Host evolutionary history and ecology shape virome composition in fishes

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

Abstract

Identifying the components of host ecology that promote virus diversity is crucial for our understanding of the drivers of virus evolution and disease emergence. As the most species-rich group of vertebrates that exhibit diverse ecologies, fish provide an ideal model system to study the impacts of host ecology on the composition of their viromes. To better understand the factors that shape virome composition in marine fishes, we characterised the viromes of 23 fish species (19 from this study and four that were sampled previously (Geoghegan et al 2018a)) using unbiased bulk RNA-sequencing (meta-transcriptomics) together with both sequence and protein structural homology searches to identify divergent viruses that often evade characterisation. These data revealed that fish virome composition – that is, viral richness, abundance and diversity – were predominantly shaped by the phylogenetic history of their hosts, as reflected in taxonomic order. In addition, preferred mean water temperature, climate, habitat depth, community diversity and whether fish swim in schools or are solitary were identified as important ecological features that shaped virome diversity and abundance in these fish. Our analysis also identified 25 new virus transcripts that could be assigned to 11 different viral families, including the first fish virus in the *Matonaviridae*. Other viruses identified fell within the *Astroviridae*, *Picornaviridae*, *Arenaviridae*, *Reoviridae*, *Hepadnaviridae*, *Paramyxoviridae*, *Rhabdoviridae*, *Hantaviridae*, *Filoviridae* and *Flaviviridae*. Our results provide a better understanding of the ecological determinants of virome diversity and support the view that fish harbour a multitude of viruses, of which the vast majority are undescribed.

Introduction

Metagenomic next-generation sequencing (mNGS) has led to a revolution in virus discovery (Shi et al 2018b, Zhang et al 2018), exposing more of the diversity, abundance and structure of the eukaryotic virosphere. However, while it is now potentially possible to reveal entire host viromes (Chang et al 2019, Geoghegan et al 2018a, Geoghegan et al 2018b, Paez-Espino et al 2016, Pettersson et al 2019, Porter et al 2019, Shi et al 2016, Shi et al 2018a, Tirosh et al 2018), we do not fully understand the factors that shape virome diversity, including the dual and interacting impact of host and virus ecology. Indeed, until recently, the study of virus ecology had largely been limited to studies of single viruses and/or single hosts and their interactions, restricting our ability to explore multifactorial impacts, including diverse aspects of host ecology, on virome diversity. Fortunately, the advent of unbiased, meta-transcriptomic RNA sequencing enables us to explore, more thoroughly, virome diversity and abundance as well as the myriad of biological and environmental factors that likely shape this diversity (Wille et al 2019, Wille 2020). Identifying the host ecological factors that promote virus diversification is also central to understanding the drivers of virus evolution and emergence. As a simple case in point, host behavioural ecology directly affects contact rates among hosts and is therefore likely to be important in shaping viral dynamics: more frequent intra- and inter-host contacts are likely to increase the potential for viral spread and diversification.

The marine environment is a rich source of viruses. It has long been known that the bacteriophage present in aquatic ecosystems outnumber that of other lifeforms 10-fold (Maranger and Bird 1995), with an estimated concentration of 10 billion virus partials 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 contrast to bacteriophage, little is known about how such environmental factors might contribute to virus diversity in natural aquatic vertebrate host populations, even though viruses can cause large-scale infection and disease in farmed fish (Crane and Hyatt 2011, Jarungsriapisit et al 2020, Whittington and Reddacliff 1995).

Fish provide an ideal model to understand how host ecology might shape the composition and abundance of those viruses that infect them. 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. As such, these animals exhibit a wide range

of ecological features that likely play an important role in shaping the diversity of their viromes, although to date there is a marked absence of work in area. Initial studies indicate that fish harbour a remarkable diversity of viruses (particularly RNA viruses) 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 appear to be the evolutionary predecessors of viruses infecting other vertebrate hosts, generally indicative of a pattern of virus-host co-divergence that can date back hundreds of millions of years. Despite the apparent diversity and ubiquity of fish viruses, they are severely under-studied when compared to mammalian and avian viruses.

To better understand how host ecology shapes virome composition we sampled viruses from a diverse range of wild-caught fish. In particular, we considered marine fish spanning 23 species across nine taxonomic orders, quantifying a variety of ecological characteristics that together may impact virome composition and abundance. We utilised unbiased bulk RNA-sequencing (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 novel viruses obtained, (iii) determine whether there are associations between virome abundance and diversity and key aspects of host ecology, including viral richness and composition, and (iv) explore whether taxonomically-related fish hosts have more similar viromes. The particular ecological characteristics 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, trophic level, and habitat depth (SI Table 1). In doing so, we provide novel insights into the evolution and ecology of fish viromes and how they are shaped by their hosts.

Methods

Ethics. Biosafety was approved by Macquarie University (ref: 52017008 56). This study involved dead fish purchased from a fish market; in these cases 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 tropical northern Queensland at approximately the same time. Fish were snapped frozen at -20°C immediately upon capture. Fish obtained from the market were purchased on the day of catch. Tissues were dissected and stored in RNALater before being transferred to a -80°C freezer. To increase the likelihood of virus discovery during metagenomic sequencing, 10 individuals from each species were pooled.

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

Transcript sequence similarity searching for viral discovery. Sequencing reads were first quality trimmed then assembled *de novo* using Trinity RNA-Seq (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 and Diamond (BLASTX) (Buchfink et al 2015), and an e-value threshold of $1x10^{-5}$ was used as a cut-off to identify positive matches. We removed non-viral hits, including host contigs with similarity to viral sequences (e.g. endogenous viral elements), as well as any contigs with high similarity to plant viruses, which were more likely to be derived from food sources. We focused our analysis on vertebrate-associated viruses by removing viral hits with high sequence similarity to invertebrate-associated viruses, which more likely originated from invertebrates within the fish rather than from the fish themselves.

Protein structure similarity searching for viral discovery. To identify highly divergent viral transcripts, including those that might be refractory to detection using similarity searching methods such as the BLAST approach described above, we also employed a protein structure-based similarity search 'orphan' contigs that did not share sequence similarity with other known sequences. Accordingly, assembled orphan contigs were translated into open reading frames (ORFs) using EMBOSS getorf program (Rice et al 2000). ORFs were arbitrarily defined as regions between two stop codons with a minimum size of 200 amino acids in length. To reduce redundancy, amino acid sequences were grouped based on sequence identity using the CD-HIT package v4.6.5

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

Inferring the evolutionary history of fish viruses. To infer the evolutionary relationships of the viruses contained in the fish samples, the translated viral contigs were combined with protein sequences obtained from NCBI RefSeq. The sequences retrieved were then aligned with those generated here again using MAFFT v7.4 (E-INS-i algorithm) as described above. Ambiguously aligned regions were removed using trimAl v.1.2 (Capella-Gutierrez et al 2009). To estimate phylogenetic trees, we selected the optimal model of amino acid substitution identified using the Bayesian Information Criterion as implemented in Modelgenerator v0.85 (Keane et al 2006) and analysed the data using the maximum likelihood approach available in IQ-TREE (Nguyen et al 2015) with 1000 bootstrap replicates. Phylogenetic trees were annotated with FigTree v.1.4.2. New viruses were named after well-known aquatic fictional characters.

