Comparative genomics of nidoviruses: towards understanding the biology and evolution of the largest RNA viruses

Gulyaeva, A.

Author: Gulyaeva, A.
Title: Comparative genomics of nidoviruses: towards understanding the biology and evolution of the largest RNA viruses
Issue Date: 2020-06-02

Comparative genomics of nidoviruses:

towards understanding the biology and evolution of the largest RNA viruses

Anastasia Gulyaeva

ISBN: 9789464022056

PhD thesis, Leiden University, 2020.

The research described in this thesis was carried out at the Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands, and was financially supported by intramural funds, an Agreement about Cooperation in Bioinformatics between Leiden University Medical Center and Lomonosov Moscow State University (MoBiLe Program), and by the European Union project EVAg 653316.

Cover image: “Forest and sky” by Anastasia Gulyaeva. Height of the vertices in the forest silhouette is equal (1 cm = 2.5 kb) to genome lengths of representatives of 109 nidovirus species delineated in 2019 ICTV proposals, ordered as tips on a phylogenetic tree (see also Figure 1 from chapter 6).

Printed by: Gildeprint, Enschede, The Netherlands.

Funding of the printing costs by the Department of Medical Microbiology, Leiden University Medical Center and by the Leiden University is gratefully acknowledged.

Comparative genomics of nidoviruses:
towards understanding the biology and evolution of the largest RNA viruses

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 2 juni 2020
klokke 13:45 uur

door

Anastasia Gulyaeva
geboren te Moskou
in 1991

Promotor
Prof. dr. A.E. Gorbalenya

Co-promotor
Dr. I.A. Sidorov

Leden promotiecommissie
Prof. dr. E.J. Snijder
Dr. ir. M. Kikkert
Prof. dr. J.J. Goeman
Dr. B.E. Dutilh (Utrecht University, The Netherlands)
Prof. dr. med. J. Ziebuhr (Justus Liebig University Giessen, Germany)

TABLE OF CONTENTS

CHAPTER 1

General Introduction

PREFACE

Viruses are ubiquitous intra-cellular parasites that account for a considerable part of the global biosphere, both in mass and diversity. Their most distinguished characteristics are a large population size, short replication cycle, interlinking high mutation rate and small genome size. Combined, these properties define a fast evolution of viruses, which facilitates virus adaptation to the host [1]. Viruses evolved an unparalleled molecular diversity of entities that use different types of DNA and RNA genomes, including dsDNA and others not found elsewhere [2].

Viruses were discovered by the end of the XIX century, and were originally described as the smallest pathogens [3]. During the subsequent century, their characterization was driven by research on infectious diseases of humans and other economically important hosts. At the time, virus research was primarily focusing on the characterization of the virus phenotype, while the characterization of genotypes was limited by the resolution of classical genetics.

About 40 years ago, the situation changed dramatically, owing to the technical advancements that introduced genome sequencing. From then on, in many cases, it became possible to trace the genetic basis of phenotypes to single nucleotides and to correlate these with replacements of amino acid residues in the virus proteome, whose entire primary structure was deduced. Genome sequencing also ushered in the age of comparative genomics that considerably accelerated and broadened our insights into the structure, function and evolution of viruses by *in silico* comparison of virus and host polynucleotides and proteins. It established a new reliable channel for the transfer of accumulated knowledge and a basis for generating new hypotheses in an evolutionary framework. Research on both previously characterized and recently discovered viruses benefited from this advancement. Subsequent years of parallel characterization of phenotype and genome proved the high quality of inferences by comparative genomics and revealed the synergy of these two approaches, notably through the use of reverse genetics. The advent of the next generation sequencing (NGS) in the XXI century made possible the high-throughput genome sequencing from miniscule quantities of biological samples. Sequencing of the entire diversity of DNA (metagenome) or RNA (metatranscriptome) molecules within a specimen became a reality. It led to a revolution in virus discovery that was no longer constrained by the characterization of pathogenicity or any other phenotypic property. Rather it became genome sequencing and comparative genomics providing sufficient evidence to recognize new viruses. Consequently, the rate of virus discovery exploded, and for the ever-increasing majority of viruses, computational analysis of their established genome sequence and deduced proteome defines what we know about these viruses [4].

Nidoviruses are one of the large monophyletic groups with a recognized societal significance, whose characterization has considerably been advanced by comparative genomics. They include deadly pathogens of animals and humans, such as porcine reproductive and respiratory syndrome virus (PRRSV), severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV) [5], and SARS-CoV-2, which caused the coronavirus disease 2019 (COVID-19) pandemic [6]. Given the recent introduction of SARS-CoVs and MERS-CoV into the human population, more nidoviruses infecting humans are feared to emerge in the future through cross-species transmission, with a group of arteriviruses causing hemorrhagic fever in nonhuman primates [7] and coronaviruses of bats [8] being of particular concern. Notably, nidoviruses include viruses with the largest known RNA genomes – entities that may offer a glimpse into the long-gone RNA world [9]. The devastating consequences of COVID-19 pandemic, high zoonotic potential of nidoviruses, negative economic impact of nidovirus infections in farm animals [10], as well as the extraordinary size of nidovirus genomes, make nidoviruses an important object of research. Our group has contributed to their characterization over many years, starting from the analysis of the first nidovirus genome sequenced, that of infectious bronchitis virus (IBV), and including SARS-CoV, MERS-CoV and many others [11-13]. This thesis describes part of the most recent studies on comparative genomics of nidoviruses, with the text below providing a background on nidoviruses and techniques of comparative genomics available by the end of 2014, when this project started.

NIDOVIRUS DIVERSITY AND TAXONOMY

Nidoviruses possess positive-sense, non-segmented linear RNA genomes in the size range of 12 – 34 kb that replicate in the cytoplasm and are packaged into enveloped virions that may vary in shape, depending on the virus lineage [13, 14]. These viruses form the order *Nidovirales* that was established by the International Committee on Taxonomy of Viruses (ICTV) in 1993 by merging two families of viruses infecting vertebrates, *Coronaviridae* (subfamilies *Coronavirinae* and *Torovirinae*) and *Arteriviridae* [15]. In subsequent years, two families of viruses infecting invertebrates, *Roniviridae* and *Mesoniviridae*, were added to the order [16, 17]. Hereafter, members of these (sub)families are referred to as coronaviruses, toroviruses, arteriviruses, roniviruses and mesoniviruses, respectively. Multiple genera were distinguished within the family *Coronaviridae*: genera *Alpha-*, *Beta-*, *Gamma-* and *Deltacoronavirus* belonging to the subfamily *Coronavirinae*, as well as genera *Torovirus* and *Bafinivirus* belonging to the subfamily *Torovirinae* [18]. The most distinguished characteristic shared by nidoviruses and recognized early in the course of research on nidoviruses, is the production of a nested set of subgenomic mRNAs. It provided a basis for the order's name: *nidus* means nest in Latin [19]. Other characteristics shared by viruses of the order include a conserved genome organization, conserved mechanism of

genome expression and a unique synteny of conserved protein domains revealed by comparative genomics [13].

With the exponential growth of the number of available nidovirus genome sequences, the number of known nidovirus species began to grow accordingly, although their formal classification within the taxonomy framework may lag behind. Likewise, the gap between the newly identified and the few experimentally characterized nidoviruses is also rapidly increasing. The latter group includes arteriviruses: equine arteritis virus (EAV) and PRRSV, and coronaviruses: transmissible gastroenteritis virus (TGEV), mouse hepatitis virus (MHV), SARS-CoV, MERS-CoV and IBV. Also, the limited characterization of several toroviruses, mesoniviruses and roniviruses, often isolated from exotic hosts, was important for understanding generalities and host- and lineage-dependent specifics of nidoviruses, and for the validation of many models of comparative genomics. Since viruses of the *Coronavirinae* and the *Arteriviridae* are most frequently sampled, they were predominantly used to characterize patterns of conservation and evolution at subfamily and family levels.

NIDOVIRUS GENOME ORGANIZATION

Nidoviruses are characterized by a conserved genome organization including multiple open reading frames (ORFs) (Fig. 1). The two largest and slightly overlapping ORFs, 1a and 1b, occupy the 5'-terminal two-thirds of the genome and encode non-structural proteins (nsps). ORF1a and ORF1b are chiefly responsible for the control of genome expression and replication, respectively [20]; together, they are referred to as the replicase gene. The 3'-terminal region of the genome contains smaller ORFs (3'ORFs), the number of which varies considerably among nidoviruses and which encode structural and, in some nidoviruses, accessory proteins. This region is chiefly responsible for virus dissemination [20]. Untranslated regions (UTRs) are present at the 5'- and 3'-ends of the genome, and may also be found between ORFs in the 3'ORFs region. The genomic 5'-end is believed to be capped [21-23], and 3'-end of the genome is polyadenylated [24, 25].

