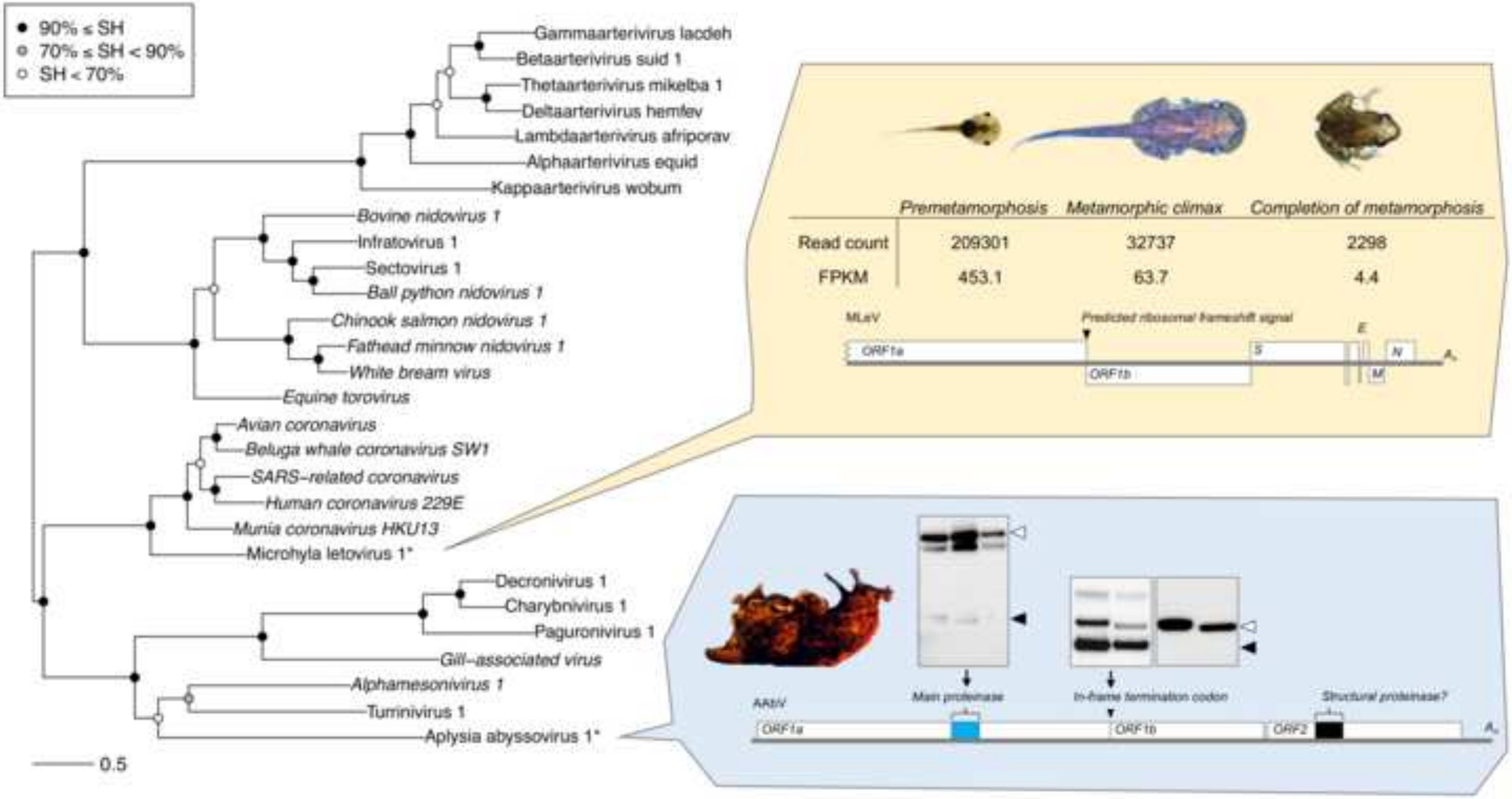
Graphical Abstract



	Premetamorphosis	*Metamorphic climax*	*Completion of metamorphosis*
Read count	209301	32737	2298
FPKM	453.1	63.7	4.4

Title: 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

Authors: Khulud Bukhari[1], Geraldine Mulley[1], Anastasia A. Gulyaeva[2], Lanying Zhao[3], Guocheng Shu[3], Jianping Jiang[3], Benjamin W. Neuman[4,5]

Affiliations

[1]University of Reading, Reading, UK

[2]Dept. Medical Microbiology, Leiden University Medical Center, Leiden, Netherlands

[3]Chengdu Institute of Biology, Chinese Academy of Science, Chengdu, China

[4]Texas A&M University-Texarkana, 7101 University Ave, Texarkana, TX 75503

[5]Address correspondence to bneuman@tamut.edu

Word count: 5956 total, 135 abstract

Abstract

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 fissipes*, which we assign to a nidovirus named Microhyla alphaletovirus 1 (MLeV). AAbV was shown to encode a functional main proteinase, and a translational readthrough signal. Phylogenetic analysis suggested that AAbV represents a new family, proposed here as Abyssoviridae. 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.

Keywords: Nidovirales; transcriptome; virus discovery; proteinase; protease; protein expression; translation; readthrough

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., 2013). 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 arthropods, and 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 fissipes* 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 proteinase, and a translational termination-suppression signal. Implications for nidovirus evolution and the origin of nidovirus structural proteins are discussed.

Results

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 (M^{pro}), 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 the 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 fissipes*, 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 nidovirus-like transcripts most likely come from RNA viruses associated with host transcriptomes.

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 mid-point 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 nidoviruses (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 *Mesoniviridae* and a recently discovered nidovirus from the marine snail *Turritella,* TurrNV. The virus-like sequence from *Microhyla fissipes* clustered with strong support as a sister group to the known *Coronavirinae*. 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 appeared on BioRxiv (Debat, 2018) from 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.

Naming and Etymology

After assigning AAbV and MLeV to nidoviruses by the above bioinformatics analysis, the genome sequences were submitted to the Nidovirus 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., 2017b, 2017a; 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 Alphaabyssovirus 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.

AAbV Genome and subgenome sequences and their potential expression

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 BioRxiv manuscript (Saberi et al., 2018).

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 all end 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 a 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 sequence 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 AAACGGTA needs to be investigated further to determine whether it functions as a transcription-regulatory sequence for viral subgenomic RNA production.

Together these data suggest that the AAbV genome is reasonably complete, robust, and represents a novel and exceptionally large nido-like virus. It has the unusual genome organization which is nonetheless consistent with the canonical nidovirus features of large replicase polyproteins 1a and 1ab, pp1a and pp1ab, respectively. They are 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.

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 polymerase where a nidovirus structure is not yet available, HHPred can sometimes detect a match to a homologous protein, such as the picornavirus polymerase.

HHPred produced confident predictions for a coronavirus-like M^{pro} (Anand et al., 2002) in pp1a (Fig. 3D). In pp1b HHPred identified a picornavirus-like RNA-dependent RNA polymerase (RdRp (te Velthuis et al., 2009)), nsp13 metal-binding helicase (Deng et al., 2014; Ivanov et al., 2004), nidovirus-specific nsp14 exonuclease (ExoN (Ma et al., 2015)) and nsp14 N7 methyltransferase (N7 MTase (Chen et al., 2009; Ma et al., 2015)). In pp2, HHPred identified a chymotrypsin-like serine proteinase (Birktoft and Blow, 1972), a feature analogous to the alphavirus capsid proteinase (Melancont and Garoff, 1987), but until now predicted in only one nidovirus, TurrNV. We have termed this the structural proteinase (S^{pro}).

Where HHPred was unable to annotate a region, a protein BLAST search was carried out to identify likely homologs among other known nidoviruses. When a match was found, both proteins were aligned using Clustal Omega (Sievers et al., 2011), and the multiple sequence alignment was used in HHPred. The most consistent matches to AAbV were from TurrNV. This identified a larger region and a more confident match to the coronavirus nsp14 ExoN-N7 MTase.

Protein BLAST was used to map the AAbV nidovirus RdRp-associated nucleotidyl transferase (NiRAN) and nsp16 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 2O 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 nsp7-10 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^{pro} 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 uridylate-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 S^{pro} in AAbV pp2. TMHMM (Krogh et al., 2001) predicted 13 transmembrane helices in pp2, which were generally arranged in pairs with large intervening domains, which we have tentatively named S^{pro}, 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. SignalP (Petersen et al., 2011) predicted an initial signal peptide at the extreme amino terminus, but after removing the predicted signal peptide and re-running the prediction with the "N-terminal truncation of input sequence" parameter set to zero, a total of six potential signal peptidase cleavage sites were detected. The identification of the nucleoprotein-like domain is based on a resemblance to the N proteins of *Bovine torovirus* and *Alphamesonivirus 1,* and to the carboxyl-terminal half of the SARS-CoV N. The features the AAbV N-like protein shares with N of other established nidoviruses are an initial glycine-rich region that may be flexibly disordered, followed by a lysine and arginine-rich region from amino acid 2869-2913 that could facilitate RNA binding, followed by a domain predicted by PSIPRED (Buchan et al., 2013) to contain a secondary structure profile similar to that of the Equine arteritis virus N and the SARS-CoV N carboxyl-terminal domain. We did not find strong evidence to support the analysis of Debat (Debat, 2018) predicting a spike-like fold in GP3, but we concur with Debat in noticing that GP2 (and we would add, GP3) have a protein secondary structure profile that resembles an alphavirus E1 protein and the E1-like protein of TurrNV.

One previous report (Prince, 2003) had noted virus-like particles described as resembling intracellular alphavirus virions, that were widespread in transmission electron micrographs of Aplysia californica tissue, which would seem to be consistent with the alphavirus-like organization of the structural polyprotein and apparent E1 homology. However, further testing is necessary to confirm whether those virus-like particles are related to AAbV.

