Identification of a *Nidovirales* Orf1a N7-guanine cap Methyltransferase signature-sequence as a genetic marker of large genome *Tobaniviridae*

François Ferron[1,3], Humberto Julio Debat[2], Etienne Decroly[1], and Bruno Canard[1,3]
[1] Architecture et Fonction des Macromolécules Biologiques, CNRS and Aix-Marseille Université, UMR 7257, Polytech Case 925, 13009 Marseille, France.
[2] Instituto de Patología Vegetal, Centro de Investigaciones Agropecuarias, Instituto Nacional de Tecnología Agropecuaria, Córdoba, Argentina.
[3] European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany.

Running title: RNA cap N7-guanine Methyltransferase in *Tobaniviridae* Orf1a

Keywords: RNA viruses, RNA capping, methyltransferase, active site, Rossmann fold, protein structure, torovirus.

Abstract

Members of the *Nidovirales* order have (+)RNA genomes amongst the largest in size in the RNA virus world. Expression of their genes is promoted through reading of genomic RNA and mRNA transcripts by the ribosome of the infected cell. The 5'-end of these RNAs is supposedly protected by an RNA-cap structure (m7GpppNm) whose most synthesis steps remain elusive. In Eukaryotes, the RNA-cap structure is methylated by RNA methyltransferases (MTases) at the RNA-cap N7-guanine position as well as the 2'-O methyl position of the first transcribed nucleotide. In *Coronaviridae*, two separate enzymes (nsp14 and nsp16) perform the N7-guanine and the 2'-OH methylation, respectively. One salient feature of the *Nidovirales* N7-guanine MTase nsp14 is that it is the only example of non-Rossman fold viral MTase known so far. Conversely, all other *Nidovirales* nsp16-like MTases have a canonical Rossman fold. Many *Nidovirales* members lack either any RNA MTase signature sequence (e.g., *Arteriviridae*), or lack a N7-guanine MTase signature sequence (e.g., *Tobaniviridae*, *Euroniviridae*, *Roniviridae*, *Medioniviridae*). Both nsp14- and nsp16-like enzyme genes are usually located in Orf1b encoding for the replication machinery. Here, we report the discovery of a putative Rossman fold RNA MTase in the Orf1a of ten *Tobaniviridae* members. Multiple sequence alignments and structural analyses identify this novel gene as a typical RNA-cap N7-guanine MTase with substrate specificity and active-site organization similar to the canonical eukaryotic RNA-cap N7-guanine MTase.

Introduction

Up to recently, four viral families, *Arteriviridae*, *Coronaviridae*, *Roniviridae*, and *Mesoniviridae* constituted the *Nidovirales* order. The *Coronaviridae* family was further divided in two subfamilies, *Coronavirinae* and *Torovirinae.* The *Torovirinae* family contained two genus *Torovirus* and *Bafinivirus*, plus few unassigned species. Recently, the ICTV proposed a novel classification into the *Nidovirales* order, with the creation of nine families [1]. In this novel ICTV proposal, ratified on February 2019, *Coronaviridae* do not encompass both *Coronavirinae* and *Toronavirinae* subfamilies. Rather, the new suborder *Tornidovirinae* (including toroviruses and bafiniviruses) contains a novel family, *Tobaniviridae*, at the same level as the *Coronaviridae,* which is now within suborder *Cornidovirineae*. This new taxonomy is used here, as depicted in Fig. 1.

Nidovirales have attracted much attention in the past decades for being causative agents of serious diseases in humans and animals. In humans, Severe Acute Respiratory Syndrome (SARS) and the Middle-East Respiratory Syndrome (MERS) coronaviruses (CoVs) have been associated with high fatality/case ratios of ~10% and ~35 %, respectively (reviewed in [2]). In animals, pathogens of significant economical and societal impact include Equine Arteritis virus (EAV, Alphaarterivirus equid) and Porcine Reproductive Respiratory Syndrome virus (PRSSV 1-2, Betaarterivirus suid 1-2).

Two salient features of these viruses are, amongst others, their complex and still obscure replication/transcription mechanism as well as their broad genome-size range. Genomes of *Arteriviridae* (12.7 -15.7 kb in length), *Mesoniviridae* (~20 kb), *Medioniviridae* (20.2 – 25 kb), *Euroniviridae* (~24.5 kb), *Roniviridae* (~26 kb) *Tobaniviridae* and *Coronaviridae* (~27-32 kb), *Abyssoviridae* (35.9 kb), and *Mononiviridae* (41.1 kb) members, share a similar genomic organisation generating viral mRNAs from a fixed genomic location, the 'nest', which has inspired the family name (*Nido* : latin for 'nest') [3,4]. *Nidovirales* genomic and mRNAs are presumably capped, although direct demonstration of the presence of an RNA cap structure is still missing for many of these *Nidovirales* members.

In Eukaryotes, the RNA cap structure is thought to be synthesized post-transcriptionally [5]. Most of (+)RNA viruses supposedly follow a conventional RNA capping pathway [6] in which nascent viral 5'-triphosphate genomic (or mRNAs) is thought to be processed through three enzymatic reactions to yield an RNA cap whose structure is indistinguishable from that of cellular mRNAs. The capping pathway implies that the 5'-triphosphate RNA is hydrolyzed into 5'-diphosphate by an RNA triphosphatase, a GMP residue (the « cap », coming from GTP) is next covalently transferred to the 5'-diphosphate RNA in the 5' to 5' orientation by a guanlylyltransferase (GTase), releasing inorganic pyrophosphate. Both the cap and the first transcribed nucleotide are then methylated at the N7-guanine (mGpppN-RNA, so-called Cap0 structure) and the 2'-oxygen position (mGpppNm-RNA, Cap1 structure) by one or two RNA methyltransferases (MTases), respectively. The MTase uses S-Adenosyl-Methionine (SAM) as the methyl donor and releases S-Adenosyl-homocysteine (SAH).

There is no high-resolution structural analysis available of *Nidovirales* RNA caps. The evidence for their presence comes from immunological detection in Torovirus genomic and mRNAs [7], and co-migration analysis in low-resolution chromatographic techniques for MHV [8] and Simian Hemorragic fever virus [9]. In other instances, *Nidovirales* cap structures supposedly conform to the canonical eukaryotic RNA capping pathway because enzymes usually deemed necessary to perform RNA capping are present in cognate genomes (reviewed in [6]). Bioinformatic analyses have been used to detect signature-sequences typically associated with RNA capping/modifying enzymes *i.e* guanlylyltransferase (GTase), polymerase, helicase, nuclease, and methyltransferase (MTase). These signature-sequences have served in many instances to isolate corresponding enzymes endowed with the predicted activity. In *Nidovirales*, this approach has been pioneered for Coronavirus enzymes, yet

regarding their RNA capping pathway, it only succeeded in predicting 2'OMTase, while a second non-canonical RNA-cap N7-guanine MTase was later structurally and functionally characterized [4,10–14]. Until then, viral MTases matched a specific structural architecture, the Rossmann fold (RF), which possess a unique interaction geometry that represents a fingerprint of common ancestry. The discovery of a non-Rossmann fold (NRF) MTase unveiled the limitation of annotating "dark proteome" data [15], lacking structural counterparts. In *Nidovirales*, enzyme activities have been experimentally evidenced only for coronaviruses and roniviruses [16]. The putative first step of the *Nidovirales* RNA capping pathway, involving a 5'-triphosphatase, is uncertain: such activity exists in the Coronavirus helicase (nsp13, [17]), but its involvement in RNA capping has not been directly demonstrated. The guanylyltransferase also remains unknown: A nidovirus RdRp-associated nucleotidyltransferase (NiRAN) activity was found to be associated to the N-terminus of the arterivirus nsp9 RdRp [18], but its role in arterivirus RNA synthesis, if any, remains to be determined.

