VIRUS EVOLUTION

Virus Evolution, 2021, 7(2), 1–10
DOI: https://doi.org/10.1093/ve/veab050
Advance access publication date: 31 May 2021
Research Article

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

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

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

[†]These authors contributed equally to this work.
*Corresponding author: E-mail: jemma.geoghegan@otago.ac.nz

Abstract

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

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

1. Introduction

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

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

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

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

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

2. Results

2.1 Identification of a novel coronavirus in the pouched lamprey

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

We detected sequence reads that shared sequence identity to coronaviruses in seven of the eight LRS-affected individuals (confirmed by reverse transcription-polymerase chain reaction (RT-PCR)). No coronaviruses or other viruses were found in the remaining LRS sample nor the four control samples. However, we were only able to generate consensus virus sequences from three of these due to the low and fragmented virus abundance in the other samples. All three samples possessed virus transcripts with sequence similarity to the open reading frames 1a (ORF1a) and 1b (ORF1b) common to the *Coronaviridae*, the latter of which contains the viral RNA-dependent RNA polymerase (RdRp) (Table 1). We also identified transcripts with marked sequence similarity to the *Nidovirales* spike (S) protein. However, the closest sequence match for the S protein transcripts were viruses from *Porcine torovirus* and *Atlantic salmon bafinivirus*, both non-coronavirus members of the *Nidovirales* (family *Tobaniviridae*), albeit these were highly divergent sequences with only ∼25–27 per cent amino acid sequence similarity (Table 1). Phylogenetic comparisons of the ORF1ab and S genomic regions of kanakana letovirus in the context of its closest relatives illustrated this striking incongruence (Fig. 1). In particular, both *Microhyla letovirus* and kanakana letovirus were more closely related to coronaviruses in the ORF1ab region but not in the S protein where they exhibited greater sequence similarity to the *Tobaniviridae*, suggestive of past recombination events (Fig. 1). In contrast, *Pacific salmon nidovirus* consistently fell as a sister group to the *Coronaviridae* in both gene regions, although with long branch lengths indicative of long periods of evolutionary divergence (Fig. 1).

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

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

Data mining the SRA for coronavirus-like sequences revealed several potentially novel virus transcripts within sequences obtained from bony fish (Table 2). We detected coronavirus-like reads in 32 sequencing libraries (of the 1279 that met our search criteria) that were further investigated. First, we identified four new viruses that again shared OFR1b sequence similarity to *Pacific salmon nidovirus* (52–56 per cent). These sequences were identified in blenny (*Acanthemblemaria sp.*), bluespotted mud hopper (*Boleophthalmus pectinirostris*), *Schizothorax waltoni*, which were sampled from China, as well as the broadnosed pipefish (*Syngnathus typhle*), sampled from Europe. The new viruses were named blenny letovirus, mud hopper letovirus, waltoni letovirus, and broadnosed pipefish letovirus, respectively. Second, we identified a viral transcript in the trout perch (*Percopsis omiscomaycus*) that shared 39 per cent amino acid sequence similarity to *Shrew coronavirus* in ORF1b and 20 per cent amino acid sequence similarity to *Bat coronavirus* in the S protein (Table 2), although it is important to note that the vast majority of viruses in the *Letovirinae* are not yet included in the non-redundant (nr) protein database. This was termed 'trout perch letovirus'. Finally, two potentially novel virus sequences were identified that, upon further inspection, shared greater sequence similarity to viruses within the *Tobaniviridae*, in a similar manner to the S protein of kanakana letovirus. These two novel viruses, found in giant mudskipper (*Periophthalmodon schlosseri*) and golden Chinese loach (*Botia superciliaris*), had >50 per cent sequence similarity in the RdRp and ∼40 per cent sequence similarity in the S protein to other fish viruses in the *Tobaniviridae* (see Table 1). These new viruses were named giant mudskipper bafinivirus and golden Chinese loach bafinivirus, respectively. The remaining 25 libraries that were further investigated did not possess transcripts with sequence similarity to the *Coronaviridae*.