AAbV Proteinases

When identifying viruses through bioinformatics, there is a risk that the sequences are either mis-assembled, contain errors, or are artifacts of the sequencing and sequence assembly processes. We tested the function of some AAbV protein features to determine if any was biologically functional, as a way to better assess whether the AAbV genome represented a replicating virus encoding functional parts.

The AAbV M^{pro} and S^{pro} plus surrounding regions up to the nearest preceding and following predicted transmembrane helix were cloned into pTriEx 1.1 and expressed with an amino-terminal herpes simplex virus epitope (HSV) tag, and a carboxyl-terminal poly-histidine (HIS) tag. Expressions were carried out by *in vitro* coupled T7 transcription and rabbit reticulocyte lysate translation. M^{pro} cleavage at an amino-terminal site was detected by the presence of an approximately 16 kDa HSV-tagged fragment (Fig. 6), which would be expected if M^{pro} cleavage occurred in the vicinity of amino acid 4375, located near the start of the region of M^{pro} homology at amino acid 4401 (Fig. 3D). S^{pro} was expressed, but did not produce any detectable cleavage products in the same assay (data not shown). From this we concluded that AAbV M^{pro} appeared to have proteinase activity in the context of our expression construct, while our S^{pro} construct did not. Further work will be needed to determine whether the failure of the putative S^{pro} to cleave was a result of the construct boundaries, assay conditions, lack of an appropriate substrate, or errors in the protein sequence.

To further characterize the activity of AAbV M^{pro}, alanine-scanning mutations were made to amino acids that appeared to match the catalytic cysteine and histidine residues of other coronavirus main proteinases. Mutation of the putative catalytic histidine H4429 did not strongly reduce proteolytic processing, while mutation of the cysteine C4538 blocked proteinase activity (Fig. 6). These data demonstrate that AAbV encodes at least one functional proteinase, but further work is needed to determine the cleavage specificity and map proteolytic processing by the AAbV M^{pro}.

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 to 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 (∆42; 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.

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. The host organism of MLeV is shown in Fig. 8B. Mapping single sequence reads onto the genome revealed a strong age dependence of MLeV detection. The number of fragments per kilobase of transcript per million mapped reads decreased by seven-fold from pre-metamorphosis to metamorphic climax, then decreased again by fourteen-fold from metamorphic climax to completion of metamorphosis. Further testing was done by reverse transcriptase polymerase chain reaction using MLeV-specific primers on the same population of adult frogs later in the year, but all the adult material tested was negative for MLeV (LZ, personal communication).

The MLeV genome is missing the 5’-end of the genome, including a 5’-untranslated region and sequences corresponding to coronavirus nsp1, nsp2 and part of nsp3. 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-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 fissipes* 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.

MLeV Protein Bioinformatics

In the pp1a region, HHPred detected matches for conserved coronavirus domains including the carboxyl-terminal domain of coronavirus nsp4, M^{pro}, 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.

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; Lokugamage et al., 2015; Mielech et al., 2014). It is worth noting that the region preceding the M^{pro} 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 M^{pro} 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 (Lokugamage 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 metazoa, 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 at taxonomic ranks above the family level.

Materials and methods

Phylogeny

Nidovirus phylogeny was reconstructed based on MSA of concatenated M^{pro}, NiRAN, RdRp, CH cluster and SF1 Helicase conserved cores (3417-3905, 5441-5866, 6095-7291, 7340-7504, 7781-8545 nt of of the Equine arteritis virus genome X53459.3), prepared with the help of Viralis 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 ModelFinder (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 visualised 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).

Protein assays

Nucleotides 12926-14176 containing the AAbV M^{pro} 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 (CCCCGAGGATCTCGAGTTGCGAATGATTTTGTCTACC) and Aby_IF_MP_R (GATGGTGGTGCTCGAGACACAGACAACACAACAAAAA) with 1x Phusion High Fidelty PCR Mastermix (Thermo Fisher Scientific). The 1283 bp PCR product was gel extracted using a QIAquick gel extraction kit (Qiagen) and cloned into pTriEx1.1 (Novagen / Merck) linearised with *Xho*I using In-Fusion HD cloning reagents (Clontech). 2 µl of the In-Fusion reaction was transformed into Stellar chemically competent cells as per the manufacturers protocol (Clontech) and selected on LB agar containing 100 ug/mL ampicillin. The final construct with a T7 RNA polymerase promoter and in-frame amino-terminal HSV and carboxyl-terminal HIS tags was verified by Sanger sequencing (Source Bioscience) of plasmid DNA purified using a QIAquick spin miniprep kit (Qiagen). Site-directed mutagenesis was carried out using the Quikchange II (Agilent) reagents and protocol. Protein expression was carried out in a 50 µl reaction volume using 0.5 µg of plasmid DNA with the TnT® Quick Coupled Transcription/Translation System (Promega) reagents and protocol. In vitro transcription and translation was carried out for 1h.

Samples containing expressed proteins were mixed with an equal volume of 2× SDS PAGE loading buffer containing 100mM Tris-HCL pH6.8, 4% w/v SDS, 20% w/v glycerol, 0.2% bromophenol blue, 2% β- mercaptoethanol. Samples were boiled at 100ºC for 10 minutes, collected by gentle centrifugation, and loaded in Mini-PROTEAN precast polyacrylamide gels (BioRad). After electrophoresis, proteins were blotted to PVDF membranes for 80 mins at 150mA using a Trans-Blot Turbo (BioRad). Membranes were blocked overnight at 4ºC with 5% (w/v) non-fat milk powder in TBST (50 mM Tris, 150 mM NaCl, 0.1% Tween 20, pH 7.5). Membranes were then washed three times for 5 min each on a rocking platform at 25 rpm with TBST buffer before addition unconjugated rabbit anti-HIS tag monoclonal antibody (Abcam) or unconjugated rabbit anti-HSV tag monoclonal antibody (Abcam) for 1 hour. Membranes were again washed three times for 5 min each with TBST buffer before addition of horseradish peroxidase-conjugated goat anti-rabbit secondary antibody for 1 hour.For detection, ChemiFast chemiluminescent reagent (Syngene) was used to detect bound secondary antibody. Samples were visualized using a Syngene Chemi XL G:Box gel documentation system. Gel images were cropped and brightness and contrast of images was adjusted using GIMP software (GIMP team).

The region from the pp1a-pp1b junction containing the putative termination-suppression signal of AAbV, nucleotides 17255-17707, was PCR amplified from a synthetic GeneArt Strings fragment (Invitrogen) using primers Aby_IF_SS_F (CCCCGAGGATCTCGAGGAGTCTTGTCGTGTGAAGT) and Aby_IF_SS_R (GATGGTGGTGCTCGAGAGGATTAATCCGTCTGTCAA). The predicted S^{pro}-containing region of AAbV, nucleotides 25918-27183, was PCR amplified from a synthetic GeneArt Strings fragment (Invitrogen) using primers Aby_IF_TryP_R (GATGGTGGTGCTCGAGCGGTTTGTTCGCATACAGA) and Aby_IF_TryP_R (GATGGTGGTGCTCGAGCGGTTTGTTCGCATACAGA). Both the S^{pro} and putative pp1a-pp1b termination-suppression signal products were cloned, expressed and detected in the same way as AAbV M^{pro}.

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 PhD student with Alexander E. Gorbalenya (A.E.G.) and her work and resources she used were partially supported by EU Horizon2020 EVAg 653316 grant and LUMC 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.

Figure Legends

Figure 1. Nidovirus phylogeny reconstructed based on concatenated MSA of five replicative domains universally conserved in nidoviruses. SH-aLRT branch support values are depicted by shaded circles. Species names that are not currently recognized by ICTV are written in plain font. Asterisks designate viruses described in this study.

Figure 2.   Sequence coverage of AAbV in public NCBI libraries. (A) Examples of the host organism *Aplysia californica* at swimming veliger, settled, metamorphic, juvenile and adult developmental stages (images not to scale, adapted from (Heyland et al., 2011; Moroz et al., 2006)). Summary of distinct sequence assemblies and reads in the TSA (B) and EST (C) matching AAbV for which the nucleotide BLAST E value was 2×10^{-70} or smaller. (D) Map of AAbV, showing the location of the replicase polyprotein genes (ORF1a, ORF1b), structural polyprotein gene (ORF2) and poly-adenosine tail (A_n). The position of sequences from the TSA (E) and EST (F) databases matching AAbV is shown.

Figure 3. Coding capacity, depth of coverage and bioinformatics of AAbV.   (A) Genome and coding capacity of AAbV and SARS-CoV are shown to scale. (B) Total depth of coverage based on a sample of 672017 aligned spots matching AAbV from

Aplysia californica RNA sequence read archives including SRR385787, SRR385788, SRR385792, SRR385793, SRR385795, SRR385800, SRR385802 and SRR385815. The putative start site of a viral subgenomic RNA species is marked with an arrow. (C) Alignment of the 5'-untranslated region and the intergenic sequence between the pp1b and pp2 genes showing a potential transcription-regulatory sequence (boxed). (D) Bioinformatic assignment of domains in AAbV. Sequence(s) used for prediction (Input) were either AAbV alone or a multiple sequence alignment containing AAbV and TurrNV. Probability score from HHPred and E value from HHPred or BLAST are shown. Accession numbers are given for sequences or protein structures identified as a match for an AAbV domain (Model).

