Virome composition in marine fish revealed by meta-transcriptomics

Jemma L. Geoghegan[1,2,3], Francesca Di Giallonardo[4], Michelle Wille[5], Ayda Susana Ortiz-Baez[6], Vincenzo A. Costa[2], Timothy Ghaly[2], Jonathon C. O. Mifsud[2], Olivia M. H. Turnbull[2], David R. Bellwood[7], Jane E. Williamson[2], Edward C. Holmes[6]

[1]Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand.

[2]Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia.

[3]Institute of Environmental Science and Research, Wellington 5018, New Zealand.

[4]The Kirby Institute, University of New South Wales, Sydney, NSW 2052, Australia.

[5]WHO Collaborating Centre for Reference and Research on Influenza, The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia.

[6]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.

[7]ARC Centre of Excellence for Coral Reef Studies and College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia.

Author for correspondence:
Jemma L. Geoghegan
jemma.geoghegan@otago.ac.nz

Keywords: fish, virome, virus evolution, metagenomics, host-jumping

Conflict of interest: the authors declare no conflict of interest.

Abstract

Revealing the determinants of virome composition is central to placing disease emergence in a broader evolutionary context. Fish are the most species-rich group of vertebrates and so provide an ideal model system to study the factors that shape virome compositions and their evolution. We characterised the viromes of 19 wild-caught species of marine fish using total RNA sequencing (meta-transcriptomics) combined with analyses of sequence and protein structural homology to identify divergent viruses that often evade characterisation. From this, we identified 25 new vertebrate-associated viruses and a further 22 viruses likely associated with fish diet or their microbiomes. The vertebrate-associated viruses identified here included the first fish virus in the *Matonaviridae* (single-strand, negative-sense RNA virus). Other viruses fell within the *Astroviridae*, *Picornaviridae*, *Arenaviridae*, *Reoviridae*, *Hepadnaviridae*, *Paramyxoviridae*, *Rhabdoviridae*, *Hantaviridae*, *Filoviridae* and *Flaviviridae* and were sometimes phylogenetically distinct from known fish viruses. We also show how key metrics of virome composition – viral richness, abundance and diversity – can be analysed along with host ecological and biological factors as a means to understand virus ecology. Accordingly, these data suggest that that the vertebrate-associated viromes of the fish sampled here are predominantly shaped by the phylogenetic history (i.e. taxonomic order) of their hosts, along with several biological factors including water temperature, habitat depth, community diversity and swimming behaviour. No such correlations were found for viruses associated with porifera, molluscs, arthropods, fungi and algae, that are unlikely to replicate in fish hosts. Overall, these data indicate that fish harbour particularly large and complex viromes and the vast majority of fish viromes are undescribed.

Introduction

Metagenomic next-generation sequencing (mNGS) has led to a revolution in virus discovery (Shi et al 2018b, Zhang et al 2018a, Zhang et al 2018b), exposing more of the diversity, scale and structure of the virosphere. However, while it is now possible to reveal host viromes en masse (Chang et al 2019, Vibin et al 2018, Geoghegan et al 2018b, Paez-Espino et al 2016, Lim et al (2015), Porter et al 2019, Roux et al 2017, Shi et al 2016, Temmam et al 2016, Tirosh et al 2018, Pettersson et al 2019), we still have an incomplete understanding of the factors that structure viromes. Until recently, studies of virus evolution were largely limited to single viruses and/or single hosts, restricting our ability to explore the diverse host and environmental factors that might structure viromes as a whole. Fortunately, this is changing with the advent of mNGS, particularly total RNA sequencing. In particular, metagenomic-based studies have shown that aspects of host biology can greatly impact virus diversification (Wille et al 2019, Wille 2020) and as such may also be key drivers of virus emergence. As a simple case in point, the behavioural ecology of host species directly affects contact rates among individuals in a population, and more frequent intra- and inter-species contacts are likely to increase the potential for viral transmission.

The marine environment is a rich source of viruses. For example, the bacteriophage in aquatic ecosystems greatly outnumber other life-forms (Maranger and Bird 1995). There is an estimated concentration of 10 billion virus particles per litre of surface water (Bergh et al 1989, Breitbart and Rohwer 2005, Middelboe and Brussaard 2017, Suttle 2005), although abundance levels vary with such factors as ocean depth (De Corte et al 2012, Lara et al 2017), temperature (Coutinho et al 2017), latitude (Gregory et al 2019) and phytoplankton bloom development (Alarcon-Schumacher et al 2019). In marked contrast to bacteriophage, little is known about the factors that contribute to virus diversity in aquatic vertebrate populations, even though viruses can cause large-scale disease outbreaks in farmed fish (Crane and Hyatt 2011, Jarungsriapisit et al 2020, Whittington and Reddacliff 1995).

Fish provide an ideal model to better understand the diversity of viruses that exist in nature as well as the range of host and environmental factors that shape virome composition and abundance. Fish are the most species-rich group of vertebrates with over 33,000 species described to date (fishbase.org), the vast majority of which (~85%) are bony fish (the Osteichthyes) (Betancur-R et al 2017). Bony fish themselves are an extremely diverse and abundant group comprising 45 taxonomic orders, exhibiting a wide range of biological features that likely play an important role in shaping the diversity of their viromes. Initial studies indicate that fish harbour a remarkable diversity of viruses, particularly those with

RNA genomes, that may exceed that seen in any other class of vertebrate (Geoghegan et al 2018a, Lauber et al 2017, Shi et al 2018a). In addition, those viruses present in fish often appear to be the evolutionary predecessors of viruses infecting other vertebrate hosts, generally indicative of a pattern of virus-host associations that can date back hundreds of millions of years, although with frequent cross-species transmission. Despite the apparent diversity and ubiquity of fish viruses, they are severely under-studied compared to mammalian and avian viruses and there is little data on the factors that determine the structure of fish viromes.

To reveal more of the unexplored aquatic virosphere we sampled wild-caught ray-finned marine fish spanning 23 species across nine taxonomic orders and quantified a variety of host characteristics that together may impact virome composition, abundance and evolution. Specifically, we utilised meta-transcriptomics together with both sequence and protein structural homology searches of known viruses to: (i) reveal the total virome composition of fish, (ii) describe the phylogenetic relationships of the novel viruses obtained, (iii) determine whether, on these data, there may be associations between virome composition, abundance, richness and diversity and particular host traits, and (iv) explore whether taxonomically-related fish hosts have more similar viromes. The host characteristics initially considered here were: fish taxonomic order, swimming behaviour (i.e. solitary or schooling fish), preferred climate, mean preferred water temperature, host community diversity (i.e. multi- or single- species community), average body length, maximum life span, trophic level, and habitat depth (SI Table 1).

Methods

Ethics. Biosafety was approved by Macquarie University, Australia (ref: 5201700856). This study involved dead fish purchased from a fish market for which no animal ethics approval was required. The pygmy goby was collected under GBRMPA permit G16/37684.1 and JCU Animal Ethics Committee #A2530.

Fish sample collection. Dead fish from 23 species were sampled for virome analysis (SI Table 1). These included 18 new species collected from a fish market in Sydney, Australia, together with four species from our previous sampling of the same fish market (Geoghegan et al 2018a). These animals were caught by commercial fisheries in coastal waters in New South Wales, Australia by several different suppliers in Autumn 2018. By way of contrast, an additional species, the pygmy goby (*Eviota zebrina*), was obtained from the coral reefs of

118 tropical northern Queensland at approximately the same time. Fish were snap frozen at -
119 20°C immediately upon capture. Fish obtained from the market were purchased on the day
120 of catch. Tissues were dissected and stored in RNALater before being transferred to a -
121 80°C freezer. To increase the likelihood of virus discovery during metagenomic sequencing,
122 10 individuals from each species were pooled.

