Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family Abysssoviridae, and from a sister group to the *Coronavirinae*, the proposed genus Alphaletovirus

Khulud Bukhari, Geraldine Mulley, Anastasia A. Gulyaeva, Lanying Zhao, Guocheng Shu, Jianping Jiang, Benjamin W. Neuman

**A R T I C L E I N F O**

**Keywords:** Nidovirales, Transcriptome, Virus discovery, Protease, Protein expression, Translation, Readthrough

**A B S T R A C T**

Transcriptomics has the potential to discover new RNA virus genomes by sequencing total intracellular RNA pools. In this study, we have searched publicly available transcriptomes for sequences similar to viruses of the **Nidovirales** order. We report two potential nidovirus genomes, a highly divergent 35.9 kb likely complete genome from the California sea hare *Aplysia californica*, which we assign to a nidovirus named Aplysia abyssovirus 1 (AAbV), and a coronavirus-like 22.3 kb partial genome from the ornamented pygmy frog *Microhyla fissingipes*, which we assign to a nidovirus named Microhyla alphaletovirus 1 (MLeV). AAbV was shown to encode a functional main protease, and a translational readthrough signal. Phylogenetic analysis suggested that AAbV represents a new family, proposed here as Abysssoviridae. MLeV represents a sister group to the other known coronaviruses. The importance of MLeV and AAbV for understanding nidovirus evolution, and the origin of terrestrial nidoviruses are discussed.

1. Introduction

Until recently, discovery of new RNA viruses proceeded slowly in a mostly hypothesis-driven manner while searching for an agent of a disease, and using antibody cross-reactivity or enough conserved motifs for successful amplification by reverse transcriptase polymerase chain reaction. With improvements in RNA transcriptome sequencing and homology-based search methods, it is now possible to capture the complete infecting RNA virome of an organism by deep-sequencing total intracellular RNA pools (Miranda et al., 2016; Shi et al., 2018, 2016).

The new sequencing methods have brought a great change to the **Nidovirales**, an order that includes viruses with complex replicase polyproteins and the largest known RNA genomes (Lauber et al., 2015). This order previously contained four family-level groups, the *Coronaviridae* which infect birds and mammals including humans, the *Arteriviridae* which infect non-human mammals, the *Mesoniviridae* which infect non-human mammals, the *Roniviridae* which infect crustaceans (Lauber et al., 2013). However, recent papers (Lauck et al., 2015; O’Dea et al., 2016; Saberi et al., 2018; Shi et al., 2018, 2016; Tokarz et al., 2015; Vasilakis et al., 2014; Wahl-Jensen et al., 2016) and our results (see below) have added to within-family diversity and revealed several highly divergent nido-like viruses which the Nidovirales Study Group proposed, pending ICTV ratification, to form four new virus families within the **Nidovirales** (Gorbalenya et al., 2017a).

In this report we describe the discovery and characterization of one of the nidoviruses prototyping a new family along with another putative nidovirus. We used BLAST searches to scan the publicly available transcriptomes and expressed sequence tag libraries available at the US National Center for Biotechnology Information, and revealed two novel nido-like virus sequences from the frog *Microhyla fissingipes* developmental transcriptome (Zhao et al., 2016) and from several transcriptome studies dealing with the marine gastropod *Aplysia californica* (Fiedler et al., 2010; Heyland et al., 2011; Moroz et al., 2006). We describe the bioinformatics of the new virus-like sequences, and demonstrate that the Aplysia virus-like sequence encodes a functional protease, and a translational termination-suppression signal. Implications for nidovirus evolution and the origin of nidovirus structural proteins are discussed.
2. Results

2.1. Virus discovery

Recent studies have identified a wide variety of virus-like sequences in intracellular RNA pools, but few new members of the Nidovirales have been reported compared to groups such as the Picornavirales. In order to determine whether additional lineages of nido-like viruses might be present, tBLASTn (Altschul et al., 1990) was used to search the transcriptome shotgun assembly (TSA) and expressed sequence tag (EST) databases for sequences encoding proteins similar to the main proteinase (Mpro), polymerase and helicase, or complete pp1b regions of the nidovirus strains Infectious bronchitis virus, Gill-associated virus, White bream virus, Cavally virus and Wobbly possum disease virus. The tBLASTn results were checked by using BLASTx to compare each result to non-redundant protein database, and results that matched back to any member of the Nidovirales were selected for further analysis. This led to the discovery of a 35.9 kb transcript and 243 other fragments from the California sea hare, Aplysia californica, and a 22.3 kb transcript from Microhyla fssipes, known as the ornamented pygmy frog. Putative virus transcripts were then compared to DNA sequences from the same organisms by nucleotide BLAST, and no evidence of either virus was found. Together, these tests suggest that both nido-like viruses most likely come from RNA viruses associated with host transcriptomes.