Figure 4. Comparison of predicted domain-level organization in polyprotein 1a of new viruses to previously described nidoviruses. Gaps have been introduced so to align predicted homologous domains. Virus naming and taxonomy conventions follow the ICTV proposals in which MLeV and AAbV were first described (Gorbalenya et al., 2017b, 2017a; Ziebuhr et al., 2017). New viruses are marked with stars, accepted taxonomic ranks are italicized and proposed taxonomic ranks are not italicized. Polyprotein processing products from SARS-CoV are shown at top. Domains are colored to indicate predicted similarity to SARS-CoV nsp1 (CoV nsp1), SARS-CoV nsp2 (nsp2-like), ubiquitin (Ub-like), macrodomains, papain-like proteinase (PL^{pro}), first section of the coronavirus Y domain (CoV Y1), first section of the arterivirus Y domain (ArV Y1) coronavirus-specific Y domain-like (CoV Y-like), carboxyl-terminal domain of coronavirus nsp4 (nsp4 CTD-like), region with PSIPRED predicted structural similarity to nsp4 CTD, main proteinase (M^{pro}), SARS-CoV nsp8-like (CoV nsp8), Equine arteritis virus nsp7α (ArV nsp7α), SARS-CoV nsp10 (CoV nsp10), protein kinase-like (Kinase), RNA methyltransferase (Mtase), potential metal ion-binding clusters with 4 cysteine or histidine residues in a 20 amino acid window (CH-cluster), transmembrane helices, hydrophobic transmembrane-like regions that may not span the membrane by analogy to coronavirus nsp4 and nsp6 (TM-like) and disordered regions (Unstructured).

Figure 5. Comparison of predicted domain-level organization in polyprotein 1b of new viruses to previously described nidoviruses. (A) Domains include the nidovirus RdRp-associated nucleotidyl transferase (NiRAN), RdRp, potential metal

ion binding clusters with four cysteine or histidine residues in a window of 20 amino acids (CH cluster), homologs of the domain of unknown function in the middle of coronavirus nsp13 (CoV nsp13b), superfamily 1 helicase (SF1 Helicase), nidovirus-specific exonuclease (ExoN) and uridylate-specific endonuclease (NEndoU), RNA cap N7 methyltransferase (N7 MTase) and RNA cap 2'-O-methyltransferase (2O MTase). (B) Domains of pp2 include the structural protease (S^{pro}), putative glycoproteins GP1, GP2 and GP3, and a nucleoprotein-like domain (N?), TMHMM-predicted transmembrane domains and SignalP-predicted signal peptidase cleavage sites.

Figure 6. Investigation of proteinase activity of AAbV M^{pro}. The AAbV main proteinase (M^{pro}; A-B) and surrounding regions were expressed as HSV and HIS-tagged constructs as shown in panel A. A white triangle marks the expected size of the 52.5 kDa uncleaved M^{pro} constructs. Black triangles mark the size of approximately 16 kDa amino-terminal cleavage products. Non-specific bands that were also present in control lanes are indicated with a star.

Figure 7. Mutational analysis of the termination-suppression signal (TSS) at the ORF1a/b junction. (A) Schematic view of the TSS expression construct and introduced HSV and HIS tags, showing only predicted RNA secondary structures that were consistent in the best six models generated by Mfold. Mutations around the stop codon (bold, producing the UAAA construct) or removing one side of the predicted stem-loops (Δ42) are shown. (B-D) Western blots showing translation of mutant TSS expression constructs in a coupled T7 polymerase rabbit reticulocyte lysate expression system. Blots were probed with anti-HSV (B, D) to detect both 25 kDa terminated and 32-35 kDa readthrough products, or with anti-HIS (C) to detect only readthrough products.

Figure 8. Coding capacity and prevalence of MLeV (A) Schematic representation of the coding capacity of MLeV compared to SARS-CoV, showing the similarities in genome organization. (B) Prevalence of MLeV transcripts in *Microhyla fissipes* by age, by total number of reads and fragments per kilobase of transcript per million mapped reads (FPKM).

Figure 9. Depth of coverage and bioinformatics of MLeV. (A) Total depth of coverage is based on 275503 aligned spots matching MLeV from Microhyla fissipes RNA sequence read archives SRR2418812, SRR2418623 and SRR2418554. The putative start sites of a viral subgenomic RNA species are marked with an arrow. Potential subgenomic RNA start sites not marked by a sharp rise in read depth are indicated with question marks. (B) Positions and usage of putative transcription-regulatory sequences. Termination codons from the preceding gene are underlined, initiation codons of the following gene are in bold. (C) Bioinformatic assignment of domains in MLeV.

Figure 10. Speculative annotation of nidovirus structural proteins. Where structures or functions were not known, proteins were categorized according to general PSIPRED secondary structure profile. Marked domains include coronavirus spike protein homologs (Spike) and structurally similar regions (β-α), alphavirus E1 homologs (E1) and structurally similar regions (βαβ), coronavirus envelope-like proteins (E-like), coronavirus membrane proteins (M-like) and structurally similar proteins (β), potential nucleoprotein (N-like), chymotrypsin-like structural proteinase (S^{pro}), similar to the bovine viral diarrhea virus structural RNAse (BVDV RNAse), proteins related to influenza A virus hemagglutinin (HA) or torovirus hemagglutinin-esterase (HE), other viral surface glycoproteins (GP-like), domains of no known function (Unknown), SignalP-predicted signal peptidase cleavage sites (SP cleavage), and potential sites cleaved by unknown proteinases by analogy to other nidovirus structural proteins.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2

Anand, K., Palm, G.J., Mesters, J.R., Siddell, S.G., Ziebuhr, J., Hilgenfeld, R., 2002. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. EMBO J. 21, 3213–3224. https://doi.org/10.1093/emboj/cdf327

Bailey-Elkin, B.A., Knaap, R.C.M., Johnson, G.G., Dalebout, T.J., Ninaber, D.K., Van Kasteren, P.B., Bredenbeek, P.J., Snijder, E.J., Kikkert, 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.M114.609644

Birktoft, J.J., Blow, D.M., 1972. Structure of crystalline α-chymotrypsin. V. The atomic structure of tosyl-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 2017.012S Expansion of the rank structure of the family Arteriviridae and renaming its taxa.

Buchan, D.W.A., Minneci, F., Nugent, T.C.O., Bryson, K., Jones, D.T., 2013. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res. 41. https://doi.org/10.1093/nar/gkt381

Chen, Y., Cai, H., Pan, J., Xiang, N., Tien, P., Ahola, T., Guo, D., 2009. Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase. Proc. Natl. Acad. Sci. 106, 3484–3489. https://doi.org/10.1073/pnas.0808790106

Chernomor, O., Von Haeseler, A., Minh, B.Q., 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst. Biol. 65, 997–1008. https://doi.org/10.1093/sysbio/syw037

Cridge, A.G., Crowe-Mcauliffe, C., Mathew, S.F., Tate, W.P., 2018. Eukaryotic translational termination efficiency is influenced by the 3′ nucleotides within the ribosomal mRNA channel. Nucleic Acids Res. 46, 1927–1944. https://doi.org/10.1093/nar/gkx1315

Debat, H.J., 2018. Expanding the size limit of RNA viruses: Evidence of a novel divergent nidovirus in California sea hare, with a ~35.9 kb virus genome. bioRxiv.

Deng, Z., Lehmann, K.C., Li, X., Feng, C., Wang, G., Zhang, Q., Qi, X., Yu, L., Zhang, X., Feng, W., Wu, W., Gong, P., Tao, Y., Posthuma, C.C., Snijder, E.J., Gorbalenya, A.E., Chen, Z., 2014. Structural basis for the regulatory function of a complex zinc-binding domain in a replicative arterivirus helicase resembling a nonsense-mediated mRNA decay helicase. Nucleic Acids Res. 42, 3464–3477. https://doi.org/10.1093/nar/gkt1310

Feng, Y.X., Yuan, H., Rein, A., Levin, J.G., 1992. Bipartite signal for read-through suppression in murine leukemia virus mRNA: an eight-nucleotide purine-rich sequence immediately downstream of the gag termination codon followed by an RNA pseudoknot. J. Virol. 66, 5127–5132.

Fiedler, T.J., Hudder, A., McKay, S.J., Shivkumar, S., Capo, T.R., Schmale, M.C., Walsh, P.J., 2010. The transcriptome of the early life history stages of the California Sea Hare Aplysia californica. Comp. Biochem. Physiol. Part D. Genomics Proteomics 5, 165–70. https://doi.org/10.1016/j.cbd.2010.03.003

Furuya, T., Macnaughton, T.B., La Monica, N., Lai, M.M.C., 1993. Natural evolution of coronavirus defective-interfering rna involves rna recombination. Virology 194, 408–413. https://doi.org/10.1006/viro.1993.1277

Gorbalenya, A.E., Brinton, M.A., Cowley, J., de Groot, R., Gulyaeva, A., Lauber, C., Neuman, B.W., Ziebuhr, J., 2017a. ICTV Pending Proposal 2017.015S. Reorganization and expansion of the order Nidovirales at the family and sub-order ranks.

Gorbalenya, A.E., Brinton, M.A., Cowley, J., de Groot, R., Gulyaeva, A., Lauber, C., Neuman, B.W., Ziebuhr, J., 2017b. ICTV Pending Proposal 2017.014S. Establishing taxa at the ranks of subfamily, genus, sub-genus and species in six families of invertebrate nidoviruses.

Gorbalenya, A.E., Lieutaud, P., Harris, M.R., Coutard, B., Canard, B., Kleywegt, G.J., Kravchenko, A.A., Samborskiy, D. V., Sidorov, I.A., Leontovich, A.M., Jones, T.A., 2010. Practical application of bioinformatics by the multidisciplinary VIZIER consortium. Antiviral Res. https://doi.org/10.1016/j.antiviral.2010.02.005

Heyland, A., Vue, Z., Voolstra, C.R., Medina, M., Moroz, L.L., 2011. Developmental transcriptome of Aplysia californica'. J. Exp. Zool. Part B Mol. Dev. Evol. 316 B, 113–134. https://doi.org/10.1002/jez.b.21383

Ivanov, K.A., Thiel, V., Dobbe, J.C., van der Meer, Y., Snijder, E.J., Ziebuhr, J., 2004. Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase. J. Virol. 78, 5619–32. https://doi.org/10.1128/JVI.78.11.5619-5632.2004

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589. https://doi.org/10.1038/nmeth.4285

King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., 2012. Togaviridae, in:

Virus Taxonomy. pp. 1103–1110.