123 **Transcriptome sequencing***.* mNGS was performed on fish tissue (liver and gill). Frozen
124 tissue was partially thawed and submerged in lysis buffer containing 1% ß-
125 mercaptoethanol and 0.5% Reagent DX before tissues were homogenized together with
126 TissueRupture (Qiagen). The homogenate was centrifuged to remove any potential tissue
127 residues, and RNA from the clear supernatant was extracted using the Qiagen RNeasy Plus
128 Mini Kit. RNA was quantified using NanoDrop (ThermoFisher) and tissues from each
129 species were pooled to 3μg per pool (250ng per individual). Libraries were constructed
130 using the TruSeq Total RNA Library Preparation Protocol (Illumina) and host ribosomal RNA
131 (rRNA) was depleted using the Ribo-Zero-Gold Kit (Illumina) to facilitate virus discovery.
132 Paired-end (100bp) sequencing of the RNA library was performed on the HiSeq 2500
133 platform (Illumina). All library preparation and sequencing were carried out by the Australian
134 Genome Research Facility (AGRF).

135 **Transcript sequence similarity searching for viral discovery***.* Sequencing reads were
136 first quality trimmed then assembled *de novo* using Trinity RNA-Seq (Haas et al 2013). The
137 assembled contigs were annotated based on similarity searches against the NCBI
138 nucleotide (nt) and non-redundant protein (nr) databases using BLASTn and Diamond
139 (BLASTX) (Buchfink et al 2015), and an e-value threshold of $1x10^{-5}$ was used as a cut-off to
140 identify positive matches. We removed non-viral hits including host contigs with similarity to
141 viral sequences (e.g. endogenous viral elements). To reduce the risk of incorrect
142 assignment of viruses to a given library due to index-hoping, those viruses with a read
143 count less than 0.1% of the highest count for that virus among the other libraries was
144 assumed to be contamination.

145 **Protein structure similarity searching for viral discovery**. To identify highly divergent
146 viral transcripts, particularly those that might be refractory to detection using similarity
147 searching methods such as the BLAST approach described above, we employed a protein
148 structure-based similarity search for 'orphan' contigs that did not share sequence similarity
149 with known sequences. Accordingly, assembled orphan contigs were translated into open
150 reading frames (ORFs) using EMBOSS getorf program (Rice et al 2000). ORFs were
151 arbitrarily defined as regions between two stop codons with a minimum size of 200 amino

152 acids in length. To reduce redundancy, amino acid sequences were grouped based on
153 sequence identity using the CD-HIT package v4.6.5 (Li and Godzik 2006). The resulting
154 data set was then submitted to Phyre2, which uses advanced remote homology detection
155 methods to build 3D protein models, predict ligand binding sites, and analyse the effect of
156 amino acid variants (Kelley et al 2015). Virus sequences with predicted structures were
157 selected on the basis of having confidence values ≥90%. Following structure prediction, we
158 used the associated annotations for preliminary taxonomic classification. To avoid false
159 positives due to the limited number of available structures in the Protein Data Bank (PDB)
160 for template modelling, the taxonomic assignment was cross-validated with the results
161 from the Diamond (BLASTX) similarity search. Subsequently, putative viruses were aligned
162 with reference viral protein sequences at the immediate higher taxonomic level (e.g. genus,
163 family), using MAFFT v7.4 (E-INS-i algorithm) (Katoh and Standley 2013). Finally, we verified
164 the similarity among sequences by careful visual inspection of the most highly conserved
165 motifs of target proteins.

166 **Inferring the evolutionary history of fish viruses***.* We inferred the evolutionary
167 relationships of the viruses contained in the fish samples and compared them with known
168 viruses to determine those that were likely associated with vertebrate or non-vertebrate
169 hosts. Specifically, we assumed that viruses that grouped with other vertebrate viruses in
170 phylogenetic trees were likely to infect the fish sampled here, while those virus that were
171 more closely related to those usually associated with other host types (such as
172 invertebrates, fungi and plants) were unlikely to infect and replicate in fish hosts. To achieve
173 this, the translated viral contigs were combined with representative protein sequences
174 within each virus family obtained from NCBI RefSeq. The sequences retrieved were then
175 aligned with those generated here again using MAFFT v7.4 (E-INS-i algorithm) as described
176 above. Ambiguously aligned regions were removed using trimAl v.1.2 (Capella-Gutierrez et
177 al 2009). To estimate phylogenetic trees, we selected the optimal model of amino acid
178 substitution identified using the Bayesian Information Criterion as implemented in
179 Modelgenerator v0.85 (Keane et al 2006) and analysed the data using the maximum
180 likelihood approach available in IQ-TREE (Nguyen et al 2015) with 1000 bootstrap
181 replicates. Phylogenetic trees were annotated with FigTree v.1.4.2. Viruses newly identified
182 here were named reflecting the host common name.

183 **Revealing virome abundance and diversity**. Transcriptomes were quantified using RNA-
184 Seq by Expectation-Maximization (RSEM) as implemented within Trinity (Li and Dewey
185 2011). We first estimated the relative abundance of a host reference gene, ribosomal
186 protein S13 (RPS13), to assess the sequencing depth across libraries. Next, we used RSEM

to estimate the relative abundance of each virus transcript in these data.

For those viruses most likely associated with fish themselves, rather than components of their diet or microbiome (see Results), we performed analyses of virome abundance and diversity using R v3.4.0 integrated into RStudio v1.0.143 and plotted using ggplot2. Both the observed virome richness and Shannon effective (i.e. alpha diversity) were calculated for each library at the virus family level using modified Rhea script sets (Lagkouvardos et al 2017, Wille et al 2019). We used generalized linear models (GLM) to initially evaluate the effect of host taxonomic order, swimming behaviour (solitary or schooling fish), preferred climate, mean preferred water temperature, host community diversity, average species length, trophic level and habitat depth on viral abundance and alpha diversity (see SI Table 1 for all variables). Models were χ^2 tested (LRT) to assess model significance. When the number of factor levels in an explanatory variable exceeded two, we conducted Tukey posthoc testing (glht) using the *multcomp* package (Hothorn et al 2008). Beta diversity (i.e. the diversity between samples) was calculated using the Bray Curtis dissimilarity matrix. Effects of variables on viral community composition were evaluated using permanova (Adonis Tests) and Mantel tests with 10,000 permutations using the *vegan* package (Oksanen 2007).

To establish connectivity (i.e. sharing) among virus families that were likely associated with non-fish hosts, we generated a cord diagram by quantifying the number of fish species harbouring each virus family identified in this study. Virus families that occur in the same fish species were represented by ribbons or links in the diagram.

Results

We used mNGS to characterise viral transcripts from 23 marine fish spanning nine taxonomic orders: 19 species from this current study together with four from our previous work (Geoghegan et al 2018a). We combined data from our previous fish sampling to expand our data set and to apply novel viral protein structural searching methods not used previously. For these reasons, individual viruses discovered in our previous study are not detailed here. Combined, the extracted total RNA was organised into 23 libraries for high-throughput RNA sequencing. Ribosomal RNA-depleted libraries resulted in a median of 45,690,996 (range 33,344,520 – 51,071,142) reads per pool.