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

Virus	LRS symptoms	Host	Gene	Sequence length (nt)	Closest match on GenBank (accession); e-value	Sequence identity (%)	Standardised viral abundance	RSP13 abundance
Kanakana letovirus (1)	Light haemorrhaging around gill slits.	*Geotria australis*	ORF1a	9,270	Pacific salmon nidovirus ORF1a (QEG08236.1); e-value = 1×10^{-155}	26.36	6.86×10^{-5}	5.52×10^{-6}
			ORF1b	6,507	Pacific salmon nidovirus ORF1b (QEG08237.1); e-value = 0	46.53	4.67×10^{-5}	
			Spike	5,022	Porcine torovirus spike protein (QBJ02013.1); e-value = 2×10^{-10}	27	4.24×10^{-5}	
Kanakana letovirus (2)	Small area of haemorrhaging near the ventral fin.	*Geotria australis*	ORF1a	9,255	Pacific salmon nidovirus ORF1a (QEG08236.1); e-value = 5×10^{-49}	27.81	1.48×10^{-5}	6.75×10^{-6}
			ORF1b	6,501	Pacific salmon nidovirus ORF1b (QEG08237.1); e-value = 4×10^{-175}	46.30	1.26×10^{-5}	
			Spike	4,989	Atlantic salmon bafinivirus spike protein (AUE23862.1); e-value = 2×10^{-4}	24.07	6.4×10^{-6}	
Kanakana letovirus (3)	Haemorrhaging on dorsal fins and gill slits. Pale liver with haemorrhaging, swollen kidney, adhesions between organs and muscles.	*Geotria australis*	ORF1a	9,270	Pacific salmon nidovirus ORF1a (QEG08236.1); e-value = 2×10^{-156}	26.22	6.38×10^{-5}	6.40×10^{-6}
			ORF1b	6,509	Pacific salmon nidovirus ORF1b (QEG08237.1); e-value = 0	43.97	4.20×10^{-5}	
			Spike	5,022	Porcine torovirus spike protein (QBJ01983.1); e-value = 1×10^{-9}	27.18	4.48×10^{-5}	

