Commission, Hong Kong Special
705 Administrative Region, China to E.C.H. We thank David Grunde-McLaughlin for suggesting
706 the name “*Styela rustica* patchyvirus” and Chris Cooney and Alex McGrath for helping with
707 sample collection.

708 **AUTHOR CONTRIBUTIONS**

709 M.E.P., E.M.M., and E.C.H. conceived of the study. E.M.M. and O.T. collected the primary
710 samples. M.E.P., J.G., R.H.P., K.V.B., J.C.O.M., S.M., M.S., and O.T. performed the analyses.
711 M.E.P. and E.C.H. drafted the manuscript. All authors reviewed and edited the manuscript.
712

REFERENCES

- 1 Licandro, P. & Brunetta, M. in *Marine Plankton: A practical guide to ecology, methodology, and taxonomy* (eds Claudia Castellani & Martin Edwards) (Oxford University Press, 2017).
- 2 Flood, P. R., Deibel, D. & Morris, C. C. Filtration of colloidal melanin from sea water by planktonic tunicates. *Nature* **355**, 630-632 (1992). <https://doi.org/10.1038/355630a0>
- 3 Kovalevskij, A. O. Entwicklungsgeschichte der einfachen Ascidien. (1866).
- 4 Satoh, N. The ascidian tadpole larva: comparative molecular development and genomics. *Nature Reviews Genetics* **4**, 285-295 (2003). <https://doi.org/10.1038/nrg1042>
- 5 Winkley, K. M., Kourakis, M. J., DeTomaso, A. W., Veeman, M. T. & Smith, W. C. in *Current Topics in Developmental Biology* Vol. 136 (ed Lilianna Solnica-Krezel) 219-242 (Academic Press, 2020).
- 6 Cracraft, J. & Donoghue, M. J. *Assembling the tree of life*. (Oxford University Press, 2004).
- 7 Winchell, C. J., Sullivan, J., Cameron, C. B., Swalla, B. J. & Mallatt, J. Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new LSU and SSU ribosomal DNA data. *Molecular Biology and Evolution* **19**, 762-776 (2002).
- 8 Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965-968 (2006). <https://doi.org/10.1038/nature04336>
- 9 Philippe, H., Lartillot, N. & Brinkmann, H. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular biology and evolution* **22**, 1246-1253 (2005).
- 10 Blair, J. E. & Hedges, S. B. Molecular phylogeny and divergence times of deuterostome animals. *Molecular biology and evolution* **22**, 2275-2284 (2005).
- 11 Creasy, A., Rosario, K., Leigh, B. A., Dishaw, L. J. & Breitbart, M. Unprecedented diversity of ssDNA phages from the family Microviridae detected within the gut of a protochordate model organism (*Ciona robusta*). *Viruses* **10**, 404 (2018).
- 12 Leigh, B. A., Djurhuus, A., Breitbart, M. & Dishaw, L. J. The gut virome of the protochordate model organism, *Ciona intestinalis* subtype A. *Virus research* **244**, 137-146 (2018).
- 13 Leigh, B., Karrer, C., Cannon, J. P., Breitbart, M. & Dishaw, L. J. Isolation and characterization of a *Shewanella* phage-host system from the gut of the tunicate, *Ciona intestinalis*. *Viruses* **9**, 60 (2017).
- 14 Jung, S.-J., Oh, M.-J., Date, T. & Suzuki, S. Isolation of marine birnavirus from sea squirts *Halocynthia roretzi*. *The biology of ascidians*, 436-441 (2001).
- 15 Zhang, C. & Suzuki, S. Aquabirnaviruses isolated from marine organisms form a distinct genogroup from other aquabirnaviruses. *Journal of Fish Diseases* **27**, 633-643 (2004).
- 16 Kitamura, S. I. *et al.* Tunic morphology and viral surveillance in diseased Korean ascidians: soft tunic syndrome in the edible ascidian, *Halocynthia roretzi* (Drasche), in aquaculture. *Journal of fish diseases* **33**, 153-160 (2010).
- 17 Kumagai, A. *et al.* Soft tunic syndrome in the edible ascidian *Halocynthia roretzi* is caused by a kinetoplastid protist. *Diseases of aquatic organisms* **95**, 153-161 (2011).
- 18 Kumagai, A. *et al.* Mass mortality of cultured ascidians *Halocynthia roretzi* associated with softening of the tunic and flagellate-like cells. *Diseases of aquatic organisms* **90**, 223-234 (2010).
- 19 Buck, C. B. *et al.* Widespread Horizontal Gene Transfer Among Animal Viruses. *eLife* (2024). <https://doi.org/10.7554/elife.97647.1>