Krogh, A., Larsson, B., Von Heijne, G., Sonnhammer, E.L.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. 305, 567–580. https://doi.org/10.1006/jmbi.2000.4315

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Lauber, C., Goeman, J.J., de Parquet, M.C., Thi Nga, P., Snijder, E.J., Morita, K., Gorbalenya, A.E., 2013. The Footprint of Genome Architecture in the Largest Genome Expansion in RNA Viruses. PLoS Pathog. 9. https://doi.org/10.1371/journal.ppat.1003500

Lauck, M., Alkhovsky, S. V, Bào, Y., Bailey, A.L., Shevtsova, Z. V, Shchetinin, A.M., Vishnevskaya, T. V, Lackemeyer, M.G., Postnikova, E., Mazur, S., Wada, J., Radoshitzky, S.R., Friedrich, T.C., Lapin, B. a, Deriabin, P.G., Jahrling, P.B., Goldberg, T.L., O'Connor, D.H., Kuhn, J.H., 2015. Historical outbreaks of simian hemorrhagic fever in captive macaques were caused by distinct arteriviruses. J. Virol. 89, 8082–7. https://doi.org/10.1128/JVI.01046-15

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Data, G.P., Sam, T., Subgroup, 1000 Genome Project Data Processing, 2009. The Sequence Alignment / Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lokugamage, K.G., Narayanan, K., Nakagawa, K., Terasaki, K., Ramirez, S.I., Tseng, C.-T.K., Makino, S., 2015. Middle East Respiratory Syndrome Coronavirus nsp1 Inhibits Host Gene Expression by Selectively Targeting mRNAs Transcribed in the Nucleus while Sparing mRNAs of Cytoplasmic Origin. J. Virol. 89, 10970–81. https://doi.org/10.1128/JVI.01352-15

Ma, Y., Wu, L., Shaw, N., Gao, Y., Wang, J., Sun, Y., Lou, Z., Yan, L., Zhang, R., Rao, Z., 2015. Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. Proc. Natl. Acad. Sci. U. S. A. 112, 9436–41. https://doi.org/10.1073/pnas.1508686112

Melancont, P., Garoff, H., 1987. Processing of the Semliki Forest virus structural polyprotein: role of the capsid protease. J. Virol. 61, 1301–1309.

Mielech, A.M., Chen, Y., Mesecar, A.D., Baker, S.C., 2014. Nidovirus papain-like proteases: Multifunctional enzymes with protease, deubiquitinating and deISGylating activities. Virus Res. 194, 184–190. https://doi.org/10.1016/j.virusres.2014.01.025

Miranda, J.A., Culley, A.I., Schvarcz, C.R., Steward, G.F., 2016. RNA viruses as major contributors to Antarctic virioplankton. Environ. Microbiol. https://doi.org/10.1111/1462-2920.13291

Moroz, L.L., Edwards, J.R., Puthanveettil, S. V., Kohn, A.B., Ha, T., Heyland, A., Knudsen, B., Sahni, A., Yu, F., Liu, L., Jezzini, S., Lovell, P., Iannucculli, W., Chen, M., Nguyen, T., Sheng, H., Shaw, R., Kalachikov, S., Panchin, Y. V., Farmerie, W., Russo, J.J., Ju, J., Kandel, E.R., 2006. Neuronal Transcriptome of Aplysia: 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, in: Advances in Virus Research. pp. 1–27. https://doi.org/10.1016/bs.aivir.2016.08.005

Neuman, B.W., Chamberlain, P., Bowden, F., Joseph, J., 2014. Atlas of coronavirus replicase structure. Virus Res. 194, 49–66. https://doi.org/10.1016/j.virusres.2013.12.004

Nguyen, L.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

O'Dea, M.A., Jackson, B., Jackson, C., Xavier, P., Warren, K., 2016. Discovery and partial genomic characterisation of a novel nidovirus associated with respiratory disease in wild shingleback lizards (Tiliqua rugosa). PLoS One 11. https://doi.org/10.1371/journal.pone.0165209

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of phylogenetics and evolution in R language. 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.1701

Prince, J.S., 2003. A presumptive alphavirus in the gastropod mollusc, Aplysia californica. Bull. Mar. Sci. 73, 673–677.

R Development Core Team, R., 2011. R: A Language and Environment for Statistical

Computing, R Foundation for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7

Saberi, A., Gulyaeva, A.A., Brubacher, J., Newmark, P.A., Gorbalenya, A., 2018. A planarian nidovirus expands the limits of RNA genome size. bioRxiv.

Sapay, N., Guermeur, Y., Deléage, G., 2006. Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. BMC Bioinformatics 7. https://doi.org/10.1186/1471-2105-7-255

Schliep, K.P., 2011. phangorn: Phylogenetic analysis in R. Bioinformatics 27, 592–593. https://doi.org/10.1093/bioinformatics/btq706

Shi, M., Lin, X.D., Chen, X., Tian, J.H., Chen, L.J., Li, K., Wang, W., Eden, J.S., Shen, J.J., Liu, L., Holmes, E.C., Zhang, Y.Z., 2018. The evolutionary history of vertebrate RNA viruses. Nature 556, 197–202. https://doi.org/10.1038/s41586-018-0012-7

Shi, M., Lin, X.D., Tian, J.H., Chen, L.J., Chen, X., Li, C.X., Qin, X.C., Li, J., Cao, J.P., Eden, 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/nature20167

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, 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. https://doi.org/10.1038/msb.2011.75

Skuzeski, 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, 365–373. https://doi.org/10.1016/0022-2836(91)90718-L

Söding, J., Biegert, A., Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33. https://doi.org/10.1093/nar/gki408

Sola, I., Almazán, F., Zúñiga, S., Enjuanes, L., 2015. Continuous and Discontinuous RNA Synthesis in Coronaviruses. Annu. Rev. Virol. 2, 265–288. https://doi.org/10.1146/annurev-virology-100114-055218

te Velthuis, A.J.W., Arnold, J.J., Cameron, C.E., van den Worm, S.H.E., Snijder, E.J., 2009. The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. Nucleic Acids Res. 38, 203–214. https://doi.org/10.1093/nar/gkp904

Tokarz, R., Sameroff, S., Hesse, R.A., Hause, B.M., Desai, A., Jain, K., Ian Lipkin, W., 2015. Discovery of a novel nidovirus in cattle with respiratory disease. J. Gen. Virol. 96, 2188–2193. https://doi.org/10.1099/vir.0.000166

Vasilakis, N., Guzman, H., Firth, C., Forrester, N.L., Widen, S.G., Wood, T.G., Rossi, S.L., Ghedin, E., Popov, V., Blasdell, K.R., Walker, P.J., Tesh, R.B., 2014. Mesoniviruses are mosquito-specific viruses with extensive geographic distribution and host range. Virol. J. 11. https://doi.org/10.1186/1743-422X-11-97

Wahl-Jensen, V., Johnson, J.C., Lauck, M., Weinfurter, J.T., Moncla, L.H., Weiler, A.M., Charlier, O., Rojas, O., Byrum, R., Ragland, D.R., Huzella, L., Zommer, E., Cohen, M., Bernbaum, J.G., Caì, Y., Sanford, H.B., Mazur, S., Johnson, R.F., Qin, J., Palacios, G.F., Bailey, A.L., Jahrling, P.B., Goldberg, T.L., O’Connor, D.H., Friedrich, T.C., Kuhn, J.H., 2016. Divergent simian arteriviruses cause simian hemorrhagic fever of differing severities in macaques. MBio 7. https://doi.org/10.1128/mBio.02009-15

Zeng, C., Wu, A., Wang, Y., Xu, S., Tang, Y., Jin, X., Wang, S., Qin, L., Sun, Y., Fan, C., Snijder, E.J., Neuman, B.W., Chen, Y., Ahola, T., Guo, D., 2016. Identification and Characterization of a Ribose 2′-O-Methyltransferase Encoded by the Ronivirus Branch of Nidovirales. J. Virol. 90, 6675–6685. https://doi.org/10.1128/JVI.00658-16

Zhao, L., Liu, L., Wang, S., Wang, H., Jiang, J., 2016. Transcriptome profiles of metamorphosis in the ornamented pygmy frog Microhyla fissipes clarify the functions of 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, B.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 Coronavirinae 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