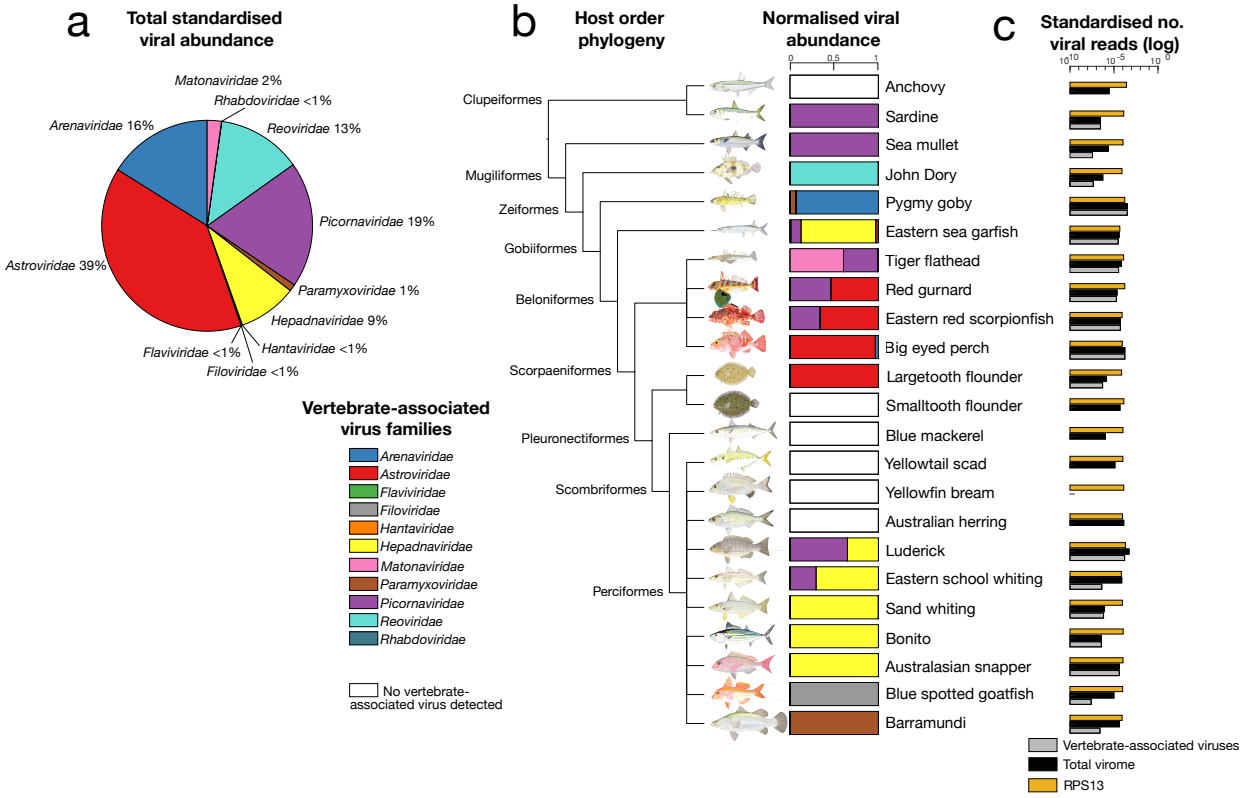


Figure 1. (A) Total standardized abundance of vertebrate-associated viruses (at the level of virus family) across the fish species examined. (B) Normalised viral abundance set out on a backbone of the fish host phylogeny at the order level. (C) Standardised number of total viral reads (black), vertebrate-associated viral reads (grey) and host reference gene ribosomal protein S13 (RPS13) (orange) in each species library.

Diversity and abundance of viruses in fish. The fish viromes characterised here contained viruses that were associated with vertebrate hosts as well as those that were more likely associated with porifera, invertebrates, fungi and algae (Figure 1). We primarily focused on the former since we assumed that the vertebrate-associated viruses were directly infecting the fish sampled, rather than being associated with the aquatic environment, diet or a co-infecting parasite, and hence are more informative in determining how host factors shape virus ecology and evolution.

Overall, we identified virus transcripts likely associated with vertebrate hosts that could be assigned to 11 viral families and present in a variety of fish species (SI Figure 1a). With the exception of the *Hepadnaviridae*, all were RNA viruses. Across all the fish sampled, those viral families found at relatively high abundances included the *Astroviridae* (representing 39% of all viruses discovered), *Picornaviridae* (19%), *Arenaviridae* (16%), *Reoviridae* (13%) and the *Hepadnaviridae* (9%) (Figure 1a). Other viral families found at lower relative

239 abundances were the *Matonaviridae* (previously the *Togaviridae*) (2%), *Paramyxoviridae*
240 (1%), as well as the *Rhabdoviridae*, *Hantaviridae*, *Filoviridae* and *Flaviviridae* (all <1%)
241 (Figure 1a). The most common vertebrate-associated viruses found in these fish were
242 picornaviruses (eight species), astroviruses (seven species) and hepadnaviruses (six
243 species) (Figure 1b). The eastern sea garfish (*Hyporhamphus australis*) harboured the most
244 diverse virome with four distinct vertebrate-associated viruses (Figure 1b). Six fish
245 contained no vertebrate-associated viruses, and we found no viral sequences in the
246 yellowfin bream (*Acanthopagrus australis*) (Figure 1c). An equivalent analysis of a host
247 reference gene, ribosomal protein S13 (RPS13) that is stably expressed in fish, revealed
248 similar abundances across species (0.004% – 0.02%), implying similar sequencing depth
249 across libraries (Figure 1c). RPS13 was, on average, ~55% more abundant than the total
250 virome.

251 We also examined viruses that were phylogenetically related to those associated with
252 porifera, molluscs, arthropods, fungi and algae, and hence were unlikely to infect the fish
253 themselves. Accordingly, we identified an additional 22 viruses across 11 virus families (SI
254 Figure 1b). These viruses were found in the *Chuviridae, Hepeviridae, Narnaviridae,*
255 *Nodaviridae, Partitiviridae, Picornaviridae, Solemoviridae, Tombusviridae, Totiviridae,*
256 *Dicistroviridae* and *Iflaviridae*, and are described in more detail below.

257 **Evolutionary relationships of fish viruses**. To infer stable phylogenetic relationships
258 among the viruses sampled and to identify those that are novel, where possible we utilised
259 the most conserved (i.e. polymerase) viral regions that comprise the RNA-dependent RNA
260 polymerase (RdRp) or the polymerase (P) ORF in the case of the hepadnaviruses. From this,
261 we identified 25 distinct and potentially novel vertebrate-associated virus species, in
262 addition to the eight novel viruses described previously (Geoghegan et al 2018a) (SI Table
263 2). All novel vertebrate-associated viruses shared sequence similarity to other known fish
264 viruses with the exception of those viruses found in the *Matonaviridae* and *Rhabdoviridae*,
265 the latter of which was found using structure similarity methods (Figure 2, SI Table 3; see
266 below). We found a further 22 viruses that clustered with viruses found in porifera,
267 molluscs, arthropods, fungi and algae (SI Figure 2, SI Figure 3, SI Figure 4).

268 Among the viruses identified was tiger flathead matonavirus (in *Neoplatycephalus*
269 *richardsoni*) – the first fish virus found in the *Matonaviridae*. This novel viral sequence
270 shared only 35% amino acid similarity with its closest relative - Guangdong Chinese water
271 snake rubivirus (Shi et al 2018a). Until recently, the only other representative of this family
272 was the distantly related human rubella virus, although additional members of this family

have recently been identified in other mammalian species (Bennett et al. 2020). Given the high levels of genetic divergence in this family, it is likely that these fish-associated viruses at least constitute a discrete and novel genus.