NIDOVIRUS LIFE CYCLE

Following virus entry into the host cell's cytoplasm and uncoating, the genome is translated by host ribosomes. Translation of ORF1a is thought to be initiated by ribosomal scanning of the genomic 5'-end [10]. In part of the cases, termination of ORF1a translation occurs at the ORF1a stop-codon, resulting in polyprotein 1a (pp1a) production. In the remaining cases, -1 programmed ribosomal frameshifting (PRF) occurs at a site located in the ORF1a 3'-terminus. PRF redirects the ribosomes to ORF1b translation, leading to production of a longer polyprotein, pp1ab [26, 27]. Polyproteins pp1a and pp1ab are co- and post-translationally cleaved by cognate protease(s), releasing intermediate precursors and

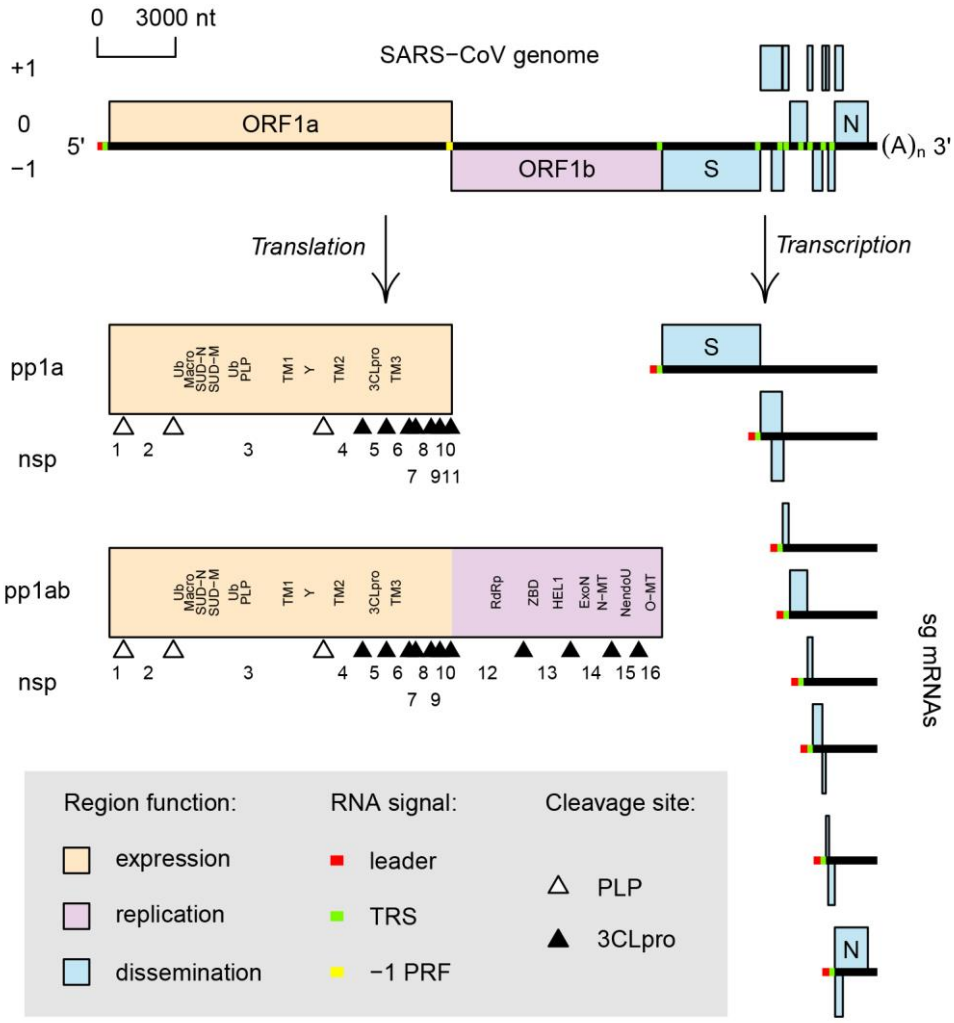


Figure 1 | SARS-CoV genome organization and expression. Genome (top), products of genome translation (left) and transcription (right) are shown. ORFs and polyprotein regions are colored according to their predominant function (see inset). Genome ORFs are depicted in their frame, with ORF1a frame set to zero. For each sg mRNA, only ORFs believed to be translated from it are shown, without indicating their frame relative to ORF1a. For genome and sg mRNAs, RNA signals are indicated by color (see inset). For polyproteins, processing scheme (see inset) and protein domains (see text for abbreviations) are specified. The NC_004718.3 record was used to prepare this figure. Note that sg mRNA 3.1 [28] is not shown; Ub and Macro domains are separated by acidic, structurally disordered region of ~70 aa [29, 30].

mature nsps (Fig. 1) [31]. This mechanism ensures nsps to be expressed early in infection, with ORF1a-encoded proteins being synthesized in higher quantities compared to ORF1b-encoded proteins [32]. Due to their location downstream of the ORF1a start-codon, the

start-codons of the 3'ORFs are inaccessible for translation initiation via canonical ribosomal scanning of the genome molecule.

Nsps assemble into a membrane-bound replication-transcription complex (RTC) that mediates replication and (subgenomic) transcription of the genome [33, 34]. Replication is amplification of genome molecules (which also serve as mRNA) using antigenome templates. Transcription is the synthesis of a nested set of subgenomic (sg) mRNAs for expression of the 3'ORFs (Fig. 1). To produce sg mRNA, minus-strand RNA synthesis on the genomic template is interrupted after a compliment of a genome motif, the body transcription-regulating sequence (TRS) located upstream of a 3'ORF, is synthesized. The nascent minus-strand RNA is then translocated to the genomic 5'-terminus, where it anneals to the leader TRS, a genome motif almost identical to the body TRS, after which minus strand synthesis resumes. The resulting subgenome-length minus-strand RNAs serve as a template for sg mRNA synthesis [19]. Most nidoviruses produce multiple sg mRNA species, each defined primarily by its body TRS. Notably, some sg mRNA species of toroviruses and all sg mRNA species of roniviruses do not share a common 5'-terminal sequence with the genomic RNA [23, 35], indicating that attenuation of the minus-strand RNA synthesis at the body TRS may be the only universal step of nidovirus transcription [19]. Most sg mRNA species are monocistronic and serve to translate only their 5'-most ORF, but some sg mRNA species are polycistronic [19, 36, 37]. Expression from separate sg mRNAs allows to regulate the abundance of the respective structural and accessory proteins relative to each other and nsps [38-40].

The assembly of a virus particle is a multistage process that includes encapsidation of viral genome by multiple copies of nucleocapsid protein, and wrapping of the nucleoprotein complex by a host membrane, carrying viral structural proteins. The wrapping is coupled with budding into the lumen of the endoplasmic reticulum (ER) or Golgi complex, and followed by transportation of the virus particles to the plasma membrane through the secretory pathway, culminating in their release from the cell [10, 41].

NIDOVIRUS PROTEOME

The virus life cycle is mediated by RNA signals of the non-coding and coding regions, including the PRF site and TRSs mentioned above, and diverse proteins that account for approximately 95% of the genome in different nidoviruses. These proteins will be described below from a genomic perspective, according to their location in one of five regions, delineated using functional considerations and sequence conservation. These regions in the order of being encoded from 5'- to 3'-end include three regions of ORF1a: pre-TM2, TM2-3CLpro-TM3, and post-TM3 (TM2 and TM3 stand for two transmembrane

domains that flank the 3C-like protease, 3CLpro); the entire ORF1b region; and the 3'ORFs region.

Pre-TM2 region of ORF1a

The N-terminal region of nidovirus polyproteins, preceding the TM2 domain, carries multiple protein domains the conservation of which varies greatly, from virus to family. Many of the domains encoded in this region remain poorly characterized even in the few well-studied nidoviruses. The predominant function of this region in vertebrate nidoviruses appears to be related to regulation of the viral life cycle and interfering with host immune defenses [10, 42-46]. Characterization of the region in invertebrate nidoviruses was limited to TM domain predictions.

In arteriviruses, from two to six nsps are produced from this region by proteolytic processing and, in certain cases, PRF: nsp1 and nsp2 in EAV; nsp1a, nsp1b, nsp1c (specific to simian arteriviruses), nsp2 and its truncated variants nsp2N and nsp2TF in other arteriviruses. Coronaviruses may encode from two to three nsps in this region: nsp1 (specific to genera *Alpha*- and *Betacoronavirus*), nsp2 and nsp3. Arterivirus nsp2 and coronavirus nsp3 are multidomain proteins, the largest among the nsps of the respective taxonomic groups.