Virome abundance and diversity. Transcriptomes were quantified using RNA-Seq by Expectation-Maximization (RSEM) as implemented within Trinity (Li and Dewey 2011). Analyses of abundance and genetic diversity were performed 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 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 and virome composition was plotted as a function of nonmetric multidimensional scaling (NMDS) ordination as well as constrained ordination (CAP) with the and *phyloseq* package (McMurdie and Holmes 2013). 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). We selected CAP in addition to nMDS because, only the variation that can be explained by the environmental variables is displayed and analysed.

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 for ecological inference and 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.

Diversity and abundance of viruses in fish. On average, fish viromes comprised more likely invertebrate-associated viruses than vertebrate-associated viruses (Figure 1). However, we focused on the latter since we assume that the vertebrate-associated viruses were directly infecting the fish sampled rather than being associated with the aquatic environment or a co-infecting parasite, and hence are more informative in determining how host factors shape virus ecology.

Overall, we identified virus transcripts that could be assigned to 11 viral families. 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* (at 39% of all viruses discovered), *Picornaviridae* (19%), *Arenaviridae* (16%), *Reoviridae* (13%) and the *Hepadnaviridae* (9%) (Figure 1a). Other viral families found at lower relative abundances were the *Matonaviridae* (previously the *Togaviridae*) (2%), *Paramyxoviridae* (1%), as well as the *Rhabdoviridae*, *Hantaviridae*, *Filoviridae* and *Flaviviridae* (all <1%) (Figure 1a). The most common vertebrate-associated viruses found in these fish were picornaviruses (eight species), astroviruses (seven species) and hepadnaviruses (six species) (Figure 1b). The eastern sea garfish (*Hyporhamphus australis*) harboured the most diverse virome with four distinct vertebrate-associated viruses (Figure 1b). Six fish contained no vertebrate-

212 associated viruses, and we found no viral sequences in the yellowfin bream (*Acanthopagrus*
213 *australis*) (Figure 1c). An equivalent analysis of a host reference gene, ribosomal protein S13 (RPS13)
214 that is stably expressed in fish, revealed similar abundances across species (0.004% – 0.02%),
215 implying similar sequencing depth across libraries (Figure 1c). RPS13 was, on average, ~55% more
216 abundant than the total virome.

217 **Evolutionary relationships of fish viruses**. To infer stable phylogenetic relationships among the
218 viruses sampled and to identify those that are novel, we utilised the most conserved (i.e.
219 polymerase) viral regions that comprise the RNA-dependent RNA polymerase (RdRp) or the
220 polymerase (P) ORF in the case of the hepadnaviruses. From this, we identified 25 distinct and
221 potentially novel vertebrate-associated virus species, in addition to the eight novel viruses
222 described previously (Geoghegan et al 2018a) (SI Table 2). All novel viruses shared sequence
223 similarity to other known fish viruses with the exception of those viruses found in the *Matonaviridae*
224 and *Rhabdoviridae* (Figure 2, SI Figure 1; see below).

225 Among the viruses identified was a fish rubivirus (fedallah virus) in the tiger flathead
226 (*Neoplatycephalus richardsoni*) - the first fish virus found in the *Matonaviridae*. This novel viral
227 sequence shared only 35% amino acid similarity with its closest relative - Guangdong Chinese water
228 snake rubivirus (Shi et al 2018a). All other viruses within this family are distantly related human
229 rubella viruses, and it is therefore likely that these non-human viruses constitute a discrete genus.
230 Another divergent virus discovered in this analysis is pip virus (*Rhabdoviridae*) in the eastern sea
231 garfish (*Hyporhamphus australis*), which was most closely related to the Fujian dimarhabdovirus
232 sampled from an amphibian host, sharing 45% amino acid RdRp sequence identity. This highly
233 divergent virus was only identified by using protein structure homology, and forms a clade that is
234 distinct from other fish rhabdoviruses (Figure 2; SI Figure 1). We also discovered two novel viral
235 sequences in the *Filoviridae* in John Dory (*Zeus faber*) and the blue spotted goatfish (*Upeneichthys*
236 *lineatus*). These viruses – termed here ahab virus and starbuck virus, respectively – shared sequence
237 similarity to the only other known fish filovirus, Wenling filefish (Shi et al 2018a). With the exception
238 of these fish viruses, all other known filoviruses including Ebola and Marburg viruses, are found in
239 mammalian hosts, notably humans, bats and primates.

240 We also found numerous viruses that cluster within established clades of fish viruses. For example,
241 aronnax virus, a member of the *Hantaviridae* discovered in the pygmy goby (*Eviota zebrina*),
242 grouped with other hantaviruses recently found in fish (Figure 2). Although they were previously
243 only thought to infect mammals, hantaviruses have now been found to infect amphibians, jawless
244 fish and ray-finned fish (Shi et al 2018a). The evolutionary history of the *Paramyxoviridae* shows two
245 distinct fish virus lineages, of which both ned virus and nemo virus in the barramundi and pygmy

goby, respectively, grouped with Pacific spade-nose shark paramyxovirus, and shared 50% and 45% amino acid L gene sequence similarity. This group of fish viruses is phylogenetically distinct from other paramyxoviruses. We also found novel fish viruses in the *Flaviviridae*, *Arenaviridae* and *Reoviridae*: although these grouped with other fish viruses, they greatly expand the known diversity of these virus families. Finally, as noted above, the most abundant viruses fell within the *Picornaviridae* and *Astroviridae*, and all shared sequence similarity to other fish viruses. Both single-stranded positive-sense RNA viruses, picornaviruses and astroviruses exist as small icosahedral capsids with no external envelope, which may aid their preservation in harsh marine environments.

The only DNA viruses we revealed were novel hepadnaviruses found in bonito (*Sarda australis*), ludrick (*Girella tricuspidata*) and eastern school whiting (*Sillago flindersi*), which fell into the divergent hepadna-like viruses, *Nackednavirus*, found in a number of fish species (Lauber et al. 2017), while daggoo virus in sand whiting (*Sillago ciliate*) expanded the fish virus clade that is more closely related to mammalian hepatitis B viruses (Dill et al 2016) (Figure 2).

Associations of host taxonomy and ecology with virome composition.

To understand the role of host ecological variables on viral ecology, we examined the role of eight host traits on shaping viral abundance (the proportion of viral reads in each sample), alpha diversity (the diversity within each sample, measured by observed richness and Shannon diversity) and beta diversity (the diversity between samples). The host features examined were: host taxonomic order, swimming behaviour (solitary or schooling fish), preferred climate, mean preferred water temperature, community diversity, average species length, trophic level and habitat depth.