- 762 20 Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* **540**, 539-543
763 (2016). <https://doi.org/10.1038/nature20167>
- 764 21 Petrone, M. E. *et al.* Evidence for an ancient aquatic origin of the RNA viral order
765 Articulavirales. *Proc Natl Acad Sci U S A* **120**, e2310529120 (2023).
766 <https://doi.org/10.1073/pnas.2310529120>
- 767 22 Mifsud, J. C. O. *et al.* Transcriptome mining extends the host range of the Flaviviridae
768 to non-bilaterians. *Virus Evolution* **9** (2022). <https://doi.org/10.1093/ve/veac124>
- 769 23 Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased
770 assessments of sequence homology and phylogeny. *Nat Commun* **13**, 6968 (2022).
771 <https://doi.org/10.1038/s41467-022-34630-w>
- 772 24 Chen, W. *et al.* A novel influenza A virus mitochondrial protein that induces cell death.
773 *Nat Med* **7**, 1306-1312 (2001). <https://doi.org/10.1038/nm1201-1306>
- 774 25 ICTV. in *Virus Taxonomy* (eds Andrew M. Q. King, Michael J. Adams, Eric B.
775 Carstens, & Elliot J. Lefkowitz) Ch. Family - Rhabdoviridae, 686-713 (Elsevier, 2012).
- 776 26 Peter J. Walker, J. F.-A., Nicolas Bejerman, Kim R. Blasdel, Ralf G. Dietzgen,
777 Anthony R. Fooks, Hideki Kondo, Gael Kurath, Pedro Luis Ramos-González, Mang
778 Shi, Robert B. Tesh, Noël Tordo, Nikos Vasilakis and Anna E. Whitfield. *Family:*
779 *Rhabdoviridae*, <<https://ictv.global/report/chapter/rhabdoviridae/rhabdoviridae>>
780 (2018).
- 781 27 Mélade, J. *et al.* Isolation of an infectious mammalian chu-like virus from tumor cells
782 of the endangered Tasmanian devil (*Sarcophilus harrisii*). *bioRxiv*,
783 2024.2011.2025.625296 (2024). <https://doi.org/10.1101/2024.11.25.625296>
- 784 28 Harvey, E., Mifsud, J. C. O., Holmes, E. C. & Mahar, J. E. Divergent hepaciviruses,
785 delta-like viruses, and a chu-like virus in Australian marsupial carnivores (dasyurids).
786 *Virus Evolution* **9** (2023). <https://doi.org/10.1093/ve/vead061>
- 787 29 Forrester, N. L. *et al.* Genome-scale phylogeny of the alphavirus genus suggests a
788 marine origin. *J Virol* **86**, 2729-2738 (2012). <https://doi.org/10.1128/jvi.05591-11>
- 789 30 Mifsud, J. C. O. *et al.* Mapping glycoprotein structure reveals Flaviviridae evolutionary
790 history. *Nature* **633**, 695-703 (2024). <https://doi.org/10.1038/s41586-024-07899-8>
- 791 31 Huang, Y., Yang, C., Xu, X.-f., Xu, W. & Liu, S.-w. Structural and functional
792 properties of SARS-CoV-2 spike protein: potential antiviral drug development for
793 COVID-19. *Acta Pharmacologica Sinica* **41**, 1141-1149 (2020).
794 <https://doi.org/10.1038/s41401-020-0485-4>
- 795 32 Ujike, M. & Taguchi, F. Recent Progress in Torovirus Molecular Biology. *Viruses* **13**
796 (2021). <https://doi.org/10.3390/v13030435>
- 797 33 Nga, P. T. *et al.* Discovery of the first insect nidovirus, a missing evolutionary link in
798 the emergence of the largest RNA virus genomes. *PLoS Pathog* **7**, e1002215 (2011).
799 <https://doi.org/10.1371/journal.ppat.1002215>
- 800 34 Firth, A. E. *et al.* Discovery of a small arterivirus gene that overlaps the GP5 coding
801 sequence and is important for virus production. *J Gen Virol* **92**, 1097-1106 (2011).
802 <https://doi.org/10.1099/vir.0.029264-0>
- 803 35 Johnson, C. R., Griggs, T. F., Gnanandarajah, J. & Murtaugh, M. P. Novel structural
804 protein in porcine reproductive and respiratory syndrome virus encoded by an
805 alternative ORF5 present in all arteriviruses. *J Gen Virol* **92**, 1107-1116 (2011).
806 <https://doi.org/10.1099/vir.0.030213-0>
- 807 36 Hou, X. *et al.* Using artificial intelligence to document the hidden RNA virosphere.
808 *Cell* (2024). <https://doi.org/10.1016/j.cell.2024.09.027>
- 809 37 van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature*
810 *Biotechnology* **42**, 243-246 (2024). <https://doi.org/10.1038/s41587-023-01773-0>