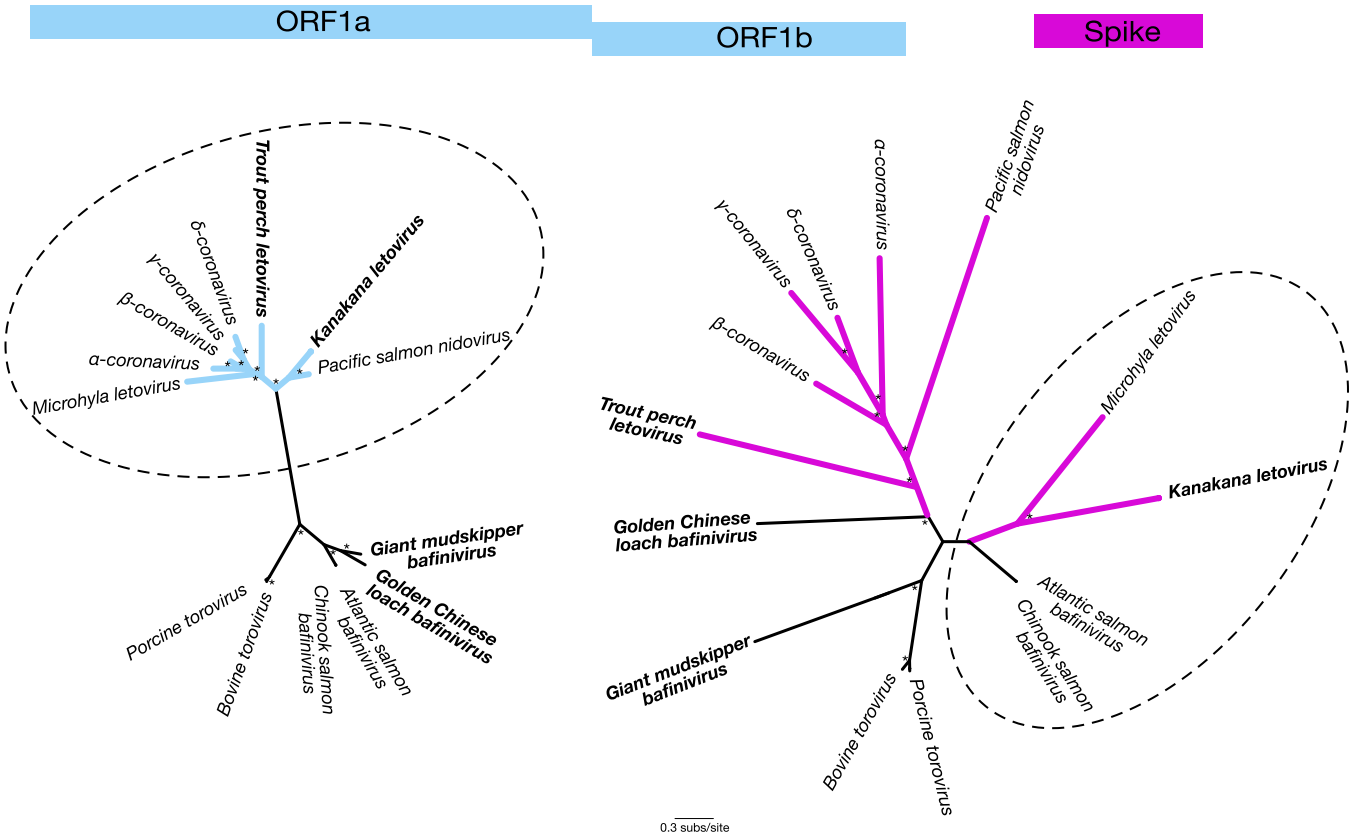

Figure 1. Unrooted phylogenetic trees for the open reading frame 1ab region (blue) and the spike protein (pink). Branches for viruses classified within the *Tobaniviridae* are shown in black while all other branches are coloured. A dashed line circles those taxa that are closest known relatives of kanakana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

2.3 The evolutionary history of novel coronaviruses in aquatic vertebrates

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

Although there are broad patterns of host clustering across the coronaviruses as a whole with, for example, most of the mammalian coronaviruses falling in the genera *Alphacoronavirus* and *Betacoronavirus*, and the genus *Deltacoronavirus* comprising avian-associated viruses only, it is notable that the phylogenetic position of the *Letovirinae* does not follow that of their hosts (Fig. 4A). Of particular note, kanakana letovirus sampled from New Zealand pouched lamprey falls as are a relative ingroup in this subfamily (Fig. 3). Similar patterns can be seen for viruses in amphibians, indicative of major cross-species transmission events. For example, letoviruses identified from amphibian hosts are highly diverse, with the axolotl letovirus more closely related to viruses in fish and the frog letovirus falling in a more divergent position, forming an outgroup to other viruses in this sub-family (Fig. 3). Indeed, phylogenetic reconciliation analysis found that host-jumping was consistently the most frequent evolutionary event, although there was also clear evidence for host-virus co-divergence that likely reflects deeper branching events in the evolutionary history of these viruses (Fig. 4B).