Despite the conservation of MTases amongst large *Nidovirales* genomes, their distribution, type, and the function of MTases in the *Nidovirales* order is not homogenous (summarized in Table 1, see text). Arterivirus or arterivirus-like genomes are not known to carry any obvious MTase signature sequence; Mesonivirus and Ronivirus genomes code for one 2'-O RF-MTase (nsp16-like) signature-sequence; The *Tobaniviridae* (*e.g.*, toroviruses and bafiniviruses, see Fig. 1) carry a single 2'-O RF-MTase (nsp16-like) signature sequence, whereas *Coronaviridae* carry both a N7-guanine non-RF-MTase (NRF-MTase, nsp14-like) and a 2'-O RF-MTase (nsp16-like) signature-sequence. A single deviation for this MTase-based grouping is that of Fathead minnow nidovirus 1 (*Tobaniviridae*) which carries both MTases (see below) as in the case of *Coronaviridae*.

The Coronavirus nsp14 is a peculiar, bi-functional enzyme combining in N-terminus a 3'-to-5' exonuclease domain (ExoN) fused to the N7-guanine NRF-MTase domain in C-terminus [4,11,12,19,20]. First, it is the only known example of a viral SAM-dependent RNA MTase (and more generally, a unique example of an RNA-Cap 0 N7-guanine MTase) that does not have a Rossman fold [11]. Since the Rossman fold is extremely widely distributed in all kingdoms of life, why this NRF-MTase fold has appeared or been conserved through evolution is puzzling. Second, the ExoN activity of nsp14 is a marker of large RNA genomes [21]. The ExoN domain confers capability of RNA mismatch repair to the replicative complex, ensuring genetic stability [11,20,22].

There is still a significant gap of general knowledge on *Nidovirales* RNA capping. For example, *i*) the guanylyltransferase step is still elusive in the whole order; *ii*) if any cap exists in *Arteriviridae*, how is RNA cap methylation achieved ? and *iii*) how is the cap N7-guanine methylated in *Tobaniviridae* and *Roniviridae*, since there is no N7-guanine MTase clearly identified in these families?

In this paper, we report the discovery of an RNA-cap N7-guanine RF-MTase sequence in the Orf1a of ten members of the *Tobaniviridae* family. Remarkably, within this family, this MTase is absent from mammalian toroviruses, but may well represent a genetic marker of non-mammalian *Tobaniviridae*.

Material and Methods

Virus genome sequences were retrieved from the NCBI database (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=439488) or Genbank (ncbi.nlm.nih.gov/genbank). The data set (given in Table 1) was that of Lauber *et al*. [21], to which novel or recently described genome sequences (retrieved in NCBI of GenBank) were added purposely. Additionally, the putative and unpublished Palaemon nidovirus and Western flower thrips nidovirus, and the proposed Botrylloides leachii nidovirus [23], were identified in the Sequence Read Archive (SRA) runs SRR5658389, SRR492945 and SRR2729873, respectively, and the virus sequences were assembled, curated and annotated as described elsewhere [24,25]. The resulting sequences are available upon request to H.D.

The conserved Nidovirales C-terminus RdRp core was used to build phylogenetic trees based on MAFFT v7.427 multiple sequence alignment with BLOSUM62 scoring matrix and G-INS-i iterative refinement method. The alignments were used as input for maximum likelihood trees generated with the FasTtree v2.1.5 software (best-fit model = JTT-Jones-Taylor-Thorton with single rate of evolution for each site = CAT). Local support values were computed using the Shimodaira-Hasegawa test (SH) with 1,000 replicates.

Structural alignment of reference or retrieved Mtases were done using EXPRESSO [26] and proofed with Chimera [27]. N7-Mtases reference motifs were inferred and defined based on visual inspection and analysis of the structures. The subsequent motifs were used as orthogonal validation and not used for the search.

Sequences were analyzed using HHblits and HHpred tools of the BioInformatics tookit ([28]). When available, authentic cleavage sites were used to predict protein gene products of the Orf1ab, Orf1a, and Orf1b polyproteins. The boundaries were otherwise approximately (+/- 10 aa) determined using structural homologies detected using HHPred, except for the N-term boundary of the Orf1b gene product: In *Nidovirales*, the absence of any structural data nor homology (outside the order) on the N-terminus of the RdRp gene (nsp9 *Arteriviridae*, nsp12 in *Coronaviridae*), which was used for phylogenic analysis, precludes precise sequence homology search in this limited area comprised between the nsp10 and nsp12 proteins (Coronavirus gene-product naming).

Multiple sequence alignments (MSA) were generated using Muscle in SeaView [29]. For each sequence of unknown structure, secondary structures were predicted using Predict Protein [30]. The predicted secondary structures were used to validate the alignment with structural references. The MSA was rendered using ESPript 3.0 [31], together with appropriate structural models as indicated, to assign secondary structures.
When possible, structural 3D models were generated using Phyre 2.0 [32]. Conserved patches of amino-acids were generated using WebLogo [33] and mapped in the structural models rendered in Chimera.

Results

All non-arterivirus Nidovirales members carry at least one 2'O-MTase

Available *Nidovirales* genome sequences were used to analyze the presence and primary structure of RNA MTases. Any RF-MTase signature sequence is detectable mainly due to the presence of a typical G-X-G-(X)n-G element part of a SAM-binding motif [34,35], whereas the detection of viral NRF-MTases is based on its homology to the MTase domain of Coronavirus nsp14. We made use of the conserved RdRp fold to build a phylogenetic tree along the whole order. From this tree, we first determined that most of *Nidovirales* code for at least one RF-MTase protein with the canonical K-D-K E catalytic tetrad of 2'O MTase (Fig.1 green & Table 1). Conversely, as previously reported, this RF-2'O MTase is lacking in EAV, LDV, and PRRSV in the *Arteriviridae* family. Interestingly, we identified two RF-MTase signature sequences in the genomes of two recently identified arteri-like viruses: the Hainan Hebius Popei Arterivirus (HHPAV, 12,496 nt) and the Nanhai gost shark arterivirus (NGSAV, 13,162 nt). Their genomes carry a *bona fide* RF-2'-O-MTase signature sequence at the end of their Orf1b, in the gene order RdRp-Hel-(ExoN*)-EndoU-2'OMTase. The NGSAV does not carry a detectable ExoN signature sequence, whereas the HHPAV genome carries a nsp14-like ExoN domain coding region but not a detectable C-terminus NRF-MTase domain, hence the ExoN* labeling (Ferron, Decroly & Canard, unpublished). All these RF-MTase with the K-D-K-E signature are localized in a conserved genomic position at the 3' end of the Orf1ab, much like their CoV and nsp16 homologues. We performed a MSA of nsp16 from *Roniviridae* and *Tobaniviridae* (Fig. 1, in green), followed by modeling a typical representative of these nsp16 (not shown). All these nsp16-like enzymes appear to be predicted as canonical RNA 2'-O MTases with a typical K-D-K-E tetrad. As noted by others in the Ronivirus nsp16 model [16], small structural differences are observed across the *Tobaniviridae* family, such as the absence of β3 strand and a shorter loop upstream helix αD. Thus, we conclude that non-arterivirus *Nidovirales* code for a RF-MTase with the canonical K-D-K E catalytic tetrad of 2'-O MTase, and that these enzymes are distributed in similar position along genomes.