2.2. Phylogenetic analysis

Phylogenetic analysis was performed by IQ Tree 1.5.5 (Nguyen et al., 2015) using five protein domains universally conserved in known and proposed nidoviruses plus the virus-like sequences described in this study (see below). The produced maximum-likelihood tree was midpoint rooted to reveal two strongly-supported super-clades, consisting of four strongly-supported major clades corresponding to arteri-like viruses, toro-like viruses, corona-like viruses, and invertebrate nido-viruses (Fig. 1). A Bayesian rooted tree (not shown) was also constructed using the same viral sequences, and it yielded the same four major clades, but with weaker support values on some branches and a basal position of the arteri-like major clade. Together these results suggest that the novel virus-like sequences likely represent distantly related members of the Nidovirales, but the tree branch uncertainty also demonstrates the limitations of these phylogenetic approaches in dealing with the extreme diversity of the sparsely sampled nido-like viruses.

The virus-like sequence from Aplysia californica formed a relatively long and moderately supported branch that clustered with other invertebrate nidoviruses, forming a sister group to a clade consisting of the Coronaviridae and a recently discovered nidovirus from the marine snail Turritella, TurrNV. The virus-like sequence from Microhyla fssipes clustered with strong support as a sister group to the known Coronavirusae. We named these putative viruses Aplysia abyssovirus (AAbV) and Microhyla letovirus (MLeV), respectively. While we were expressing viral proteins to biologically validate the new sequences and preparing this manuscript, a second manuscript (Debat, 2018) appeared on BioRxiv, Humberto Debat who was describing the same Aplysia virus from the same source material, posted April 24th, 2018, where it is called Aplysia californica nido-like virus. That report covers the tissue tropism and age-dependent prevalence of the Aplysia virus thoroughly, so in this manuscript we will focus on bioinformatics analysis and biological validation of this virus. It is our opinion that the name Aplysia californica nido-like virus should be regarded as an alternate name to Aplysia abyssovirus.

2.3. Naming and etymology

After assigning AAbV and MLeV to nidoviruses by the above bioinformatics analysis, the genome sequences were submitted to the NCBI database. After assigning AAbV and MLeV to nidoviruses by the above bioinformatics analysis, the genome sequences were submitted to the NCBI database.

Nicovirus Study Group (NSG) of the International Committee on the Taxonomy of Viruses (ICTV) for their accommodation in the nidovirus taxonomy; BN, senior author of this manuscript, is a member of the NSG and AAG assisted NSG with analysis of these viruses. Classification of these and other viruses were described in several taxonomic proposals that were made publicly available in the pending proposals section of ICTV on June 23rd, 2017, revised on November 26th, 2017 (Gorbalenya et al., 2017a, 2017b; Ziebuhr et al., 2017) and August 12, 2018. They were approved by the ICTV Executive Committee in July 2018 and will be placed for ratification by ICTV in 2018. Throughout this report, we will follow the taxa naming and taxonomy from the pending ICTV taxonomic proposals cited above, which we interpret to establish priority in discovering and naming these viruses and establishing the respective taxa.

The etymology of the name abyssovirus is from the word abyss, a reference to the aquatic environment where Aplysia lives, to the Sumerian god of watery depths Abzu, and to its discovery in an RNA transcriptome obtained by “deep” sequencing technology. Based on relatively low amino acid identity to the other families in the Nidovirales, it is our opinion that AAbV prototypes a new nidovirus family, which was confirmed in the analysis described in the pending proposal. The NSG has also accepted our proposal to name the new family Abyssoviridae, the new genus Alphabysiovirus and the new species Aplysia abyssovirus 1.

The etymology of the name letovirus is in reference to the source of the virus in frogs, and their connection to the mythological Leto, daughter of the titans Coeus and Phoebe. In the story, Leto turned some inhospitable peasants into frogs after they stirred up the mud at the bottom of a pool so that she could not drink from it. Based on the low sequence identity but high conservation of domains found in the Coronavirinae, it is our opinion that MLeV is a member of a sister group to all known coronaviruses, but still within the Coronavirinae. Based on our input, the NGS named the new genus Alphaletovirus in the pending proposal.
The host of AAbV is shown in Fig. 2A. The virus was recovered from a variety of adult tissues, and from several developmental stages of the host organism, as described elsewhere (Debat, 2018). Fragments of AAbV were detected in 9 TSA and 9 EST databases, compiled over several years by three labs working in Florida and the UK (Fig. 2B-C).

The AAbV genome is represented in its longest and most complete available form by the transcriptome shotgun assembly sequence GBBW01007738 which represents a reverse-complementary genomic sequence. Remarkably, the organization of the AAbV genome has several features typical for viruses of the Alphavirus genus of the Togaviridae family (King et al., 2012) that could be contrasted with those conserved in the nidoviruses. They include: a) two in-frame open reading frames (ORFs; ORF1a and ORF1b) of the replicase gene that are separated by a stop codon rather than overlapping and including a nidovirus-like ribosomal frameshift signal in the overlap, and b) a single structural polyprotein gene (ORF2) rather than several ORFs encoding structural proteins. The 35913 nt long AAbV genome has a 74 nt 5′-untranslated region, a 964 nt 3′-untranslated region, and a short poly-A tail (Fig. 2D). Despite these alphavirus-like features, BLASTx analysis confirmed that the AAbV replicase polyprotein clusters with the Nidovirales, as depicted in Fig. 1. Each part of the genome is represented in 3–20 independent sequences from the TSA and EST databases available at www.ncbi.nlm.nih.gov as of November 26th, 2017 (Fig. 2E-F). The AAbV genome (Fig. 3A) is the second-largest currently reported RNA virus genome, behind a new 41.1 kb planarian nidovirus described in a different reference. The longest sequence match between the 5′-untranslated region and homologous transcription-regulatory sequence in the viral 5′-untranslated region to produce negative-stranded RNAs of subgenomic size (Sola et al., 2015). The longest sequence match between the 5′-untranslated region and intergenic region of AAbV is shown in Fig. 3C. It consists of six of eight identical nucleotides, which could form eight base pairs with a reverse-complementary viral minus strand due to the possibility of both A-U and G-U wobble base pairing.