This analysis revealed that fish phylogenetic relationships (as reflected in taxonomic order), played the most important role in shaping the composition of fish viromes. This pattern was consistent when assessing viral abundance, alpha diversity and beta diversity (Figures 3 and 4). In addition, fish order (χ^2=0.003, df=8, p=0.0049) and mean preferred water temperature (χ^2=0.008, df=1, p=0.035) were important predictors of viral abundance, such that Scopaeniformes (i.e. bigeye ocean perch, red gurnard, tiger flathead, and eastern red scorpionfish) had significantly higher viral abundance compared to Pleuronectiformes (i.e. largetooth flounder) (Tukey: z=3.766, p=0.00479), while viral abundance had a negative relationship to mean preferred water temperature (Figure 3).

We used two measures of alpha diversity: 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

279 single species communities. As with viral abundance, there was a significant difference in alpha
280 diversity between Scopaeniformes compared to Pleuronectiformes (Tukey Richness z=3.039,
281 p=0.0495; Tukey Shannon z=2.845, p=0.05). Notably, mid-water fish had decreased viral richness
282 compared to benthic fish (Tukey z=-2.452, p=0.0338), and fish that reside in multispecies
283 communities had a larger Shannon diversity compared to single species communities (χ^2=0.17089,
284 df=1, p=0.05) (Figure 3).

285 Our analysis also revealed that fish order (R^2=0.57215, p=0.003), swimming behaviour (R^2=0.09904,
286 p=0.005), climate (R^2=0.13315, p=0.012) and mean preferred water temperature (R^2=0.1005, p=0.05)
287 are significant predictors of beta diversity. A conical constrained ordination (CAP) model developed
288 using these factors was significant (F_{11}=2.37, p=0.002) (Figure 4). In this ordination analyses, only
289 the variation that can be explained by the environmental variables is displayed and analysed (Figure
290 4).

291

292 Discussion

293 The metagenomic revolution is enabling us to uncover more of a largely unknown virosphere,
294 including highly divergent viruses that often elude characterisation. Here, we utilised mNGS to
295 reveal new viruses in fish and used these data to better understand the host ecological factors that
296 have had the greatest impact on shaping virus diversity and abundance. To do so we characterised
297 the viromes of 23 species of marine fish that spanned nine taxonomic orders, with our analysis
298 revealing that host phylogeny (taxonomy) plays a central role in shaping fish viromes. We also
299 found that several ecological features were also important determinants of virus abundance and/or
300 diversity, namely preferred mean water temperature, climate, habitat depth, community diversity
301 and whether fish swim in schools or are solitary. In addition, we have identified 25 novel viruses
302 spanning 11 different virus families, including the first fish virus in the *Matonaviridae*.

303 Many of the viruses identified in this study were phylogenetically related to other, recently
304 discovered, fish viruses (Dill et al 2016, Geoghegan et al 2018a, Lauber et al 2017, Shi et al 2018a).
305 However, there were a few notable exceptions. Fedallah virus in the tiger flathead in the
306 *Matonaviridae* represents the only fish viral species in this family, which forms a distinct clade with a
307 rubivirus discovered in a Chinese water snake. Human rubella virus is the only other virus previously
308 known within this family. The discovery of this phylogenetically distinct fish virus tentatively
309 suggests the possibility of a fish host origin for this family, although it is clear that a wider set of
310 hosts need to be sampled. Indeed, this might also be the case for other virus families such as the

311 *Hantaviridae* and *Filoviridae*, as fish viruses often fall basal to viruses in other vertebrate hosts such
312 as birds and mammals (also see Shi et al 2018a). In contrast, in some other virus families such as the
313 *Astroviridae, Picornaviridae, Flaviviridae* and *Rhabdoviridae*, fish viruses are found throughout the
314 phylogeny which is suggestive of a past history of host-jumping. Regardless, available data
315 suggests that fish viruses harbour more diversity compared to the better studied mammalian and
316 avian viruses within these families, and that the discovery of novel viruses in fish has expanded our
317 knowledge of the diversity, evolutionary history and host range of RNA viruses in general.

318 As well as identifying new viruses, we investigated host ecological features may have shaped the
319 overall composition of fish viruses. A key result from this analysis was that fish virome composition,
320 reflected in measures of viral richness, abundance and diversity, is predominantly shaped by the
321 phylogenetic relationships (i.e. taxonomy) of the host in question. This in turn suggests that fish
322 viruses might have co-diverged with their hosts over evolutionary time-scales (Geoghegan et al
323 2017), a pattern supported by the general relationship between vertebrate host class and virus
324 phylogeny observed for RNA viruses as a whole (Shi et al 2018a). Alternatively, it is possible that the
325 strong association of host taxonomy and virome composition is indicative of preferential host
326 switching among fish species, otherwise known as the 'phylogenetic distance effect' (Longdon et al
327 2014), perhaps because viruses spread more often between phylogenetically closely related hosts
328 due to the use of similar cell receptors (Charleston and Robertson 2002).

329 Combined with host order, virus abundance was also negatively associated with the hosts' preferred
330 water temperature. Specifically, our analysis revealed that viruses were more abundant in fish that
331 preferred cooler temperatures compared to those fish preferring warmer temperatures. Indeed,
332 virus transmission and disease outbreaks have been shown to be influenced by temperature and
333 seasonality in farmed fish (Crane and Hyatt 2011). Moreover, for some viruses, host mortality is
334 water temperature dependent. For example, a highly infectious disease in fish, nervous necrosis
335 virus, is more pathogenic at higher temperatures (Toffan et al 2016) while infectious hematopoietic
336 necrosis virus, which causes disease in salmonid fish such as trout and salmon, causes mortality only
337 at low temperatures (Dixon et al 2016). As the oceans continue to warm, it is crucial to understand
338 the impact of increased temperatures on both marine life and virus evolution and emergence,
339 especially as it is projected that outbreaks of marine diseases are likely to increase in frequency and
340 severity (Dallas and Drake 2016, Karvonen et al 2010).

341 Also, of note was that fish living in a diverse community harboured more diverse viromes at a higher
342 abundance compared to fish that live in less diverse, single-species communities. Previously, host
343 community diversity has been hypothesised to lead to a decrease in infectious disease risk through
344 the theory of the 'dilution effect' (Schmidt and Ostfeld, 2001). This theory views an increase in host

345 species' community diversity as likely to reduce disease risk on the basis that alternative host
346 species would negatively influence the preferred host as reservoirs, and both experimental and field
347 studies have shown this phenomenon to occur across many host systems, particularly those
348 involving vector-borne disease (Keesing et al 2006, LoGiudice et al 2003, Ostfeld and Keesing 2012).
349 Although it might be reasonable to assume that increased virus abundance and diversity is directly
350 correlated with disease risk, the association between host community diversity with that of virus
351 diversity and abundance has not previously been tested. Our results indicated that high community
352 diversity in fish increases virus diversity and abundance. It might be the case that increases in
353 community diversity in fish simply increases the total number of hosts in the system in turn
354 increasing viral diversity, particularly since host jumping between fish appears to be common in fish
355 viruses (Geoghegan et al 2018a).

