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. Grimwood³, Jane Kitson⁴, Cindy Baker⁵, Cara L. Brosnahan⁶, Anjali Pande^{5,6}, Edward C. Holmes⁷, Neil J. Gemmell¹, Jemma L. Geoghegan^{3,8*}

¹Department of Anatomy, University of Otago, Dunedin, New Zealand.

²Department of Biological Sciences, Macquarie University, Sydney, Australia.

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

⁴Kitson Consulting Ltd, Invercargill/Waihopai, New Zealand.

⁵National Institute of Water and Atmospheric Research, New Zealand.

⁶Animal Health Laboratory and Diagnostic and Surveillance Directorate, Ministry for Primary Industries, Upper Hutt, New Zealand.

⁷Marie 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.

⁸Institute of Environmental Science and Research, Wellington 5018, New Zealand.

#authors contributed equally

*Correspondence: jemma.geoghegan@otago.ac.nz

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

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 relationships 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.

45 Introduction

46 The *Coronaviridae* are a family of enveloped, positive-strand RNA viruses that have a broad host
 47 range, although they are predominantly associated with mammals and birds. Several
 48 coronaviruses have emerged as highly pathogenic viruses in humans and other animals, the most
 49 recent and notable of which is *Severe acute respiratory syndrome coronavirus-2* (SARS-CoV-2)
 50 (Wu et al. 2020) which, at the time of writing, has infected over 100 million people since its initial
 51 identification at the end of 2019. Coronaviruses, like other families within the order *Nidovirales*,
 52 possess the longest genomes seen in RNA viruses (27-34 kb in the case of the coronaviruses),
 53 and seemingly replicate with higher fidelity than other known positive-sense RNA viruses due to
 54 RNA proofreading associated with an exonuclease domain (Denison et al. 2011).

55 Most coronaviruses are classified within the sub-family *Orthocoronavirinae* that infect mammals
 56 and birds, as well as a single virus identified in a reptilian host (Shi et al. 2018). However, our
 57 understanding of the host range of the *Coronaviridae* has recently been enhanced by large-scale
 58 metagenomic studies, including the identification of a distinct second sub-family - the *Letovirinae*.
 59 Specifically, an analysis of published transcriptome data identified a novel coronavirus in the
 60 ornamented pygmy frog (*Microhyla fissipes*) termed *Microhyla letovirus* (Bukhari et al. 2018). A
 61 subsequent virological survey of dead and moribund cultured Chinook salmon (*Oncorhynchus*
 62 *tshawytscha*) identified *Pacific salmon nidovirus* (Mordecai et al. 2019). Phylogenetic analysis
 63 revealed that *Microhyla letovirus* and *Pacific salmon nidovirus* formed a sister group to all other
 64 known coronaviruses, with the large genetic divergence between them suggesting that this group
 65 is likely far larger and under-sampled (Bukhari et al. 2018; Mordecai et al. 2019). It is therefore
 66 unsurprising that more recent probes of public sequencing data have identified additional
 67 coronavirus-like sequences in non-mammalian aquatic vertebrate hosts, namely in fish, reptile and
 68 amphibian transcriptomes (Edgar et al. 2020). Indeed, this extensive data mining study identified
 69 six novel viruses within the *Letovirinae*: four in bony fish, one in the axolotl (*Ambystoma*
 70 *mexicanum*) and a further, more divergent virus in the loggerhead sea turtle (*Caretta caretta*)
 71 (Edgar et al. 2020).

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 bony and jawless fish sampled in New Zealand and Australia, combined 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.

Results

Identification of a novel coronavirus in the common carp

As part of a large virological survey of freshwater fish across the Murray-Darling Basin in Australia, ten species of native and invasive bony fish were caught and subjected to meta-transcriptomic analysis (Costa et al. 2021). This led to the identification of a novel member of the *Coronaviridae*, termed *Murray-Darling carp letovirus* (Costa et al. 2021). This partial virus sequence was obtained from common carp sampled from Edward River, New South Wales. This divergent viral sequence shared ORF1b sequence similarity with *Pacific salmon nidovirus* (51%), as well as with various gammacoronaviruses, including *Bottlenose dolphin coronavirus* and *Beluga whale coronavirus* (46%), and had a standardised abundance of 1.2×10^{-8} (Table 1). The phylogenetic position of this virus is discussed in more detail below and elsewhere (Costa et al. 2021).