The distribution of NRF-MTase is uneven along Nidovirales genomes

As indicated above, many *Nidovirales* members, such as SARS-CoV, stand out because they possess a NRF-MTase, participating to cap N7-guanine methylation, embedded at the C-terminus of their nsp14 gene product. It have been previously reported that the N7-guanine-MTase domain is not uniformly present in nsp14-containing nidoviruses [21]. The phylogenetic tree of Figure 1 reports our findings, and confirms that coronaviruses and most mesoniviruses only possess signature-sequences lining the SAM binding site seen in the nsp14 MTase domain structure. Conversely, the other members of the *Nidovirales* order (*Arteriviridae, Medionivirdae, Roniviridae, Euroniviridae, Abyssoviridae, Mononiviridae* and *Tobaniviridae*) do not seem to embed any nsp14-like NRF-MTase. Two notable exceptions are apparent: First, the unique members of *Abyssoviridae, Mononiviridae* do possess a MTase signature-sequence at the C-terminus of their nsp14-like gene, (i.e., fused to the ExoN domain). Curiously, though, this MTase is readily detectable using HH-Pred as a RF-MTase which does qualify neither as a K-D-K-E 2'-O MTase nor as a N7-guanine MTase (see below). Second, the Fathead Minnow nidovirus 1, belonging to the *Tobaniviridae,* is also an exception as we detected a nsp14-like NRF MTase at the expected genomic Orf1b position. Against, based on the homology with the nsp14-like MTase domain characterized by the

absence of K-D-K-E catalytic residue and NRF folding, we hypothesize that these enzymes might play a role in guanine N7 methylation of the RNA cap structure.
Apart from coronaviruses and most mesoniviruses, how nidoviruses methylate their RNA-cap N7-guanine position is thus still unclear.

What defines N7-guanine MTase features in Nidovirales?

Unlike 2'-O MTases and their well-defined K-D-K-E tetrad, and apart from NRF nsp14-like MTases, signature sequences of RNA cap N7-guanine cap N7-guanosine MTases are much less obvious. The wealth of structural data and mechanistic insight on the N7-guanine MTases enzymes has been acquired from the microsporidian parasite *Encephalitozoon cuniculi* [36], the poxvirus MTase D1:D12 heterodimer [37] and *Reoviridae* [38,39] in the case of RNA viruses. There are only three crystal structure of RF-MTase known to methylate N7-guanine RNA caps of RNA viruses : those of *Reoviridae* (Rotavirus VP4 and Reovirus Lambda2), and Flavivirus, whose NS5 N-terminus domain is a bi-functional N7-guanine and 2'O-MTase [40]. Taking together these data, and combining structural information with the human RNA N7-guanine methyltransferase crystal structure (PDB ID: 3BGV), one cannot define a specific N7-guanine MTase signature (Fig. S1).
From this analysis and others [35] it is clear that the structure conservation prevailed over sequence conservation. No amino acid motifs typical of an RNA N7-guanine MTase can be clearly identified (Fig. S1). A narrower structural comparison including only *Encephalitozoon cuniculi,* poxvirus and the human RNA N7-guanine MTase allows detection of 6 amino acids or motifs K / G/D/ HY / E / Y (Fig. S2). Typical SAM binding motifs are defined by a combination of 3 motifs placed at specific structural positions: K within an α helix, followed by a GxGxG motif within a loop and conserved D. The 3 latter motifs are located in the MTase structural pocket where the N7 methylation supposedly occurs.
In order to use an independent method, we performed a structural search using HHpred on PDB and SCOPe databases [41], using each sequence as a query. This allowed to refine the boundaries of the MTase domain. The structure with the closest hit is a SAM-dependent MTase from *Pectobacterium atrosepticum* and then *Encephalitozoon cuniculi*. A quick superimposition of the retrieved structures with N7-guanine MTase shows that these structures can be structurally aligned but with very limited sequence conservation. Surprisingly, apart from a conserved leucine residue 2 amino acid upstream the poorly conserved SAM binding motif (only the third G is strictly conserved) and a glycine-rich motifs in the fifth beta sheet, nothing else is conserved nor defines an active site (Fig. S3).
Taking these criteria into account, there was no obvious indication of RF N7-guanine MTase features in *Nidovirales* other than the two mentioned above in *Abyssoviridae* and *Mononiviridae*. Currently, based on the detection of RF-MTase fold, one can infer a 2' O-MTase sequence signature when the K-D-K-E tetrad is present, but not a RNA-cap N7-guanine MTase.

Selected Tobaniviridae members possess a RF-MTase signature-sequence in Orf1a lacking the canonical 2'O catalytic K-D-K-E tetrad

Then, how do *Roniviridae* and *Tobaniviridae* methylate the RNA-cap N7-guanine position? One possibility would be that these nsp16-like were bi-functional, carrying both 2'-O and N7-guanine methylation activity such as evidenced in Flavivirus NS5 protein, with no obvious signature sequence of this bi-functionality. Another possibility would be that another gene

would code for an enzyme performing this methylation, and had escaped detection by standard bioinformatic methods.

We thus performed a larger search of MTase signature-sequences along the whole Orf1ab in all *Nidovirales*: surprisingly, a RF-MTase signature-sequence was detected in Orf1a of 10 members of the *Tobaniviridae* family (Fig. 2). Strictly conserved amino-acids in these new viral MTases define three motifs: 3 glycines of the SAM binding site (G54, 56, and 58 in WBV) 2 residues downstream of a 3 amino acid hydrophobic patches in a β-strand structure, a histidine (H117 immediately followed by either F or Y, and Glu175 (Fig. S4). Interestingly the catalytic K-D-K-E tetrad associated to 2'O-MTase is lacking, suggesting that they could correspond to the missing RNA-cap N7-guanine MTase.

Remarkably, the novel Orf1a MTase gene was associated with reptiles and fish. Nonetheless, in two cases (Xinzhou Nematode virus 6 and Xinzhou toro-like virus 1, from the *Sectovirus 1* and *Infratovirus 1* species respectively), the viruses were isolated from snake-associated nematodes, although their proper host and life style has not been clearly assessed. The mammalian toroviruses (EToV, BToV, Bovine TCH5 nidovirus, GToV and PToV) do not carry the MTase signature sequence, and thus the latter is associated so far to non-mammalian *Tobaniviridae*.