The sequence of the genomic 5′-terminus is supported by the five assemblies (GBBW01007738, GAZL01021275, GBDA01037198, GBCZ01030948, and GBCZ01030949) that end within one nucleotide of each other. The EST sequence EB188990 contains the same sequence with an additional 5′-GGCTCGAG-3′ that may represent part of the 5′-terminal region missing from GBBW01007738. However, we prefer to side with the preponderance of sequence data and consider GBBW01007738 the most complete AAbV genome available until further biological evidence emerges.

The sequence of the 3′-terminus is supported by 6 TSA sequence assemblies and 1 EST sequence that end all within one nucleotide of each other. Every part of the genome is represented in at least three TSA sequence assemblies. Genome coverage is more abundant at the 3′-end, which could be evidence of 3′-coterminal subgenomic RNA species, or could be a result of the method used to prepare cDNA.

Genetic variation among these sequences is as follows. There are four short EST sequences which appear to join different discontinuous regions of the genome together, but the joins occur at different positions in the middle of genes and cannot be explained by nidovirus-like discontinuous transcription. These oddly joined sequence fragments likely represent either defective RNA species (Furuya et al., 1993), or artifacts of the EST preparation process. Two sequence assemblies differed from the others, with A replacing G at nucleotide 1627, and in another assembly A replacing the consensus G at position 28005, both of which could be attributed to natural mutations or the actions of host cytidine deaminase on the viral minus strand. There is also some variation in the preserved poly-A tail sequences, presumably from the difficulty of accurately reading long stretches of a single nucleotide.

In order to test whether there was support for AAbV subgenomic RNA species in the raw sequence data, individual sequence reads were mapped to the AAbV genome using Bowtie 2.3.4.1 (Langmead and Salzberg, 2012) and SAMtools 1.9 (Li et al., 2009). There was no noticeable change in read depth at the junction between ORF1a and ORF1b, but there was a sudden increase of about seven-fold in read depth immediately before the start of ORF2 (Fig. 3B), suggesting that ORF2 may be expressed from a subgenomic mRNA produced in relative abundance compared to the genomic RNA, as would be expected for a member of the Nidovirales. Numerous low-frequency AAbV sequence variants were identified in the raw sequence data, but none were consistent across all datasets, and no indels were consistently present within 1000 nucleotides of the start of ORF2. This was interpreted to indicate that either the viral subgenomic mRNA did not contain the expected nidovirus-like leader-body structure, or that any potential 5′-terminal leader sequences were not captured in the raw data.

Nidoviruses express their structural and accessory proteins via a set of 3′-coterminal nested subgenomic RNAs, which are produced by discontinuous transcription on the genomic template. In this process, the polymerase is thought to pause at transcription-regulatory sequences located upstream of each gene, occasionally resulting in a template switch to homologous transcription-regulatory sequence in the viral 5′-untranslated region to produce negative-stranded RNAs of subgenomic size (Sola et al., 2015). The longest sequence match between the 5′-untranslated region and intergenic region of AAbV is shown in Fig. 3C. It consists of six of eight identical nucleotides, which could form eight base pairs with a reverse-complementary viral minus strand due to the possibility of both A-U and G-U wobble base pairing. However, none of the available TSA or EST sequences showed direct evidence of a subgenomic RNA species, such as a consistently-spliced transcript, or a large number of sequence reads that stop at the putative transcription-regulatory sequence. This sequence AAACGATG or AAA transcribed transcript, or a large number of sequence reads that stop at the putative transcription-regulatory sequence. This sequence AAACGATG or AAA transcribed transcript, or a large number of sequence reads that stop at the putative transcription-regulatory sequence. This sequence AAACGATG or AAA transcribed transcript, or a large number of sequence reads that stop at the putative transcription-regulatory sequence. This sequence AAACGATG or AAA...
expressed via a translational readthrough rather than frameshift mechanism, while potential structural protein genes are presumably expressed from a single subgenomic RNA to produce structural polyprotein pp2.

2.5. AAbV protein bioinformatics

To annotate the functional protein domains encoded in the AAbV genome, a series of bioinformatics tools were used. Wherever possible, we have followed the convention of SARS-associated coronavirus (SARS-CoV) species in naming domains and polyprotein processing products (Ref?). When run against the PDB database, HHPred (Söding et al., 2005) predicts function based on structure. For domains like the picornavirus polymerase.