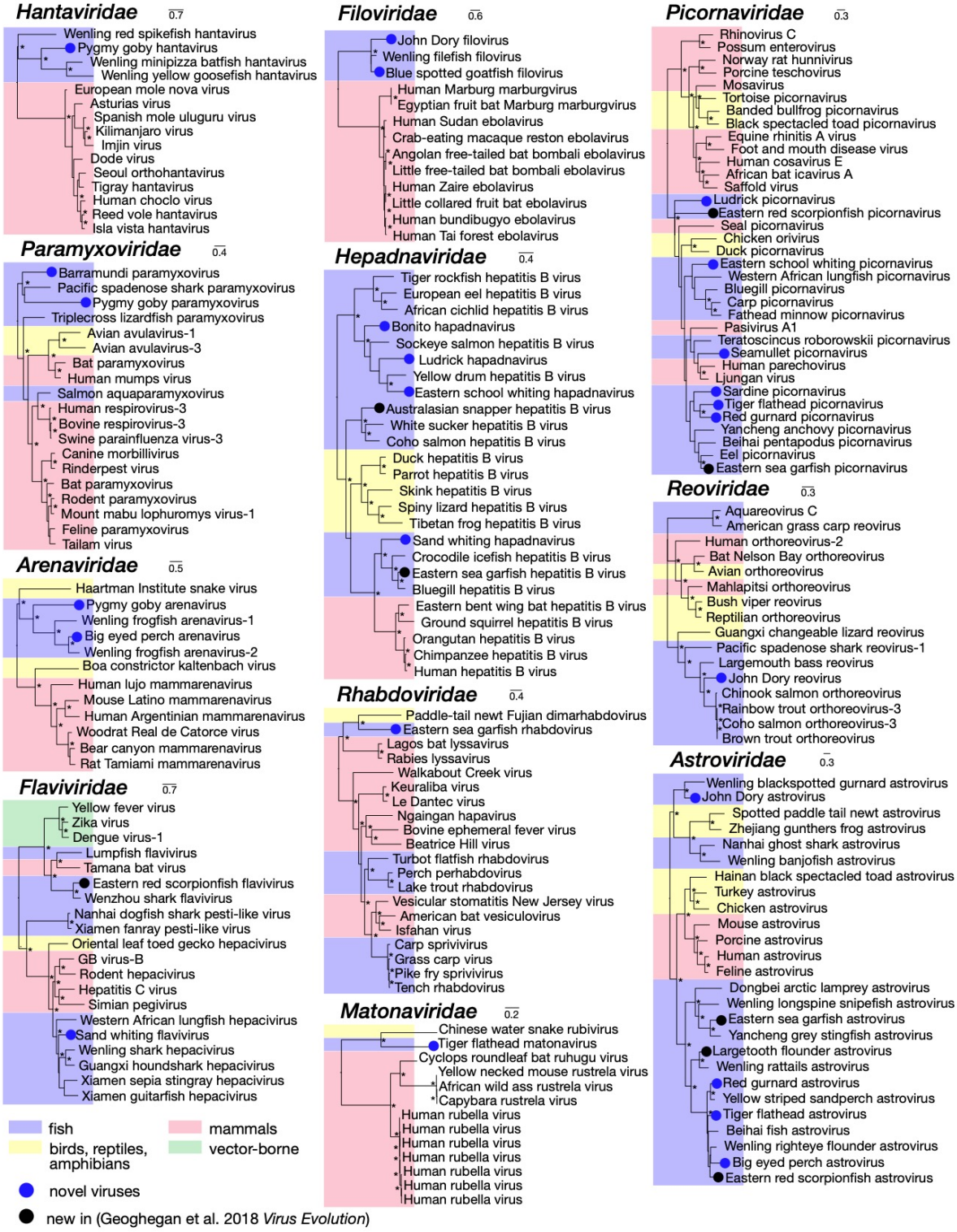


Figure 2. Phylogenetic relationships of likely vertebrate-associated viruses identified here. The maximum likelihood phylogenetic trees show the topological position of the newly

279 discovered viruses (blue circles) and those identified in an earlier study (Geoghegan et al.
280 2018), in the context of their closest phylogenetic relatives. Branches are highlighted to
281 represent host class (fish = blue; mammals = red; birds, reptiles and amphibians = yellow;
282 vector-borne (mammals and arthropods) = green). All branches are scaled according to the
283 number of amino acid substitutions per site and trees were mid-point rooted for clarity only.
284 An asterisk indicates node support of >70% bootstrap support. See SI Table 3 for all
285 accession numbers.

286

287 Another divergent virus discovered in this analysis is eastern sea garfish rhabdovirus (in
288 *Hyporhamphus australis*), which was most closely related to Fujian dimarhabdovirus
289 sampled from an amphibian host, sharing 45% amino acid RdRp sequence identity.
290 Notably, this highly divergent virus was only identified by using protein structure homology,
291 and forms a clade that is distinct from other fish rhabdoviruses (Figure 2). We also identified
292 two novel viral sequences in the *Filoviridae* in John Dory (*Zeus faber*) and the blue spotted
293 goatfish (*Upeneichthys lineatus*). These viruses shared sequence similarity to the only other
294 known fish filovirus, Wenling filefish filovirus (Shi et al 2018a). With the exception of these
295 fish viruses, all other known filoviruses including Ebola and Marburg viruses, are found in
296 mammalian hosts, notably humans, bats and primates.

297 We also found numerous viruses that cluster within established clades of fish viruses. For
298 example, pygmy goby hantavirus (in *Eviota zebrina*) grouped with other hantaviruses
299 recently found in fish (Figure 2). Although they were previously only thought to infect
300 mammals, hantaviruses have now been found to infect amphibians, jawless fish and ray-
301 finned fish (Shi et al 2018a). The evolutionary history of the *Paramyxoviridae* shows two
302 distinct fish virus lineages, of which both barramundi and pygmy goby paramyxoviruses
303 grouped with Pacific spade-nose shark paramyxovirus and shared 50% and 45% amino
304 acid L gene sequence similarity, respectively. This group of fish viruses is phylogenetically
305 distinct from other paramyxoviruses. We also found novel fish viruses in the *Flaviviridae*,
306 *Arenaviridae* and *Reoviridae*: although these grouped with other fish viruses, they greatly
307 expand the known diversity of these virus families. Finally, as noted above, the most
308 abundant viruses fell within the *Picornaviridae* and *Astroviridae*, and all shared sequence
309 similarity to other fish viruses. Notably, both picornaviruses and astroviruses are single-
310 stranded positive-sense RNA viruses that possess small icosahedral capsids with no
311 external envelope, which may aid their preservation in harsh marine environments.

312 The only DNA viruses we identified were novel hepadnaviruses. Those found in bonito

313 (*Sarda australis*), ludrick (*Girella tricuspidata*) and eastern school whiting (*Sillago flindersi*),
314 fell into the divergent group of hepadna-like viruses, the nackednaviruses, that have been
315 identified in a number of fish species (Lauber et al. 2017). In contrast, sand whiting
316 hepadnavirus (in *Sillago ciliate*) fell into the fish virus clade that is more closely related to
317 mammalian hepatitis B viruses (Dill et al 2016) (Figure 2).