Table 1. Novel viruses identified in this study from bony fish in Australia and by mining the Sequence Read Archive (SRA), their detection method and sampling locations.

Virus	Host	Detection method (NCBI accession of SRA data)	Sampling location	Contig length (nt) matching RdRp (ORF1b) of coronavirus	Closest amino acid match; GenBank accession
<i>Murray-Darling carp letovirus</i>	<i>Cyprinus carpio</i>	Fish sampling + meta-transcriptomics	Australia (Edward River, New South Wales)	225	<i>Pacific salmon nidovirus</i> ORF1b (50%); QEG08237.1
<i>Blenny letovirus</i>	<i>Acanthemblemaria</i> sp.	SRA search (SRR5997671)	Caribbean	4995	<i>Pacific salmon nidovirus</i> ORF1b (56%); QEG08237.1
<i>Mud hopper letovirus</i>	<i>Boleophthalmus pectinirostris</i>	SRA search (SRR5012117)	China	4170	<i>Pacific salmon nidovirus</i> ORF1b (56%); QEG08237.1
<i>Waltoni letovirus</i>	<i>Schizothorax waltoni</i>	SRA search (SRR6233344)	China	906	<i>Pacific salmon nidovirus</i> ORF1b (52%); QEG08237.1
<i>Broadnosed pipefish letovirus</i>	<i>Syngnathus typhle</i>	SRA search (SRR663248)	Europe	225	<i>Pacific salmon nidovirus</i> ORF1b (55%); QOQ03052.1
<i>Trout perch letovirus</i>	<i>Percopsis omiscomaycus</i>	SRA search (SRR5997814)	North America	2046	<i>Shrew coronavirus</i> ORF1b (39%); YP_009755838.1
				3942	<i>Bat coronavirus</i> spike protein (20%); QBP43256.1
<i>Giant mudskipper bafinivirus</i>	<i>Periophthalmodon schlosseri</i>	SRA search (SRR5012122)	China	6786	<i>White bream virus</i> ORF1b (53%); YP_803213.1
				3519	<i>Fathead minnow nidovirus</i> spike protein (40%); YP_009505583
	<i>Botia supercilialis</i>		China	6996	<i>White bream virus</i>

Golden Chinese loach bafinivirus		SRA search (SRR5997844)			ORF1b (57%) YP_803213.1
				3618	White bream virus Spike protein (42%) YP_803215.1

100

101 **Identification of a novel coronavirus in the pouched lamprey**

102 Following a disease investigation into Lamprey Reddening Syndrome (LRS) in pouched lamprey
103 from New Zealand (*Geotria australis*) – te reo Māori: kanakana (South Island), or piharau (North
104 Island) – by the New Zealand Ministry for Primary Industries (MPI) in 2012, 12 individuals (eight
105 from this investigation and four more recently collected control samples) were subjected to meta-
106 transcriptomic analysis. This led to the identification of a novel and divergent member of the
107 *Coronaviridae* in several individuals. We tentatively named this new virus *Kanakana letovirus*, and
108 as each virus shared >97% sequence similarity at the nucleotide level they are likely genomic
109 variants of the same virus species.