HHPred produced confident predictions for a coronavirus-like M

(Anand et al., 2002) in pp1a (Fig. 3D). In pp1b HHPred identified a picornavirus-like RNA-dependent RNA polymerase (RdRp) (Debat, 2018) of the Mpro domain from both sides.

Protein BLAST was used to map the AAbV nidovirus RdRp-associated nucleotidyl transferase (NiRAN) and nspl6 2O- MTase domains to homologous domains from other nidoviruses. The corresponding regions of AAbV and the top protein BLAST match were then submitted to HHPred in align mode, which uses predicted structure and primary sequence data to compare proteins. This led to confident identifications of the NiRAN and a match for the divergent but functional 20 MTase domain of Gill-associated virus (Zeng et al., 2016). One other uncharacterized domain was also identified in both AAbV and TurrNV by protein BLAST, in the position where the coronavirus conserved replication accessory proteins nspl7–10 were expected (Fig. 3D). However, there was not enough similarity between the AAbV-TurrNV conserved domain and other nidovirus domains to confidently assign a function to this region.

We also looked for transmembrane regions which are typically clustered in three regions in nidovirus pp1a. Domain-level maps of new and known nidoviruses pp1a and pp1b are shown in Figs. 4 and 5A, respectively. Nidoviruses typically have a cluster of an even number of transmembrane helices near the midpoint of pp1a, equivalent to nsp3 of SARS coronavirus. Nidoviruses also have two other clusters of 2–8 transmembrane helices flanking the M

domain from both sides.

AAbV is also missing some common but not universally conserved nidovirus domains. AAbV does not appear to encode a homolog of the uridyylate-specific nidovirus endonuclease (NendoU), nor is there enough un-annotated protein sequence in pp1b to accommodate an NendoU. This result is in line with the lack of this domain in other invertebrate nidoviruses (Nga et al., 2011). We were also not able to corroborate the prediction (Debat, 2018) of a papain-like proteinase domain situated among the predicted transmembrane regions of the first transmembrane cluster, or of a potential S-like domain of the structural polyprotein.

The pp2 gene of AAbV encodes a putative structural polyprotein of 3224 amino acids. HHPred and BLAST were not able to detect matches for any domains except M

in AAbV pp2. TMHMM (Krogh et al., 2001)
predicted 13 transmembrane helices in pp2, which were generally ar- ranged in pairs with large intervening domains, which we have tenta- tively named Spro, predicted surface glycoproteins GP1–3 and a possible nucleoprotein (Fig. 5B). Included in pp2 are additional smaller domains that have not been named yet, pending a better understanding of pp2 proteolytic processing. Signalf (Petersen et al., 2011) predicted an ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- initial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- tial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the ini- i
AAbV encodes at least one functional proteinase, but further work is needed to determine the cleavage specificity and map proteolytic processing by the AAbV Mpro.

2.7. AAbV pp1ab expression

Another unusual feature of AAbV was the presence of an in-frame stop codon separating the pp1a and pp1b genes, rather than the expected ribosomal frameshift signal found in most other nidoviruses. We note that an in-frame stop codon separates the putative pp1a and pp1b of the molluscan nidovirus Tunninivirus 1, which was phylogenetically grouped with AAbV and Alphamesonivirus 1 (Fig. 1). This suggested that AAbV may use a translational termination-suppression signal as a way to control expression of the pp1b region. Termination-suppression signals are found in several other viruses including alphaviruses and some retroviruses, and typically consist of a UAG or UGA stop codon followed by an RNA secondary structure element, and the efficiency of suppression normally depends on the stop codon, the nucleotides immediately following the stop codon, and the free energy of the RNA secondary structure element (Feng et al., 1992). The pp1a gene of AAbV ends in a UGA stop codon, and the region that follows was predicted by MFold (Zuker, 2003) to be capable of forming several related RNA secondary structure elements, of which the most consistently predicted is shown in Fig. 7A. A potential pseudoknot-like conformation in the same region is shown by Debat (Debat, 2018).

To investigate protein expression at the pp1a-pp1b region, nucleotides 17255–17707 were cloned into pTriex 1.1 with amino-terminal HSV and carboxyl-terminal HIS tags. This construct would allow detection and quantification of the 25 kDa proteins that stopped at the natural UGA stop codon that would have an HSV tag only, and 35 kDa readthrough products that would have both HSV and HIS tags. Expression of this construct produced the expected 25 kDa termination product and 35 kDa readthrough product (Fig. 7B-D). Based on densitometry analysis (not shown), it was estimated that 25–30% of translation events resulted in readthrough.

The choice of stop codon and elements of the two codons that follow have been shown to affect the efficiency of translational termination (Cridge et al., 2018; Skuzeski et al., 1991). To further investigate the AAbV termination-suppression signal, constructs were made in which the region around the pp1a stop codon was perturbed from the wild-type UGAC, predicted to produce near optimal termination, to UAAA, predicted to produce much less than optimal termination. In another construct, 42 nucleotides predicted to form one side of the predicted RNA stem-loops were deleted (Δ42; Fig. 7A). Mutation of the AAbV pp1a stop codon had little effect on readthrough efficiency (Fig. 7B), but deletion of 42 nucleotides predicted to be involved in RNA secondary structures appeared to decrease readthrough, and led to a smaller readthrough product as predicted. Together these results indicate that the pp1b region of AAbV is probably expressed by readthrough of a UGA stop codon, mediated by a functional termination-suppression signal that is dependent on sequences following the stop codon.