356 To compare virome diversity between fish species, we measured beta diversity. CAP analysis
357 demonstrated that beta diversity in fish is multifactorial, with previously significant factors such as
358 temperature and host order being significant. Additionally, we found that swimming behaviour (i.e.
359 swimming in schools or solitary) and climate (subtropical, temperature and tropical) to further
360 contribute to the variation in virome diversity. That is, virome composition was typically more
361 similar among fish that exhibited the same ecological traits, such as swimming behaviour. We have
362 previously shown that schooling fish harbour more viruses compared to their solitary counterparts
363 (Geoghegan et al 2018a) since close contact while shoaling likely facilitates virus transmission
364 between hosts (Johnson et al. 2011). Although here we found no difference in virus species richness
365 between schooling and solitary fish, our results indicate that swimming behaviour is nevertheless
366 important in shaping virome composition.

367 Finally, it is noteworthy that since these fish species were market-bought rather than being directly
368 sampled during fishing trips (with the exception of the pygmy goby), it is possible that viruses with
369 short durations of infection were not detected. This notwithstanding, the diversity of viruses
370 discovered here provides further support for the proposition that fish harbour a very large number
371 of viruses (Shi et al. 2018; Lauber et al, 2017). Even the pygmy goby, one of the shortest-lived
372 vertebrates on earth (Depczynski and Bellwood 2005), harboured novel viruses that were assigned
373 to three distinct virus families.

374 In sum, the new viruses discovered here greatly expand our knowledge of the evolutionary history
375 of many virus families, with viruses identified in fish species that span highly diverse taxonomic
376 orders. More broadly, the use of metagenomics coupled with a diverse multi-host, tractable system
377 such as fish has enabled us to reveal some of the host ecological factors that shape virome
378 composition.

Data Availability

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

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 a Macquarie University Grant awarded to JLG.

Figures

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.

Figure 2. Phylogenetic relationships of likely vertebrate-associated viruses identified here (see SI Figure 1 for taxon labels). The maximum likelihood phylogenetic trees show the topological position of the newly discovered viruses (blue circles) and those identified in an earlier study (Geoghegan et al. 2018), in the context of their closest phylogenetic relatives. Branches are highlighted to represent host class (fish = blue; mammals = red; birds, reptiles and amphibians = yellow; vector-borne (mammals and arthropods) = green). All branches are scaled according to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. An asterisk indicates node support of >70% bootstrap support.

Figure 3. Significant explanatory variables in generalized linear models (GLM) for viral 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.

410 **Figure 4**. Constrained (canonical) ordination (CAP) using the bray Curtis dissimilarity matrix for
411 viromes of fish species. Vectors indicate direction and strength (length) of relationships between
412 species and significant explanatory variables. Colour, shape and fill correspond to host species
413 order, climate and schooling behaviour, respectively.

414

Supplementary Information

SI Table 1. Fish species sampled and the host ecological features used in this analysis, obtained from fishbase.org. These 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, 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 Figure 1. Phylogenetic relationships of likely vertebrate-associated viruses identified here. The maximum likelihood phylogenetic trees show the topological position of the newly discovered viruses (blue circles) and those identified in an earlier study (Geoghegan et al. 2018), in the context of their closest phylogenetic relatives. Branches are highlighted to represent host class (fish = blue; mammals = red; birds, reptiles and amphibians = yellow; vector-borne (mammals and arthropods) = green). All branches are scaled according to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. An asterisk indicates node support of >70% bootstrap support.

References

Alarcon-Schumacher T, Guajardo-Leiva S, Anton J, Diez B (2019). Elucidating viral communities during a phytoplankton bloom on the west Antarctic peninsula. *Front Microbiol* **10:** 1014.

Bergh Ø, BØrsheim KY, Bratbak G, Heldal M (1989). High abundance of viruses found in aquatic environments. *Nature* **340:** 467-468.

Betancur-R R, Wiley EO, Arratia G, Acero A, Bailly N, Miya M *et al* (2017). Phylogenetic classification of bony fishes. *BMC Evol Biol* **17:** 162.

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

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

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, Eden J-S, Hartley WJ, Shi M, Rose K, Holmes EC (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. *Syst Biol* **51:** 528-535.

Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD *et al* (2017). Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun* **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. *Ecol Evol* **6:** 7931-7938.

De Corte D, Sintes E, Yokokawa T, Reinthaler T, Herndl GJ (2012). Links between viruses and prokaryotes throughout the water column along a North Atlantic latitudinal transect. *ISME J* **6:** 1566-1577.

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

Dill JA, Camus AC, Leary JH, Di Giallonardo F, Holmes EC, Ng TF (2016). Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses. *J Virol* **90:** 7920-7933.

Dixon P, Paley R, Alegria-Moran R, Oidtmann B (2016). Epidemiological characteristics of infectious hematopoietic necrosis virus (IHNV): a review. *Vet Res* **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 Pathog* **13:** e1006215.

Geoghegan JL, Di Giallonardo F, Cousins K, Shi M, Williamson JE, Holmes EC (2018a). Hidden diversity and evolution of viruses in market fish. *Virus Evol* **4:** vey031-vey031.

Geoghegan JL, Pirotta V, Harvey E, Smith A, Buchmann JP, Ostrowski M *et al* (2018b). Virological sampling of inaccessible wildlife with drones. *Viruses* **10**.

Gregory AC, Zayed AA, Conceicao-Neto N, Temperton B, Bolduc B, Alberti A *et al* (2019). Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177:** 1109-1123.e1114.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J *et al* (2013). De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nat Protoc* **8:** 10.1038/nprot.2013.1084.

Hothorn T, Bretz F, Westfall P (2008). Simultaneous inference in general parametric models. *Biom J* **50:** 346-363.

Jarungsriapisit J, Nuñez-Ortiz N, Nordbø J, Moore LJ, Mæhle S, Patel S (2020). The effect of temperature on the survival of salmonid alphavirus analysed using in vitro and in vivo methods. *Aquac* **516:** 734647.

Karvonen A, Rintamaki P, Jokela J, Valtonen ET (2010). Increasing water temperature and disease risks in aquatic systems: climate change increases the risk of some, but not all, diseases. *Int J Parasitol* **40:** 1483-1488.

Katoh K, Standley DM (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30:** 772-780.

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McLnerney JO (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol* **6:** 29.

Keesing F, Holt RD, Ostfeld RS (2006). Effects of species diversity on disease risk. *Ecol Lett* **9:** 485-498.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10:** 845-858.

Lagkouvardos 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.

Lara E, Vaqué D, Sà EL, Boras JA, Gomes A, Borrull E *et al* (2017). Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Sci Adv* **3:** e1602565.

Lauber C, Seitz S, Mattei S, Suh A, Beck J, Herstein J *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*.