Figure 1

Click here to download high resolution image

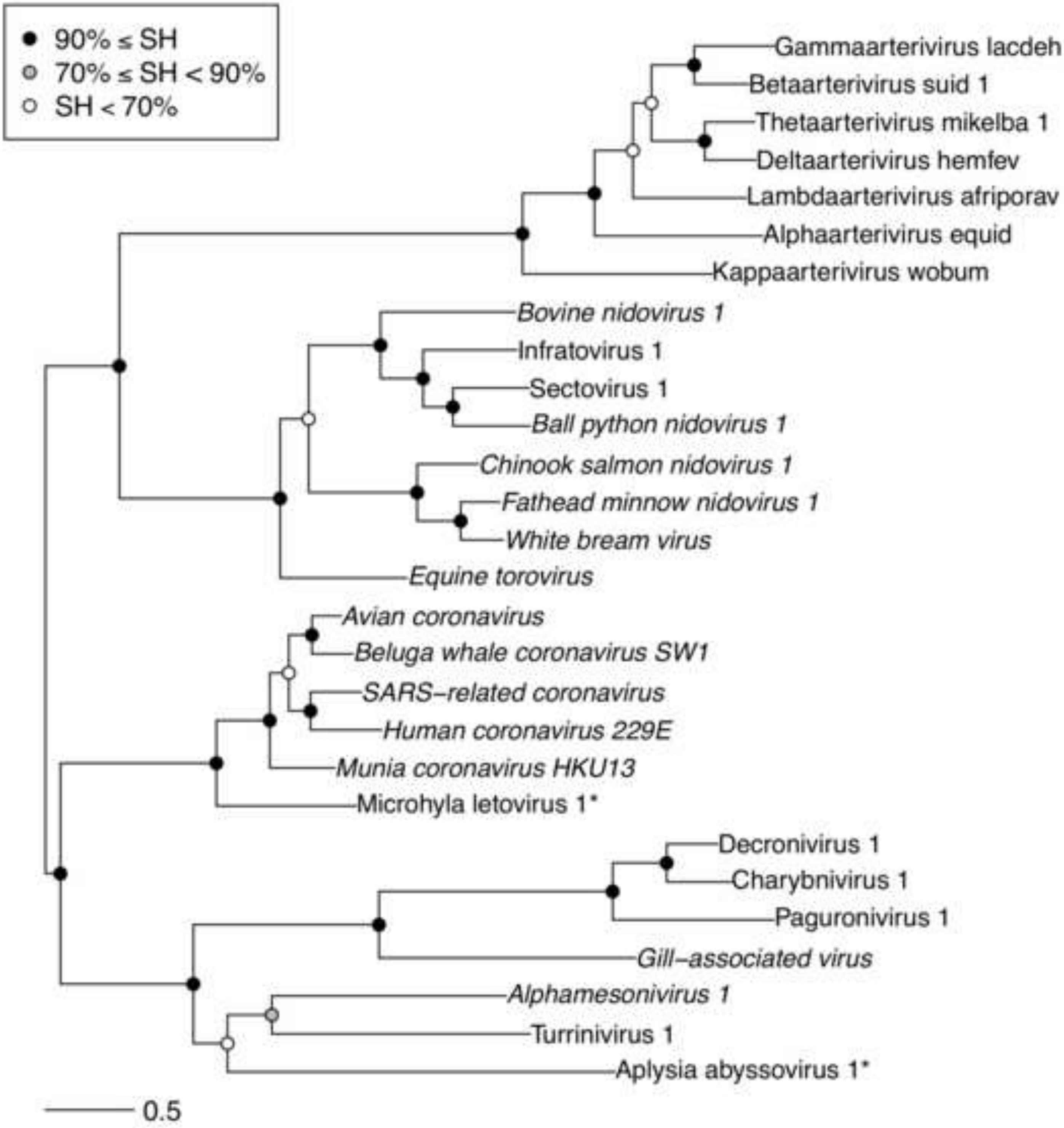
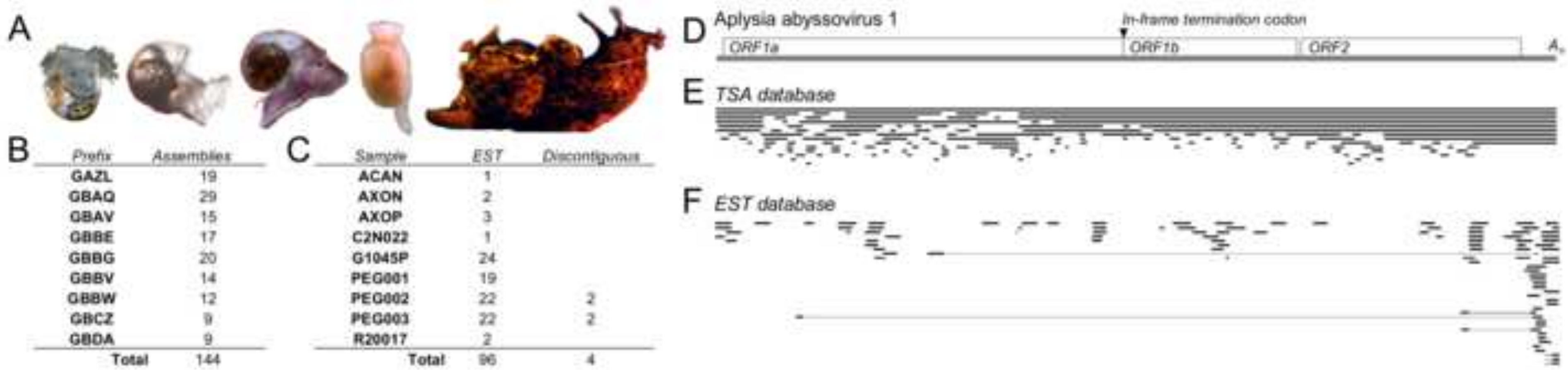


Figure 2



B

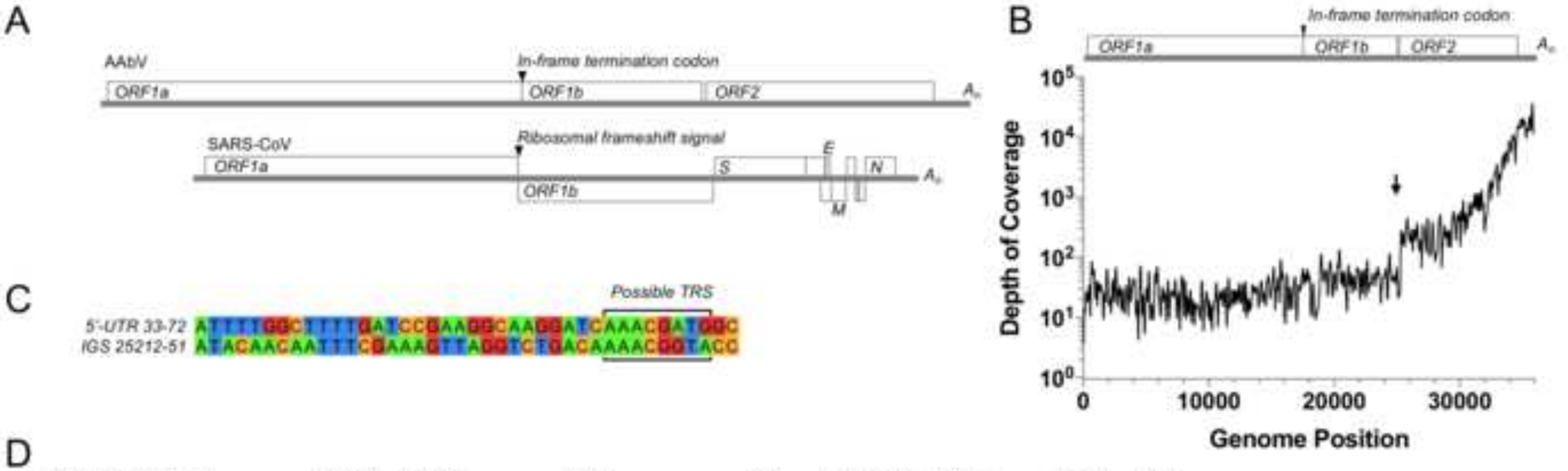
Prefix	Assemblies
GAZL	19
GBAQ	29
GBAV	15
GBBE	17
GBBG	20
GBBV	14
GBBW	12
GBCZ	9
GBDA	9
Total	144

C

Sample	EST	Discontiguous
ACAN	1	
AXON	2	
AXOP	3	
C2N022	1	
G1045P	24	
PEG001	19	
PEG002	22	2
PEG003	22	2
R20017	2	
Total	96	4

Figure 3

Click here to download high resolution image



D

Aplysia abyssovirus 1	Protein	Region	Input	Tool	Probability	E value	Model	Note
Main protease	pp1a	4401-4618	Abyssovirus 1 only	HHPred	96%	0.044	3D23	Main protease, Human coronavirus HKU1
Conserved pp1a	pp1a	5386-5629	Abyssovirus 1 only	Protein BLAST		1.0×10^{-7}	NC_032496	Unknown function, Turrinivirus 1
NiRAN	pp1ab	5953-6087	Abyssovirus 1 only	HHPred align	96%	9.4×10^{-8}	NP_828869	nsp12 NiRAN & RdRP, SARS-CoV
RNA dependent RNA polymerase	pp1ab	6497-6893	Abyssovirus 1 only	HHPred	99%	3.9×10^{-13}	2CKW	RNA dependent RNA polymerase, Sapporo virus
Metal binding helicase/NTPase	pp1ab	6963-7596	Abyssovirus 1 only	HHPred	100%	2.6×10^{-40}	5WWP	nsp13, Middle east respiratory syndrome-related-CoV
Exonuclease/N7 Methyltransferase	pp1ab	7628-7940	Alignment with Turrinivirus 1	HHPred	100%	1.4×10^{-89}	5C8T	nsp14 Exonuclease & N7 Methyltransferase, SARS-CoV
Exonuclease region	pp1ab	7655-7913	Abyssovirus 1 only	HHPred	87%	0.39	5C8T	nsp14 Exonuclease & N7 Methyltransferase, SARS-CoV
N7 Methyltransferase region	pp1ab	7891-8056	Abyssovirus 1 only	HHPred	32%	41	2P8J	SAM-dependent methyltransferase, Clostridium acetobutylicum
2O Methyltransferase	pp1ab	8248-8353	Abyssovirus 1 only	HHPred align	67%	3.2×10^{-5}	YP_001661452	nsp16 2'-O-methyltransferase, Gill-associated virus
Structural Protease	ppS	276-471	Abyssovirus 1 only	HHPred	99%	1.0×10^{-13}	2R0L	HGFA family S1 serine protease, Human