110 We detected sequence reads that shared sequence identity to coronaviruses in seven of the eight
111 LRS-affected individuals (confirmed by RT-PCR). No coronaviruses or other viruses were found in
112 the remaining LRS sample nor the four control samples. However, we were only able to generate
113 consensus virus sequences from three of these due to the low and fragmented virus abundance in
114 the other samples. All three samples possessed virus transcripts with sequence similarity to the
115 open reading frames 1a (ORF1a) and 1b (ORF1b) common to members of the *Coronaviridae*, the
116 latter of which contains the viral RNA-dependent RNA polymerase (RdRp) (Table 2). We also
117 identified transcripts with marked sequence similarity to the *Nidovirales* spike (S) protein. However,
118 the closest sequence match for the S protein transcripts were viruses from *Porcine torovirus* and
119 *Atlantic salmon bafinivirus*, both non-coronavirus members of the *Nidovirales* (family
120 *Tobamiviridae*), albeit these were highly divergent sequences with only ~25-27% amino acid
121 sequence similarity (Table 2). Phylogenetic comparisons of the ORF1ab and S genomic regions of
122 *Kanakana letovirus* in the context of its closest relatives illustrated this striking incongruence
123 (Figure 1). In particular, both *Microhyla letovirus* and *Kanakana letovirus* were more closely related
124 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 (Figure 1). In contrast, *Pacific salmon nidovirus* consistently fell as a sister-group the *Coronaviridae* in both gene regions, although with long branch lengths indicative of long periods of evolutionary divergence (Figure 1)

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% are indicated with an asterisk (*).

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* (Figure 2). However, the *Kanakana letovirus* in samples 1 and 3 were the

141 most abundant across the ORF1ab and S proteins. The presence of *Kanakana letovirus* in seven
142 of the individual lampreys was confirmed by RT-PCR (Supplementary Figure 1). No other viruses
143 were detected in these data.

144

145

146 **Figure 2.** Standardised abundance of *Kanakana letovirus* transcripts (yellow) in comparison to the
147 lamprey host gene, RSP13 (green), across the three lamprey samples that contained the highest
148 abundance of the novel coronavirus.

149

150

151

152

153

154

Table 2. Novel virus transcripts identified in this study from pouched lamprey (*Geotria australis*), their corresponding closest genetic match on GenBank, and standardised abundances.

Virus	LRS symptoms	Host	Gene	Sequence length (nt)	Closest match on GenBank (accession); e-value	Sequence identity	Standardised viral abundance	RSP13 abundance
<i>Kanakana letovirus</i> (1)	Light haemorrhaging around gill slits.	<i>Geotria australis</i>	ORF1a	9270	Pacific salmon nidovirus ORF1a (QEG08236.1); e-value=1x10 ⁻¹⁵⁵	26.36%	6.86x10 ⁻⁵	5.52x10 ⁻⁶
			ORF1b	6507	Pacific salmon nidovirus ORF1b (QEG08237.1); e-value=0	46.53%	4.67x10 ⁻⁵	
			Spike	5022	Porcine torovirus spike protein (QBJ02013.1); e-value=2x10 ⁻¹⁰	27%	4.24x10 ⁻⁵	
<i>Kanakana letovirus</i> (2)	Small area of haemorrhaging near the ventral fin.	<i>Geotria australis</i>	ORF1a	9255	Pacific salmon nidovirus ORF1a (QEG08236.1); e-value=5x10 ⁻⁴⁹	27.81%	1.48x10 ⁻⁵	6.75x10 ⁻⁶
			ORF1b	6501	Pacific salmon nidovirus ORF1b (QEG08237.1); e-value=4x10 ⁻¹⁷⁵	46.30%	1.26x10 ⁻⁵	
			Spike	4989	Atlantic salmon bafinivirus spike protein (AUE23862.1); e-value=2x10 ⁻⁴	24.07%	6.4x10 ⁻⁶	
<i>Kanakana letovirus</i> (3)	Haemorrhaging on dorsal fins and gill slits. Pale liver with haemorrhaging, swollen kidney, adhesions between organs and muscles.	<i>Geotria australis</i>	ORF1a	9270	Pacific salmon nidovirus ORF1a (QEG08236.1); e-value=2x10 ⁻¹⁵⁶	26.22%	6.38x10 ⁻⁵	6.40x10 ⁻⁶
			ORF1b	6509	Pacific salmon nidovirus ORF1b (QEG08237.1); e-value=0	43.97%	4.20x10 ⁻⁵	
			Spike	5022	Porcine torovirus spike protein (QBJ01983.1); e-value=1x10 ⁻⁹	27.18%	4.48x10 ⁻⁵	