2.8. MLeV genome

Microhyla letovirus is represented by a single assembly (accession number GECV01031551) of 22304 nucleotides that potentially encodes a partial corona-like virus from near the end of a protein equivalent to SARS-CoV nsp3 to the 3′-end (Fig. 8A). No other matches for this sequence were found in the TSA or EST databases by nucleotide BLAST.
nsp1, nsps2 and part of nsps3. The size of the missing part of the genome can be estimated at 1500–4000 nucleotides based on comparison to complete genomes from the relatively small deltacoronaviruses or the relatively large alphacoronaviruses. The MLeV genome contains a 572 nucleotide 3′-untranslated region and an 18-nucleotide poly-A adenosine tail.

The genome organization of MLeV was similar to that of coronaviruses, with a predicted -1 ribosomal frameshift signal. Usually, a programmed -1 ribosomal frameshift signal consists of three elements: a slippery sequence that is most commonly UUUAAAC in coronaviruses, a stop codon for the upstream coding region, and a strong RNA secondary structure or pseudoknot. MLeV encodes a potential slippery sequence at nucleotide 6085 (UUUAAAC) followed immediately by a UAA stop codon for pp1a. The region following the putative frameshift signal was predicted by Mfold to adopt a stem-loop conformation which may be part of an RNA pseudoknot (not shown), but further biological characterization is needed to determine the boundaries of the frameshifting region and test its frameshifting efficiency.

The 3′-end of the MLeV genome contains six ORFs that could encode proteins of 50 or more amino acids, which presumably include the viral structural proteins. Five of the six 3′-end ORFs are preceded by a sequence UCUAAHA (where H is any nucleotide except G), that resembles the UCUAAAC transcription regulatory sequence of the coronavirus mouse hepatitis virus. These candidate transcription-regulatory sequences start 6–66 nucleotides before the AUG start codon of the next ORF. Without the 5′-end or any evidence of viral subgenomic RNAs, it is not possible to be certain how the 3′-end ORFs are expressed, but these repeated sequences are evidence that MLeV may express its structural proteins from subgenomic RNAs in the manner of coronaviruses. Unfortunately, the original RNA sample that was used for Microhyla fssipes transcriptomic analysis was completely consumed, and could not be further tested by RT-PCR.

The first of these downstream ORFs encodes a large S-like protein of 1526 amino acids with an amino-terminal signal peptide predicted by SignalP and a carboxyl-terminal transmembrane region predicted by TMHMM. The second and third ORFs appear to encode a unique single-pass transmembrane protein of 55 amino acids (ORF 2b) and a unique soluble 157 (ORF 3) amino acid protein, respectively, which are likely strain-specific accessory proteins. The fourth ORF encodes an E-like protein of 77 amino acids, with an amino-terminal predicted transmembrane region followed by a potential amphipathic helix predicted by Amphipaseek (Sapay et al., 2006). The fifth ORF encodes a 241 amino acid long three-pass transmembrane protein that resembles the coronavirus M protein, and the sixth ORF encodes a putative N protein of 459 amino acids. Together, these 3′-ORFs appear to encode a
complete coronavirus functional repertoire, and are present in the same order found on all other currently known coronavirus genomes (Neuman and Buchmeier, 2016). The start codons of the putative S and M ORFs appear to overlap with the stop codons of preceding ORFs, indicating a relatively compact genome.

To test whether there was support for MLeV subgenomic RNA species in the raw sequence data, individual sequence reads were mapped to the MLeV genome using the same method used for AAbV above (Fig. 9A).

There was not a noticeable change in read depth at the junction between ORFs 1a and 1b of MLeV, suggesting that polyprotein 1b is expressed by a translational rather than transcriptional mechanism. However, there were two sudden increases of about eight-fold in read depth immediately before the start of the N ORF and near the beginning of the adjacent E and M ORFs (Fig. 9B). Expected increases in read depth before the putative S gene and the largest putative accessory gene were not detected. As with AAbV, many low-frequency sequence variants were detected in the raw sequence.
data, but no indels were consistently present in the region surrounding the putative transcription-regulatory sequences. These data suggest that at least the M and N genes of MLeV are expressed via subgenomic mRNAs.

2.9. MLeV protein bioinformatics

In the pp1a region, HHpred detected matches for conserved coronavirus domains including the carboxyl-terminal domain of coronavirus nsp4, Mβ, nsp7, nsp8, nsp9 and nsp10 (Fig. 8C). In the pp1b region, HHpred detected matches for a picornavirus-like RdRp, the nsp13 metal-binding helicase, the nsp14 ExoN-N7 MTase, the nsp15 NEndoU, and the nsp16 2O MTase. In the structural protein region, HHpred detected a match for the amino-terminal domain of coronavirus N in the putative MLeV N protein.