The RF-MTase signature-sequence in Orf1a is a putative RNA cap N7-guanine MTase

We performed a structure-based alignment of this Orf1a MTase with the integration of several structurally-defined eukaryotic RNA-cap N7-guanine MTases (Fig. 3A), of which Ecm1 can be considered as prototypic [36]. The conserved amino-acids in these new MTase set remain essentially the same: two glycines of the SAM binding site (G54 and 56 in WBV, G74-76 in Ecm1), a histidine (H117 in WBV, H144 in Ecm1) immediately followed by either F or Y. Glu175 is also very conserved, and in the Ecm1 structure, its homosteric counterpart Glu225 is positioned close to His144 and interacts also with the guanine base, strongly suggesting that this MTase is an N7-guanine MTase. The alignment allows to define 5 conserved motifs which are represented in Fig. 3B, and structurally aligned onto the vaccinia N7-guanine MTase structure D12 (PDB: 4KCB).

Discussion

RNA virus cap-MTases perform two types of reactions: the methylation of the N7-guanine of the RNA cap, and the methylation of the adenosine 2'-O ribose of the first transcribed nucleotide [6]. Regarding the guanine methylation, this reaction can also be achieved on a GTP molecule before capping, in some virus families, and the N7-methyl GTP is then used to cap the RNA [42]. This is the case with the alphavirus nsp1 enzyme, whose structure is currently unknown. Conversely, several crystal structures of viral enzymes involved in N7-guanine methylation of RNA caps have been reported, such as that of flaviviruses (NS5, [43]), rotaviruses and reoviruses (VP4 and lambda2, respectively [38,39]), and coronaviruses (nsp14, [11,12]).

The viral 2'-O MTases have been better defined than N7-guanine MTases at the structural and functional level. They all use a K-D-K-E catalytic tetrad, and crystal structures have been determined for several virus groups, such as Reovirus, Flavivirus, Coronavirus, and *Mononegavirales* (reviewed in [6], [44]). All these enzymes (N7- and 2'O–MTase), with SARS-CoV nsp14 as a notable exception (see below), belong to the Rossmann-fold (RF) type MTase (reviewed in [34,35]). They bind the methyl donor (SAM), and are characterized by several structural features: their Rossmann fold is a seven-strands β-sheet surrounded with 6 α-helices; the seventh β-strand is inserted in an anti-parallel orientation between the 5th and 6th strand. The SAM cofactor binds to the first structural motif βαβ, which bears the amino acid sequence motif G-x-G-$(x)_n$-G. The RF is an evolutionary ancient fold, which has been widely evolved to perform a variety of chemical reactions. Its structural plasticity is well illustrated by the Flavivirus NS5 MTase, which is able to perform both N7-guanine and 2'-O ribose methylation with the same ~33 kDa domain fused at the N-terminus of the viral RdRp domain [40].

The presence of a MTase in *Nidovirales* Orf1a is unexpected, and had not been reported before, although signature sequences of an uncharacterized MTase had been detected in Ball Python nidovirus, Sectovirus 1 and Infratovirus 1 [23]. Both protein and enzyme functions in *Nidovirales* have been classified into a functional triangle [21], each side of the triangle representing a carrier of the following functions: Orf1a: host defence; Orf1b: genome maintenance; Orfs at the 3'-end: structural products. In this triangle, the MTase-type of activity involved in genome maintenance would usually map to Orf1b [21]. Perhaps indeed this MTase is involved in tasks other than RNA capping, which remain to be determined. Accordingly, it was recently reported that viral or cellular MTases can be recruited by viruses in order to induce internal methylation of their genome [45] and escape to the antiviral response mediated by MDA5 [46]. Out of the gain of this (still putative) enzyme activity in this virus group, the copy number of Orf1a products is expected to be 3-6 time higher than those of Orf1b, raising the following questions: why only these *Tobaniviridae* members need such additional MTase ? Why do they need it in high amounts, provided that the Orf1a stoichiometry is thought to be higher than that of Orf1b products? Are activity levels provided by this enzyme low? What is its actual substrate specificity?

Also, it is legitimate to question the specificity of single copy MTases in *Nidovirales*, such as mammalian *Tobaniviridae* or *Roniviridae* (Fig. 2). Since they do not have any obvious RNA-cap N7-guanine MTase, how is RNA-cap methylation achieved, if any? Are their nsp16-like enzymes of broad or promiscuous specificity?

Our work also uncovers an unexpected 2'-O MTase distribution in viruses seemingly blurring the contour of *Arteriviridae,* a discovery that, in turn, may have a significant impact on our understanding of genome size evolution in *Nidovirales*. Two newly identified arteri-like virus genomes (HHPA and NGSA) provide food for thought regarding RNA genome

evolution: Phylogenetic branching of both HHPA and NGSA suggests that arteriviruses might not be primitive small version of larger Coronavirus genomes, but may rather originate from size-reduction of a large Nidovirus ancestor genome.

In any case, the type, diversity, and distribution of RNA MTases along the *Nidovirales* deserves a closer look : SAM-dependent MTases are ancient folds associated with RNA stability and evolution [47]. Their presence and properties in a phylogenetic tree may well give interesting clues regarding RNA genome evolution and its associated issue of host defense mechanisms.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program through the ANTIVIRALS project under the Marie Sklodowska-Curie grant agreement N° 642434.