- 811 38 Ammayappan, A. & Vakharia, V. N. Nonvirion protein of novirhabdovirus suppresses
812 apoptosis at the early stage of virus infection. *J Virol* **85**, 8393-8402 (2011).
813 <https://doi.org/10.1128/jvi.00597-11>
- 814 39 Petrone, M. E. *et al.* A ~40-kb flavi-like virus does not encode a known error-
815 correcting mechanism. *Proc Natl Acad Sci U S A* **121**, e2403805121 (2024).
816 <https://doi.org/10.1073/pnas.2403805121>
- 817 40 Nanglu, K., Lerosey-Aubril, R., Weaver, J. C. & Ortega-Hernández, J. A mid-
818 Cambrian tunicate and the deep origin of the ascidiacean body plan. *Nature*
819 *Communications* **14**, 3832 (2023). <https://doi.org/10.1038/s41467-023-39012-4>
- 820 41 Braun, K., Leubner, F. & Stach, T. Phylogenetic analysis of phenotypic characters of
821 Tunicata supports basal Appendicularia and monophyletic Ascidiacea. *Cladistics* **36**,
822 259-300 (2020). <https://doi.org/10.1111/cla.12405>
- 823 42 Marcelino, V. R. *et al.* CCMetagen: comprehensive and accurate identification of
824 eukaryotes and prokaryotes in metagenomic data. *Genome Biol* **21**, 103 (2020).
825 <https://doi.org/10.1186/s13059-020-02014-2>
- 826 43 Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw
827 reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).
828 <https://doi.org/10.1186/s12859-018-2336-6>
- 829 44 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**,
830 421 (2009). <https://doi.org/10.1186/1471-2105-10-421>
- 831 45 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing
832 reads. *2011* **17**, 3 (2011). <https://doi.org/10.14806/ej.17.1.200>
- 833 46 Andrews, S. *FastQC*, <<https://github.com/s-andrews/FastQC>> (2023).
- 834 47 Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of
835 ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211-3217 (2012).
836 <https://doi.org/10.1093/bioinformatics/bts611>
- 837 48 Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast
838 single-node solution for large and complex metagenomics assembly via succinct de
839 Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015).
840 <https://doi.org/10.1093/bioinformatics/btv033>
- 841 49 Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M. & Khan, A. S. A Reference
842 Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput
843 Sequencing for Novel Virus Detection. *mSphere* **3** (2018).
844 <https://doi.org/10.1128/mSphereDirect.00069-18>
- 845 50 Bigot, T., Temmam, S., Pérot, P. & Eloit, M. RVDB-prot, a reference viral protein
846 database and its HMM profiles. *F1000Research* **8**, 530 (2020).
847 [https://doi.org:https://doi.org/10.12688/f1000research.18776.2](https://doi.org/https://doi.org/10.12688/f1000research.18776.2)
- 848 51 Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life
849 scale using DIAMOND. *Nature Methods* **18**, 366-368 (2021).
850 <https://doi.org/10.1038/s41592-021-01101-x>
- 851 52 Bushnell, B., Rood, J. & Singer, E. BBMerge - Accurate paired shotgun read merging
852 via overlap. *PLoS One* **12**, e0185056 (2017).
853 <https://doi.org/10.1371/journal.pone.0185056>
- 854 53 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
855 *Bioinformatics* **30**, 1236-1240 (2014). <https://doi.org/10.1093/bioinformatics/btu031>
- 856 54 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data
857 with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
858 <https://doi.org/10.1186/1471-2105-12-323>
- 859 55 Wang, J. *et al.* The conserved domain database in 2023. *Nucleic Acids Res* **51**, D384-
860 d388 (2023). <https://doi.org/10.1093/nar/gkac1096>