Li B, Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform* **12:** 323.

Li W, Godzik A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658-1659.

LoGiudice K, Ostfeld RS, Schmidt KA, Keesing F (2003). The ecology of infectious disease: Effects of host diversity and community composition on Lyme disease risk. *PNAS* **100:** 567-571.

Longdon B, Brockhurst MA, Russell CA, Welch JJ, Jiggins FM (2014). The evolution and genetics of virus host shifts. *PLoS Pathog* **10:** e1004395.

Maranger R, Bird DF (1995). Viral abundance in aquatic systems: a comparison between marine and fresh waters. *Mar* **121:** 217-226.

McMurdie PJ, Holmes S (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8:** e61217.

Middelboe M, Brussaard CPD (2017). Marine viruses: key players in marine ecosystems. *Viruses* **9:** 302.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32:** 268-274.

Oksanen J, Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. and Wagner, H. (2007). Vegan: community ecology package. R v.2.2-0.

Ostfeld RS, Keesing F (2012). Effects of host diversity on infectious disease. *Annu Rev Ecol Evol Syst* **43**: 157-182.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N *et al* (2016). Uncovering Earth's virome. *Nature* **536:** 425-430.

Pettersson JHO, Piorkowski G, Mayxay M, Rattanavong S, Vongsouvath M, Davong V *et al* (2019). Meta-transcriptomic identification of hepatitis B virus in cerebrospinal fluid in patients with central nervous system disease. *Diagn Microbiol Infect Dis* **95:** 114878-114878.

Porter AF, Shi M, Eden J-S, Zhang Y-Z, Holmes EC (2019). Diversity and evolution of novel invertebrate DNA viruses revealed by meta-transcriptomics. *Viruses* **11:** 1092.

Rice P, Longden I, Bleasby A (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet* **16:** 276-277.

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.

Shi M, Lin X-D, Chen X, Tian J-H, Chen L-J, Li K *et al* (2018a). The evolutionary history of vertebrate RNA viruses. *Nature* **556:** 197-202.

Shi M, Zhang YZ, Holmes EC (2018b). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Res* **243:** 83-90.

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

Tirosh O, Conlan S, Deming C, Lee-Lin S-Q, Huang X, Barnabas BB *et al* (2018). Expanded skin virome in DOCK8-deficient patients. *Nat Med* **24:** 1815-1821.

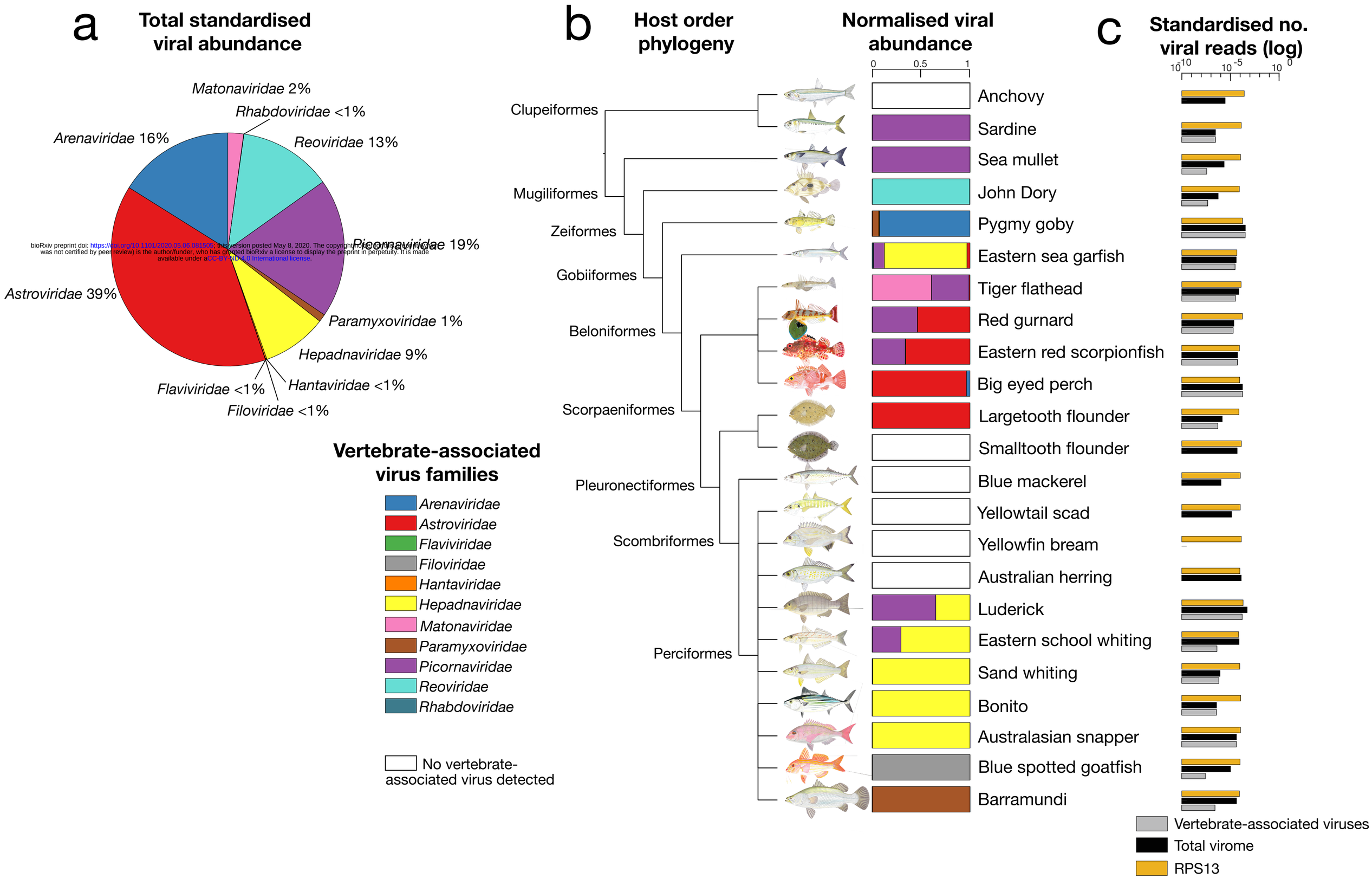
Toffan A, Panzarin V, Toson M, Cecchettin K, Pascoli F (2016). Water temperature affects pathogenicity of different betanodavirus genotypes in experimentally challenged Dicentrarchus labrax. *Dis Aquat Organ* **119:** 231-238.

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. *Aust Vet J* **72:** 421-424.