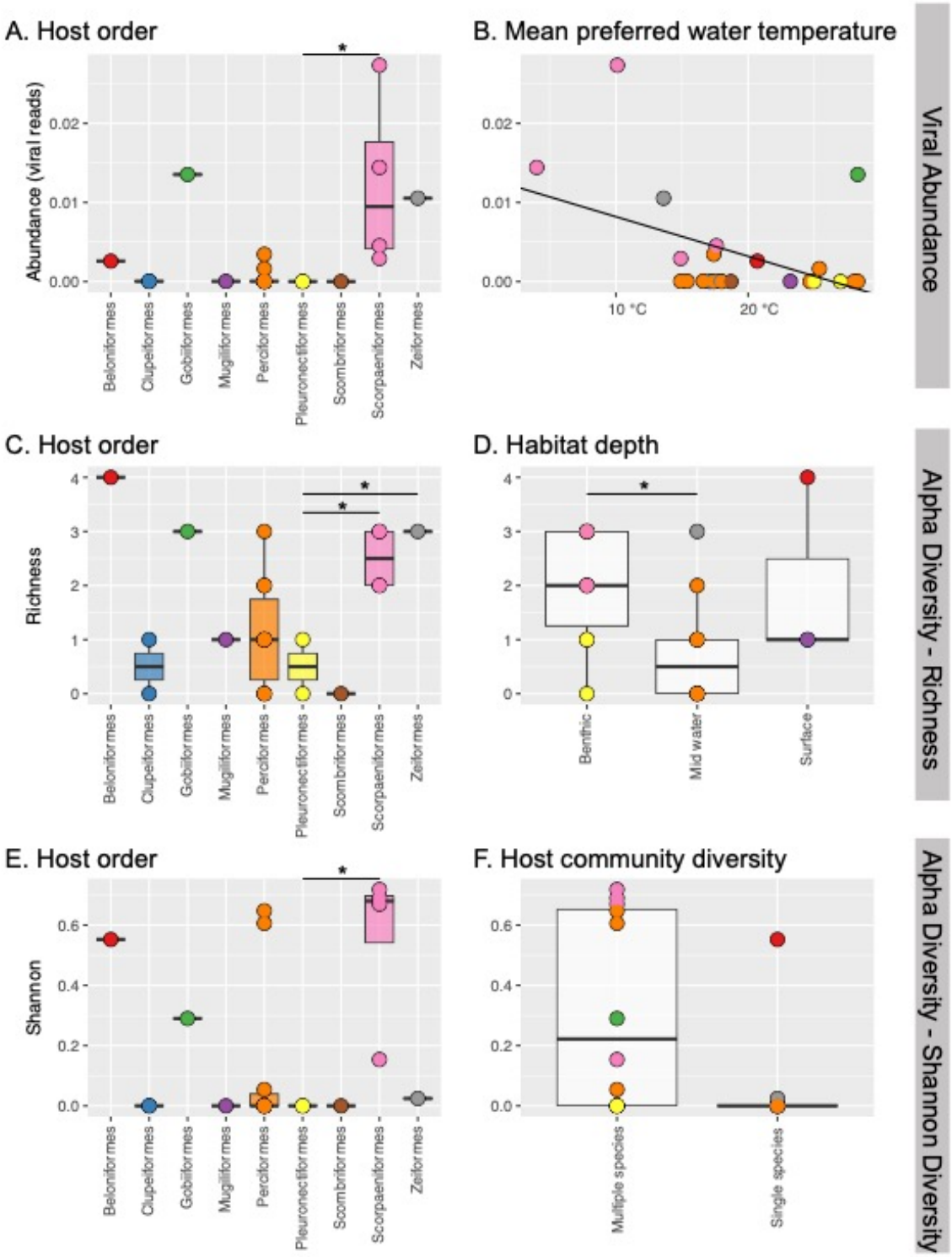
318 As expected, many of the viruses identified here were associated with marine hosts
319 belonging to invertebrates (including porifera, molluscs and arthropods; n = 20), fungi (n =
320 1) and algae (n = 1) as determined by their phylogenetic position and sequence similarity to
321 viruses previously described in these taxa (SI Figure 2, SI Figure 3, SI Figure 4). This implies
322 that these viruses more likely originated from host species that are associated with fish diet,
323 fish microbiomes or the surrounding environment, rather than from the fish themselves.
324 None of these viruses are highly divergent from other known viruses, but do help fill gaps in
325 the phylogenetic diversity of these groups.

326 **Assessing the impact of host biology on virome composition**. Our relatively small
327 sample of 23 fish species precluded us from performing a detailed statistical analysis of the
328 relationship between host traits and virome composition. Rather, we provide an initial
329 analysis that should be regarded as a framework for understanding how key host variables
330 might impact viral ecology and evolution, and that can be extended as more species are
331 analysed.

332 To this end we examined the possible association between eight host traits and viral
333 abundance (the proportion of viral reads in each sample), alpha diversity (the diversity
334 within each sample, measured by observed richness and Shannon diversity) and beta
335 diversity (the diversity between samples). The host traits initially considered here were: host
336 taxonomic order, swimming behaviour (solitary or schooling fish), preferred climate, mean
337 preferred water temperature, community diversity, average species length, maximum life
338 span, trophic level and habitat depth.

339 We first focused on the vertebrate-associated virome. This initial analysis revealed that the
340 phylogenetic relationships of the fish studied, as reflected in their taxonomic order,
341 seemingly had the strongest association with the overall composition of fish viromes. This
342 pattern was consistent when assessing viral abundance, alpha diversity and beta diversity
343 (Figure 3). That is, fish order (χ^2=0.003, df=8, p=0.0049) and mean preferred water
344 temperature (χ^2=0.008, df=1, p=0.035) were important predictors of viral abundance, such
345 that Scopaeniformes (i.e. bigeye ocean perch, red gurnard, tiger flathead, and eastern red
346 scorpionfish) had significantly higher viral abundance than Pleuronectiformes (i.e. largetooth

347 and smalltooth flounder) (Tukey: z=3.766, p=0.00479), while viral abundance had a negative
348 relationship to mean preferred water temperature (Figure 3). It is worth noting, however,
349 that virus abundance within the Scopaeniformes were widely distributed and that their
350 overall high abundance might only be due to a few species or individuals.



351 **Figure 3**. Significant explanatory variables in generalized linear models (GLM) for viral
352 abundance and two measures of alpha diversity. Viral abundance is best explained by (A)

fish host order and (B) mean preferred water temperature. Alpha diversity is best explained by (C) host order and (D) preferred habitat (Observed Richness) and by (E) host order and (F) host community diversity (Shannon Diversity). Stars indicate significant differences between groups determined by posthoc Tukey tests. Points represent different fish species and are coloured by host order.

We applied two measures of alpha diversity to our sample set: observed richness, a count of the number of viral families, and Shannon diversity, which also incorporates abundance. Observed richness was best explained by fish order (χ^2=22.839, df=8, p=3.8^{-6}) and habitat depth (χ^2=3.914, df=2, p=0.032), while Shannon diversity was best explained by fish order (χ^2=0.96, df=8, p=0.016) and community diversity (χ^2=0.41, df=1, p=0.05), with a larger Shannon diversity in multispecies communities compared with single species communities. As with viral abundance, there was a significant difference in alpha diversity between Scopaeniformes compared to Pleuronectiformes (Tukey Richness z=3.039, p=0.0495; Tukey Shannon z=2.845, p=0.05). Notably, in these data mid-water fish had decreased viral richness compared to benthic fish (Tukey z=-2.452, p=0.0338), and fish that reside in multispecies communities had a larger Shannon diversity compared to single species communities (χ^2=0.17089, df=1, p=0.05) (Figure 3). Our analysis also revealed that fish order (R^2=0.57215, p=0.003), swimming behaviour (R^2=0.09904, p=0.005), climate (R^2=0.13315, p=0.012) and mean preferred water temperature (R^2=0.1005, p=0.05) were significant predictors of beta diversity.

Importantly, we repeated the above analysis on the factors associated with virome composition on those viruses (n = 22) that likely infected hosts other than fish. Because we can assume that these viruses do not replicate in fish (for example, because they are related to host diet), and hence should be not shaped by aspects of fish biology and ecology, this analysis effectively constitutes an internal negative control. Indeed, this analysis revealed no association between virome composition and host ecological traits (viral abundance: p=0.0; alpha diversity: p=0.3; Shannon diversity: p=0.9; and beta diversity: p=0.3), thereby adding weight to the biological associations described above in the fish viruses.

Discussion

The metagenomic revolution is enabling us to uncover more of a largely unknown

386 virosphere. Here, we utilised mNGS to identify new viruses associated with fish,
387 characterising the viromes of 23 species of marine fish that spanned nine taxonomic orders
388 and identifying 47 novel viruses spanning 22 different virus families. This included 25 new
389 vertebrate-associated viruses and a further 22 viruses associated with protozoans, plants,
390 arthropods and fungi. Interestingly, the novel viruses included the first fish virus in the
391 *Matonaviridae* that are the closest phylogenetic relatives of the mammalian rubella viruses.
392 We also used these data to provide an initial assessment of how aspects of host biology
393 might impact virus diversity and evolution. Although our study was limited to 23 fish
394 species, on these data we found that host phylogeny (taxonomy) was strongly associated
395 with the composition of fish viromes. We also identified several other host traits that were
396 also associated with virus abundance and/or diversity, particularly preferred mean water
397 temperature, climate, habitat depth, community diversity and whether fish swim in schools
398 or are solitary. That these traits were not correlated with the composition of diet and
399 microbiome-associated viruses that do not actively replicate in fish suggests that the
400 patterns observed in marine fish are real, although it will clearly be important to test these
401 initial conclusions using larger numbers of fish species sampled from a diverse set of
402 environments.