As with AAbV, we then widened our search to include conserved coronavirus domains that do not yet have known protein structures. This led to a match for the carboxyl-terminal region of nsp3, amino-terminal region of nsp4, nsp6, the nsp12 NiRAN domain, and a match between coronavirus M and the proposed MLeV M protein. Neither the proposed MLeV S nor E protein could be further corroborated by bioinformatics tools. Together, this indicated that MLeV appears to encode a complete set of conserved coronavirus-like proteins from the carboxyl-terminal region of nsp3 through the end of the genome.

3. Discussion and conclusions

With the addition of MLeV, AAbV and a host of other recently-published highly divergent nidoviruses, the field of nidovirus evolution is due for a revision, which will require a detailed approach and that will fit best in another study. However, a few tentative conclusions can be drawn from these new viruses.

Firstly, the new viruses confirm that the region of pp1a up to the SARS-CoV nsp4 equivalent, which seems to contain a variety of anti-host countermeasures in the viruses where this region has been studied (Neuman et al., 2014), is highly variable and does not appear to contain any universally-conserved domains. As previously noted (Lauber et al., 2013), this part of the genome appears to have the most genetic flexibility, even within viral genera, and likely has great relevance to those studying interactions between viruses and innate immunity (Bailey-Elkin et al., 2014; Lukugamage et al., 2015; Mielech et al., 2014). It is worth noting that the region preceding the Mβ in AAbV is over 13 kb – larger than most other complete RNA virus genomes.

Secondly, two elements of genome architecture seem to be conserved throughout the Nidovirales: a Mβ flanked by multi-pass transmembrane regions, and the block containing NiRAN-RNA polymerase-metal binding-Helicase. Knowledge of these apparent nidovirus genetic synapomorphies should make it possible to design searches to detect even more divergent nido-like viruses in transcriptomes.

Thirdly, the NendoU domain appears to be found only in viruses infecting vertebrate animals, and is lacking in every known nidovirus-like genome from an invertebrate host. This suggests that the function of NendoU may have evolved as a countermeasure to conserved metazoan viral RNA recognition machinery involved in innate immunity (Lukugamage et al., 2015).

Fourthly, while most currently known nidovirus species are associated with terrestrial hosts, the greatest phylogenetic diversity of nidoviruses is now associated with hosts that live in aquatic environments. Since terrestrial metazoan transcriptomes are relatively well-sampled in comparison to aquatic and particularly marine metazoans, we would predict this trend is likely to continue. Of the eight proposed nidovirus families shown in Figs. 4 and 5, four contain only viruses associated with aquatic hosts, two (Arteriviridae (Shi et al., 2018) and the proposed Tobaniviridae) are found in a mix of strictly aquatic and strictly terrestrial animals, and two (Coronaviridae, Mesoniviridae) are in part associated with hosts such as mosquitoes and frogs that have an obligate aquatic larval phase. Taken together, this data suggests that it may be useful to consider potential routes of interspecies transmission between marine, freshwater and terrestrial hosts in future studies of nidovirus evolution, as more data becomes available.

Lastly, the structural protein repertoire of nidoviruses appears to be quite broad compared to other known virus orders. There do not appear to be any conserved nidovirus structural proteins with the possible exception of the nucleoprotein (discussed elsewhere (Neuman and Buchmeier, 2016)), and even that homology can only be regarded as hypothetical until more structures of putative nucleoproteins are solved. A tentative categorization of nidovirus structural proteins, based on size, predicted transmembrane regions, and predicted protein secondary structure is shown in Fig. 10. If correct, this would indicate that nidoviruses have a diverse set of structural proteins that includes a variety of possibly unrelated spike-like proteins plus components shared with Orthomyxoviridae (HA and HE), Togaviridae (E1 and the E3 structural serine proteinase), Flaviviridae (the capsid RNAse). This structural repertoire appears to be variously expressed from subgenomic RNAs encoding a single gene (as proposed for MLeV), giant polyproteins such as that of AAbV, and a mix of intermediate-sized polyproteins and single genes, as in the Roniviridae. Taken together,
these observations suggest that structural proteins are widely shared and exchanged among RNA viruses, and that conserved elements of the replicase will be more useful than structural proteins for anyone trying to construct trees that connect viruses taxonomic ranks above the family level.

4. Materials and methods

4.1. Phylogeny

Nidovirus phylogeny was reconstructed based on MSA of concatenated Mpro, N, NiRAN, RdRp, CH cluster and SF1 Helicase conserved cores (3417–3905, 5441–5866, 6095–7291, 7340–7504, 7781–8545 nt of the Equine arteritis virus genome X53459.3), prepared with the help of Viralas platform (Gorbalenya et al., 2010). Representatives of 28 nidovirus species (Supplementary table 1) delineated in recent ICTV proposals (Brinton et al., 2017; Gorbalenya et al., 2017b, 2017a; Ziebuhr et al., 2017) were used. Phylogeny was reconstructed by IQ Tree 1.5.5 using a partition model where the evolutionary model for each of the five domains was selected by ModellFinder (Chernomor et al., 2016; Kalyaanamoorthy et al., 2017; Nguyen et al., 2015). To estimate branch support, Shimodaira-Hasegawa-like approximate likelihood ratio test (SH-aLRT) with 1000 replicates was conducted. The tree was midpoint rooted and visualized with the help of R packages APE 3.5 and phangorn 2.0.4 (Paradis et al., 2004; R Development Core Team, 2011; Schliep, 2011).