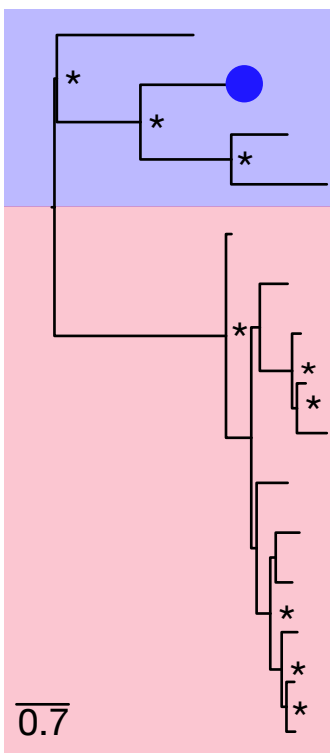
Wille M, Shi M, Klaassen M, Hurt AC, Holmes EC (2019). Virome heterogeneity and connectivity in waterfowl and shorebird communities. *ISME J* **13:** 2603-2616.

Wille M (2020). Unravelling virus community ecology in bats through the integration of metagenomics and community ecology. *Mol Ecol* **29:** 23-25.

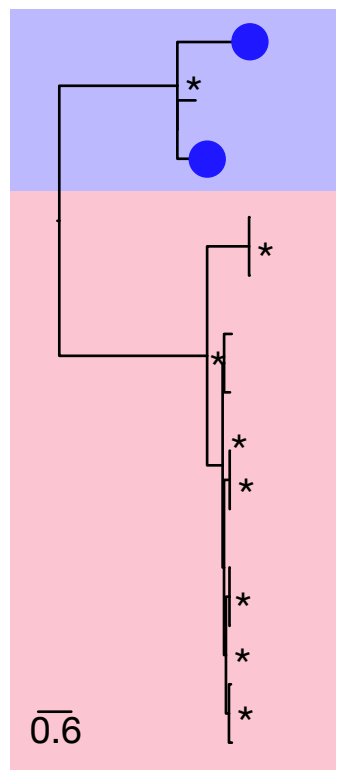
Zhang Y-Z, Shi M, Holmes EC (2018). Using metagenomics to characterize an expanding virosphere. *Cell* **172:** 1168-1172.