All nidoviruses encode at least one transmembrane domain, TM1, in the pre-TM2 genome region (Fig. 2). It resides in nsp2 and nsp3 of arteriviruses and coronaviruses, respectively (Fig. 1). TM1 together with other ORF1a-encoded TM domains may anchor RTC to cellular membranes [47], and were shown to induce cellular membrane rearrangements, such as double membrane vesicles formation [48, 49]. The precise role of the latter is subject to active research and may include local enrichment of particular viral proteins, compartmentalization and facilitating virus-specific processes, and protection of virus RNAs from host cell defenses [50].

Another ubiquitous domain of this region is the papain-like protease (PLP) that was (tentatively) identified in all vertebrate nidoviruses. The number of PLPs varies, one is encoded by toroviruses, from one to two – by coronaviruses, and from three to four – by arteriviruses [31, 45, 51]. To distinguish between multiple PLPs of a single nidovirus, their names are supplemented with indices: 1a, 1b, 1c (specific to simian arteriviruses) and 2 for arteriviruses; 1 and 2 for coronaviruses (Fig. 2). In arteriviruses, PLP1a (covalently linked to an N-terminal zinc-finger domain), PLP1b, PLP1c and PLP2 reside in nsp1a, nsp1b, nsp1c and nsp2, respectively; the only exception is EAV, where proteolytically inactive PLP1a and active PLP1b both reside in nsp1 [52-58]. Coronavirus PLPs reside in nsp3 [59].

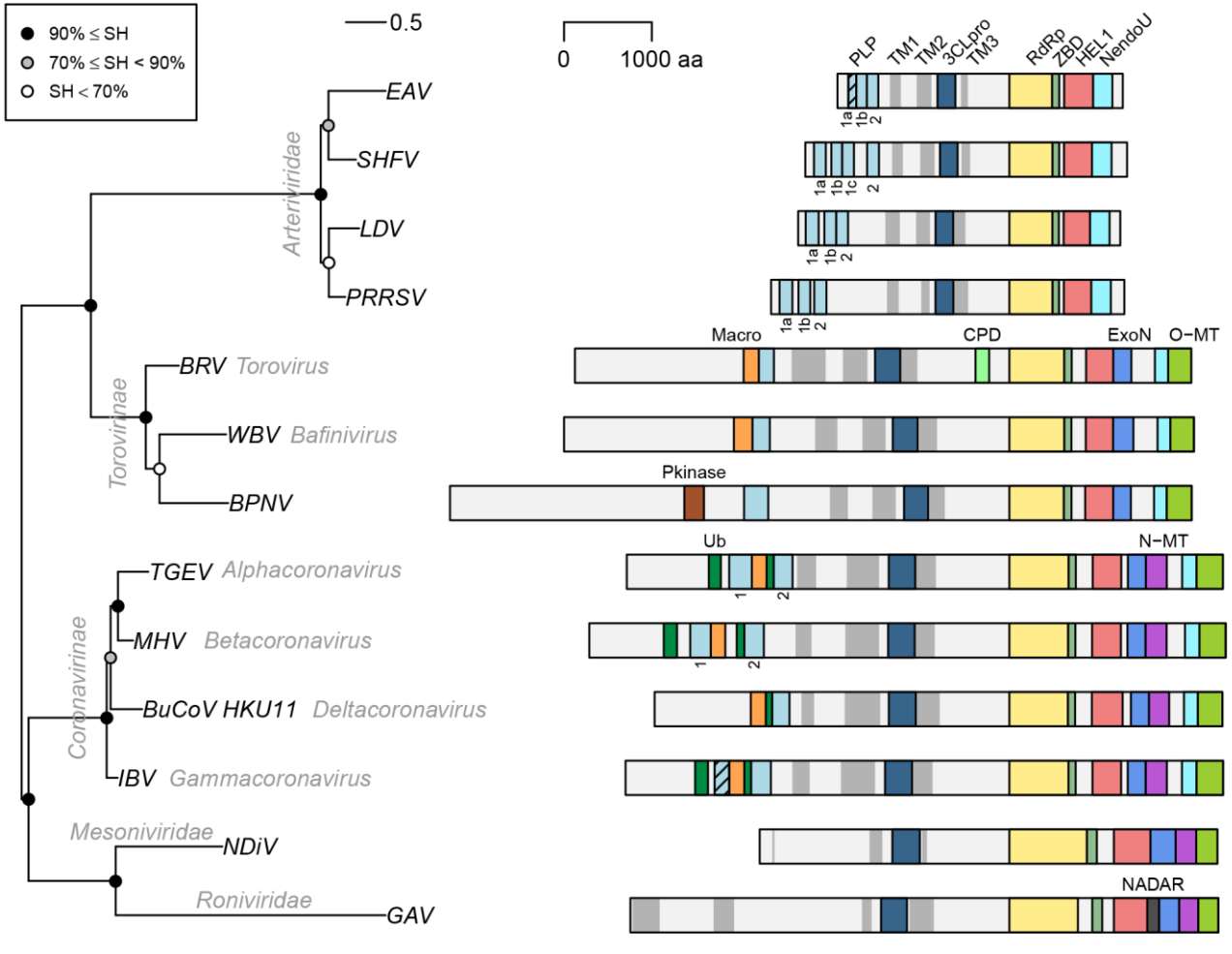

90% ≤ SH
70% ≤ SH < 90%
SH < 70%
0.5
0
1000 aa
PLP
TM1
TM2
3CLpro
TM3
RdRp
ZBD
HEL1
NendoU
1a
1b
1c
2
Macro
CPD
ExoN
O-MT
Pkinase
Ub
N-MT
1
NADAR
EAV
SHFV
LDV
PRRSV
Arteriviridae
BRV Torovirus
WBV Bafinivirus
BPNV
Torovirinae
TGEV Alphacoronavirus
MHV Betacoronavirus
BuCoV HKU11 Deltacoronavirus
IBV Gammacoronavirus
Coronavirinae
Mesoniviridae
NDiV
Roniviridae
GAV

Figure 2 | Midpoint-rooted phylogeny and pp1ab domain organization of nidoviruses representing (sub)families and genera recognized by ICTV as of 2014 [18], and BPNV [14]. Names of taxonomic groups are indicated in grey italic font. Phylogeny was reconstructed based on Viralis MSA [60] of the conserved core of RdRp, using IQ-Tree 1.5.5 [61] with automatically selected rtREV+F+I+G4 evolutionary model. To estimate branch support, SH-like approximate likelihood ratio test with 1000 replicates was conducted. Polyproteins are shown as light grey bars. TM domains are shown as dark grey bars; TM helices were predicted by TMHMM2.0c [62] and clustered if separated by less than 300 aa (less than 180 aa for arteri- and toroviruses). Other domains, whose coordinates were obtained from the Viralis database [60], are shown as colored bars; proteolytically inactive PLP domains are indicated by stripes on bars; indices of PLP domains are specified below the bars. SHFV, simian hemorrhagic fever virus; LDV, lactate dehydrogenase-elevating virus; BRV, Breda virus; WBV, white bream virus; BuCoV_HKU11, bulbul coronavirus HKU11; NDiV, Nam Dinh virus; GAV, gill-associated virus.

Arterivirus PLP1a, PLP1b and PLP1c are more similar to each other than to the arterivirus PLP2 [31, 56], which has a distinct fold with a zinc-finger embedded in it [63, 64]. Coronavirus PLPs share sequence and structural similarity, including a zinc-finger connecting two sub-domains of the protease [31, 65-67]. Torovirus PLP exhibits the strongest similarity to picornavirus leader protease and appears to lack a zinc-finger [51].

Arteri- and coronavirus PLPs, whose proteolytic activity was characterized experimentally, cleave N-terminal regions of pp1a and pp1ab at 1 to 3 sites to release their own, and, in case of coronaviruses, also upstream nsp(s) [53-56, 59]. In addition to the autoproteolytic activity mediated by its PLP domain, arterivirus nsp1/nsp1a couples translation of genomic RNA to transcription and, probably, particle formation [40, 68-70]. Arterivirus PLP2 and coronavirus PLPs possess deubiquitinating and deISGylating activities in surrogate systems; they are believed to inhibit cellular responses to viral infection by removing ubiquitin and ubiquitin-like molecule ISG15 from proteins of innate immune signaling pathways [63, 71-73]. Interestingly, ubiquitin-like (Ub) domains are part of coronavirus nsp3: one is positioned in the very N-terminus of the protein, and another – immediately upstream of the most C-terminal PLP (Fig. 1,2) [42, 74], both were initially identified in structural studies of SARS-CoV nsp3 [29, 66].