4.2. Protein assays

Nucleotides 12926–14176 containing the AaBV Mpro and flanking regions extending to the preceding and following predicted transmembrane regions was produced as a synthetic GeneArt Strings DNA fragment (Invitrogen). This was used as the template in a 50 µl PCR reaction using primers Aby_IF_MP_F (CAGACGAGTCCTGAGAGTGGTGTTGTTGGAAGT) and Aby_IF_MP_R (GATGGTGGTGCTCGAGGCGGTTCGTCGAGA) with 1x Phusion High Fidelity PCR Mastermix (Thermo Fisher Scientific). The 1283 bp PCR product was gel extracted using a QIAquick gel extraction kit (Qiagen) and cloned into a TOPO TA Cloning vector (Invitrogen). This was used as the template in a 50 µl PCR reaction using primers Aby IF TryP_F (GATGGTGGTGCTCGAGGCGGTTCGTCGAGA) and Aby IF TryP_R (GATGGTGGTGCTCGAGGCGGTTCGTCGAGA). The predicted Spro-containing region of AaBV, nucleotides 25918–27183, was PCR amplified from a synthetic GeneArt Strings fragment (Invitrogen) using primers Aby IF TryP_F (GATGGTGGTGCTCGAGGCGGTTCGTCGAGA) and Aby IF TryP_R (GATGGTGGTGCTCGAGGCGGTTCGTCGAGA). Both the Spro and putative pp1a-pp1b termination-signal products were cloned, expressed and detected in the same way as AaBV Mpro.

4.3. Microhyla prevalence

Data for the MLeV prevalence study comes from a published report (Zhao et al., 2016). Briefly, nine tadpoles were sacrificed, using three individuals from each of the three developmental stages as independent biological replicates. One microgram of mRNA of each stage sample was sequenced on an Illumina HiSeq. 2000 platform by NovoGene (Beijing), and paired-end reads were generated.