157

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 1). We detected coronavirus-like reads

in 32 sequencing libraries (of the 1279 that met our search criteria) that were further investigated. First, we identified four new viruses that again shared ORF1b sequence similarity to *Pacific salmon nidovirus* (52-56%). 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 tentatively named *Blenny letovirus*, *Mud hopper letovirus*, *Waltoni letovirus* and *Broadnosed pipefish letovirus*, respectively. Second, we identified a viral transcript that shared 39% amino acid sequence similarity to *Shrew coronavirus* in ORF1b and 20% amino acid sequence similarity to *Bat coronavirus* in the S protein (Table 1), although it is important to note that the vast majority of viruses in the *Letovirinae* are not yet included in the nr database. 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% sequence similarity in the RdRp and ~40% sequence similarity in the S protein to other fish viruses in the *Tobaniviridae* (see Table 1). These new viruses were tentatively 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*.

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) (Figure 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 and a loggerhead sea turtle (*Caretta caretta*) (Figure 3c).

189

190 **Figure 3.** (a) Schematic phylogram showing where bony and jawless fish fall in the greater animal
 191 phylogeny and colour-coded to signify the host class in (b) and (c). (b) An order-level phylogenetic
 192 tree showing the relationships between families within the *Nidovirales*. Branches are colour-coded
 193 to signify the host class shown in part a. Branches in bold represent those viruses identified in this
 194 study. All branches are scaled according to the number of amino acid substitutions per site. (c)
 195 Maximum likelihood phylogenetic tree of the open reading frame 1b (ORF1b), which contains the
 196 viral RdRp, showing the topology of the nine newly discovered coronaviruses (bold font; blue and
 197 cyan), which fall within *Letovirinae*. The order-level tree in (b) was used to root the tree. All
 198 branches are scaled according to the number of amino acid substitutions per site. Percentage
 199 node labels with bootstrap support >70% are indicated with an asterisk (*).

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 (Figure 4a). Of particular note, *Kanakana letovirus* sampled from New Zealand pouched lamprey falls as an ingroup in this subfamily. Similar patterns can be seen for viruses in amphibians and reptiles, indicative of major cross-species transmission events. For example, the two known coronaviruses from reptiles fall across the two different sub-families, while the letoviruses identified from amphibian hosts are highly divergent, 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. Indeed, phylogenetic reconciliation analysis found that host-jumping was consistently the most frequent 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 (Figure 4b).

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 Figure 2. The ‘untangle’ function was used in TreeMap3.0 to maximise the congruence between the host (left) and virus (right) phylogenies. (b) Reconciliation analysis of each virus family using eMPress (Santichaivekin et al. 2020) illustrates the proportion of 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: 0 for co-divergence, 1 for duplication, 1 for host-jumping and 1 for extinction.

227 Amino acid conservation across the Coronaviridae.

228 Compatible with their phylogenetic relationships, the new viruses identified here possessed several
229 motifs that are conserved among the coronaviruses, particularly within the RdRp domain in ORF1b
230 (Figure 5). These included the amino acid sequences 'GWDYPKCD' and 'ILSDDGV' that are
231 mostly conserved among the *Coronaviridae* but not the *Nidovirales* as a whole (Figure 5). These
232 conserved domains are essential for metal ion chelation and binding of the primer 3'-end template
233 complex (de Souza Luna et al. 2007).

234

235 **Figure 5.** (a) Genome structure of *Pacific salmon nidovirus* and the corresponding transcripts
236 found in the New Zealand pouched lamprey, highlighting the more conserved region, ORF1b (in
237 blue), which includes the RdRp. A bar graph shows the sequence similarity among all viruses
238 within the *Coronaviridae* included in Figure 3. (b) Multiple virus amino acid sequence alignment of
239 the open reading frame 1b (ORF1b), containing the RdRp, showing motifs conserved throughout
240 all known genera within the *Coronaviridae*, including the newly discovered viruses within the
241 *Letovirinae* for which virus transcripts in this region were identified.