403 Many of the viruses identified in this study were phylogenetically related to other, recently
404 discovered, viruses of fish (Dill et al 2016, Geoghegan et al 2018a, Lauber et al 2017, Shi et
405 al 2018a). However, there were some notable exceptions. Tiger flathead matonavirus
406 represents the only fish viral species in the *Matonaviridae* and forms a distinct clade with a
407 rubivirus discovered in a Chinese water snake. The discovery of this phylogenetically
408 distinct fish virus tentatively suggests the possibility of a fish host origin for this family,
409 although it is clear that confirmation will require the sampling of a far wider set of hosts.
410 Indeed, it is notable that additional rubella-like viruses have recently been identified in a
411 range of mammalian hosts, including bats (Bennett et al. 2020). A fish origin might also be
412 the case for other virus families such as the *Hantaviridae* and *Filoviridae*, as the fish viruses
413 in these families often fall basal to viruses in other vertebrate hosts such as birds and
414 mammals (also see Shi et al 2018a). In contrast, in some other virus families such as the
415 *Astroviridae, Picornaviridae, Flaviviridae* and *Rhabdoviridae*, viruses associated with fish are
416 distributed throughout the phylogeny suggestive of a past history of common host-jumping.
417 Regardless, available data suggests that fish viruses harbour more phylogenetic diversity
418 than the better studied mammalian and avian viruses within these families. It is also clear
419 that the discovery of novel viruses in fish has expanded our knowledge of the diversity,
420 evolutionary history and host range of RNA viruses in general.

421 Although there is often a clear phylogenetic division between those viruses likely to infect
422 fish and those associated with diet or microbiome, in some cases this separation can be
423 nuanced. For instance, although totiviruses were thought to only infect unicellular fungi,
424 their known host range has now expanded to include arthropods and fish (Mikalsen et al.
425 2016, Mor et al. 2016, Lovoll et al. 2010). In particular, piscine myocarditis virus is a totivirus
426 shown by *in situ* hybridisation to infect Atlantic salmon and is associated with
427 cardiomyopathy syndrome in salmon (Haugland et al. 2011). Similarly, viruses within the
428 *Narnaviridae* are widespread in fungi, and have now been extended to include both
429 invertebrates (Shi et al. 2016) and protist (Charon et al. 2019). Due to their phylogenetic
430 position, we assume the narna-like viruses identified here are associated with fungal
431 parasites in these samples.

432 As well as identifying new viruses, we sought to provisionally identify associations between
433 host traits and the overall composition of fish viruses, although this analysis was clearly
434 limited by the available sample size. A notable observation was that fish virome
435 composition, reflected in measures of viral richness, abundance and diversity, is most
436 impacted by the phylogenetic relationships (i.e. taxonomy) of the host in question. This in
437 turn suggests that fish viruses might have co-diverged with fish hosts over evolutionary
438 time-scales, a pattern supported by the general relationship between vertebrate host class
439 and virus phylogeny observed for RNA viruses as a whole (Shi et al 2018a). However, it is
440 also clear that cross-species is also a common occurrence in virus evolution (Geoghegan et
441 al 2017). Indeed, it is possible that the strong association of host taxonomy and virome
442 composition in some cases reflects preferential host switching among fish species
443 (otherwise known as the 'phylogenetic distance effect'; Longdon et al 2014), perhaps
444 because viruses spread more often between phylogenetically closely related hosts due to
445 the use of similar cell receptors (Charleston and Robertson 2002). These competing
446 theories could be tested by more detailed co-phylogenetic comparisons among fish
447 species that exhibit no ecological overlap thereby precluding cross-species transmission.

448 Our analysis also provided some evidence that virus abundance was negatively associated
449 with the preferred water temperature of the fish species in question. Specifically, viruses
450 were more abundant in fish that preferred cooler temperatures compared to those that
451 prefer warmer temperatures. In this context it is noteworthy that virus transmission and
452 disease outbreaks have been shown to be influenced by temperature and seasonality in
453 farmed fish (Crane and Hyatt 2011). Moreover, for some viruses, host mortality is water
454 temperature-dependent. For example, a highly infectious disease in fish, nervous necrosis
455 virus, is more pathogenic at higher temperatures (Toffan et al 2016), while infectious

456 hematopoietic necrosis virus, which causes disease in salmonid fish such as trout and
457 salmon, causes mortality only at low temperatures (Dixon et al 2016). As the oceans
458 continue to warm, it is crucial to understand the impact of increased temperatures on both
459 marine life and virus evolution and emergence, especially as it is projected that outbreaks
460 of marine diseases are likely to increase in frequency and severity (Dallas and Drake 2016,
461 Karvonen et al 2010).

462 Also of note was that on these data, fish living in diverse, multi-fish species communities
463 harboured more diverse viromes at a higher abundance than fish that live in less diverse,
464 single-species communities. Previously, host community diversity has been hypothesised
465 to lead to a decrease in infectious disease risk through the theory of the 'dilution effect'
466 (Schmidt and Ostfeld, 2001). This theory views an increase in host species' community
467 diversity as likely to reduce disease risk, because encounter rates among preferred hosts
468 are decreased, and both experimental and field studies have shown this phenomenon to
469 occur across many host systems, particularly those involving vector-borne disease
470 (Keesing et al 2006, LoGiudice et al 2003, Ostfeld and Keesing 2012). Although it might be
471 reasonable to assume that increased virus abundance and diversity is directly correlated
472 with disease risk, the association between host community diversity with that of virus
473 diversity and abundance has not previously been tested. Our results, although preliminary,
474 indicated that high multi-species community diversity in fish may be associated with
475 increased virus diversity and abundance. It is possible that elevated community diversity in
476 fish simply increases the total number of hosts in the system, in turn increasing viral
477 diversity, particularly since host jumping appears to be common in fish viruses (Geoghegan
478 et al 2018a).

479 Finally, it is noteworthy that since these fish species were market-bought rather than being
480 directly sampled during fishing trips (with the exception of the pygmy goby), it is possible
481 that viruses with short durations of infection were not detected. In addition, the relatively
482 small number of individuals sampled here, and that samples were necessarily pooled to aid
483 virus discovery, unavoidably limits some of the conclusions drawn. In particular, the host
484 traits summarised here, such as life span, were taken at the overall species level rather than
485 for the individuals sampled. It is therefore important to broaden sampling of fish and their
486 viruses both geographically and seasonally, and include phenotypic data for the individuals
487 sampled. This notwithstanding, our data again shows that fish harbour a very large number
488 of diverse viruses (Shi et al. 2018; Lauber et al, 2017). Indeed, even the pygmy goby, one of
489 the shortest-lived vertebrates on earth that lives for a maximum of 59 days on the reef
490 (Depczynski and Bellwood 2005), harboured novel viruses that were assigned to three

distinct virus families.

The new viruses discovered here greatly expand our knowledge of the evolutionary history of many virus families, particularly those with RNA genomes, with viruses identified in fish species that span highly diverse taxonomic orders. More broadly, the use of metagenomics coupled with a diverse multi-host, tractable system such as fish has the potential to reveal how host factors can shape the composition of viromes and that might ultimately lead to cross-species transmission and virus emergence.

Data Availability

All sequence reads generated in this project are available under the NCBI Short Read Archive (SRA) under BioProject PRJNA637122 and all consensus virus genetic sequences have been deposited in GenBank under accession MT579871-MT579895

Acknowledgements

We thank the New South Wales Department of Primary Industries for help sourcing fish samples. We thank efishalbum.com for fish images in Figure 1, which were used with permission. ECH and DRB are funded by ARC Australian Laureate Fellowships (FL170100022 and FL190100062, respectively). This work was partly funded by ARC Discovery grant DP200102351 to ECH and JLG and a Macquarie University Grant awarded to JLG.