Figure 4

Click here to download high resolution image

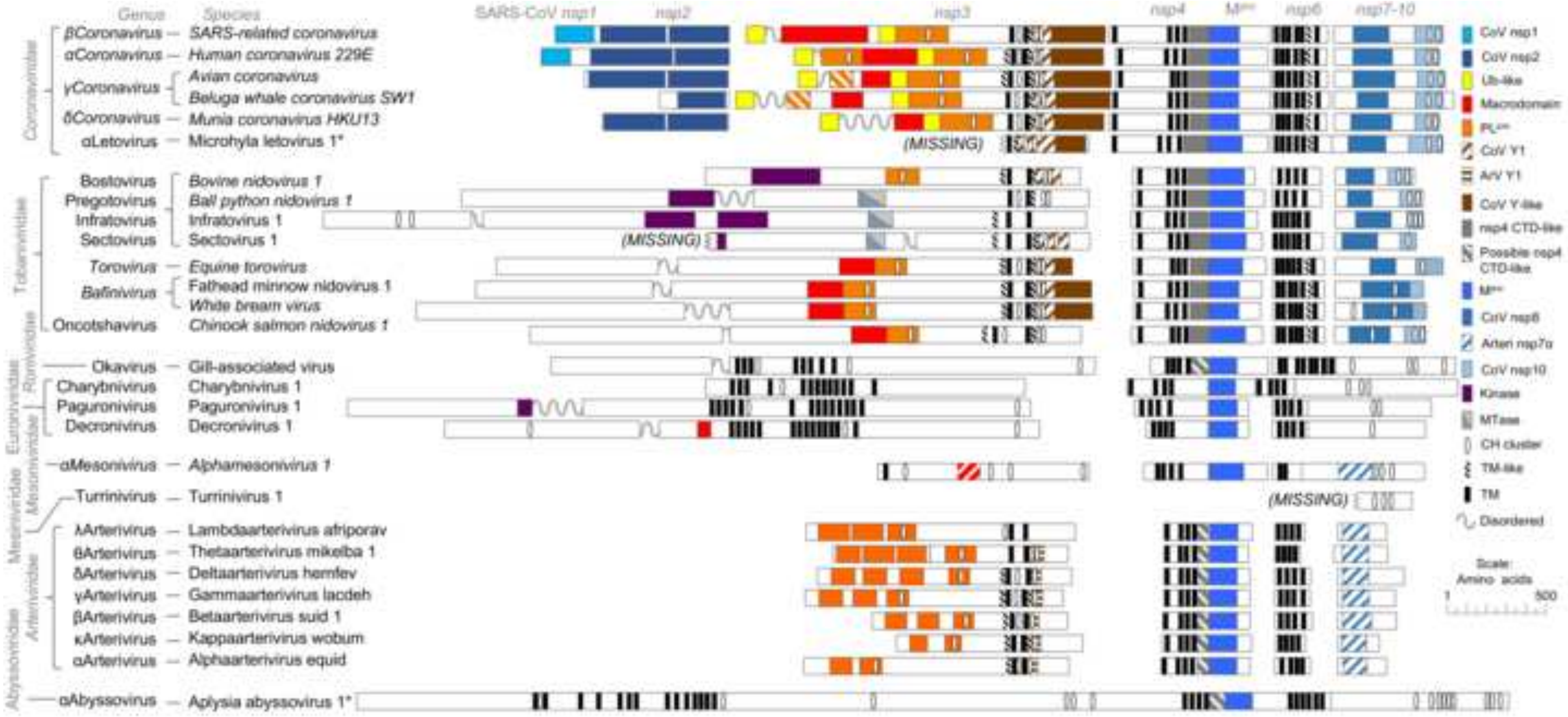

Genus
Species
SARS-CoV nsp1
nsp2
nsp3
nsp4
Mpro
nsp6
nsp7-10
Coronaviridae
βCoronavirus — SARS-related coronavirus
αCoronavirus — Human coronavirus 229E
γCoronavirus
Avian coronavirus
Beluga whale coronavirus SW1
δCoronavirus — Munia coronavirus HKU13
αLetovirus — Microhyla letovirus 1*
(MISSING)
Tobaniviridae
Bostovirus
Bovine nidovirus 1
Pregotovirus
Ball python nidovirus 1
Infratovirus
Infratovirus 1
Sectovirus
Sectovirus 1
(MISSING)
Torovirus — Equine torovirus
Bafinivirus
Fathead minnow nidovirus 1
White bream virus
Oncotshavirus
Chinook salmon nidovirus 1
Roniviridae
Euroniviridae
Okavirus — Gill-associated virus
Charybnivirus
Charybnivirus 1
Paguronivirus
Paguronivirus 1
Decronivirus
Decronivirus 1
Mesoniviridae
Mesnidovirineae
αMesonivirus — Alphamesonivirus 1
Turrinivirus — Turrinivirus 1
(MISSING)
Arteriviridae
λArterivirus — Lambdaarterivirus afriporav
θArterivirus — Thetaarterivirus mikelba 1
δArterivirus — Deltaarterivirus hemfev
γArterivirus — Gammaarterivirus lacdeh
βArterivirus — Betaarterivirus suid 1
κArterivirus — Kappaarterivirus wobum
αArterivirus — Alphaarterivirus equid
Abyssoviridae
αAbyssovirus — Aplysia abyssovirus 1*
CoV nsp1
CoV nsp2
Ub-like
Macrodomain
PLpro
CoV Y1
A/V Y1
CoV Y-like
nsp4 CTD-like
Possible nsp4 CTD-like
Mpro
CoV nsp6
Arteri nsp7α
CoV nsp10
Kinase
MTase
CH cluster
TM-like
TM
Disordered
Scale:
Amino acids
1
500

Figure 5

Click here to download high resolution image

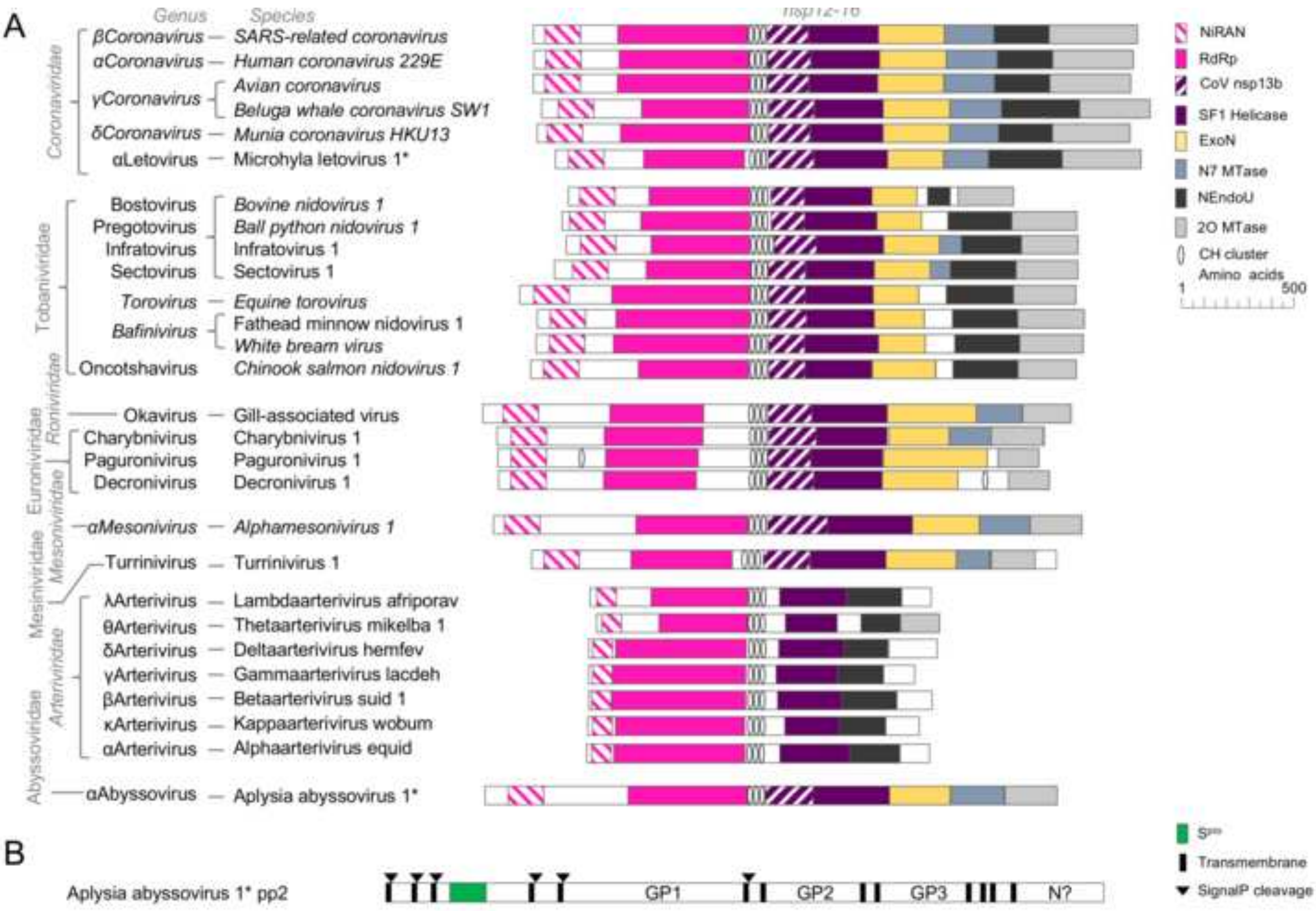

A
Genus
Species
nsp12-16
Coronaviridae
βCoronavirus — SARS-related coronavirus
αCoronavirus — Human coronavirus 229E
γCoronavirus
Avian coronavirus
Beluga whale coronavirus SW1
δCoronavirus — Munia coronavirus HKU13
αLetovirus — Microhyla letovirus 1*
Tobaniviridae
Bostovirus
Bovine nidovirus 1
Pregotovirus
Ball python nidovirus 1
Infratovirus
Infratovirus 1
Sectovirus
Sectovirus 1
Torovirus — Equine torovirus
Bafinivirus
Fathead minnow nidovirus 1
White bream virus
Oncotshavirus
Chinook salmon nidovirus 1
Roniviridae
Okavirus — Gill-associated virus
Euroniviridae
Charybnivirus
Charybnivirus 1
Paguronivirus
Paguronivirus 1
Decronivirus
Decronivirus 1
Mesoniviridae
αMesonivirus — Alphamesonivirus 1
Mesiniviridae
Turrinivirus — Turrinivirus 1
Arteriviridae
λArterivirus — Lambdaarterivirus afriporav
θArterivirus — Thetaarterivirus mikelba 1
δArterivirus — Deltaarterivirus hemfev
γArterivirus — Gammaarterivirus lacdeh
βArterivirus — Betaarterivirus suid 1
κArterivirus — Kappaarterivirus wobum
αArterivirus — Alphaarterivirus equid
Abyssoviridae
αAbyssovirus — Aplysia abyssovirus 1*
NiRAN
RdRp
CoV nsp13b
SF1 Helicase
ExoN
N7 MTase
NEndoU
2O MTase
CH cluster
Amino acids
1
500
B
Aplysia abyssovirus 1* pp2
GP1
GP2
GP3
N?
S2'''
Transmembrane
SignalP cleavage