References

1. Siddell SG, Walker PJ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, Gorbalenya AE, Harrach B, Harrison RL, Junglen S, et al.: Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018). *Arch Virol* 2019, 164:943–946.
2. Song Z, Xu Y, Bao L, Zhang L, Yu P, Qu Y, Zhu H, Zhao W, Han Y, Qin C: From SARS to MERS, Thrusting Coronaviruses into the Spotlight. *Viruses* 2019, 11.
3. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ: Nidovirales: evolving the largest RNA virus genome. *Virus Res* 2006, 117:17–37.
4. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LLM, Guan Y, Rozanov M, Spaan WJM, Gorbalenya AE: Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J Mol Biol* 2003, 331:991–1004.
5. Ramanathan A, Robb GB, Chan S-H: mRNA capping: biological functions and applications. *Nucleic Acids Res* 2016, 44:7511–7526.
6. Decroly E, Ferron F, Lescar J, Canard B: Conventional and unconventional mechanisms for capping viral mRNA. *Nat Rev Microbiol* 2011, 10:51–65.
7. van Vliet ALW, Smits SL, Rottier PJM, de Groot RJ: Discontinuous and non-discontinuous subgenomic RNA transcription in a nidovirus. *EMBO J* 2002, 21:6571–6580.
8. Lai MM, Patton CD, Stohlman SA: Further characterization of mRNA's of mouse hepatitis virus: presence of common 5'-end nucleotides. *J Virol* 1982, 41:557–565.
9. Sagripanti JL, Zandomeni RO, Weinmann R: The cap structure of simian hemorrhagic fever virion RNA. *Virology* 1986, 151:146–150.
10. Decroly E, Imbert I, Coutard B, Bouvet M, Selisko B, Alvarez K, Gorbalenya AE, Snijder EJ, Canard B: Coronavirus nonstructural protein 16 is a cap-0 binding enzyme possessing (nucleoside-2'O)-methyltransferase activity. *J Virol* 2008, 82:8071–8084.
11. Ferron F, Subissi L, Silveira De Morais AT, Le NTT, Sevajol M, Gluais L, Decroly E, Vonrhein C, Bricogne G, Canard B, et al.: Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. *Proc Natl Acad Sci U S A* 2018, 115:E162–E171.
12. Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, Lou Z, Yan L, Zhang R, Rao Z: Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proc Natl Acad Sci U S A* 2015, 112:9436–9441.
13. Chen Y, Su C, Ke M, Jin X, Xu L, Zhang Z, Wu A, Sun Y, Yang Z, Tien P, et al.: Biochemical and structural insights into the mechanisms of SARS coronavirus RNA ribose 2'-O-methylation by nsp16/nsp10 protein complex. *PLoS Pathog* 2011, 7:e1002294.
14. Decroly E, Debarnot C, Ferron F, Bouvet M, Coutard B, Imbert I, Gluais L, Papageorgiou N, Sharff A, Bricogne G, et al.: Crystal structure and functional analysis of the SARS-coronavirus RNA cap 2'-O-methyltransferase nsp10/nsp16 complex. *PLoS Pathog* 2011, 7:e1002059.
15. Perdigão N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, Signal B, Gloss BS, Hammang CJ, Rost B, et al.: Unexpected features of the dark proteome. *Proc Natl Acad Sci U S A* 2015, 112:15898–15903.
16. Zeng C, Wu A, Wang Y, Xu S, Tang Y, Jin X, Wang S, Qin L, Sun Y, Fan C, et al.: Identification and Characterization of a Ribose 2'-O-Methyltransferase Encoded by the Ronivirus Branch of Nidovirales. *J Virol* 2016, 90:6675–6685.
17. Ivanov KA, Ziebuhr J: Human coronavirus 229E nonstructural protein 13: characterization of duplex-unwinding, nucleoside triphosphatase, and RNA 5'-triphosphatase

activities. *J Virol* 2004, 78:7833–7838.
18. Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, Janssen GMC, Ruben M, Overkleeft HS, van Veelen PA, Samborskiy DV, Kravchenko AA, Leontovich AM, et al.: Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. *Nucleic Acids Res* 2015, 43:8416–8434.
19. Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, Ziebuhr J: Discovery of an RNA virus 3'->5' exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc Natl Acad Sci U S A* 2006, 103:5108–5113.
20. Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, Decroly E: RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. *Proc Natl Acad Sci U S A* 2012, 109:9372–9377.
21. Lauber C, Goeman JJ, Parquet M del C, Nga PT, Snijder EJ, Morita K, Gorbalenya AE: The footprint of genome architecture in the largest genome expansion in RNA viruses. *PLoS Pathog* 2013, 9:e1003500.
22. Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR: High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *J Virol* 2007, 81:12135–12144.
23. Bukhari K, Mulley G, Gulyaeva AA, Zhao L, Shu G, Jiang J, Neuman BW: Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family Abyssoviridae, and from a sister group to the Coronavirinae, the proposed genus Alphaletovirus. *Virology* 2018, 524:160–171.
24. Debat HJ: 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* 2018, doi:10.1101/307678.
25. Debat HJ: An RNA Virome Associated to the Golden Orb-Weaver Spider Nephila clavipes. *Front Microbiol* 2017, 8.
26. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C: Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res* 2006, 34:W604–608.
27. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 2004, 25:1605–1612.
28. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V: A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* 2018, 430:2237–2243.
29. Gouy M, Guindon S, Gascuel O: SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010, 27:221–224.
30. Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, Hönigschmid P, Schafferhans A, Roos M, Bernhofer M, et al.: PredictProtein--an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res* 2014, 42:W337–343.
31. Gouet P, Robert X, Courcelle E: ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res* 2003, 31:3320–3323.
32. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE: The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 2015, 10:845–858.
33. Crooks GE, Hon G, Chandonia J-M, Brenner SE: WebLogo: a sequence logo generator. *Genome Res* 2004, 14:1188–1190.

34. Martin JL, McMillan FM: SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold. *Curr Opin Struct Biol* 2002, 12:783–793.
35. Byszewska M, Śmietański M, Purta E, Bujnicki JM: RNA methyltransferases involved in 5' cap biosynthesis. *RNA Biol* 2014, 11:1597–1607.
36. Fabrega C, Hausmann S, Shen V, Shuman S, Lima CD: Structure and mechanism of mRNA cap (guanine-N7) methyltransferase. *Mol Cell* 2004, 13:77–89.
37. De la Peña M, Kyrieleis OJP, Cusack S: Structural insights into the mechanism and evolution of the vaccinia virus mRNA cap N7 methyl-transferase. *EMBO J* 2007, 26:4913–4925.
38. Sutton G, Grimes JM, Stuart DI, Roy P: Bluetongue virus VP4 is an RNA-capping assembly line. *Nat Struct Mol Biol* 2007, 14:449–451.
39. Tao Y, Farsetta DL, Nibert ML, Harrison SC: RNA synthesis in a cage--structural studies of reovirus polymerase lambda3. *Cell* 2002, 111:733–745.
40. Ray D, Shah A, Tilgner M, Guo Y, Zhao Y, Dong H, Deas TS, Zhou Y, Li H, Shi P-Y: West Nile virus 5'-cap structure is formed by sequential guanine N-7 and ribose 2'-O methylations by nonstructural protein 5. *J Virol* 2006, 80:8362–8370.
41. Fox NK, Brenner SE, Chandonia J-M: SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 2014, 42:D304–309.
42. Ahola T, Kääriäinen L: Reaction in alphavirus mRNA capping: formation of a covalent complex of nonstructural protein nsP1 with 7-methyl-GMP. *Proc Natl Acad Sci U S A* 1995, 92:507–511.
43. Egloff M-P, Benarroch D, Selisko B, Romette J-L, Canard B: An RNA cap (nucleoside-2'-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. *EMBO J* 2002, 21:2757–2768.
44. Paesen GC, Collet A, Sallamand C, Debart F, Vasseur J-J, Canard B, Decroly E, Grimes JM: X-ray structure and activities of an essential Mononegavirales L-protein domain. *Nat Commun* 2015, 6:8749.
45. Martin B, Coutard B, Guez T, Paesen GC, Canard B, Debart F, Vasseur J-J, Grimes JM, Decroly E: The methyltransferase domain of the Sudan ebolavirus L protein specifically targets internal adenosines of RNA substrates, in addition to the cap structure. *Nucleic Acids Res* 2018, 46:7902–7912.
46. Ringeard M, Marchand V, Decroly E, Motorin Y, Bennasser Y: FTSJ3 is an RNA 2'-O-Methyltransferase recruited by HIV to avoid innate immunity sensing. *Nature* 2019,
47. Rana AK, Ankri S: Reviving the RNA World: An Insight into the Appearance of RNA Methyltransferases. *Front Genet* 2016, 7:99.
48. Varshney D, Petit A-P, Bueren-Calabuig JA, Jansen C, Fletcher DA, Peggie M, Weidlich S, Scullion P, Pisliakov AV, Cowling VH: Molecular basis of RNA guanine-7 methyltransferase (RNMT) activation by RAM. *Nucleic Acids Res* 2016, 44:10423–10436.
49. Reinisch KM, Nibert ML, Harrison SC: Structure of the reovirus core at 3.6 A resolution. *Nature* 2000, 404:960–967.
50. Hausmann S, Zheng S, Fabrega C, Schneller SW, Lima CD, Shuman S: Encephalitozoon cuniculi mRNA cap (guanine N-7) methyltransferase: methyl acceptor specificity, inhibition BY S-adenosylmethionine analogs, and structure-guided mutational analysis. *J Biol Chem* 2005, 280:20404–20412.
51. Kyrieleis OJP, Chang J, de la Peña M, Shuman S, Cusack S: Crystal structure of vaccinia virus mRNA capping enzyme provides insights into the mechanism and evolution of the capping apparatus. *Struct Lond Engl 1993* 2014, 22:452–465.