242 Discussion

243 Virological sampling of freshwater bony and jawless fish, combined with mining publicly available
244 sequence data, enabled us to identify nine new members of the *Nidovirales*, seven of which fell
245 within the *Coronaviridae* and a further two fell within the related *Tobamiviridae*. The seven new
246 coronavirus-like transcripts were close genetic relatives of viruses classified within the *Letovirinae*
247 in the RdRp, including *Pacific salmon nidovirus*, *Microhyala letovirus* and other viruses recently
248 identified in fishes, amphibians and reptiles, forming a sister clade to all other known
249 coronaviruses.

250 The discovery of these new viruses not only expands the phylogenetic diversity of the *Letovirinae*,
251 but it also sheds important new light on the evolution and host range in the *Coronaviridae* as a
252 whole. In particular, our study revealed that coronavirus hosts include both freshwater and marine
253 fish spanning multiple taxonomic orders. This includes an ancient lineage of jawless fish, herein
254 represented by the pouched lamprey from New Zealand, that represents one of the evolutionary
255 oldest vertebrates still to exist, and one that has remained morphologically unchanged for around
256 360 million years (Gess et al. 2006), although its true age has recently been debated (Miyashita et
257 al. 2021). These jawless and bony fish hosts were also sampled across a range of geographic
258 localities: Australia, New Zealand, China, Europe and North America. Although divergent in their
259 sequences, conserved motifs across the entire virus family strongly supports the inclusion of these
260 new viruses within the *Coronaviridae* and more specifically the *Letovirinae*. With the expansion of
261 virological surveys of wildlife, it is likely that additional viruses within the *Coronaviridae* will be
262 discovered. Indeed, the long branch lengths separating the divergent families within the *Nidovirales*
263 is indicative of very limited sampling across this order of RNA viruses.

264 The genomic structure of both *Pacific salmon nidovirus* and *Ambystoma mexicanum nidovirus*
265 includes a short, non-coding region between the ORF1ab and the remaining genes, including the S
266 protein (Edgar et al. 2020; Mordecai et al. 2019). It has been suggested that this homologous gap
267 represents possible genome segmentation in this clade (Edgar et al. 2020). We were,
268 unfortunately, unable to determine the complete genomic structure of the viruses identified here
269 and so cannot directly address this issue. It is noteworthy, however, that the S protein of *Kanakana*

270 *letovirus* shared more sequence similarity with toroviruses rather than coronaviruses, while *Trout*
271 *perch letovirus* and *Pacific salmon nidovirus* consistently fell within the *Coronaviridae*. This pattern
272 of phylogenetic incongruence between these genomic regions is indicative of past recombination
273 within the *Nidovirales*. At present, the occurrence of this process seems to be limited to only some
274 aquatic vertebrate hosts, although given the propensity for coronaviruses to recombine (Denison et
275 al. 2011), it is likely that this will be observed in additional species.

276 First recorded in 2011, lamprey populations in New Zealand's Southland region have been affected
277 by LRS, causing reddening along the length of the body with increased mortalities (Brosnahan et
278 al. 2019). We cannot assign the novel *Kanakana letovirus* discovered here as the causative agent
279 of this syndrome. Nevertheless, since lamprey are classed as a nationally vulnerable threatened
280 species (Dunn et al. 2018), it is important that further investigation is undertaken to determine the
281 true prevalence of coronavirus in lamprey populations and whether the virus is associated with
282 LRS.

283 That *Letoviruses* form a sister clade to all other coronaviruses suggests that the backbone of the
284 *Coronaviridae* phylogeny overall mirrors that of their vertebrate hosts, with different viral genera
285 associated with different vertebrate classes, and the clear transition from aquatic to land, and back
286 again (Zhang et al. 2018). Nevertheless, lamprey represent an ancient lineage of jawless fish that
287 fall sister to all other vertebrates, including jawed fish, amphibians and reptiles. In this context it is
288 particularly notable that the novel coronavirus identified here in the pouched lamprey does not fall
289 basal to other fish coronaviruses, in turn revealing that cross-species transmission events have
290 played a major role in the evolutionary history of these viruses and which is also apparent from our
291 co-phylogenetic reconciliation analysis. Hence, as is the case with many families of RNA viruses
292 (Geoghegan et al. 2017), the broad association between different coronaviruses and their
293 vertebrate host classes suggests that the evolutionary history of this group likely dates to at least
294 the origin of the vertebrates some 530 million years ago. However, on top of this overall pattern of
295 long-term virus-host co-divergence, there have also been frequent instances of cross-species virus
296 transmission. As the case of the pouched lamprey demonstrates, such host jumping events may
297 be particularly commonplace in hosts that reside in aquatic environments, perhaps reflecting a