Figure 6

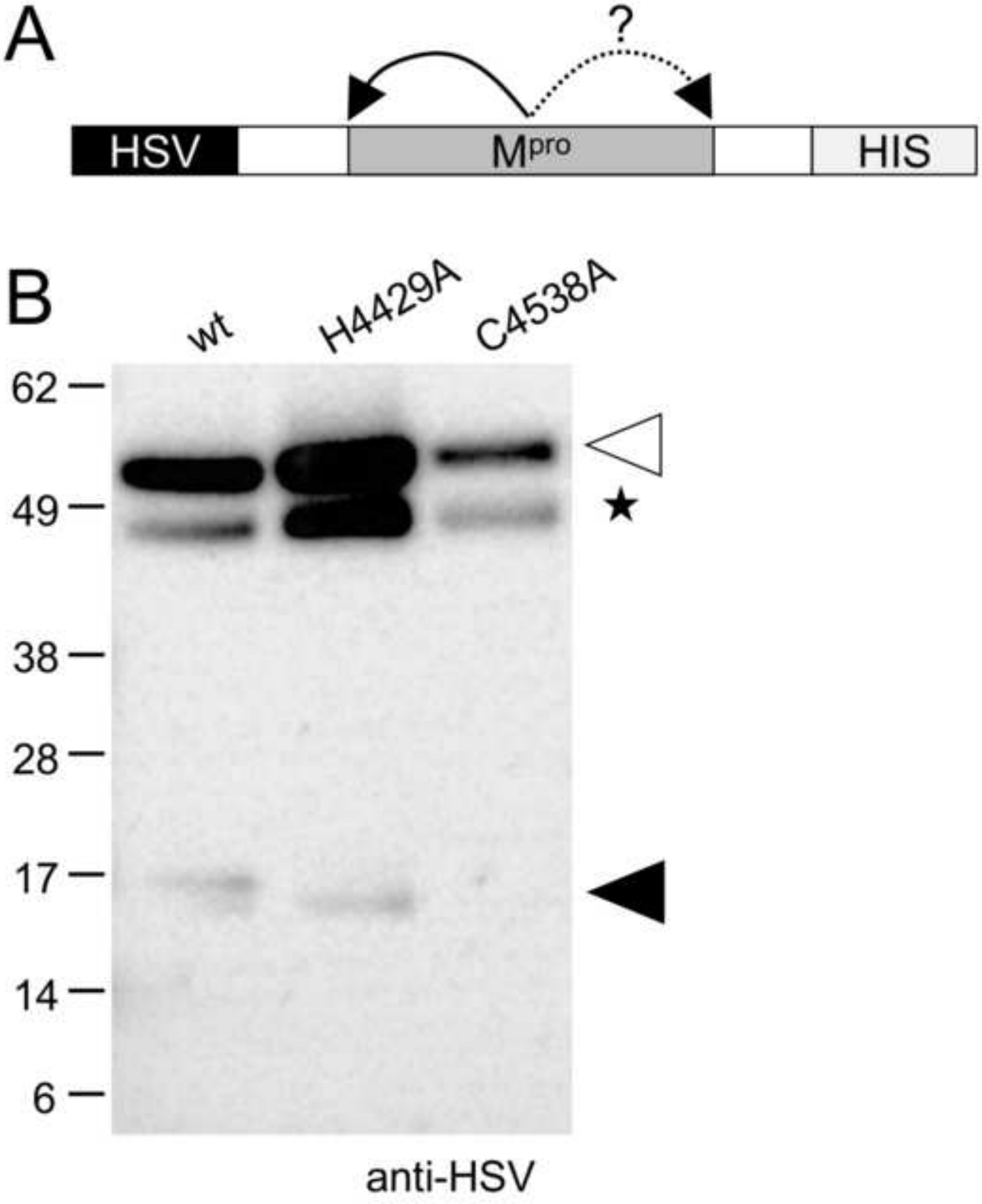


Figure 7

Click here to download high resolution image

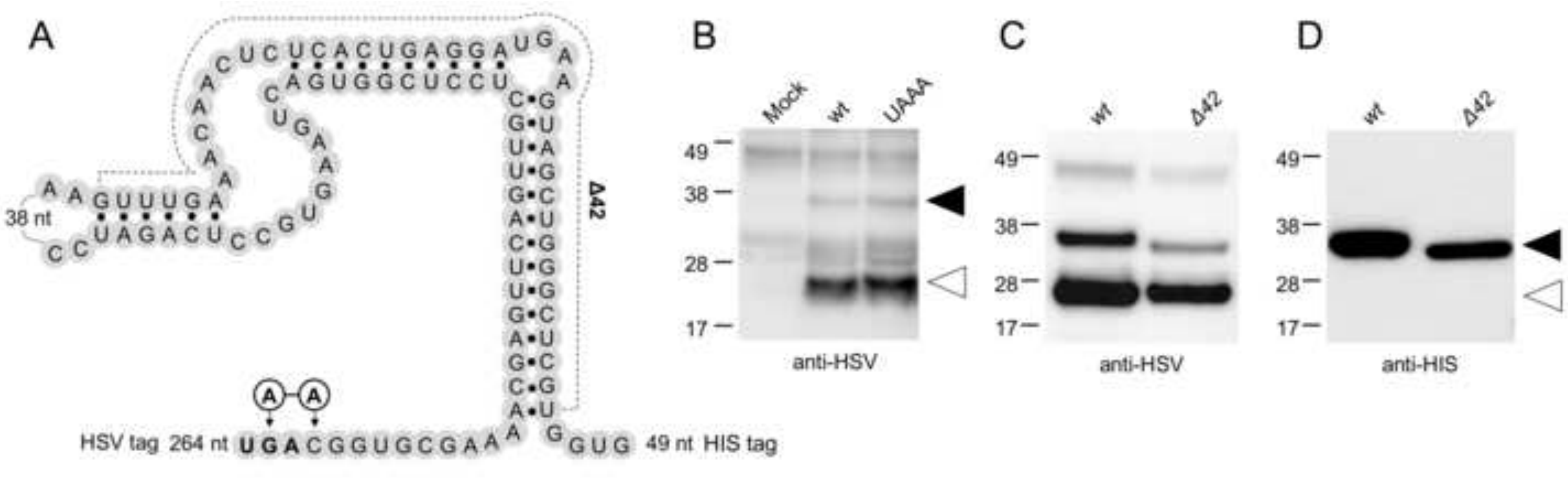
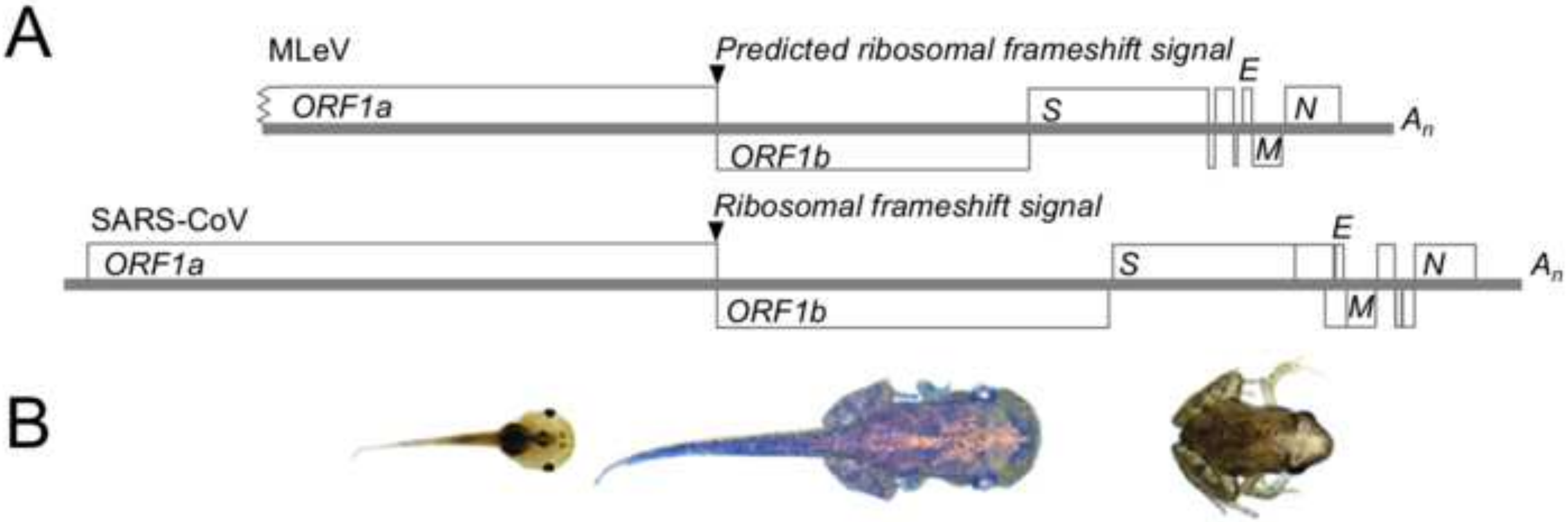