Table 1. Distribution and putative substrate specificities of MTases in the four families of the *Nidovirales* Order. Only the *Tobaniviridae* is expanded and genera indicated in italic plus unassigned viruses (see Fig. 1). *: Arteriviruses do not usually carry any MTase, except the two newly identified Hainan Hebius Popei arterivirus and Nanhai gost shark arterivirus (see text).**: although in the nsp14 locus immediately downstream the ExoN gene, the MTase has a Rossmann fold. Virus species names are indicated and genome size corresponds to exemplar virus sequence used in this study.

Virus	Genome size (nt)	MTase Orf1a	N7-MTase nsp14-like	2'OMTase nsp16
Arteriviridae	~12,5 -15,7000	-	-	(+/-)*
Tobaniviridae:				
Torovirus				
Porcine torovirus	28,301	-	-	+
Bovine torovirus	28,479	-	-	+
Equine torovirus	27,992	-	-	+
Bostovirus				
Bovine nidovirus 1	20,261	-	-	+
Bafinivirus				
White bream virus	26,660	+	-	+
Fathead minnow nidovirus 1	27,318	+	+	+
Oncotshavirus				
Chinook *salmon nidovirus 1*	27,004	+	-	+
Infratovirus				
Infratovirus 1	30,353	+	-	+
Sectovirus				
Sectovirus 1	25,960	+	-	+
Tiruvirus				
Shingleback nidovirus 1	23,832	+	-	+
Pregotovirus				
Ball python nidovirus 1	33,452	+	-	+
New/unnassigned genus				
Bellinger river virus	30,742	+	-	+
Guangdong red-banded snake torovirus	30,859	+	-	+
Python nidovirus	32,399	+	-	+

Goat torovirus	28,487	-	-	+
Coronaviridae	~27-32,000	-	+	+
Medioni/ Mesoniviridae	~20-25,000	-	+	+
Roni/Euroniviridae	~26,000	-	-	+
Abyssoviridae	35,906	-	+**	+
Mononiviridae	41,178	-	+**	+

Figure legends

Figure 1. Presence, type, and loci of signature-sequences of *Nidovirales* RNA MTases : RF MT of unknown specificity (light red), non-RF « nsp14-like » MT (yellow), and RF « nsp16-like » MT (green). Both light red and yellow MTases map to the nsp14 C-ter locus, ie., immediately downstrean the ExoN domain, while « nsp16-like » MT map to the nsp16 locus, at the end of Orf1b. The tree was made based on MAFFT v7.427 multiple sequence alignment with BLOSUM62 scoring matrix and G-INS-i iterative refinement method. The alignments were used as input for maximum likelihood trees generated with the FasTtree v2.1.5 software (best-fit model = JTT-Jones-Taylor-Thorton with single rate of evolution for each site = CAT). Local support values were computed using the Shimodaira-Hasegawa test (SH) with 1,000 replicates. Numbers at the nodes represent FasTtree support values and scale var substitutions per site. The tree included two novel Ronivirus-like and Mesonivirus-like genome sequences: Western Flower Thrips Mesonivirus, and Palaemon Nidovirus, respectively (WFTV, PAN, unpublished).
When present in the NCBI viral genomes database in the 92 *Nidovirales* complete genomes repository (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=76804), no accession number is indicated. When an accession number is given in parenthesis, it is referring to the GenBank accession number (https://www.ncbi.nlm.nih.gov/genbank/).
From top to bottom of the figure:
Arteriviridae :
HHPAV : Hainan hebius popei arterivirus (MG600021) ; NGSAV: Nanhai ghost shark arterivirus (MG600024); SHFV : simian hemorrhagic fever virus ; EAV : Equine arteritis virus; LDV, Lactate elevating virus ; PRRSV-1 and 2, porcine reproductive and respiratory syndrome virus.
Tobaniviridae :
BToV : Breda virus ; GToV : Goat torovirus ; EToV : Berne virus (CAA36747); PToV : Porcine torovirus ; CSV : Chinook salmon bafinivirus ; FHMNV : Fathead Minnow nidovirus ; WBV : white bream virus ; BoNV: Bovine nidovirus TCH5 ; INToV: Xinzhou nido-like virus 6 ; SECToV; Xinzhou toro-like virus 1; GRBS-ToV : Guangdong red banded snake torovirus (MG600030); ShNV : shingleback nidovirus 1 (KX184715) ; BPNV : Ball python nidovirus 1 ; MVNV : Morelia viridis nidovirus ; BRV : Bellinger River virus (MF685025) ;
Coronaviridae:
IBV : Infectious Bronchitis Virus ; BWCoV : Beluga Whale coronavirus SW1 ; HCoV_HKU1 : Human coronavirus HK1 ; MHV : Mouse hepatitis virus ; BatCoV_HKU9 : Rousettus bat Coronavirus HKU9 ; SARS-CoV : Severe acute respiratory syndrome coronavirus ; MERS_CoV : middle-east respiratory syndrome coronavirus ; Ty-BatCoV_HKU4 : Tylonycteris bat coronavirus HKU4 ; Pi-BatCoV_HKU5 : pipistrellus bat coronavirus HKU5 ; FIPV : Feline infectious peritonitis virus ; Rh-BatCoV-HKU2 : Rhinolophus bat coronavirus HKU2 ; HCoV_229E : Human coronavirus 229E ; HCoV_OC43 : Human coronavirus OC43 (YP_003766) ; HCoV_NL63 : Human coronavirus NL63 ; PEDV : Porcine epidemic diarrhea virus ; Sc-BatCoV_512 : Scotophilus bat coronavirus 512 ; Mi-BatCoV-1A : bat coronavirus 1A ; Mi-BatCoV_HKU8 : Miniopterus bat coronavirus HKU8.
Medioniviridae:
BlNV: Botrylloides leachii nidovirus; TurrNV: Turrinivirus 1 ; WFTV : Western Flower Thrips virus.
Mesoniviridae :

MenoV : Meno virus ; NseV : Nse virus ; HanaV : Hana virus ; KSaV : Karang sari virus; DKNV : Dak Nong virus ; CASV : Casuarina virus ; NDiV : Nam Dinh virus ; CavV : Cavally virus ; DMV: Dianke mesonivirus.
Roniviridae:
PAN : Palaemon nidovirus ; GAV : Gill-associated virus; YHV : Yellow head virus (EU487200).
Euroniviridae:
PagRV: Paguronivirus 1; CharNV: Charybnivirus 1; DecNV: Decronivirus 1.
Abyssoviridae:
AAbV: Aplysia abyssovirus 1
Mononiviridae:
PSCNV: Planidovirus 1

Figure 2. Distribution of the Orf1a MTase along the A) *Tobaniviridae* phylogentic tree (see Fig. 1). B) approximate genome organisation and gene content along the family described in A. The genomes are not drawn to scale. Only the Bostovirus (*Remotovirinae*) genome is represented smaller than its fellow members to account for its ~20 kb genome size.