298 combination of evolutionary antiquity (and hence more long-term opportunity for cross-species
299 transmission) and greater physical interactions between multi-species assemblages, although this
300 clearly needs to be investigated in greater detail.

301

302 Materials & Methods

303 ***Bony fish sample collection in Australia and RNA sequencing.*** Samples of freshwater bony
304 fish (Osteichthyes) were collected from the Murray-Darling Basin river system in Australia between
305 January and March 2020 and screened for coronaviruses using meta-transcriptomic sequencing
306 (see Costa et al. (2021) for details of the broader study). Seven species of native Australian fish
307 species were collected for this purpose: bony breem (*Nematalosa erebi*), spangled perch
308 (*Leiopotherapon unicolor*), Australian smelt (*Retropinna semoni*), Murray-Darling rainbowfish
309 (*Melanotaenia fluviatilis*), flat-headed gudgeon (*Philypnodon grandiceps*), western carp gudgeon
310 (*Hypseleotris* sp.) and unspotted hardyhead (*Craterocephalus fulvus*). Three species of invasive
311 fish were also collected: common carp (*Cyprinus carpio*), goldfish (*Carassius auratus*) and eastern
312 mosquitofish (*Gambusia holbrooki*). Importantly, all fish species were apparently healthy, with no
313 obvious clinical signs of disease. Fish were caught using boat electrofishing procedures and
314 dissected immediately upon capture. Tissue specimens (liver and gills) were placed (in the same
315 tube) in RNALater and stored in a portable -80°C freezer, then later in a -80°C freezer in the
316 laboratory at Macquarie University, Sydney. To facilitate virus discovery, multiple individuals (1–10)
317 were pooled according to species and the location in which they were captured. Sampling was
318 conducted under the approval of Macquarie University Animal Ethics Committee, approval number
319 2019/035.

320 Frozen samples of liver and gill tissue were placed together in 600µl of lysis buffer containing 0.5%
321 foaming reagent (Reagent DX, Qiagen) and 1% of β-mercaptoethanol (Sigma-Aldrich). Submerged
322 tissue samples were homogenized with TissueRuptor (Qiagen) for one minute at 5000 rpm. To
323 further homogenise tissue samples and remove tissue residues, the homogenate was centrifuged
324 at full speed for three minutes. The homogenate was carefully removed and RNA from the clear
325 supernatant was extracted using the RNeasy Plus Mini Kit (Qiagen) following the manufacturer's

protocol. Extracted RNA was quantified using NanoDrop (ThermoFisher) and RNA from each species was pooled corresponding to the site in which they were captured, resulting in 36 sample libraries. RNA libraries were constructed using the Truseq Total RNA Library Preparation Protocol (Illumina). To enhance viral discovery and reduce the presence of non-viral reads, host ribosomal RNA (rRNA) was depleted using the Ribo-Zero-Gold Kit (Illumina) and paired-end sequencing (100bp) was performed on the NovaSeq 500 platform (Illumina). Sample library construction, rRNA depletion and RNA sequencing were performed at the Australian Genome Research Facility (AGRF).

Lamprey sample collection in Aotearoa New Zealand and RNA sequencing.

Lamprey reddening syndrome (LRS) describes 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 – 2013, the New Zealand Ministry of Primary Industries (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 (*Geotria australis*) (te reo Māori: kanakana - South Island, or piharau - North Island), 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 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 which 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 freezer. Sampling was conducted in consultation with local Māori (rūnanga and kaumatua).