References

Alarcon-Schumacher T et al. (2019) 'Elucidating viral communities during a phytoplankton bloom on the west Antarctic peninsula', *Frontiers in Microbiology* 10: 1014.

Bennett AJ et al. (2020) 'Relatives of rubella virus in diverse mammals', *Nature* in press.

Bergh Ø, et al. (1989) 'High abundance of viruses found in aquatic environments', *Nature* 340: 467-468.

Betancur-R R et al. (2017) 'Phylogenetic classification of bony fishes', *BMC Evolutionary Biology* 17: 162.

Breitbart M, Rohwer F (2005) 'Here a virus, there a virus, everywhere the same virus?', *Trends in Microbiology* 13: 278-284.

Buchfink B, Xie C, Huson DH (2015) 'Fast and sensitive protein alignment using DIAMOND', *Nature Methods* 12: 59-60.

Charon J, Grigg MJ, Eden J-S, et al. (2019) 'Novel RNA viruses associated with Plasmodium vivax in human malaria and Leucocytozoon parasites in avian disease', *PLoS Pathog* 15, e1008216.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics* 25: 1972-1973.

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', *BMC One Health Outlook* 1: 5.

Charleston MA, Robertson DL (2002) 'Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny', *Systematic Biology* 51: 528-535.

Coutinho FH et al. (2017) 'Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans', *Nature Communications* 8: 15955.

Crane M, Hyatt A (2011) 'Viruses of fish: an overview of significant pathogens', *Viruses* 3: 2025-2046.

Dallas T, Drake JM (2016) 'Fluctuating temperatures alter environmental pathogen transmission in a Daphnia-pathogen system', *Ecology and Evolution* 6: 7931-7938.

De Corte D et al. (2012) 'Links between viruses and prokaryotes throughout the water column along a North Atlantic latitudinal transect', *The ISME Journal* 6: 1566-1577.

Depczynski M, Bellwood DR (2005) 'Shortest recorded vertebrate lifespan found in a coral reef fish', *Current Biology* 15: R288-289.

Dill JA et al. (2016) 'Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses', *Journal of Virology* 90: 7920-7933.

Dixon P et al. (2016) 'Epidemiological characteristics of infectious hematopoietic necrosis virus (IHNV): a review', *Veterinary Research* 47: 63.

Geoghegan JL, Duchêne S, Holmes EC (2017) 'Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families', *PLoS Pathogens* 13: e1006215.

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

Geoghegan JL et al. (2018b) 'Virological sampling of inaccessible wildlife with drones', *Viruses* 10.

Gregory AC et al. (2019) 'Marine DNA viral macro- and microdiversity from pole to pole', *Cell* 177: 1109-1123.e1114.

579 Haas BJ et al. (2013) 'De novo transcript sequence reconstruction from RNA-Seq: reference
580 generation and analysis with Trinity', *Nature Protocols* 8: 10.1038/nprot.2013.1084.
581
582 Haugland Ø (2011) 'Cardiomyopathy syndrome of Atlantic salmon (Salmo salar L.) is
583 caused by a double-stranded RNA virus of the *Totiviridae* family', *Journal of Virology*
584 85 (11) 5275-5286.
585
586 Hothorn T, Bretz F, Westfall P (2008) 'Simultaneous inference in general parametric
587 models', *Biometrical Journal* 50: 346-363.
588
589 Jarungsriapisit J et al. (2020) 'The effect of temperature on the survival of salmonid
590 alphavirus analysed using in vitro and in vivo methods', *Aquaculture* 516: 734647.
591
592 Karvonen A et al. (2010) 'Increasing water temperature and disease risks in aquatic
593 systems: climate change increases the risk of some, but not all, diseases', *International*
594 *Journal for Parasitology* 40: 1483-1488.
595
596 Katoh K, Standley DM (2013) 'MAFFT multiple sequence alignment software version 7:
597 improvements in performance and usability', *Molecular Biology and Evolution* 30: 772-780.
598
599 Keane TM et al. (2006) 'Assessment of methods for amino acid matrix selection and their
600 use on empirical data shows that ad hoc assumptions for choice of matrix are not justified',
601 *BMC Evolutionary Biology* 6: 29.
602
603 Keesing F, Holt RD, Ostfeld RS (2006) 'Effects of species diversity on disease risk', *Ecology*
604 *Letters* 9: 485-498.
605
606 Kelley LA et al. (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis',
607 *Nature Protocols* 10: 845-858.
608
609 Lagkouvardos I et al. (2017) 'Rhea: a transparent and modular R pipeline for microbial
610 profiling based on 16S rRNA gene amplicons', *PeerJ* 5: e2836.
611
612 Lara E et al. (2017) 'Unveiling the role and life strategies of viruses from the surface to the
613 dark ocean', *Science Advances* 3: e1602565.
614
615 Lauber C et al. (2017) 'Deciphering the origin and evolution of hepatitis B viruses by means
616 of a family of non-enveloped fish viruses', *Cell Host and Microbe*.
617
618 Li B, Dewey CN (2011) 'RSEM: accurate transcript quantification from RNA-Seq data with
619 or without a reference genome', *BMC Bioinformatics* 12: 323.
620
621 Li W, Godzik A (2006) 'Cd-hit: a fast program for clustering and comparing large sets of
622 protein or nucleotide sequences', *Bioinformatics* 22: 1658-1659.
623
624 Lim ES et al. (2015) 'Early life dynamics of the human gut virome and bacterial microbiome
625 in infants', *Nature Medicine* 21, 1228–1234.
626
627 LoGiudice K et al. (2003) 'The ecology of infectious disease: Effects of host diversity and
628 community composition on Lyme disease risk', *Proceedings of the National Academy of*
629 *Science USA* 100: 567-571.
630

631 Longdon B et al. (2014) ‘The evolution and genetics of virus host shifts’, *PLoS Pathogens*
632 10: e1004395.
633
634 Løvoll M, Wiik-Nielsen J, Grove S et al. (2010) ‘A novel totivirus and piscine reovirus (PRV)
635 in Atlantic salmon (*Salmo salar*) with cardiomyopathy syndrome (CMS)’, *Virology*
636 *Journal*, 7, 309.
637
638 Maranger R, Bird DF (1995) ‘Viral abundance in aquatic systems: a comparison between
639 marine and fresh waters’, *Marine Ecology* 121: 217-226.
640
641 McMurdie PJ, Holmes S (2013) ‘phyloseq: an R package for reproducible interactive
642 analysis and graphics of microbiome census data’, *PLoS One* 8: e61217.
643
644 Middelboe M, Brussaard CPD (2017) ‘Marine viruses: key players in marine ecosystems’,
645 *Viruses* 9: 302.
646
647 Mikalsen AB, Haugland Ø, Evensen Ø (2016) ‘Chapter 16 - Totiviruses of Fish’, *Aquaculture*
648 *Virology*, 251-258.
649
650 Mor SK, Phelps NB (2016) ‘Molecular detection of a novel totivirus from golden shiner
651 (*Notemigonus crysoleucas*) baitfish in the USA’, *Archives of Virology*, 161(8), 2227-34.
652
653 Nguyen L-T et al. (2015) ‘IQ-TREE: a fast and effective stochastic algorithm for estimating
654 maximum-likelihood phylogenies’, *Molecular Biololgy and Evolution* 32: 268-274.
655
656 Oksanen J et al. (2007) ‘Vegan: community ecology package’, R v.2.2-0.
657
658 Ostfeld RS, Keesing F (2012) ‘Effects of host diversity on infectious disease’, *Annual*
659 *Reviews of Ecology, Evolution and Systematics* 43: 157-182.
660
661 Paez-Espino D et al. (2016) ‘Uncovering Earth's virome’, *Nature* 536: 425-430.
662
663 Pettersson JHO et al. (2019) ‘Meta-transcriptomic identification of hepatitis B virus in
664 cerebrospinal fluid in patients with central nervous system disease’, *Diagnostic*
665 *Microbiology and Infectious Disease* 95: 114878-114878.
666
667 Porter AF et al. (2019) ‘Diversity and evolution of novel invertebrate DNA viruses revealed
668 by meta-transcriptomics’, *Viruses* 11: 1092.
669
670 Rice P, Longden I, Bleasby A (2000) ‘EMBOSS: the European molecular biology open
671 software suite’, *Trends in Genetics* 16: 276-277.
672
673 Roux S et al. (2017) ‘Ecogenomics of virophages and their giant virus hosts assessed
674 through time series metagenomics’, *Nature Communications* 8, 858.
675
676 Shi M, Lin X-D et al. (2016) ‘Redefining the invertebrate RNA virosphere’, *Nature* 540: 539-
677 543.
678
679 Shi M, Lin X-D et al. (2018a) ‘The evolutionary history of vertebrate RNA viruses’, *Nature*
680 556: 197-202.
681
682 Shi M, Zhang YZ, Holmes EC (2018b) ‘Meta-transcriptomics and the evolutionary biology of
683 RNA viruses’, *Virus Research* 243: 83-90.