Figure 3. A) Structural alignment of *Tobaniviridae* methyltransferases together with various prokaryotic, eukaryotic and viral members; B) Vaccinia Virus N7-guanine methyltransferases structure (PDB : 4CKB) with highlighted motifs 1 to 5 and their corresponding amino-acid frequencies as determined by WebLogo analysis.

Figure 1

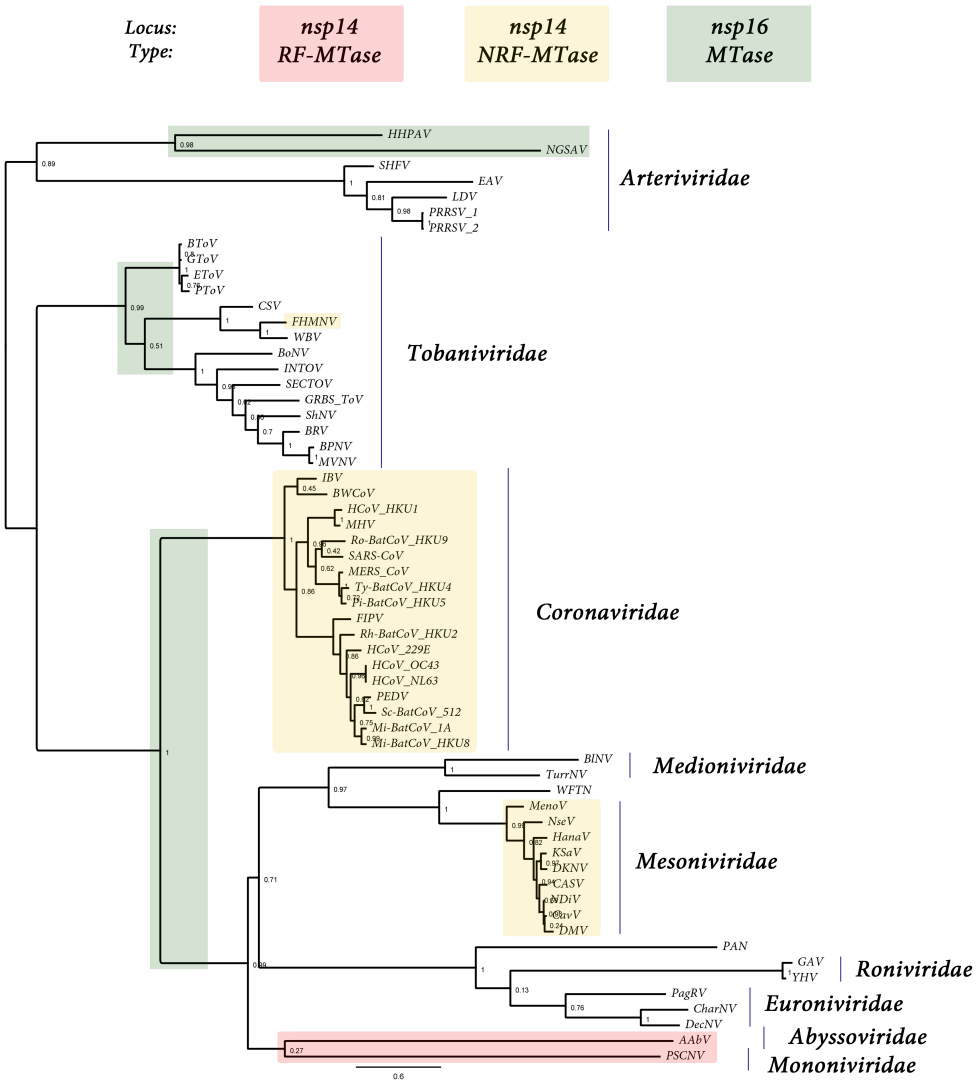
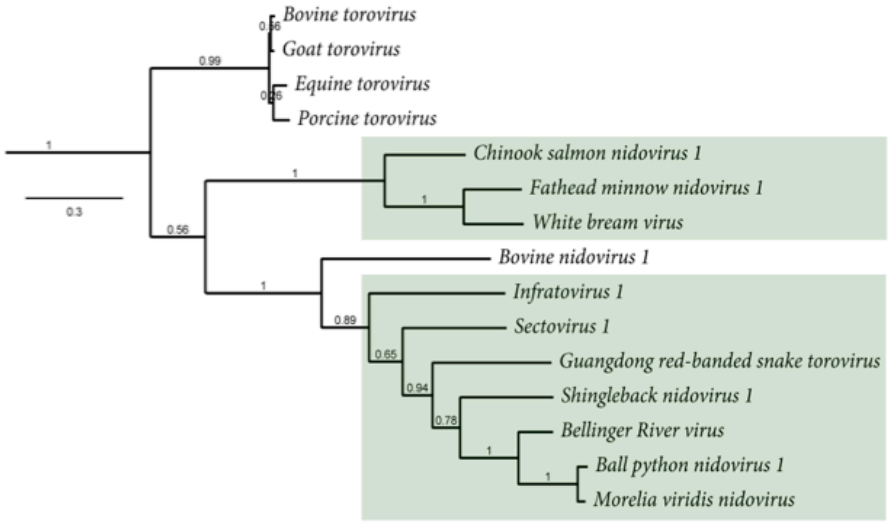

Locus:
Type:
nsp14
RF-MTase
nsp14
NRF-MTase
nsp16
MTase
HHPAV
NGSAV
SHFV
EAV
LDV
PRRSV_1
PRRSV_2
Arteriviridae
BToV
CToV
EToV
PToV
CSV
FHMNV
WBV
BoNV
INTOV
SECTOV
GRBS_ToV
ShNV
BRV
BPNV
MVNV
Tobaniviridae
IBV
BWCoV
HCoV_HKU1
MHV
Ro-BatCoV_HKU9
SARS-CoV
MERS_CoV
Ty-BatCoV_HKU4
Pi-BatCoV_HKU5
FIPV
Rh-BatCoV_HKU2
HCoV_229E
HCoV_OC43
HCoV_NL63
PEDV
Sc-BatCoV_512
Mi-BatCoV_1A
Mi-BatCoV_HKU8
Coronaviridae
BlNV
TurrNV
WFTN
Medioniviridae
MenoV
NseV
HanaV
KSaV
DKNV
CASV
DMV
Mesoniviridae
PAN
GAV
YHV
Roniviridae
PagRV
CharNV
DecNV
Euroniviridae
AAbV
PSCNV
Abyssoviridae
Mononiviridae