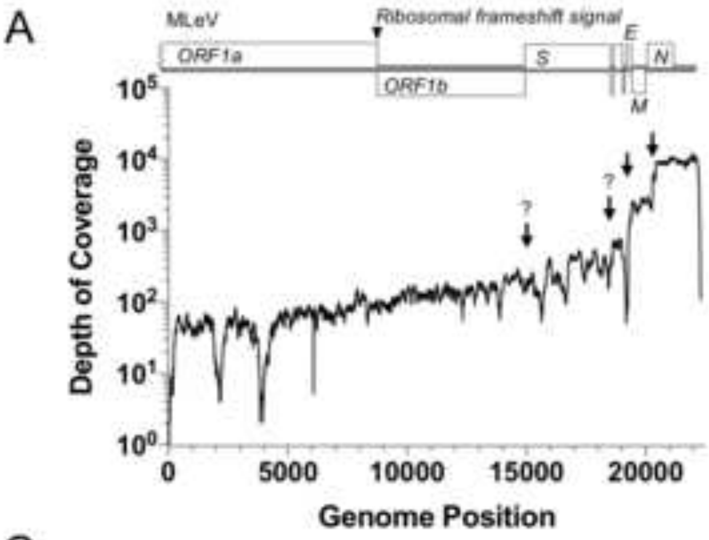
Figure 8


A
MLeV
Predicted ribosomal frameshift signal
ORF1a
ORF1b
S
E
M
N
A_n
SARS-CoV
Ribosomal frameshift signal
ORF1a
ORF1b
S
E
M
N
A_n
B

	Premetamorphosis	*Metamorphic climax*	*Completion of metamorphosis*
Read count	209301	32737	2298
FPKM	453.1	63.7	4.4

Figure 9

Click here to download high resolution image



B

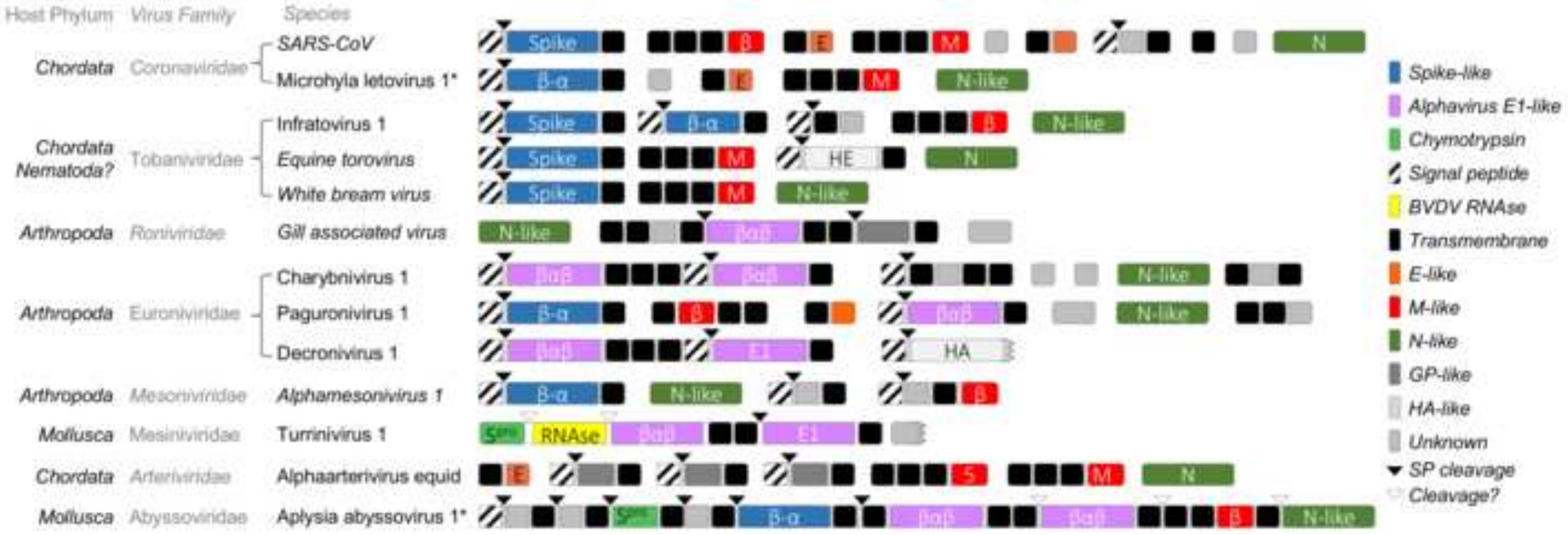
Preceding Gene			Genome Position	Putative TRS?			Following Gene	Read Depth Increases
pp1b			13911	TTCAAC**ATGA**			S-like	no
S-like	TAG	(40)	18538	TTCAATA	(55)	**ATG**	ORF 3a	no
ORF 3b	TAA	(138)	19212	TTCAAAA	(48)	**ATG**	E-like	yes
E-like			19439	TTCAATA**ATG**			M-like	
M-like	TAG	(41)	20272	TTCAAAA	(59)	**ATG**	N-like	yes

C

Microhyla letovirus 1	*Protein*	*Region*	*Input*	*Tool*	*Probability*	*E value*	*Model*	*Note*
nsp3 Y domain	pp1a	104-287	MLeV only	HHPred align	96.6%	1.3 × 10^{-9}	NP_828862	nsp3 Y1 and CoV-Y domains, SARS-CoV
nsp4 N terminal region	pp1a	547-736	MLeV only	Protein BLAST		4.0 × 10^{-4}	APX44046	IBV nsp4
nsp4 C-terminal domain	pp1a	801-907	MLeV only	HHPred	100%	3.8 × 10^{-37}	3VC8	FCoV nsp4 C-terminal domain
nsp5 M^{pro}	pp1a	924-1211	MLeV only	HHPred	100%	3.0 × 10^{-119}	3D23	HCoV-HKU1 M^{pro}
nsp6	pp1a		MLeV only	Protein BLAST		4.0 × 10^{-8}	APD51497	229E-related bat CoV nsp6
nsp7	pp1a	1508-1597	MLeV only	HHPred	99.9%	3.1 × 10^{-29}	1YSY	SARS-CoV nsp7
nsp8	pp1a	1598-1783	MLeV only	HHPred	100%	3.7 × 10^{-73}	3UB0	FCoV nsp8
nsp9	pp1a	1784-1894	MLeV only	HHPred	99.9%	3.1 × 10^{-30}	2J98	HCoV-229E nsp9
nsp10	pp1a	1901-2021	MLeV only	HHPred	100%	1.0 × 10^{-4}	5C8S	SARS-CoV nsp10
nsp12 NiRAN	pp1ab	2105-2252	MLeV only	Protein BLAST		3.0 × 10^{-13}	YP_005352845	NiRAN, Sparrow CoV HKU17
nsp12 middle domain	pp1ab	2285-2445	MLeV only	HHPred	97.4%	8.9 × 10^{-6}	4S1T	Fragmentary second NEndoU, HCoV-229E
nsp12 RdRp	pp1ab	2461-2909	MLeV only	HHPred	99.5%	3.5 × 10^{-17}	2CJQ	BVDV RdRp
nsp13 ZF/Helicase	pp1ab	2954-3549	MLeV only	HHPred	99.9%	9.4 × 10^{-32}	4N0N	EAV nsp10 ZF/Helicase
nsp14 ExoN/N7-Methyltransferase	pp1ab	3568-4009	MLeV only	HHPred	100%	4.0 × 10^{-109}	5C8S	SARS-CoV nsp14
nsp15 NEndoU	pp1ab	4010-4327	MLeV only	HHPred	100%	2.4 × 10^{-84}	2GTI	MHV nsp15
nsp16 2O-Methyltransferase	pp1ab	4328-4631	MLeV only	HHPred	100%	4.8 × 10^{-45}	2XYQ	SARS-CoV nsp16
Putative Spike homolog	S	1-1526	MLeV only	TMHMM 2.0, SignalP 4.1, PSIPRED 3.3		n/a	n/a	Signal peptidase cleavage, N-terminal beta region, C-terminal alpha region, 1 TM predicted – S-like by inference
Putative E protein homolog	E	1-77	MLeV only	TMHMM 2.0, AmphipaSeek		n/a	n/a	1 TM, amphipathic region – E-like by inference
Membrane protein	M	1-130	MLeV only	Protein BLAST		0.36	ADX59454	MERS-CoV N protein N-terminal domain
Nucleoprotein N-terminal domain	N	8-133	MLeV only	HHPred	88.5%	0.4	4UD1	Cardioderma bat CoV Kenya/KY43/2006 M

Figure 10

Click here to download high resolution image


Host Phylum
Virus Family
Species
Chordata
Coronaviridae
SARS-CoV
Microhyla letovirus 1*
Chordata
Nematoda?
Tobaniviridae
Infratovirus 1
Equine torovirus
White bream virus
Arthropoda
Roniviridae
Gill associated virus
Arthropoda
Euroniviridae
Charybnivirus 1
Paguronivirus 1
Decronivirus 1
Arthropoda
Mesoniviridae
Alphamesonivirus 1
Mollusca
Mesiniviridae
Turrinivirus 1
Chordata
Arteriviridae
Alphaarterivirus equid
Mollusca
Abyssoviridae
Aplysia abyssovirus 1*
Spike
β-α
βαβ
E1
β
E
M
S
N
N-like
HE
HA
RNAse
S^pro
Spike-like
Alphavirus E1-like
Chymotrypsin
Signal peptide
BVDV RNAse
Transmembrane
E-like
M-like
N-like
GP-like
HA-like
Unknown
SP cleavage
Cleavage?