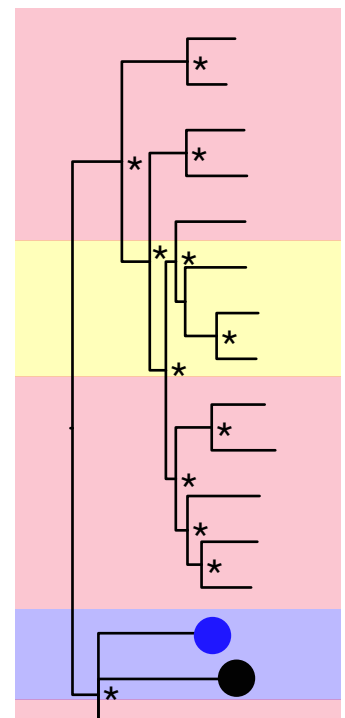
Hantaviridae



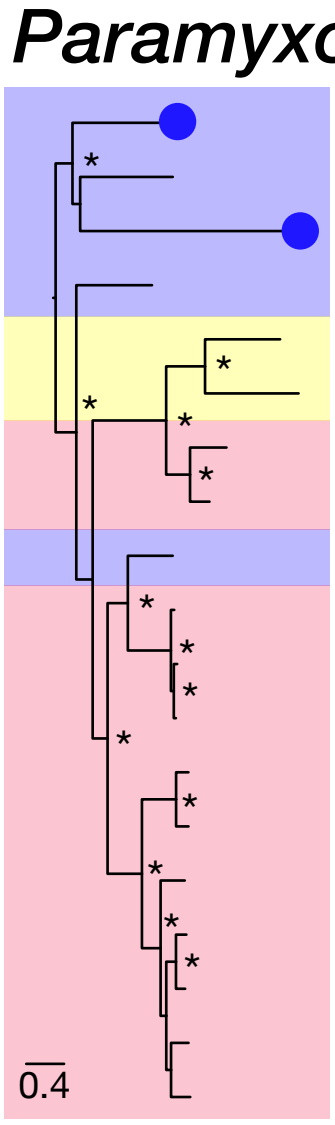
Filoviridae



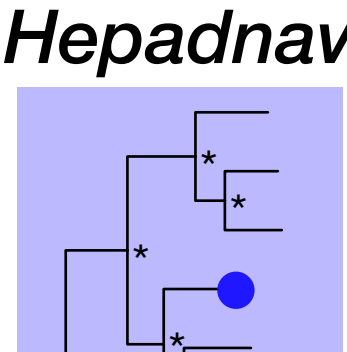
Picornaviridae



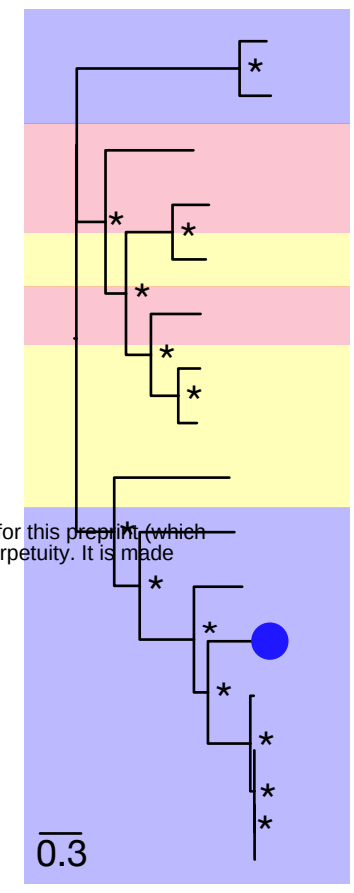
Paramyxoviridae



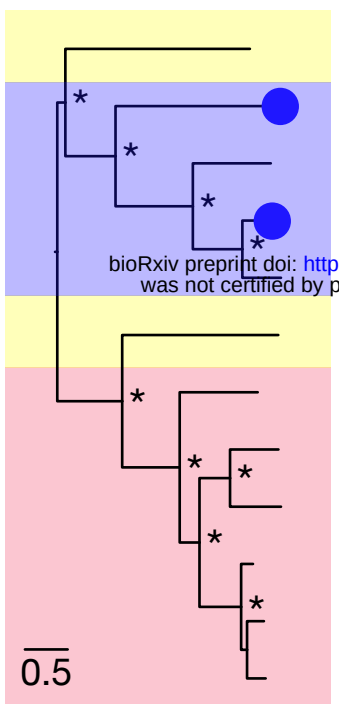
Hepadnaviridae



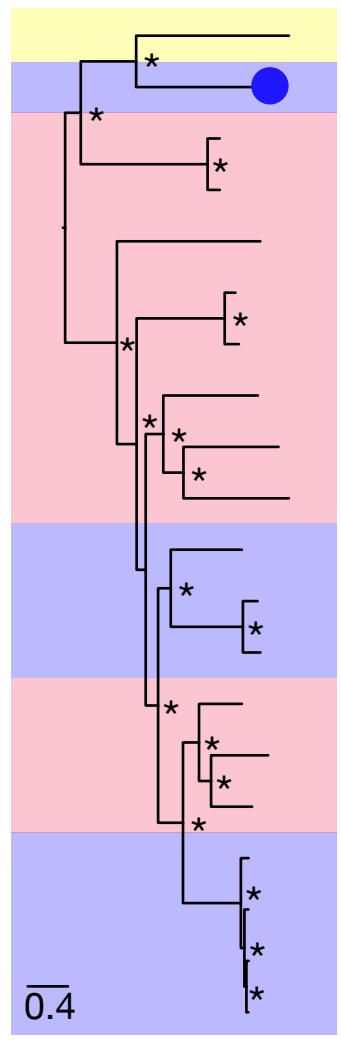
Reoviridae



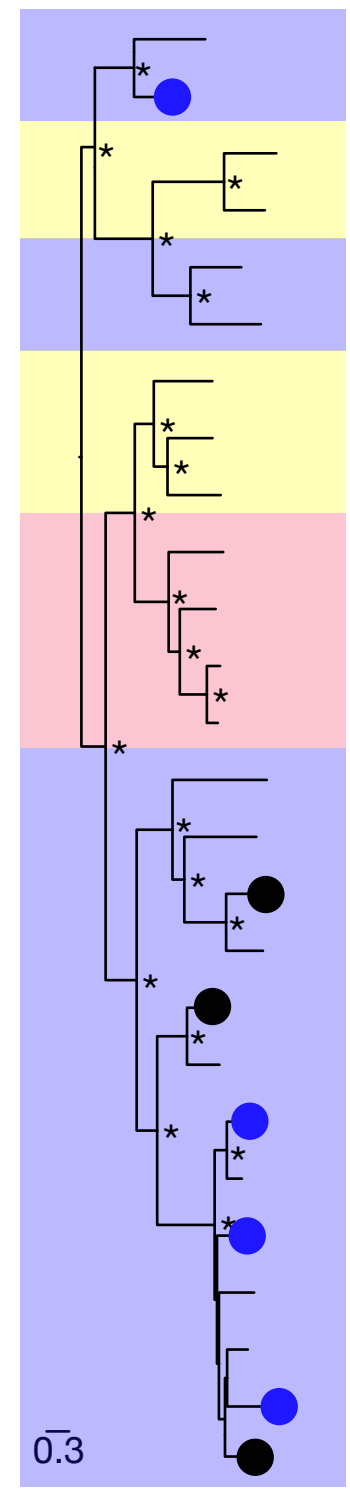
Arenaviridae



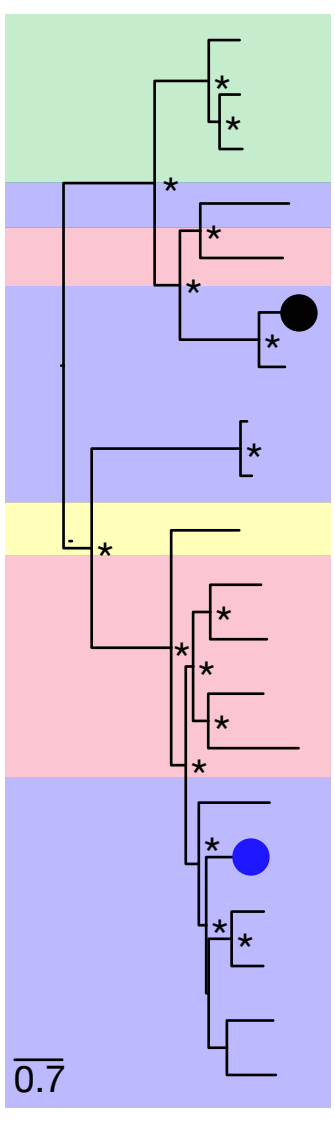
Rhabdoviridae



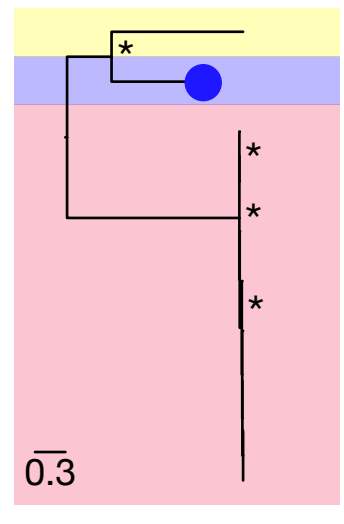
Astroviridae



Flaviviridae



Matonaviridae



fish

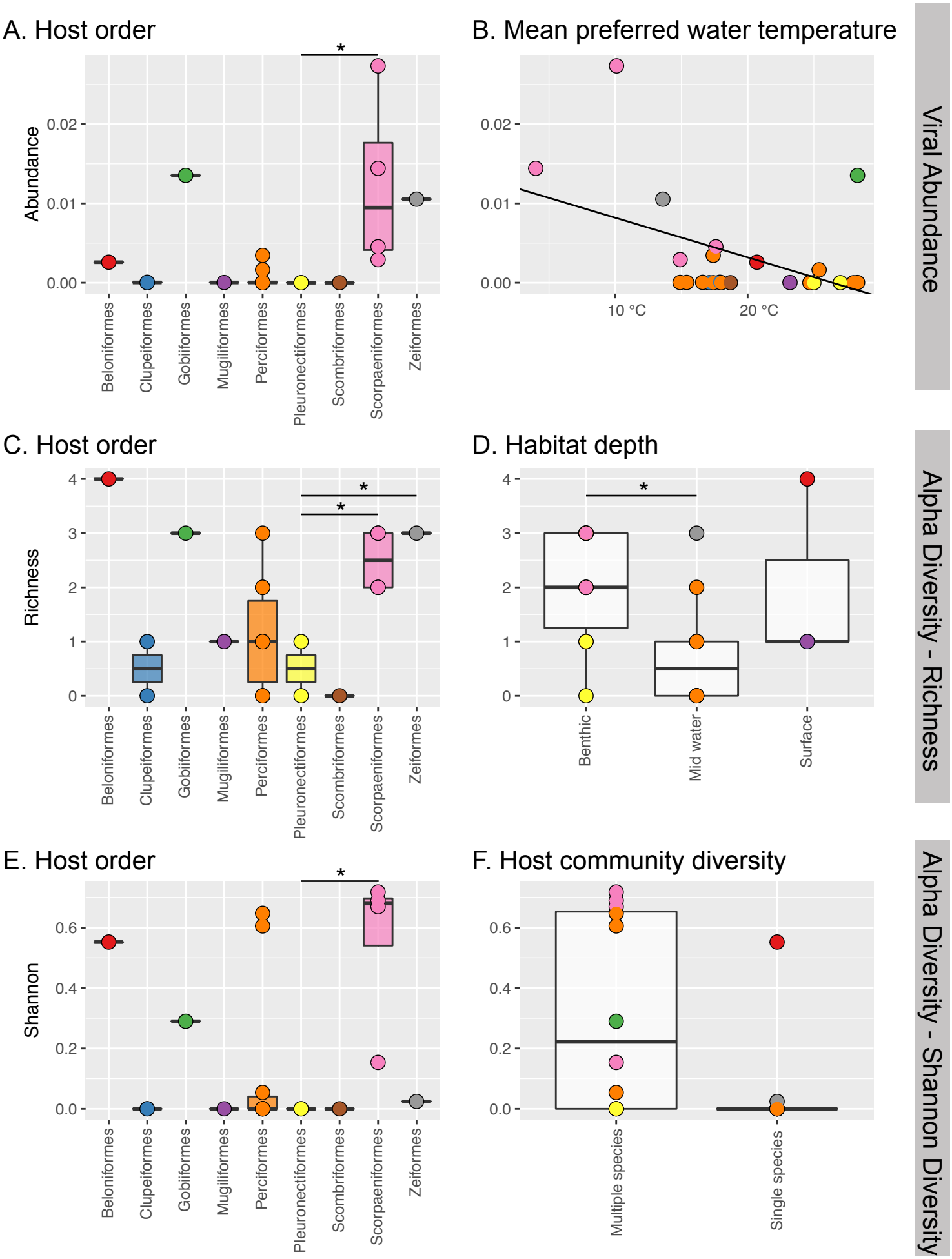
mammals

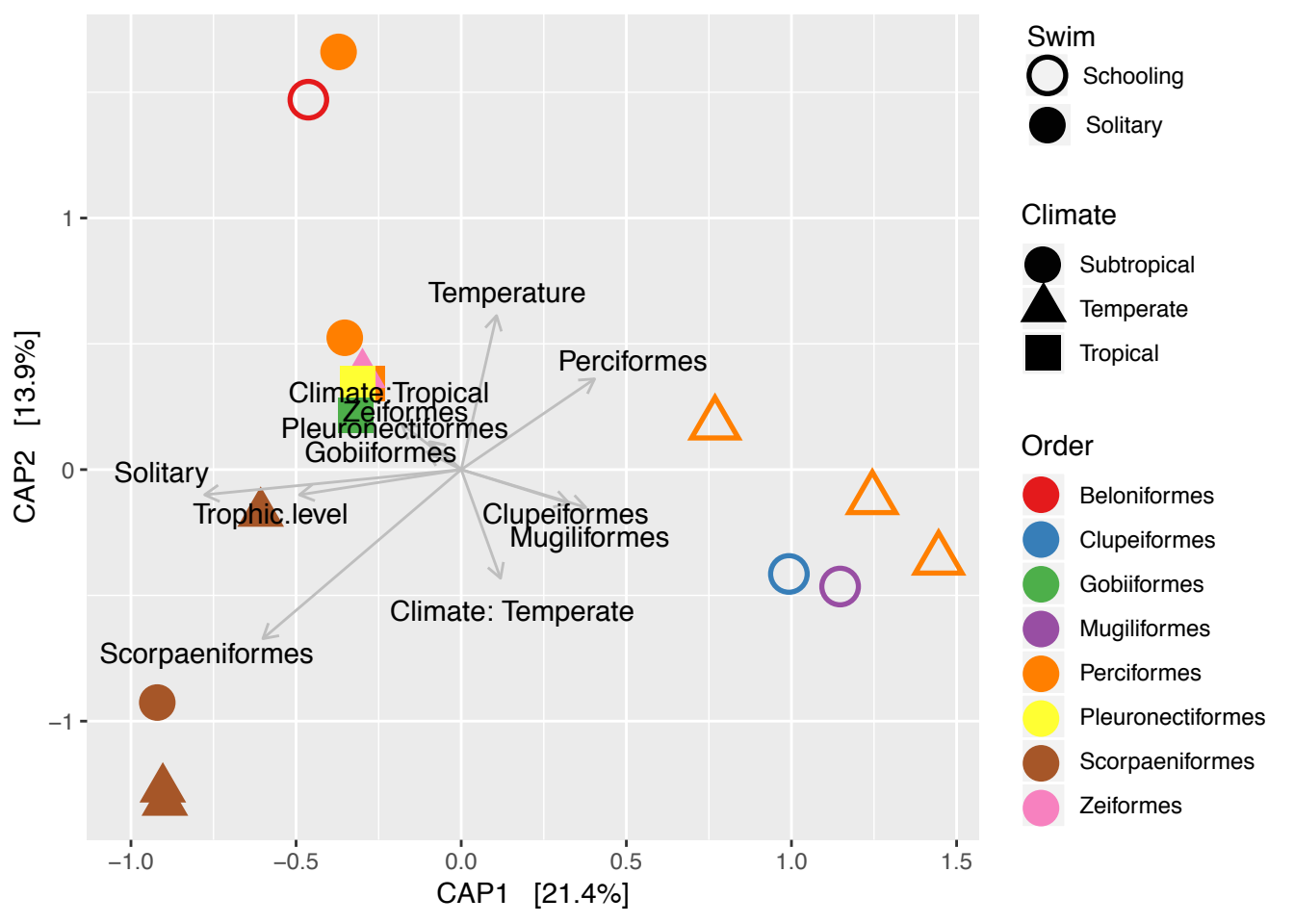
vector-borne

birds, reptiles, amphibians

novel viruses

new in (Geoghegan et al. 2018 *Virus Evolution*)


A. Host order
Abundance
0.02
0.01
0.00
*
Beloniformes
Clupeiformes
Gobiiformes
Mugiliformes
Perciformes
Pleuronectiformes
Scombriformes
Scorpaeniformes
Zeiformes
B. Mean preferred water temperature