Figure 2

A


Bovine torovirus
Goat torovirus
Equine torovirus
Porcine torovirus
Chinook salmon nidovirus 1
Fathead minnow nidovirus 1
White bream virus
Bovine nidovirus 1
Infratovirus 1
Sectovirus 1
Guangdong red-banded snake torovirus
Shingleback nidovirus 1
Bellinger River virus
Ball python nidovirus 1
Morelia viridis nidovirus
0.99
1
0.3
0.56
1
1
0.89
0.65
0.94
0.78
1
1

B

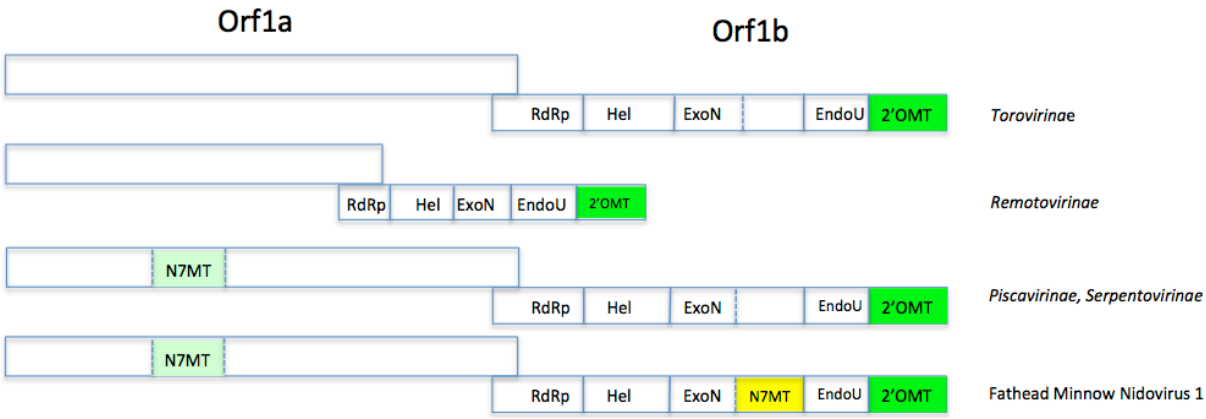
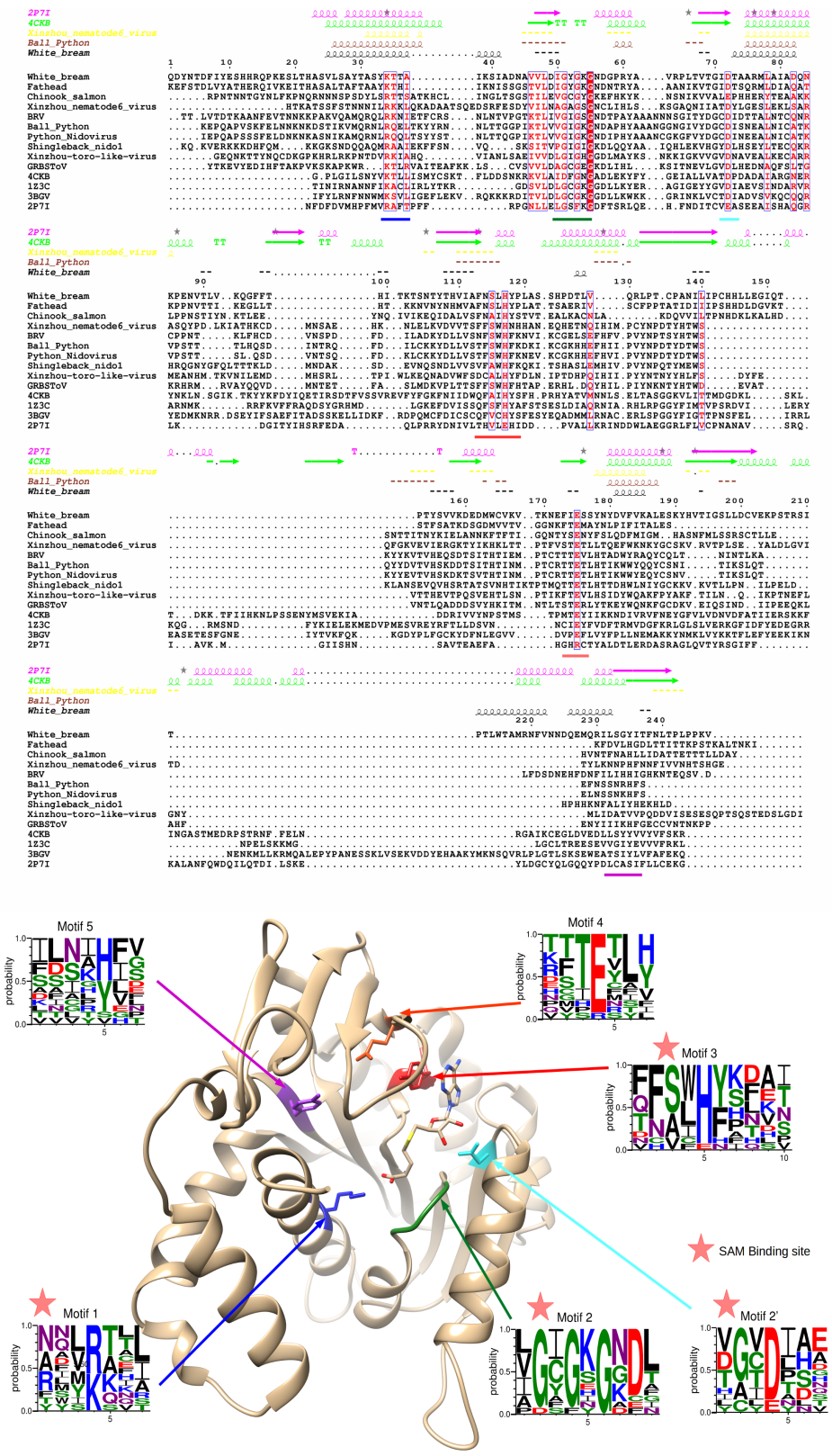

Orf1a
Orf1b
RdRp
Hel
ExoN
EndoU
2'OMT
Torovirinae
RdRp
Hel
ExoN
EndoU
2'OMT
Remotovirinae
N7MT
RdRp
Hel
ExoN
EndoU
2'OMT
Piscavirinae, Serpentovirinae
N7MT
RdRp
Hel
ExoN
N7MT
EndoU
2'OMT
Fathead Minnow Nidovirus 1

Figure 3


2P7I
4CKB
Xinzhou_nematode6_virus
Ball_Python
White_bream
Fathead
Chinook_salmon
BRV
Python_Nidovirus
Shingleback_nido1
Xinzhou-toro-like-virus
GRBSToV
1Z3C
3BGV
Motif 5
Motif 4
Motif 3
SAM Binding site
Motif 1
Motif 2
Motif 2'
probability