Acknowledgements

A.A.G. is a Ph.D. student with Alexander E. Gorbalenya (A.E.G.) and her work and resources she used were partially supported by European Union Horizon2020 EVAg 653316 grant and Leiden University Medical Center, United Kingdom MoBiLe Program to A.E.G. She thanks Igor A. Sidorov, Dmitry V. Samborskiy, and A.E.G. for help with the dataset used in her analysis. The work of L.Z., G.S. and J.J. was supported by a key project from Chinese Academy of Sciences (KJZD-EW-L13) and the National Natural Science Foundation of China (No. 31471964). K.B. was supported by a studentship from the Ministry of Education in Saudi Arabia (S13280). K.B. thanks Ian M. Jones of the University of Reading for assistance in planning and carrying out protein expression and detection studies.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.virol.2018.08.010.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1006/jmbi.2000.3082.
Anand, K., Palm, G.J., Mesters, J.R., Siddell, S.G., Ziebuhr, J., Hilgenfeld, R., 2002. Structure of the middle east respiratory syndrome coronavirus (MERS-CoV) papain-like protease bound to ubiquitin facilitates targeted disruption of deubiquitinating activity to demonstrate its role in innate immune suppression. J. Mol. Biol. 68, 187 332637. https://doi.org/10.1093/jmb/68.1.322.
Bailey-Ellkin, B.A., Knapo, R.C.M., Johnson, G.G., Ninaber, D.K., Van Kasteren, P.B., Breidenbeck, P.J., Snijder, E.J., Kiekert, M., Mark, B.L., 2014. Crystal structure of the middle east respiratory syndrome coronavirus (MERS-CoV) papain-like protease bound to ubiquitin facilitates targeted disruption of deubiquitinating activity to demonstrate its role in innate immune suppression. J. Biol. Chem. 289, 34667–34682. https://doi.org/10.1074/jbc.M114.569644.
Birktoft, J.J., Blow, D.M., 1972. Structure of crystalline α-chymotrypsin. V. The atomic structure of trypt-α-chymotrypsin at 2 Å resolution. J. Mol. Biol. 68, 187–240. https://doi.org/10.1016/0022-2836(72)90210-0.
Brinton, M.A., Gulyaeva, A., Balasuriya, U.B.R., Dunowska, M., Faaberg, K.S., Goldberg, T., Leung, F.-C., Nauwynck, H.J., Snijder, E.J., Stadejek, T., Gorbalenya, A.E., 2017. ICTV Pending proposal 2017.0125. Expansion of the rank structure of the family Arteriviridae and renaming its taxa.
Buchan, D.W.A., Minneci, F., Nugent, T.C.O., Breyson, K., Jones, D.T., 2013. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res. 41. https://
proteases: multifunctional enzymes with protease, deubiquitinating and deSylating activities. Virus Res. 194, 184-190. https://doi.org/10.1016/j.virusres.2014.01.025.
Miranda, J.A., Culely, A.L., Schwarcz, C.R., Steward, G.F., 2016. RNA viruses as major contributors to Antarctic viroplankton. Environ. Microbiol. https://doi.org/10.1111/1462-2920.13291.
Moroz, L.L., Edwards, J.R., Puthanveetil, S.V., Kohn, A.B., Ha, T., Heyland, A., Knudsen, B., Sahni, A., Yu, F., Liu, L., Jerzinski, S., Lovell, P., Iannuccelli, W., Chen, M., Nguyen, T., Sheng, H., Shaw, R., Kalachikov, S., Panchin, Y.V., Farmere, W., Russo, J.J., Yu, K., Li, Z., Chung, E.R., 2006. Neuronal transcription of aplycna: neuronal compartments and circuitry. Cell 127, 1453-1467. https://doi.org/10.1016/j.cell.2006.09.052.
Neuman, B.W., Buchmeier, M.J., 2016. Supramolecular architecture of the coronavirus particle. Adv. Virus Res. 1-27. https://doi.org/10.1016/bs.avrs.2016.08.004.
Neuman, B.W., Chamberlain, D.P., Rot, F., Joseph, J., 2014. Analysis of arterivirus re- plisome structure. Virus Res. 194, 46-66. https://doi.org/10.1016/j.virusres.2013.12.004.
Nguyen-T.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268-274. https://doi.org/10.1093/molbev/msu300.
Nga, P.T., Parquet, M.Dc., Lauber, C., Paradia, M., Nabehtesmib, T., Yu, F., et al., 2011. Discovery of the First Insect Borna Virus: Exploring Evolutionary Lineage of the Largest RNA Virus Genomes. PLoS Pathog 7 (9), e1002215. https://doi.org/10.1371/journal.ppat.1002215.
O’Dea, M.A., Jackson, B., Jackson, C., Xavier, P., Warren, K., 2016. Discovery and partial genomic characterisation of a novel arterivirus associated with respiratory disease in wild shingleback lizards (Tiliqua rugosa). PLoS One 11. https://doi.org/10.1371/journal.pone.0156299.
Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R. Bioinformatics 20, 289–290. https://doi.org/10.1093/bioinformatics/btg412.
Petersen, T.N., Brunak, S., Von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods. https://doi.org/10.1038/nmeth.710.
Prince, J.S., 2003. A presumptive alphavirus in the gastropod mollusc, Aplysia californica. Comp. Biochem. Physiol. Part D Genom. Proteom. 5, 165.
Prince, J.S., Buchmann, J., Wang, W., Xu, J., Holmes, E.C., Zhang, Y.Z., 2016. Redefining the invertebrate RNA virosphere. Nature 540, 539-543. https://doi.org/10.1038/nature21672.
Sifers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 508. https://doi.org/10.1038/msb.2011.75.
Suzekni, J.M., Nichols, L.M., Gesteland, R.F., Atkins, J.F., 1991. The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. J. Mol. Biol. 218, 363–375. https://doi.org/10.1016/0022-2836(91)90767-B.
Söding, J., Biegert, A., Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucl. Acids Res. 33, 69–72. https://doi.org/10.1093/nar/gki408.
Sola, I., Almazán, F., Zúñiga, S., Enjuanes, L., 2015. Continuous and discontinuous RNA synthesis in coronavirus. Annu. Rev. Virol. 2, 265–288. https://doi.org/10.1146/annurev-virology-010414-013609.
Sorensen, H.S., Rasmussen, J.M., Christensen, N., Nielsen, M.K., 2019. The RNA Replication Complex in the Coronavirus Nidovirus Replication Complex. Viruses 11, 1467. https://doi.org/10.3390/v110901467.
Tebbe, L., Rao, K., Jackson, C.B., Ranganathan, S., Veldman, P.J., 2018. The polyprotein is the central component of the SARS coronavirus replication complex. J. Virol. 92, 5461–5473. https://doi.org/10.1128/jvi.00583-17.
Tebus, B., Jiang, Z., Zuckerman, J.H., 2012. RNA dependence of RNA RNA virus replication. J. Cell Biochem. 113, 1655–1663. https://doi.org/10.1002/jcb.24791.
Tebus, B., Jiang, Z., Zuckerman, J.H., 2012. RNA dependence of RNA virus replication. J. Cell Biochem. 113, 1655–1663. https://doi.org/10.1002/jcb.24791.
thyroid hormone receptors in metamorphosis. Sci. Rep. 6. https://doi.org/10.1038/srep27310.
Ziebuhr, J., Baric, R.S., Baker, S., de Groot, R.J., Drosten, C., Gulyaeva, A., Haagmans, B. L., Neuman, R.W., Perlman, S., Poon, L.L.M., Sola, I., Gorbalenya, A.E., 2017. ICTV Pending Proposal 2017.013S. Reorganization of the family Coronaviridae into two families, Coronaviridae (including the current subfamily Coronavirus and the new subfamily Letovirinae) and the new family Tobaniviridae (accommodating the current subf).
Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415. https://doi.org/10.1093/nar/gkg595.