Another pre-TM2 domain conserved in multiple nidovirus lineages is the macrodomain, originally named X domain [75] and subsequently ADRP domain, due to its homology with cellular adenosine diphosphate ribose 1''-phosphotase [12]. The domain resides in nsp3 of all coronaviruses and a collinear pp1a/1ab position of toroviruses belonging to genera *Torovirus* and *Bafinivirus* (Fig. 1,2). The macrodomain of several coronaviruses was shown to possess ADRP activity *in vitro* [76, 77], and to bind mono- and poly-ADP-ribose (MAR and PAR) [78]. It was also proposed to bind adenosine monophosphate (AMP) ribose based on structural conservation in study of alphavirus macrodomains [79].

A cellular ADRP catalyzes the second reaction of the tRNA splicing pathway metabolite (ADP-ribose 1",2"-cyclic phosphate) processing, with the first reaction being catalysed by cyclic phosphodiesterase, as was demonstrated in *in vitro* experiments [80, 81]. Based on

analogy with this pathway, ADRP activity of nidoviral macrodomain was suggested to modulate the pace of a similar yet-to-be identified pathway by processing its metabolites [12]. Tagging proteins with PAR, PARylation, is a signal used by the cell to trigger antiviral defenses. PAR-binding activity of nidoviral macrodomain was suggested to counteract these defenses by acting on PARylated proteins [78], which became the leading hypothesis in the field.

In nsp3 of SARS-CoV, a large insertion named "**S**ARS-**u**nique" **d**omain (SUD) was identified immediately downstream of the conserved macrodomain (Fig. 1) [12]. It includes two divergent and adjacent copies of macrodomain, SUD-N and SUD-M, which bind G-quadruplexes rather than ADP-ribose [82, 83]. In addition to SARS-CoV, the SUD domain was shown to be conserved in other coronaviruses belonging to a monophyletic group 2b within the genus *Betacoronavirus* [74]. SUD-M-like domain was identified in several *Betacoronovirus* species outside of the 2b group [82]. Likewise, several *Alphacoronavirus* species were reported to contain another divergent macrodomain homolog in analogous nsp3 position [82]. Furthermore, macrodomain was not identified in the torovirus ball python nidovirus (BPNV) which encodes a homolog of protein kinase (Pkinase) in a similar polyprotein location, ~450 aa upstream of the PLP domain (Fig. 2) [12].

C-terminal regions of arterivirus nsp2 and coronavirus nsp3 have a similar domain organization: the C-terminal PLP domain is followed by a region of low conservation, which is called hypervariable region (HVR) in arteriviruses, TM1 domain and a unique conserved domain of unknown function: cysteine-rich domain of arteriviruses and Y domain of coronaviruses [10, 30, 42, 74]. The region is rich in zinc-binding modules, one was predicted to be embedded in the TM1 domain, another was tentatively identified in the arterivirus cysteine-rich domain and two in the coronavirus Y domain [42, 50]. Notably, non-EAV arteriviruses also express two truncated versions of nsp2 with alternative C-terminal regions, nsp2N and nsp2TF. Truncated proteins are expressed via -1 and -2 PRF at a genome site corresponding to HVR C-terminus; the PRFs redirect ribosome to translation of small ORFs in alternative frames [84].

TM2-3CLpro-TM3 region of ORF1a

This region includes three proteins, nsp3-nsp5 and nsp4-nsp6, in arteri- and coronaviruses, respectively [31]. A similar organization may be found in other nidoviruses due to the observed sequence conservation. The middle protein in this layout includes the 3CLpro, which was named after the 3C protease of picornaviruses. They share sequence, structural and functional similarity that includes a narrow substrate specificity towards (commonly) Glu/Gln and a small residue in the P1 and P1' subsites of the cleavage site, respectively [11, 31, 85-89]. Several key residues of 3C/3CLpros substrate-binding pocket include a hallmark His residue downstream of the nucleophile in the primary structure. The flanking

of 3C-like protease by TM2 and TM3 in a common precursor is a distinguishing feature of nidoviruses. Another specific characteristic of nidovirus 3CLpro is that its enzymatic domain with a chymotrypsin-like fold, universally conserved in all 3C/3CLpros, is fused with a variable accessory C-terminal domain [31, 90, 91]. The nucleophilic residue of 3CLpro varies depending on the nidovirus lineage: arteri- and toroviruses employ serine as part of a catalytic triad, while corona-, roni- and mesoniviruses employ a cysteine residue as part of a catalytic dyad [31, 86, 87, 92, 93]. The 3CLpro autocatalytically releases its nsp, which is nsp4 and nsp5 in arteri- and coronaviruses, respectively, as well as all downstream nsps, from pp1a and pp1ab polyproteins [31].

Post-TM3 region of ORF1a

The post-TM3 region of ORF1a encodes small proteins with no reported sequence or structure conservation across nidoviruses. In arteriviruses, the region encodes four proteins: nsp6, nsp7a, nsp7b, and nsp8; in coronaviruses – five proteins: nsp7 to nsp11 (Fig. 1) [10, 42]. They appear to serve as replication cofactors to enzymes encoded in ORF1b, although their exact function remains poorly understood and contested.

The region is best characterized in SARS-CoV. Purified nsp7 and nsp8 subunits of SARS-CoV were shown to form a hexadecameric cylinder-like complex (in contrast, feline coronavirus nsp7:nsp8 complex is a 2:1 heterotrimer [94]) proposed to serve as a processivity factor of RNA replication [95]). This hypothesis was later corroborated in a functional study [96]. Nsp8 was shown to possess *de novo* RdRp activity *in vitro*, and was proposed to synthesize primers for the main RNA polymerase of the virus [97]. The complex of nsp7 and nsp8 was shown to possess primer extension RdRp activity *in vitro*, and was hypothesized to function as a second, independent RNA polymerase [98]. Nsp10 is a cofactor essential for efficient proofreading and capping during virus replication and transcription [99, 100].

The post-TM3 region was not characterized in toroviruses or invertebrate nidoviruses, although sequence conservation between toro- and coronaviruses was documented for nsp7 and nsp8 [96]. Toroviruses of the *Torovirus* genus encode an extra lineage-specific domain in the 3’-terminus of ORF1a (Fig. 2) [101]. Because of its similarity to a better characterized cellular enzyme, it was named cyclic phosphodiesterase (CPD) domain [12]. Nidovirus CPD, like the ADRP/macrodomain (see above), was proposed to influence the pace of a yet-to-be identified pathway by processing its ADP-ribose 1”,2”-cyclic phosphate metabolites [12]. Homologs of CPD are also encoded in the 3’ORFs region (ns2 protein) of coronaviruses belonging to the monophyletic group 2a within the *Betacoronavirus* genus [12, 13, 101, 102]. Characterization of MHV ns2 revealed no CPD activity but demonstrated cleavage of 2’,5’-linked oligoadenylates, common cofactors of an interferon-induced antiviral pathway [103]. Accordingly, this domain is also called 2’,5’-phosphodiesterase (2’-PDE).