2.4 Amino acid conservation across the *Coronaviridae*

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

3. Discussion

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

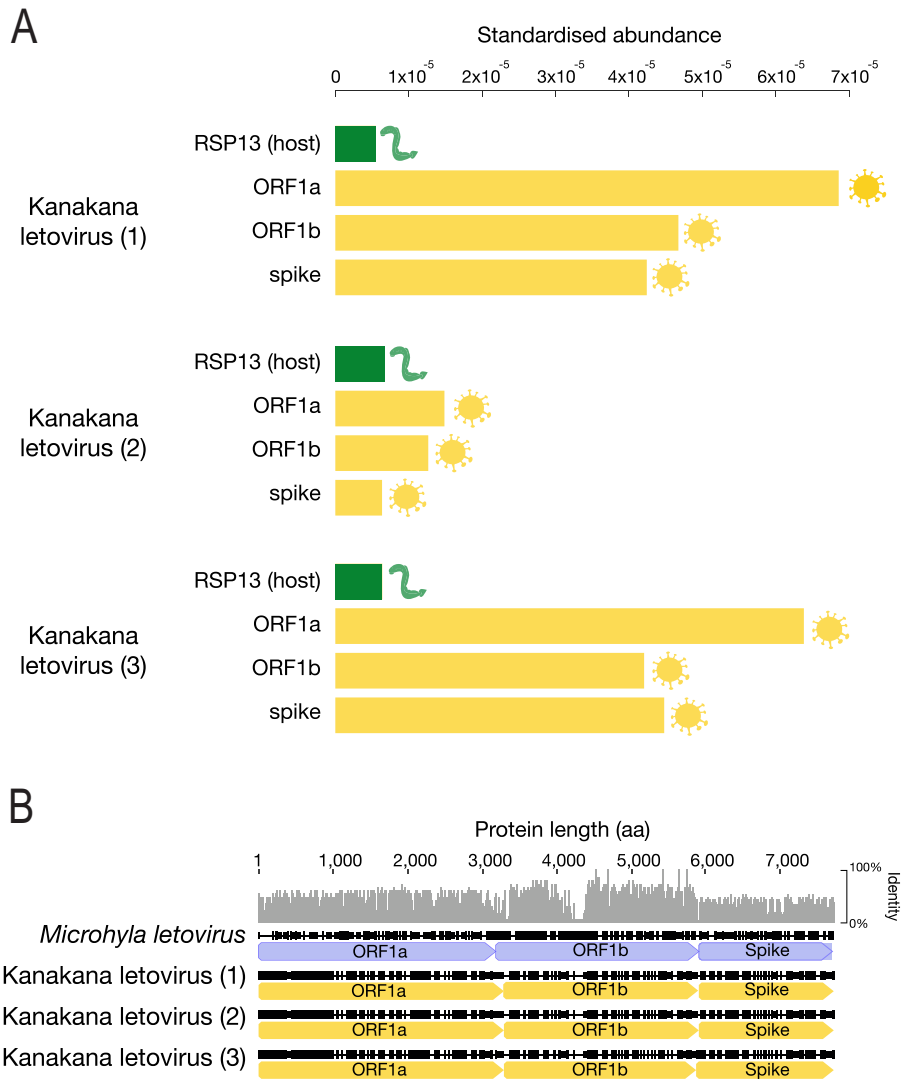

Figure 2. (A) Standardised abundance of kanakana letovirus transcripts (yellow) in comparison to the lamprey host gene, RSP13 (green), across the three lamprey samples that contained the highest abundance of the novel coronavirus. (B) Schematic alignment of all three transcripts against *Microhyla letovirus*,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

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

The discovery of these new viruses not only expands the phylogenetic diversity of the *Letovirinae*, but it also sheds important new light on the evolution and host range in the *Coronaviridae* as a whole. In particular, our study revealed that coronavirus hosts include both freshwater and marine fish spanning multiple taxonomic orders. This includes an ancient lineage of jawless fish, herein represented by the pouched lamprey from New Zealand, that represents one of the evolutionary oldest vertebrates still to exist and one that has experienced little morphological change for around 360 million years (Gess, Coates, and Rubidge 2006), although its true age has recently been debated (Miyashita et al. 2021). These jawless and bony fish hosts were also sampled across a range of geographic localities: New Zealand, China, Europe and North America. Although divergent in their sequences, conserved motifs across the entire virus family strongly support the inclusion of these new viruses within the *Coronaviridae* and more specifically the *Letovirinae*. With the expansion of virological surveys of wildlife, it is likely that additional viruses within the *Coronaviridae* will be discovered. Indeed, the long branch lengths separating the divergent families within the *Nidovirales* are indicative of very limited sampling across this order of RNA viruses.

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

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

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

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

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

That *Letoviruses* form a sister clade to all other coronaviruses suggests that the backbone of the *Coronaviridae* phylogeny overall mirrors that of their vertebrate hosts, with different viral genera associated with different vertebrate classes, and the clear transition from aquatic to land (Zhang et al. 2018). Nevertheless, lamprey represent an ancient lineage of jawless fish that fall sister to all other vertebrates, including jawed fish, amphibians, and reptiles. In this context, it is particularly notable that the novel coronavirus identified here in the pouched lamprey does not fall basal to other fish coronaviruses, in turn revealing that cross-species transmission events have played a major role in the evolutionary history of these viruses and which is also apparent from our co-phylogenetic reconciliation analysis. Hence, as is the case with many families of RNA viruses (Geoghegan, Duchêne, and Holmes 2017), the broad association between different coronaviruses and their vertebrate host classes suggests that the evolutionary history of this group likely dates to at least the origin of the vertebrates some 530 million years ago. However, on top of this overall pattern of long-term virus-host co-divergence, there have also been frequent instances of cross-species virus transmission. As the case of the pouched lamprey demonstrates, such host jumping events may be particularly commonplace in hosts that reside in aquatic environments, perhaps reflecting a combination of evolutionary antiquity (and hence more long-term opportunity for cross-species transmission) and greater physical interactions between multi-species assemblages, although this clearly needs to be investigated in greater detail. Indeed, the parasitic behaviour of lamprey might facilitate viral transmission.