Suttle CA (2005) ‘Viruses in the sea’, *Nature* 437: 356-361.

Temmam S et al (2016) ‘Characterization of viral communities of biting midges and identification of novel thogotovirus species and rhabdovirus genus’, *Viruses* 8, 77.

Tirosh O et al. (2018) ‘Expanded skin virome in DOCK8-deficient patients’, *Nature Medicine* 24: 1815-1821.

Toffan A et al. (2016) ‘Water temperature affects pathogenicity of different betanodavirus genotypes in experimentally challenged Dicentrarchus labrax’, *Diseases of Aquatic Organisms* 119: 231-238.

Vibin J et al. (2018) ‘Metagenomics detection and characterisation of viruses in faecal samples from Australian wild birds’, *Scientific Reports* 8, 8686.

Whittington RJ, Reddacliff GL (1995) ‘Influence of environmental temperature on experimental infection of redfin perch (*Perca fluviatilis*) and rainbow trout (*Oncorhynchus mykiss*) with epizootic haematopoietic necrosis virus, an Australian iridovirus’, *Australaian Veterinary Journal* 72: 421-424.

Wille M, Shi M et al. (2019) ‘Virome heterogeneity and connectivity in waterfowl and shorebird communities’, *The ISME Journal* 13: 2603-2616.

Wille M (2020) ‘Unravelling virus community ecology in bats through the integration of metagenomics and community ecology’, *Molecular Ecology* 29: 23-25.

Zhang Y-Z, Shi M, Holmes EC (2018) ‘Using metagenomics to characterize an expanding virosphere’, *Cell* 172: 1168-1172.

Zhang Y-Z, Wu WC, Shi M, Holmes EC (2018) ‘The diversity, evolution and origins of vertebrate RNA viruses’, *Current Opinion in Virology*, 31: 9-16.

716 # Figures

717 **Figure 1**. (A) Total standardized abundance of vertebrate-associated viruses (at the level of
718 virus family) across the fish species examined. (B) Normalised viral abundance set out on a
719 backbone of the fish host phylogeny at the order level. (C) Standardised number of total
720 viral reads (black), vertebrate-associated viral reads (grey) and host reference gene
721 ribosomal protein S13 (RPS13) (orange) in each species library.

722 **Figure 2**. Phylogenetic relationships of likely vertebrate-associated viruses identified here.
723 The maximum likelihood phylogenetic trees show the topological position of the newly
724 discovered viruses (blue circles) and those identified in an earlier study (Geoghegan et al.
725 2018), in the context of their closest phylogenetic relatives. Branches are highlighted to
726 represent host class (fish = blue; mammals = red; birds, reptiles and amphibians = yellow;
727 vector-borne (mammals and arthropods) = green). All branches are scaled according to the
728 number of amino acid substitutions per site and trees were mid-point rooted for clarity only.
729 An asterisk indicates node support of >70% bootstrap support. See SI Table 3 for all
730 accession numbers.

731 **Figure 3**. Significant explanatory variables in generalized linear models (GLM) for viral
732 abundance and two measures of alpha diversity. Viral abundance is best explained by (A)
733 fish host order and (B) mean preferred water temperature. Alpha diversity is best explained
734 by (C) host order and (D) preferred habitat (Observed Richness) and by (E) host order and
735 (F) host community diversity (Shannon Diversity). Stars indicate significant differences
736 between groups determined by posthoc Tukey tests. Points represent different fish species
737 and are coloured by host order.

738

Supplementary Information

SI Table 1. Fish species sampled and the host features used in this analysis, with the latter obtained from fishbase.org. These features comprised fish taxonomic order, swimming behaviour (i.e. solitary or schooling fish), preferred climate, mean preferred water temperature, host community diversity (i.e. multi- or single- species community), average species length, maximum life span, trophic level and habitat depth.

SI Table 2. Amino acid identity, contig length and relative frequency of the viruses identified in this study. This does not include viruses described in Geoghegan et al (2018a).

SI Table 3. Accession numbers of viruses used to construct virus phylogenetic trees in Figure 2.

SI Figure 1. Graphical representation of virus families shared between fish species in this study. Fish species are shown in grey at the edges. The coloured connecting lines illustrate cases where viruses are present in both hosts for (A) viruses associated with fish hosts and (B) viruses likely associated with non-vertebrate host taxa (arthropods, fungi, plants, and protozoa). Caution must be met while interpretating this figure since virus abundance is not visually displayed. For example, sequence reads of reovirus were detected in the sand whiting at very low abundance (A) and this was insufficient to be included in the phylogenetic analysis shown in Figure 2. In addition, connecting lines depict hosts that share viruses from the same family and do not represent direct virus transmission between hosts. For both (A) and (B) only those fish species carrying viruses are shown.

SI Figure 2. Maximum likelihood phylogenetic trees depicting relationships among newly identified viruses associated with non-fish hosts that are members of the *Totiviridae*, *Narnaviridae* and *Partitiviridae*. Orange-coloured tips indicate the novel viruses within each family, while the relevant clades are highlighted with green-shaded boxes. Highly supported clades corresponding to SH-aLRT ≥ 80% and UFboot ≥ 95% are shown with white circles at the tree nodes. The number of amino acid substitutions per site is represented with a scale bar below each tree. The non-vertebrate host taxa (arthropods, fungi, plants, and protozoa) associated to each virus family are shown in the inset.

SI Figure 3. Maximum likelihood phylogenetic trees depicting relationships among newly identified viruses associated with non-fish hosts that are members of the families *Solemoviridae* and *Tombusviridae*. Orange-coloured tips indicate the novel viruses within each family, while the relevant clades are highlighted with purple-shaded boxes. Highly supported clades corresponding to SH-aLRT ≥ 80% and UFboot ≥ 95% are shown with

white circles at the tree nodes. The number of amino acid substitutions per site is represented with a scale bar below each tree. The non-vertebrate host taxa (arthropods, fungi, plants, and protozoa) associated to each virus family are shown in the inset.

SI Figure 4. Maximum likelihood phylogenetic trees depicting relationships among newly identified viruses associated with non-fish hosts that are members of the families *Hepeviridae, Chuviridae, Nodaviridae, Iflaviridae, Dicistroviridae* and *Picornaviridae*. Orange-coloured tips indicate the novel viruses within each family, while the relevant clades are highlighted with blue-shaded boxes. Highly supported clades corresponding to SH-aLRT ≥ 80% and UFboot ≥ 95% are shown with white circles at the tree nodes. The number of amino acid substitutions per site is represented with a scale bar below each tree. The non-vertebrate host taxa (arthropods, fungi, plants, and protozoa) associated to each virus family are shown in the inset.