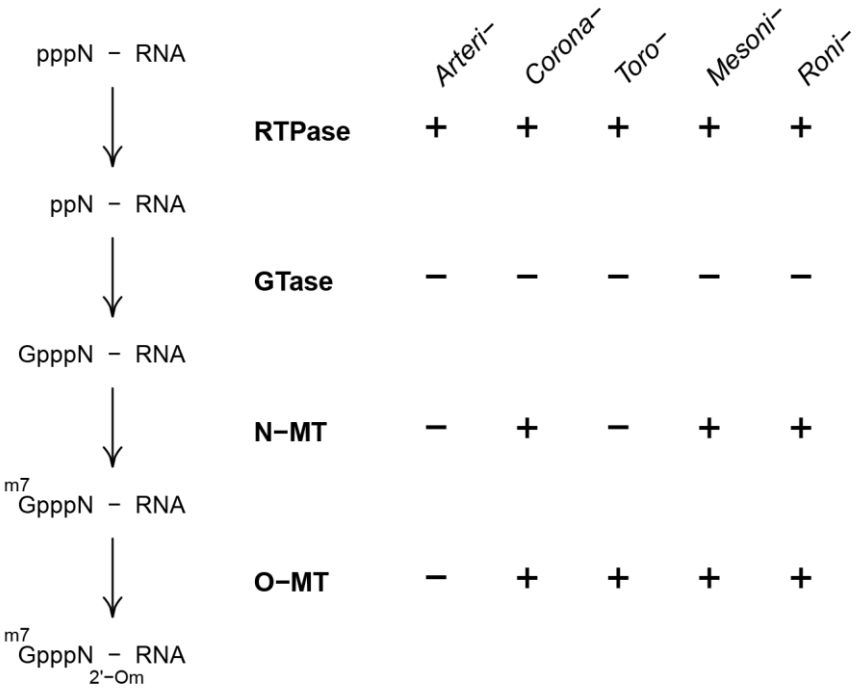


Figure 3 | Capping pathway and enzymes in relation to the proteome of nidoviruses. The conventional mRNA capping pathway is shown on the left, with the enzymes catalyzing the respective four reactions listed in bold. Further to the right, presence of these enzymes in viruses of five nidovirus (sub)families, each designated by its prefix, is listed. RTPase, 5'-triphosphotase; GTase, guanylyl transferase; N-MT, guanine-N7-methyltransferase; O-MT, 2'-O-methyltransferase. In $^{m7}GpppN_{2'\text{-}Om}$ notation, ^{m7}G stands for 7-methylguanosine, p stands for phosphate, $N_{2'\text{-}Om}$ stands for the 5'-terminal nucleoside of the RNA molecule, methylated at the ribose-2'-O position. For details, see text.

ORF1b region

The ORF1b region encodes key components of nidoviral RTC, most of which are found in either all or multiple nidovirus lineages. Accordingly, the region is the most conserved in the nidovirus genome, both in respect to its amino acid sequence, and the order of protein domains. The key and essential, nidovirus-wide conserved domains of the region, listed in N- to C-terminus order, include: RNA-dependent RNA polymerase (RdRp), zinc-binding domain (ZBD), and helicase of superfamily 1 (HEL1) [13].

RdRp catalyzes the synthesis of nascent RNAs on viral templates and mediates both genome replication and transcription [104]. Comparative sequence analysis and protein modelling mapped the RdRp to the C-terminal portion of the most N-terminal nsp encoded by ORF1b, that is nsp9 in arteriviruses and nsp12 in coronaviruses (Fig. 1) [11, 105]. On the RdRp tree, the nidovirus lineage is a sister group to the distantly related and

better characterized RdRps of the Picorna-like supergroup, which use protein primers to initiate RNA synthesis. The nidovirus RdRps differ from their distant homologs of the Picorna-like supergroup and other ssRNA+ viruses through the Gly-to-Ser replacement in the GDD tripeptide (C motif) that includes two catalytic aspartate residues [13]. Thus, the RdRp SDD tripeptide is a signature of nidoviruses, although it could also be found in RdRps of ssRNA- viruses [106].

ZBD and HEL1 reside in the N- and C-terminal parts of a single nsp, which is nsp10 in arteriviruses and nsp13 in coronaviruses (Fig. 1). ZBD includes twelve Cys and His residues that coordinate three zinc ions, and is thought to regulate HEL1 activity [107]. HEL1 is a helicase, NTP-dependent enzyme capable of dissociating nucleic acid base pairs. This activity may assist RdRp by unwinding double-stranded RNA duplex and/or a secondary structure of a single-stranded RNA during viral genome replication and transcription [108]. In addition, nidovirus HEL1 possesses RNA 5'-triphosphotase (RTPase) activity that may catalyze the first reaction of the RNA capping pathway [109, 110]. No homologs of ZBD were found in other viruses, making it a marker of the order *Nidovirales* [107]. In contrast, the closest homologs of the HEL1 domain are encoded by plant and animal viruses of the Alpha-like supergroup [111]. The ZBD-HEL1 organization was found also in cellular helicases involved in nonsense-mediated mRNA decay [107, 112], which may have functional implications and indicates a possible common origin (see below).

In addition to HEL1 RTPase activity, two other ORF1b-encoded enzymes, guanine-N7-methyltransferase (N-MT) and 2'-O-methyltransferase (O-MT), may catalyze the third and fourth reactions of the conventional mRNA capping pathway (Fig. 3) [113-116]. N-MT and O-MT reside in coronaviruses nsp14 and nsp16 (Fig. 1), respectively, and they are colinear in the pp1ab polyproteins of mesoni- and roniviruses, whose nsps are yet to be described fully (Fig. 2) [12, 93, 115]. However, contrary to their essential involvement in the mRNA capping, these enzymes are not conserved in all nidoviruses (Fig. 2, 3). Specifically, toroviruses encode O-MT, but appear to lack N-MT, while both N-MT and O-MT are missing in arteriviruses [93]. Additionally, the enzyme catalyzing the second reaction of the capping pathway, guanylyltransferase (GTase), has not been identified in any nidovirus [116]. Since nidoviruses are unlikely to subvert the capping machinery of eukaryotic hosts that functions in the nucleus, it remains unresolved how they synthesize the 5'-end cap [21-23], which controls translation initiation and protects the RNA molecule from degradation [117]. This uncertainty leaves open also the question about the natural targets of N-MT and O-MT, and methylation of other substrates than the 5'-terminal nucleotides remains a valid option [12].

Nidoviruses with genomes larger than 20 kb encode an exoribonuclease of the DEDD superfamily (ExoN) downstream of HEL1 [12, 93]. ExoN and N-MT reside in the N- and

C-terminal regions, respectively, of the same nsp (nsp14 in coronaviruses) [12, 115]. ExoN was shown to cleave RNA in the 3'-to-5' direction [39], and specifically to hydrolyze a single mismatched nucleotide at the 3'-end of an RNA molecule in a duplex [100]. Compared to other RNA viruses, mutation rates of ExoN-containing nidoviruses were shown to be lower, while ExoN inactivation increased the mutation rate [118, 119]. Based on these results, the genomic co-localization of ExoN with RdRp and HEL1, and ExoN homology to the DNA proofreading enzymes, ExoN must be a unique RNA proofreading enzyme that ensures the fidelity of the replication machinery in nidoviruses with large genomes [12].

Unlike other viruses, all vertebrate nidoviruses encode a uridylate-specific endonuclease (NendoU) in the 3'-terminal region of ORF1b (nsp11 and nsp15 of arteri- and coronaviruses, respectively) [12, 93]. NendoU cleaves RNA after U nucleotides and its inactivation compromises RNA replication, although its substrate(s) and function(s) in the nidovirus life cycle remain unknown [120-123]. Coronavirus nsp15, containing the NendoU domain, may counteract host innate immune response [124]. Cellular homologs of NendoU were shown to release certain small nucleolar RNAs from pre-mRNA introns [125], and to play a role in the shaping of ER [126].

Besides the above domains found in either all or multiple nidovirus lineages, ORF1b encodes lineage-specific domains. One of these domains, which is totally uncharacterized and apparently unrelated to others, resides in most C-terminal nsp12 of arteriviruses [10]. Another lineage-specific domain resides between HEL1 and ExoN of roniviruses *Gill-associated virus* and *Yellow head virus* (Fig. 2). Its poorly characterized homologs were found in diverse cellular organisms and viruses. Based on the position of some of these homologs within bacterial nicotinamide adenine dinucleotide (NAD) biosynthesis operons, the domain was named NADAR (after NAD and ADP-ribose) and implicated in regulation of NAD metabolism [127].

3'ORFs region

The ORFs located downstream of ORF1a/1b encode structural and accessory proteins. Structural proteins are proteins forming virus particles. Coronaviruses universally employ four structural proteins that are encoded in the order from 5' to 3': large multidomain spike (S) glycoprotein, transmembrane envelope (E) and matrix (M) proteins, and nucleocapsid (N) protein. Multiple copies of the N protein bind the virus genome to form a nucleoprotein complex, M protein is abundant in the envelope, and S protein forms the structures that protrude from the envelope and interact with cellular receptors during virus entry [41]. In addition, the N protein may stimulate replication, indicating a cross-talk between structural and non-structural proteins [128]. S, M and N proteins are essential for the formation of infectious virus particles [129]. E protein has ion channel activity, and

may be dispensable for virus replication [130, 131]. Arteri-, toro- and mesoniviruses encode equivalents of S, M and N proteins of coronaviruses, which may be designed differently and whose intergroup similarity is either weak or uncertain [13, 132, 133]. For example, a complex of several small arterivirus proteins may correspond to coronavirus S protein [134]. An equivalent of the coronavirus N protein was also identified in roniviruses [135].