4. Materials and methods

4.1 Lamprey sample collection in Aotearoa New Zealand and RNA sequencing

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

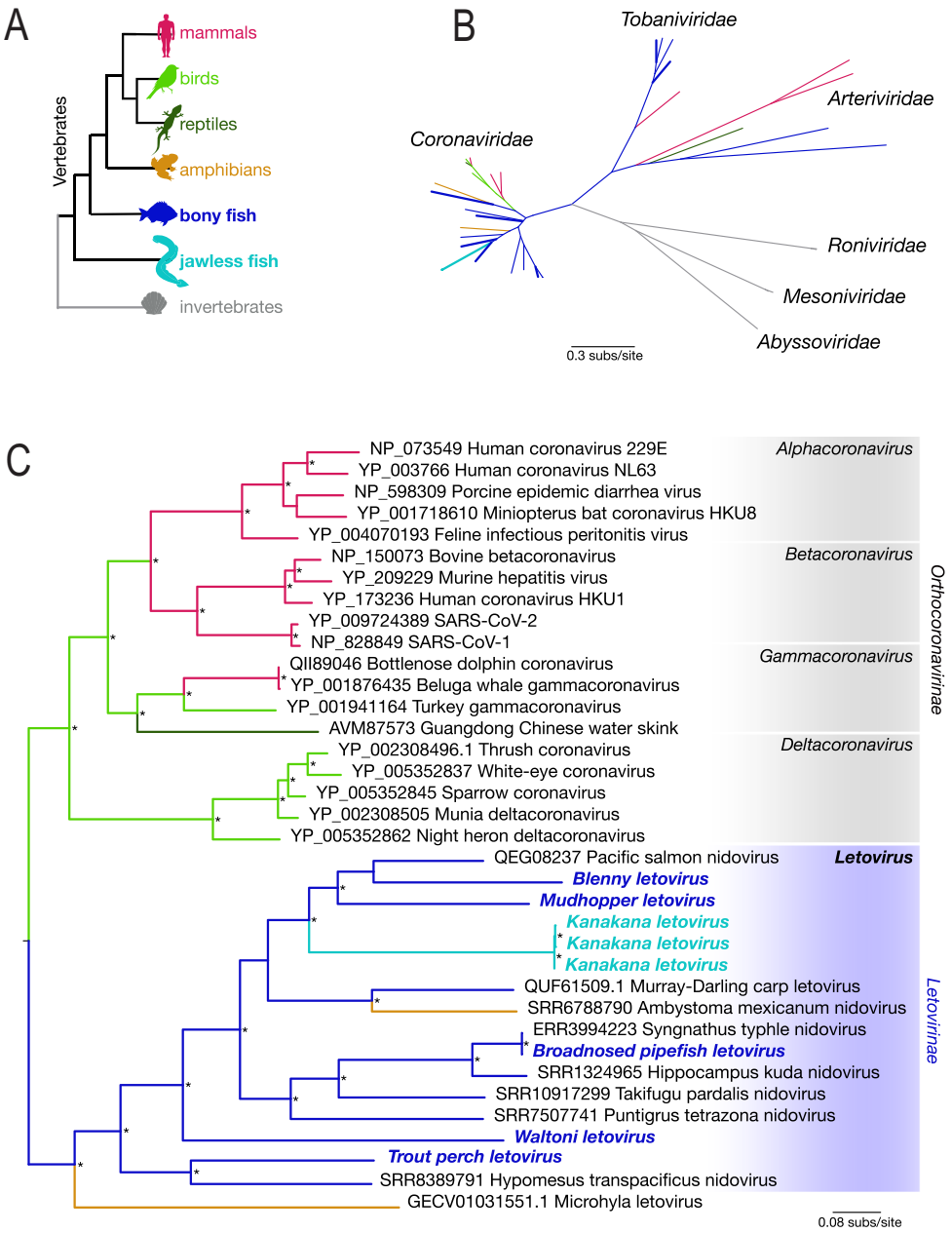

Figure 3. (A) Schematic phylogram showing where bony and jawless fish fall in the greater animal phylogeny and colour-coded to signify the host class in (B) and (C), informed by the relevant literature on vertebrate evolution. (B) An order-level phylogenetic tree showing the relationships between families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