0.02
0.01
0.00
10 °C
20 °C
Viral Abundance
C. Host order
Richness
4
3
2
1
0
*
*
Beloniformes
Clupeiformes
Gobiiformes
Mugiliformes
Perciformes
Pleuronectiformes
Scombriformes
Scorpaeniformes
Zeiformes
D. Habitat depth
4
3
2
1
0
*
Benthic
Mid water
Surface
Alpha Diversity - Richness
E. Host order
Shannon
0.6
0.4
0.2
0.0
*
Beloniformes
Clupeiformes
Gobiiformes
Mugiliformes
Perciformes
Pleuronectiformes
Scombriformes
Scorpaeniformes
Zeiformes
F. Host community diversity
0.6
0.4
0.2
0.0
Multiple species
Single species
Alpha Diversity - Shannon Diversity


Swim
Schooling
Solitary
Climate
Subtropical
Temperate
Tropical
Order
Beloniformes
Clupeiformes
Gobiiformes
Mugiliformes
Perciformes
Pleuronectiformes
Scorpaeniformes
Zeiformes
Temperature
Perciformes
Climate:Tropical
Zeiformes
Pleuronectiformes
Gobiiformes
Solitary
Trophic.level
Clupeiformes
Mugiliformes
Climate: Temperate
Scorpaeniformes
CAP2 [13.9%]
CAP1 [21.4%]
1
0
−1
−1.0
−0.5
0.0
0.5
1.0
1.5