Accessory proteins are defined as those that are dispensable for virus replication in tissue culture [136]. Some accessory proteins, such as SARS-CoV 3a protein, were identified in virus particles in apparently low molar quantities; their role remain uncertain [137, 138]. Nidoviruses differ considerably in respect to the number, type, and gene location of accessory proteins encoded in the 3'ORFs region. Generally, arteriviruses do not encode accessory proteins in the 3'ORFs region, while coronaviruses may encode from a few to multiple accessory proteins, many of which are small but some exceed 300 aa in size. Among the best characterized are the CPD/2'-PDE proteins, encoded by group 2a coronaviruses [13, 103], and the hemagglutinin-esterase (HE), encoded by group 2a coronaviruses and members of the genus *Torovirus* [101, 139, 140]. Accessory proteins are believed to play a role in virus-host interactions, such as interfering with cellular metabolic pathways and evading host immune defenses [138].

NIDOVIRUS MACROEVOLUTION

Due to their large genome size, nidoviruses may have the largest and most diverse proteome among RNA viruses. The origins of nidovirus proteome are numerous and its evolution is complex, and both are barely understood. As mentioned above, nidoviruses diverged considerably and unevenly in different genome regions, with only a small portion of the genome – mostly in ORF1b – being conserved across the entire virus order. In other non- and protein-coding regions, sequence conservation is phylogenetically restricted to lineages at different taxonomy levels, from species to families.

Only key replicative domains are conserved across the order *Nidovirales*, both in respect to their sequence and genome location. Two of these domains, RdRp and HEL1, are the largest and least diverged, which makes them favorable for the reconstruction of a nidovirus-wide phylogeny. Five (sub)families of nidoviruses invariably form distinct clades on the tree reconstructed based on these domains, either individually or in combination with each other and 3CLpro [14, 93, 133]. However, the root of the tree and relative position of the clades remain uncertain, as they vary depending on virus sampling, choice of domains, outgroup and algorithm.

Two conventional mechanisms, point mutation and recombination, shape proteome composition and evolution under the pressure of selection. As is typical for RNA viruses, mutation rates of nidoviruses are high, they were estimated as $2.5x10^{-6}$ and $9.0x10^{-7}$ mutations per nucleotide per replication cycle in MHV and SARS-CoV, respectively [118, 141]. Combined with a short replication cycle and large progeny, this results in nidoviruses of different families accepting multiple substitutions at almost every genome position upon divergence. In these viruses, replacements are observed even in the most conserved positions, such as catalytic residues of replicative proteins [13]. Accordingly, the amount of substitutions, accumulated in conserved proteins of nidoviruses and organisms of the tree of life since the time their respective most recent common ancestors (MRCAs) existed, is considered comparable [20]. It is conceivable that the actual number of substitutions per position upon nidovirus divergence may have been underestimated, due to limitations of the existing techniques, difficulties of reconstructing the chain of replacements at large evolutionary distances, and paucity of sampling of virus genome sequences available for analysis. Because of these complications, there is a considerable uncertainty about the timeframes of nidovirus evolution from species to the order levels. Based on general considerations, it was proposed that nidovirus lineages might have an ancient origin [105]. This hypothesis is supported by an estimation of divergence time of coronaviruses as 55.8 million years ago using a state-of-the art evolutionary model. Also, this timeframe is compatible with the separation of all invertebrate nidoviruses in a large monophyletic clade [14, 93], indicative of nidovirus-host coevolution, although a different topology was observed in some studies [133].

Besides single residue replacements, genetic changes may involve many residues or even domains as a result of recombination between two or more genomes (parents). Recombinant progeny has a distinct phylogenetic signal that separates it from the parents and is used to identify recombination, which is yet to be studied directly at the molecular level. Similarly to other RNA viruses, nidovirus recombination is believed to occur when the RdRp switches from one template (donor) to another (acceptor) in the course of genome replication [142]. Three types of recombination are distinguished based on the nature of the donor and acceptor templates. Homologous recombination occurs when the RdRp switches between orthologous regions of closely related viral genomes. Aberrant homologous recombination occurs when the RdRp switches between non-orthologous regions of closely related viral genomes (the same viral genome may serve as both donor and acceptor). Non-homologous recombination occurs when the RdRp switches between a viral genome and an RNA molecule of a different origin [143, 144]. RNA secondary structure, and sequence similarity between donor and acceptor templates in and around the crossover site were suggested to guide recombination [145, 146]. Besides, alternative mechanisms of recombination including biphasic recombination with imprecise

intermediates [147] and nonreplicative recombination [148] were reported for other RNA viruses.

Homologous recombination is a major mechanism of nidovirus microevolution, responsible for a considerable fraction of natural intra-species variation [13]. In contrast, two other types of recombination are detected less frequently. Both of these mechanisms mediate major genome innovations, domain acquisition and loss, the fixation of which may occur only if there is no counteracting purifying selection pressure.

Aberrant homologous recombination is the mechanism behind gene duplications and losses. In nidoviruses paralogous domains bearing hallmarks of duplication, such as tandem location and similarity clustering, were documented [13]. The variable numbers of PLP domains encoded in ORF1a of vertebrate nidoviruses are believed to have been generated as a result of gene duplications and losses [30, 31, 149]. As all PLPs of coronaviruses are associated with an N-terminal Ub domain (Fig. 2), it was suggested that a duplication of the Ub-PLP cassette may have occurred in an ancestral coronavirus [74]. The coronavirus macrodomain is thought to have given rise to SUD-N and SUD-M domains through duplication in an ancestral betacoronavirus which was followed by domains diversification (Fig. 1) [42, 82, 83]; a similar duplication must have occurred independently in an ancestral alphacoronavirus [82]. Coronavirus nsp2 appears to consist of a duplicated fold [42], nsp3 of human coronavirus HKU1 (HCoV-HKU1) harbours multiple short tandem repeats upstream of PLP1 (different isolates possess from 2 to 17 perfect, and from 1 to 4 imperfect copies of the acidic NDDEDVVTGD repeat) [150, 151] thought to be a result of duplication, while coronavirus nsp8 and nsp9 were suggested to have emerged as a result of RdRp and 3CLpro duplication, respectively, accompanied with a profound divergence and specialization to a new function [97, 152]. Also, the N-MT might have originated by duplication of the O-MT domain early in evolution of nidoviruses (Gorbalenya, personal communication). Duplication of a cluster of 3'ORFs may have led to the emergence of structural genes unique for the clade of simian arteriviruses [153]. In all documented cases, except the case of the HCoV-HKU1 tandem repeats, the similarity between duplicates is low or very low, and in a number of cases duplications have been recognized only upon analysis of tertiary structures, indicating that the actual number of duplications may be underreported. Duplications appear to have occurred in all three major nidovirus regions, ORF1a, ORF1b and 3'ORFs, and in different lineages at different scales of divergence, indicating that they have been common throughout nidovirus evolution. Another notable feature of this process is that similar duplications seem to have occurred independently (or in parallel) in several lineages. This appears to be the case with the duplication of PLPs in arteri- and coronaviruses, and macrodomains in alpha- and betacoronaviruses. This observation is indicative of pervasive selection pressure and common constraints in different nidovirus lineages.

Non-homologous recombination is the mechanism behind gene acquisition from hosts and other viruses co-infecting host cells together with nidoviruses. It has been invoked for nidovirus domains that have homologs in other biological entities, and is equivalent to lateral or horizontal gene transfer (LTG or HTG), a major mechanism of biological evolution [154]. A major challenge in the analysis of this type of events is assigning the donor and recipient species for domain transfer, which requires placing this event in a broader evolutionary context that may not be readily reconstructed. The hemagglutinin esterase, encoded in the 3'ORFs genome region of group 2a coronaviruses and the genus *Torovirus*, was probably the first RNA virus domain proposed to have been acquired using this mechanism [101, 139]. Influenza virus (InfV) C seemed to be a plausible donor of HE since it relies on this enzyme for cell entry while nidoviruses use it to bind a secondary receptor [155]. Since the respective HE-containing corona- and toroviruses are separated by a large evolutionary distance in replicative proteins and both groups are closely related to viruses that are HE-free, it is unlikely that HE was acquired by their common ancestor. Instead, either of the two nidovirus groups might have acquired HE from an external source, possibly InfV C, and then that HE might have been captured by the other group through a recombination event [140]. Raoul J. de Groot also suggested that HE might have been acquired by the two nidovirus groups independently [140]. Another element that has scattered phylogenetic distribution is mobile RNA module s2m. It is a stem-loop module that is present in the genomic 3'-terminus of a number of corona-, picorna-, calici- and astroviruses [156]. The module is characterized by a high level of conservation on primary, secondary and tertiary structure levels, and is believed to have been acquired by various groups of viruses through non-homologous recombination, while its function remains obscure [156].