and these tissues were then archived in formalin-fixed paraffin-embedded (FFPE) blocks and stored at room temperature until they were sent to the University of Otago for RNA sequencing in 2020. Fresh tissues were obtained from four additional *Geotria australis* individuals, collected on separate field expeditions from the Mokau River, Aria, in 2015 ($n = 2$) and Okuti River, Banks Peninsula, in 2019 ($n = 2$), both of which are separate catchments areas to locations sampled in 2012. These individuals displayed no signs of LRS that has never been recorded from the Mokau or Okuti rivers. These samples were used as controls. Muscle and skin subsamples from the control individuals were taken in the field and were preserved in RNALater (Thermo Fisher Scientific) and in a −80°C freezer. Sampling was conducted in consultation with local Māori (rūnanga and kaumatua).

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

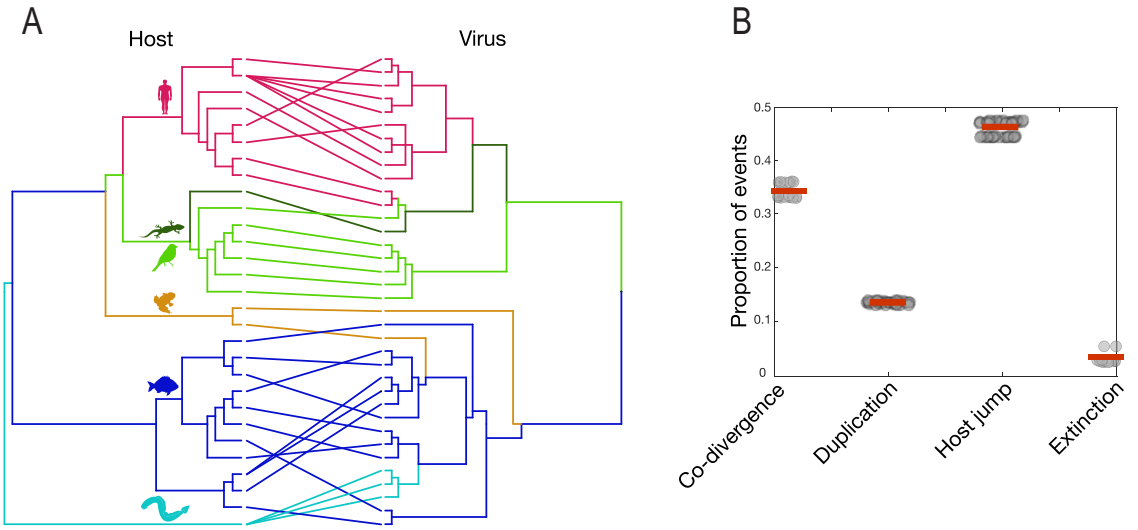

Figure 4. (A) Tanglegram of rooted phylogenetic trees for the *Coronaviridae* and their hosts. Lines and branches are coloured to represent host class, as in Fig. 2. The 'untangle' function was used in TreeMap3.0 to maximise the congruence between the host (left) and virus (right) phylogenies. Supplementary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