354 Skin tissue was dissected from 26 FFPE blocks, representing 26 individuals. Total RNA was
355 extracted using a Macherey-Nagel NucleoSpin® totalRNA FFPE XS kit following the NucleoSpin
356 totalRNA FFPE XS protocol. Control samples from the Mokau and Okuti rivers were extracted
357 separately using a Zymo Research Direct-zol RNA Miniprep Plus kit. The purity of the extracted
358 RNA from all sample types was evaluated (NanoDrop and Qubit Fluorometer) and 12 of the
359 highest scoring samples (eight LRS and four controls) were sent to Otago Genomics Facility for
360 sequencing using the Truseq Total RNA Library Preparation Protocol (Illumina). Host ribosomal
361 RNA (rRNA) was depleted using the Ribo-Zero-Gold Kit (Illumina) to enhance viral discovery.
362 Libraries were quality checked and sequenced in one lane on the HiSeq 2500 V4 Illumina platform.

363 ***Data mining the Sequence Read Archive (SRA) for coronaviruses.***

364 To identify additional coronavirus-like sequences we screened bony fish transcriptomes present in
365 the SRA. We focused on a subset of available SRA libraries, particularly those RNA-seq libraries
366 without sequencing selection (e.g., poly(A)+ -selected). We excluded zebrafish (*Danio rerio*) runs
367 from the screens due to the large number of experimental data sets. To reduce redundancy, we
368 selected the first six runs for a given species in each experimental data set. In total, 1279 libraries
369 (both single- and paired-end reads) were selected. SRA runs were obtained through the European
370 Nucleotide Archive (<https://www.ebi.ac.uk/ena/browser/home>) that mirrors the SRA but allows for
371 reads to be downloaded directly in a FASTQ. Reads were then analysed using our coronavirus
372 discovery pipeline. The pipeline consisted of an initial screen that aligned the reads against a
373 custom database of *Coronaviridae* and other protein sequences within the *Nidovirales* using
374 Diamond BLASTx (v.2.0.2) (Buchfink et al. 2015) with an e-value threshold of 1×10^{-5} . Results were
375 manually inspected in Geneious Prime (www.geneious.com) to exclude spurious alignments to
376 low-complexity regions. Where coronavirus-like reads were detected (n=32) we quality trimmed
377 and assembled *de novo* raw reads into contigs using Trinity RNA-Seq (v2.9.1) (Haas et al. 2013).
378 The assembled contigs were then compared to NCBI non-redundant protein database using
379 Diamond BLASTx with an e-value threshold of 1×10^{-5} . To identity highly divergent sequences we
380 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 dataset for the given species.

Transcript sequence similarity searching for novel coronaviruses.

Sequencing reads were first quality trimmed 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 et al. 2015).

Phylogenetic analysis.

To infer the evolutionary relationships of the coronaviruses, newly discovered translated viral contigs were combined with representative protein 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 et al. 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 1000 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 minimize 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 cost of zero. Following the parsimony principle, the reconciliation proceeds by minimising the total event cost.

Estimation of virus abundance.

Transcriptomes were quantified using RNA-Seq by Expectation-Maximization (RSEM) 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).

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 [IDT]). The Maxima Reverse Transcriptase (Thermo Scientific) protocol was followed and did not include the GC-rich template options: 1 μ l template total RNA, 1 μ l ReadyMade Random Hexamers (10mM), 1 μ l dNTP mix (10mM), 11.5 μ l nuclease-free water, 4 μ l 5X RT Buffer, 0.5 μ l (20U) RiboLock RNase Inhibitor (Thermo Scientific), 1 μ l Maxima Reverse Transcriptase. The final volume (20 μ l) was stored at -20°C until used in polymerase chain reactions (PCR).

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*.

436 **Data availability.**

437 Sequencing reads and virus sequences from pouched lamprey samples are available online via
438 the Genomics Aotearoa database to maintain Māori data sovereignty of native species (taonga).
439 All other data are available on NCBI (accession numbers available soon).

440

441 **Supplementary Material**

442 **Supplementary Figure 1.** Agarose gels electrophoresis showing PCR products from primers that
443 targeted regions in coronavirus ORF1ab (including the RdRp) for 12 individuals of pouched
444 lamprey.