Besides the HE-protein domain mentioned above, many other domains might have been acquired via a non-homologous recombination mechanism, although the exact origins of these domains are less clear, and their number and identity require reconstruction of the proteome composition of ancestral nidoviruses. There is little doubt that domains found in many organisms and/or viruses but identified only in few nidoviruses are of external origin. These include CPD/2'-PDE of toro- and coronaviruses [12, 13, 101, 102], Pkinase of BPNV [14], and uridine kinase of beluga whale coronavirus SW1 [157]. If the ancestral nidovirus evolved from an astro-like virus [13], all known conserved ORF1ab enzymes and domains, except for TM2-3CLpro-TM3 and RdRp, must have been acquired at some point of nidovirus evolution. In several cases, the viral origin of nidovirus domains was suggested based on sequence affinity: ADRP/macrodomain and HEL1 might have been acquired from viruses of an alpha-like supergroup [78, 158], while O-MT might have been transferred from a flavivirus or a virus of the order *Mononegavirales* [159]. On the other hand, sequence affinity between the respective domains of different origins listed above

as well as between corona- and torovirus PLPs and foot-and-mouth disease virus leader protease [51, 75] could be explained by other evolutionary scenarios, including domain transfer in the opposite direction and/or involvement of host homologs. For example, the unique association of the superfamily 1 helicase domain with an N-terminal zinc-binding module, observed only in nidovirus helicases and eukaryotic Upf1-like helicases, as well as the structural similarity between these helicases, may point to the cellular origin of nidovirus ZBD and HEL1 [107, 160]. Likewise, the acquisition of ExoN and NendoU from a host by ancestors of nidoviruses seems likely due to the phyletic distribution of their homologs, restricted (with the exception of the arenavirus exoribonuclease that does not exhibit specific sequence affinity to nidovirus ExoN) to cellular organisms [12, 93].

The acquisition of ExoN, which mediates RNA proofreading, was most decisive for nidoviruses. The domain is encoded by all nidovirus families except arteriviruses, a group characterized by genome sizes that do not exceed 16 kb, 4 kb smaller than the next smallest nidovirus [20]. These two characteristics – lack of ExoN and small genome size – are tightly interconnected, due to the inverse correlation between mutation rate and genome size in viruses and prokaryotes [161, 162]. Accordingly, arteriviruses (and all other RNA viruses lacking proofreading activity) are believed to be locked in a state of “Eigen trap”: due to the low fidelity of RNA synthesis, their genome sizes must remain small to avoid “error catastrophe”, systemic abortion of viral infection after the number of accumulated mutations reaches a critical threshold [163-165]. Acquisition of ExoN by an ancestral nidovirus was proposed to have allowed an escape from “Eigen trap”, leading to unprecedented genome expansion and emergence of nidoviruses with the largest RNA genomes known [93].

The results discussed above and others implicate a continuous accumulation of substitutions, duplication, and horizontal gene transfer in shaping evolution of the entire genome and proteome of nidoviruses. However, the available reconstructions of nidovirus macroevolution are alignment-based and hence involve only a small fraction of the genome – few universally conserved replicative domains. To address this challenge, a new alignment-free approach to reconstruction of virus evolution was proposed in our group [20]. It models dynamics of genome-size change during evolution of a monophyletic group of extant viruses under the assumption that fundamental constraints acting on nidovirus genomes remain unchanged in the course of evolution, and hence both extant and extinct nidoviruses belong to the same evolutionary trajectory. Functionally equivalent genome regions of the extant viruses are delineated using few orthologous residues, and their sizes are noted. Spline regression is then used to approximate the relationship between the size of each region and the genome, later differentiated to produce a model of relative contribution of each region to genome expansion. The approach was applied to the genomes of nidoviruses which were split into five regions, three of which – ORF1a, ORF1b,

and 3'ORFs – accounted for >95% of the genome size. The resulting model had a three-wave shape, predicting that genome expansion in 15-20, 20-26 and 26-32 kb size ranges was dominated by ORF1b, ORF1a and 3'ORFs expansion, respectively. This order can be explained by the predominant functions of the regions: modification of replication machinery (ORF1b) might have required adaptation of the expression mechanisms (ORF1a), and an increase in virion size to accommodate growing viral genomes (3'ORFs). Notably, according to the model, a new wave of ORF1b domination in genome expansion is starting in the 26-32 kb genome size range, indicating a possibility of a second cycle of genome expansion [20].

TOOLS OF NIDOVIRUS COMPARATIVE GENOMICS

Comparative genomics has been instrumental in the characterization of nidoviruses since the first nidovirus genome was sequenced. With only a single nidovirus genome sequence and a few dozen others available [166], and with no prior knowledge about the function of non-structural proteins of nidoviruses, bioinformaticians identified an array of six key replicative domains of nidoviruses: TM2-3CLpro-TM3-RdRp-ZBD-HEL1, as well as correctly predicted nine out of eleven 3CLpro cleavage sites in the IBV replicase [11, 167-169]. These studies utilized three approaches to domain mapping and functional assignment: i) analysis of aa residue distribution to reveal enrichment indicative of structural and functional significance; ii) identification of distant homology to a better characterized protein through profile comparison and motif recognition; iii) enhancement of weak sequence signals by making them conditional on other information available. Subsequent experimental characterization verified and corroborated these tentative assignments [170]. In addition to the domains listed above, comparative genomics identified diverse PLPs [51, 149, 171, 172], macrodomain [12, 75], Pkinase [14], ExoN [12], NendoU [12], O-MT [12], HE [101, 139], and CPD/2'-PDE [12, 13, 101, 102] as well as many Zn-binding modules, and the deubiquitinating activity of the coronavirus PLP [173]. Importantly, these analyses were also accompanied with a few "false positives" when tentative assignments and functional interpretations were refuted later (for details see [170] and also [93]). Likewise, comparative genomics missed some distant relationships ("false negatives") which were either revealed by comparative structural analysis (Ub domain [29, 66]), or required experimental characterization, as was the case with the identification of N-MT domain [115]. This experience indicated limitations and challenges of comparative genomics of distant relationships in the so-called twilight and midnight zones [174, 175], central to which are the tools and databases available, and the divergence of the domains in question. These aspects are briefly discussed below.

The most basic comparative genomics approach is to look for sequence patterns that are unlikely to have emerged by chance – a signature of selection indicative of functional

importance. These include anomalies in residue distribution. For example, fluctuations of G+C content along a nidovirus genome sequence may point to a region subject to transcription [176], while in a protein sequence, regions rich in Cys and His are potential zinc-binding modules, elevated concentration of Cys may be associated with a secreted protein whose structure is maintained by disulphide bridges, regions rich in basic residues may serve for nucleic acid binding, and domains enriched with hydrophobic residues are likely to be transmembrane. Identification of sequence motifs, such as cleavage sites of specific proteases and sites of post-translational modification of a protein [177, 178], are other examples of bioinformatic input to experimental research. Importantly, sequence patterns are not restricted to a primary structure. A predicted secondary structure of a protein can offer clues about its fold and function, whereas a predicted secondary and tertiary structure of an RNA genome can help to identify functional elements regulating its expression. For example, the PRF site at the ORF1a/1b junction of nidoviruses can be recognized as a characteristic combination of a slippery sequence, where the frameshifting takes place, and a downstream pseudoknot structure stalling the ribosome and prompting the frameshifting [27, 179].

Many approaches of comparative genomics involve obtaining sequence alignment that seeks to maximize similarity between input sequences. When similarity is considered statistically significant (see below), it is interpreted using biological reasoning, typically as aligned sequences being homologous, i.e. descendants of a common ancestral sequence. Technically, alignment is a matrix where rows correspond to sequences, while columns contain aligned residues and may include gaps introduced to maximize residue similarity and representing insertions and deletions that happened during evolution. Upon alignment of homologous sequences, residue variation in a column reflects structure, function, and selection pressure on a residue of a biomolecule. It is used in various analyses including prediction of secondary structure, identification of functionally important residues and evolutionary inferences. Also, alignment facilitates transfer of knowledge: if a functional role was experimentally established for residues in an aligned sequence, homologous residues of other aligned sequences can be readily identified, allowing to predict their function.

Depending on whether two or more sequences are included into the alignment, pairwise and multiple sequence alignments (MSAs) are distinguished. Optimal (characterized by the highest possible sequence similarity score) pairwise sequence alignment can be built by dynamic programming algorithms, which produce a solution by gradually finding optimal sub-solutions. The computational complexity of building optimal MSA is extremely high, and heuristic techniques, such as progressive alignment, where an approximate phylogenetic tree is reconstructed to guide gradual building of an MSA, are used instead [180, 181].