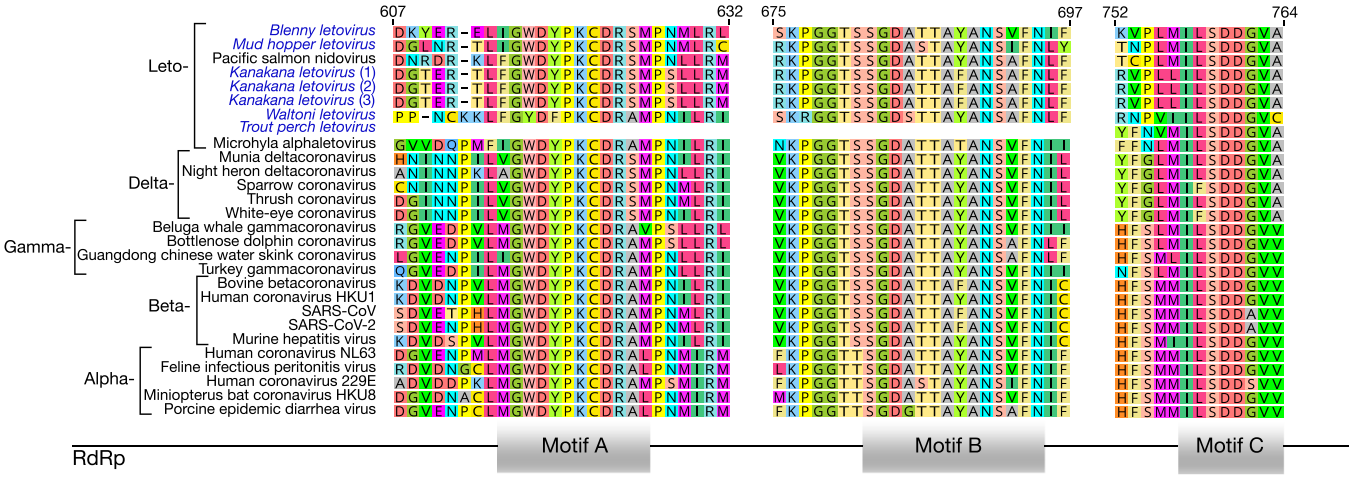

Figure 5. Multiple virus amino acid sequence alignment of the RdRp showing three motifs (A, B and C) conserved throughout all known genera within the *Coronaviridae*, including the newly discovered viruses within the *Letovirinae* for which virus transcripts in this region were identified.

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

4.2 Data mining the Sequence Read Archive for coronaviruses

To identify additional coronavirus-like sequences, we screened bony fish transcriptomes present in the SRA. We focused on a subset of available SRA libraries, particularly those RNA-seq libraries without sequencing selection (e.g. poly(A)+ -selected). We excluded zebrafish (*Danio rerio*) runs from the screens due to the large number of experimental data sets. To reduce redundancy, we selected the first six runs for a given species in each experimental data set. In total, 1,279 libraries (both single- and paired-end reads) were selected. SRA runs were obtained through the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) that mirrors the SRA but allows for reads to be downloaded directly in a FASTQ. Reads were then analysed using our coronavirus discovery pipeline. The pipeline consisted of an initial screen that aligned the reads against a custom database of *Coronaviridae* and other protein sequences within the *Nidovirales* using Diamond BLASTx (v.2.0.2) (Buchfink, Xie, and Huson 2015) with an e-value threshold of 1×10^{-5}. Results were manually inspected in Geneious Prime (www.geneious.com) to exclude spurious alignments to low-complexity regions. Where coronavirus-like reads were detected ($n = 32$), we quality trimmed

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

4.3 Transcript sequence similarity searching for novel coronaviruses

Sequencing reads were first quality trimmed and then assembled *de novo* using Trinity RNA-Seq (v.2.11.0) (Haas et al. 2013). The assembled contigs were annotated based on similarity searches against the NCBI nucleotide (nt) and non-redundant protein (nr) databases using BLASTn (Altschul et al. 1990) and Diamond BLASTx (v.2.0.2) (Buchfink, Xie, and Huson 2015).

4.4 Phylogenetic analysis

To infer the evolutionary relationships of the coronaviruses, newly discovered translated viral contigs were combined with representative amino acid sequences from both the *Coronaviridae* and the order *Nidovirales*. All these sequences were obtained from NCBI GenBank. The sequences retrieved were then aligned with those generated here using MAFFT (v7.4) (Katoh et al. 2002) employing the E-INS-i algorithm. Ambiguously aligned regions were removed using trimAl (v.1.2) (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009). To estimate phylogenetic trees, we utilised the maximum likelihood approach available in IQ-TREE (v 1.6.8) (Nguyen et al. 2015), selecting the best-fit model of amino acid substitution, LG, with ModelFinder (Kalyaanamoorthy et al. 2017), and using 1,000 bootstrap replicates. Phylogenetic trees were annotated with FigTree (v.1.4.2).