445 **Supplementary Table 1.** Primer set used for the RT-PCR confirmation of *Kanakana letovirus* in
446 specimens of pouched lamprey.

447

448 **Acknowledgements**

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

453

454 **Funding**

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

460

References

- Altschul, S. F., et al. (1990), 'Basic local alignment search tool', *Journal of Molecular Biology*, 215 (3), 403-10.
- Brosnahan, Cara 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 (3), 416-36.
- Buchfink, B., Xie, C., and Huson, D. H. (2015), 'Fast and sensitive protein alignment using DIAMOND', *Nature Methods*, 12 (1), 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, Irene, 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 (5), 578.
- Capella-Gutiérrez, Salvador, Silla-Martínez, José M., and Gabaldón, Toni (2009), 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics*, 25 (15), 1972-73.
- Charleston, M. A. 'TreeMap', (<https://sites.google.com/site/cophylogeny/home>), accessed 18 March 2021.
- Costa, Vincenzo A., et al. (2021), 'Metagenomic sequencing reveals a lack of virus exchange between native and invasive freshwater fish across the Murray-Darling Basin, Australia', *bioRxiv*, 2021.02.25.432824.
- de Souza Luna, Luciano Kleber, et al. (2007), 'Generic Detection of Coronaviruses and Differentiation at the Prototype Strain Level by Reverse Transcription-PCR and Nonfluorescent Low-Density Microarray', *Journal of Clinical Microbiology*, 45 (3), 1049-52.
- Denison, Mark R., et al. (2011), 'Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity', *RNA Biology*, 8 (2), 270-79.
- Dunn, N.R., et al. (2018), 'Conservation status of New Zealand freshwater fishes, 2017. New Zealand Threat Classification Series 24', *Department of Conservation*.
- Edgar, Robert C., et al. (2020), 'Petabase-scale sequence alignment catalyses viral discovery', *bioRxiv*, 2020.08.07.241729.
- Geoghegan, Jemma L., Duchêne, Sebastián, and Holmes, Edward C. (2017), 'Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families', *PLoS Pathogens*, 13 (2), e1006215.
- Gess, Robert W., Coates, Michael I., and Rubidge, Bruce S. (2006), 'A lamprey from the Devonian period of South Africa', *Nature*, 443 (7114), 981-84.
- Haas, Brian J., et al. (2013), 'De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis', *Nature Protocols*, 8 (8), 1494-512.
- Kalyaanamoorthy, S., et al. (2017), 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods*, 14 (6), 587-89.

502 Katoh, Kazutaka, et al. (2002), 'MAFFT: a novel method for rapid multiple sequence alignment
503 based on fast Fourier transform', *Nucleic Acids Research*, 30 (14), 3059-66.

504 Li, Bo and Dewey, Colin N. (2011), 'RSEM: accurate transcript quantification from RNA-Seq data
505 with or without a reference genome', *BMC Bioinformatics*, 12 (1), 323.

506 Miyashita, T., et al. (2021), 'Non-ammocoete larvae of Palaeozoic stem lampreys', *Nature* 591,
507 408–412.

508 Mordecai, Gideon J., et al. (2019), 'Endangered wild salmon infected by newly discovered
509 viruses', *eLife*, 8, e47615.

510 Nguyen, Lam-Tung, et al. (2015), 'IQ-TREE: a fast and effective stochastic algorithm for
511 estimating maximum-likelihood phylogenies', *Molecular Biology and Evolution*, 32 (1), 268-
512 74.

513 Santichaivekin, Santi, et al. (2020), 'eMPress: a systematic cophylogeny reconciliation tool',
514 *Bioinformatics*.

515 Shi, M., et al. (2018), 'The evolutionary history of vertebrate RNA viruses', *Nature*, 556 (7700), 197-
516 202.

517 Wu, Fan, et al. (2020), 'A new coronavirus associated with human respiratory disease in China',
518 *Nature*, 579 (7798), 265-69.

519 Zhang, Yong-Zhen, et al. (2018), 'The diversity, evolution and origins of vertebrate RNA viruses',
520 *Current Opinion in Virology*, 31, 9-16.

521