861 56 Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with
862 a New HHpred Server at its Core. *Journal of Molecular Biology* **430**, 2237-2243
863 (2018). <https://doi.org/10.1016/j.jmb.2017.12.007>
864 57 Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2
865 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858
866 (2015). <https://doi.org/10.1038/nprot.2015.053>
867 58 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods*
868 **19**, 679-682 (2022). <https://doi.org/10.1038/s41592-022-01488-1>
869 59 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*
870 **596**, 583-589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>
871 60 Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a
872 language model. *Science* **379**, 1123-1130 (2023).
873 <https://doi.org/10.1126/science.adc2574>
874 61 Kirchdoerfer, R. N. *et al.* Pre-fusion structure of a human coronavirus spike protein.
875 *Nature* **531**, 118-121 (2016). <https://doi.org/10.1038/nature17200>
876 62 Goodwin, E. *et al.* Infants Infected with Respiratory Syncytial Virus Generate Potent
877 Neutralizing Antibodies that Lack Somatic Hypermutation. *Immunity* **48**, 339-349.e335
878 (2018). <https://doi.org/10.1016/j.immuni.2018.01.005>
879 63 Liu, Y. *et al.* Prefusion structure of human cytomegalovirus glycoprotein B and
880 structural basis for membrane fusion. *Sci Adv* **7** (2021).
881 <https://doi.org/10.1126/sciadv.abf3178>
882 64 Roche, S., Rey, F. A., Gaudin, Y. & Bressanelli, S. Structure of the prefusion form of
883 the vesicular stomatitis virus glycoprotein G. *Science* **315**, 843-848 (2007).
884 <https://doi.org/10.1126/science.1135710>
885 65 Basore, K. *et al.* Cryo-EM Structure of Chikungunya Virus in Complex with the Mxra8
886 Receptor. *Cell* **177**, 1725-1737.e1716 (2019). <https://doi.org/10.1016/j.cell.2019.04.006>
887 66 Renner, M. *et al.* Flavivirus maturation leads to the formation of an occupied lipid
888 pocket in the surface glycoproteins. *Nat Commun* **12**, 1238 (2021).
889 <https://doi.org/10.1038/s41467-021-21505-9>
890 67 Du, S. *et al.* Cryo-EM structure of severe fever with thrombocytopenia syndrome virus.
891 *Nat Commun* **14**, 6333 (2023). <https://doi.org/10.1038/s41467-023-41804-7>
892 68 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-
893 generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
894 <https://doi.org/10.1093/bioinformatics/bts565>
895 69 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
896 improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
897 <https://doi.org/10.1093/molbev/mst010>
898 70 Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for
899 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**,
900 1972-1973 (2009). <https://doi.org/10.1093/bioinformatics/btp348>
901 71 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
902 Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).
903 <https://doi.org/10.1093/molbev/msaa015>
904 72 Chen, Y. M. *et al.* RNA viromes from terrestrial sites across China expand
905 environmental viral diversity. *Nat Microbiol* **7**, 1312-1323 (2022).
906 <https://doi.org/10.1038/s41564-022-01180-2>
907 73 Neri, U. *et al.* Expansion of the global RNA virome reveals diverse clades of
908 bacteriophages. *Cell* **185**, 4023-4037.e4018 (2022).
909 <https://doi.org/10.1016/j.cell.2022.08.023>

- 910 74 Wolf, Y. I. *et al.* Doubling of the known set of RNA viruses by metagenomic analysis
911 of an aquatic virome. *Nat Microbiol* **5**, 1262-1270 (2020).
912 <https://doi.org/10.1038/s41564-020-0755-4>
- 913 75 Zayed, A. A. *et al.* Cryptic and abundant marine viruses at the evolutionary origins of
914 Earth's RNA virome. *Science* **376**, 156-162 (2022).
915 <https://doi.org/doi:10.1126/science.abm5847>
- 916 76 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood
917 phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321 (2010).
918 <https://doi.org/10.1093/sysbio/syq010>
919