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

4.5 Estimation of virus abundance

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

4.6 PCR confirmation

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

4.7 Etymology of virus names

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

Data availability

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

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

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

Funding

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

Conflict of interest: None declared.

References

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

Babaian, A., and Edgar, R. C. (2021) 'Ribovirus Classification by a Polymerase Barcode Sequence', *bioRxiv*. 10.1101/2021.03.02.433648

Brosnahan, C. L. et al. (2019) 'Lamprey (Geotria australis; Agnatha) Reddening Syndrome in Southland Rivers, New Zealand 2011–2013: Laboratory Findings and Epidemiology, Including the Incidental Detection of an Atypical Aeromonas salmonicida', *New Zealand Journal of Marine and Freshwater Research*, 53: 416–36.

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

Bukhari, K. et al. (2018) 'Description and Initial Characterization of Metatranscriptomic Nidovirus–Like Genomes from the Proposed New Family Abyssoviridae, and from a Sister Group to the Coronavirinae, the Proposed Genus Alphaletovirus', *Virology*, 524: 160–71.

Cano, I. et al. (2020) 'Isolation of a Chinook Salmon Bafinivirus (CSBV) in Imported Goldfish Carassius auratus L. in the United Kingdom and Evaluation of Its Virulence in Resident Fish Species', *Viruses*, 12: 578.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009) 'trimAl: A Tool For Automated Alignment Trimming in Large-Scale Phylogenetic Analyses', *Bioinformatics*, 25: 1972–3.

Charleston, M. A. (2011), *TreeMap* <https://sites.google.com/site/cophylogeny/home> accessed 18 Mar 2021.

Costa, V. A. et al. (2021) 'Metagenomic Sequencing Reveals a Lack of Virus Exchange between Native and Invasive Freshwater Fish across the Murray-Darling Basin, Australia', *Virus Evolution*, 7(1): veab034.

Denison, M. R. et al. (2011) 'Coronaviruses: An RNA Proofreading Machine Regulates Replication Fidelity and Diversity', *RNA Biology*, 8: 270–9.

Dunn, N. R. et al. (2018) *Conservation Status of New Zealand Freshwater Fishes, 2017. New Zealand Threat Classification Series 24*, Department of Conservation, New Zealand.

Edgar, R. C. et al. (2020) 'Petabase-Scale Sequence Alignment Catalyses Viral Discovery', *bioRxiv*. 10.1101/2020.08.07.241729

Geoghegan, J. L., Duchêne, S., and Holmes, E. C. (2017) 'Comparative Analysis Estimates the Relative Frequencies of Co-divergence and Cross-species Transmission within Viral Families', *PLoS Pathogens*, 13(2): e1006215.

Gess, R. W., Coates, M. I., and Rubidge, B. S. (2006) 'A Lamprey from the Devonian Period of South Africa', *Nature*, 443: 981–4.

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, Gorbalenya, A. E. et al. (2020) 'The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-nCoV and Naming It SARS-CoV-2', *Nature Microbiology*, 5: 536–44.

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

Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods*, 14: 587–9.

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

Li, B., and Dewey, C. N. (2011) 'RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome', *BMC Bioinformatics*, 12: 323.

Miyashita, T. et al. (2021) 'Non-ammocoete Larvae of Palaeozoic Stem Lampreys', *Nature*, 591: 408–12.

Mordecai, G. J. et al. (2019) 'Endangered Wild Salmon Infected by Newly Discovered Viruses', *eLife*, 8: e47615.

Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.

Santichaivekin, S. et al. (2020) 'eMPRess: A Systematic Cophylogeny Reconciliation Tool', *Bioinformatics*, btaa978.

Shi, M. et al. (2018) 'The Evolutionary History of Vertebrate RNA Viruses', *Nature*, 556: 197–202.

Snijder, E. J. et al. (2003) 'Unique and Conserved Features of Genome and Proteome of SARS-coronavirus, an Early Split-off from the Coronavirus Group 2 Lineage', *Journal of Molecular Biology*, 331: 991–1004.

Wu, F. et al. (2020) 'A New Coronavirus Associated with Human Respiratory Disease in China', *Nature*, 579: 265–9.

Zhang, Y.-Z. et al. (2018) 'The Diversity, Evolution and Origins of Vertebrate RNA Viruses', *Current Opinion in Virology*, 31: 9–16.