The processes of establishing sequence homology and building sequence alignment are often intertwined. A new sequence (query) is usually compared to a database of known sequences, in a process that produces its alignment with every entry of the database (targets). The sequence similarity score of each alignment is used to calculate a measure of its statistical significance, and those alignments that satisfy a selected statistical significance threshold are considered nonrandom, reflecting either genuine homology or occasional convergence [182]. Heuristic algorithms such as BLAST are often used to decrease computational intensity of the search [183]. To increase the sensitivity of the search and facilitate detection of distant homology, query and/or targets can be represented by profiles instead of individual sequences. A profile is a statistical model that allows to comprehensively describe a family of homologous sequences by accommodating information about the nature and frequency of residues in each column of their MSA. The two most popular profile types are the position-specific scoring matrix (PSSM) and the hidden Markov model (HMM) [184-186].

Distinguishing weak similarity from chance events, a common challenge in studies of proteins of nidoviruses, requires the proper use of statistical significance measures and thresholds. The most widely used measure, employed by various software packages and characterizing an alignment between a query and a database target, is the E-value. This is the number of alignments, characterized by the same or a greater degree of similarity between query and target, that are expected to be found in a database of the same size for a query of the same size just by chance. Conventionally, alignments characterized by E-value < 0.001 are given serious consideration as potentially reflecting genuine homology between the aligned sequences. Importantly, the E-value depends on the size of the database: with the growth of the database, the E-value characterizing an alignment between a query and a database target would increase, even though the alignment itself would remain unchanged [181]. Individual software packages may employ unique statistical significance measures, specific for the underlying algorithm. For example, the HH-suite software package designed to perform HMM-HMM comparisons employs Probability, a measure estimating the probability of the target HMM to be homologous to the query HMM on a scale from 0 to 100%. The measure takes secondary structure similarity into account, does not depend on the size of the database, but is sensitive to the size of the query. When Probability exceeds 95%, homology between the aligned profiles is believed to be nearly certain [186, 187].

Confident recognition of weak similarity may require applying the most sensitive tools of homology and motif detection while taking other considerations into account. One of the powerful approaches is to limit the search space based on biological reasoning, and thus facilitate the detection of weak signals by increasing the signal-to-noise ratio. For instance, identification of 3CLpro cleavage sites was facilitated by considering only small regions

around tentative domain borders in large nidovirus polyproteins instead of the entire polyproteins [11]. Another important consideration is to use protein rather than nucleotide sequences whenever possible when dealing with distant homology, as protein sequences are unaffected by synonymous nucleotide substitutions and hence diverge slower than nucleic acid sequences. Finally, it is important to use databases where sequences from diverse organisms and viruses are represented, as it expands coverage of the sequence spaces occupied by various protein families, and hence increases the chances of detecting distant homology [175].

Analysis of evolutionary relationships between homologous sequences can be facilitated by building a phylogenetic tree [188]. Trees reflecting interspecies nidovirus relationships are considered "deep" because of their considerable branch lengths reflecting the high number of substitutions. Two preferred methodologies for "deep" phylogeny reconstruction are Maximum Likelihood (ML) and Bayesian inference [188]. Both methodologies are centered around the likelihood function, $L(D|t,v,\Theta)$, probability of the data (sequence alignment) D given tree topology t, branch lengths v, and substitution model parameters Θ. ML algorithms reconstruct phylogeny by finding, with the help of various heuristics, values of t, v, and Θ parameters that maximize the likelihood function [189]. Bayesian algorithms employ Bayes' theorem to estimate probability distribution of parameter values given the data, $P(t,v,\Theta|D)$, based on the likelihood function and prior knowledge about the values of parameters. The estimation relies on the Markov chain Monte Carlo sampling procedure [190]. Phylogeny can serve as a basis for ancestral state reconstruction analysis inferring the state of a phenotypic trait for extinct viruses corresponding to its internal nodes, given that the state of the trait is known for extant viruses represented by its tips [191]. This analysis can be applied to a broad range of traits, from the nature of a catalytic residue to a host habitat [92].

Comparative genomics analysis can be facilitated by the taxonomic classification of viruses under consideration. Taxonomic classification offers a framework to organize the existing knowledge about virus biology. It also helps to design comparative genomics experiments, e.g. by allowing to represent each of the species in a virus group by a single genome – a technique that is appropriate in analyses on a macroevolutionary scale, and helps to account for the existing sampling bias, with a disproportionately large number of available virus genomes belonging to a few species of high societal significance. A classification assigning a newly discovered virus to a taxonomic group may immediately offer clues about its biology, as the virus is likely to share biological properties characteristic for the group. However, devising virus taxonomy is a challenging task, as it requires building multi-level hierarchical classification while dealing with large evolutionary distances separating fast-evolving viruses. Several tools, including PASC (PAirwise Sequence Comparison; [192]), DEmARC (DivErsity pArtitioning by hieRarchical Clustering; [193]), and

SDT (Sequence Demarcation Tool; [194]) were designed to build taxonomic classifications of viruses.

SCIENTIFIC QUESTIONS

Studies included in the next three chapters of this thesis focused on scientific questions about the composition and evolution of the nidovirus genome and proteome, and their connection to the biology of nidoviruses. All of the studies included in the next three chapters extensively used methods of comparative genomics. These studies benefited from previous comparative genomics research on nidoviruses (see above), and were facilitated by an explosive growth in nidovirus genome discovery by NGS (including also genome sequences provided by collaborators of our group), as well as by a steady advancement of tools and databases employed in comparative sequence analyses. In **chapter 2**, the characterization of arterivirus pp1ab N-terminus encoding multiple PLPs, that included three times more species than the published reports and employed advanced toolkits for homology and evolutionary analyses, provided insight into the role and contribution of duplication in virus adaptation. In **chapter 3**, a collaboration between bioinformaticians and experimental researchers allowed to analyze a protein domain adjacent to the RdRp (the only ORF1b region that remained uncharacterized in all nidovirus lineages despite decades of prior research) in respect to its conservation in nidoviruses, evolutionary origin, biochemical activity and potential function. Finally, **chapter 4** presents an analysis of a highly divergent nidovirus with the largest known RNA 41.1 kb genome. Its analysis was insightful for advancing our existing understanding of limits and mechanisms of RNA genome expansion, linkage between major characteristics defining nidoviruses, and evolutionary plasticity of the nidovirus proteome and its expression. That study also prompted research that addresses an important technical challenge of RNA virus comparative genomics: **chapter 5** describes a tool, LArge Multidomain Protein Annotator (LAMPA), developed for homology recognition and annotation of large and divergent multidomain proteins of RNA viruses.

REFERENCES

1. Geoghegan JL, Holmes EC: **Evolutionary Virology at 40**. *Genetics* 2018, **210**(4):1151-1162.
2. Baltimore D: **Expression of animal virus genomes**. *Bacteriol Rev* 1971, **35**(3):235-241.
3. Beijerinck MW: **Concerning a contagium vivum fluidum as cause of the spot disease of tobacco leaves.** *Verh Kon Akad Wetensch* 1898, **6**:3-21.
4. Zhang YZ, Shi M, Holmes EC: **Using Metagenomics to Characterize an Expanding Virosphere**. *Cell* 2018, **172**(6):1168-1172.
5. Fehr AR, Perlman S: **Coronaviruses: an overview of their replication and pathogenesis**. *Methods Mol Biol* 2015, **1282**:1-23.
6. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL *et al*: **A pneumonia outbreak associated with a new coronavirus of probable bat origin**. *Nature* 2020, **579**(7798):270-273.
7. Bailey AL, Lauck M, Sibley SD, Friedrich TC, Kuhn JH, Freimer NB, Jasinska AJ, Phillips-Conroy JE, Jolly CJ, Marx PA *et al*: **Zoonotic Potential of Simian Arteriviruses**. *J Virol* 2015, **90**(2):630-635.
8. Graham RL, Donaldson EF, Baric RS: **A decade after SARS: strategies for controlling emerging coronaviruses**. *Nat Rev Microbiol* 2013, **11**(12):836-848.
9. Joyce GF: **The antiquity of RNA-based evolution**. *Nature* 2002, **418**(6894):214-221.
10. Snijder EJ, Kikkert M, Fang Y: **Arterivirus molecular biology and pathogenesis**. *J Gen Virol* 2013, **94**(Pt 10):2141-2163.
11. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM: **Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis**. *Nucleic Acids Res* 1989, **17**(12):4847-4861.
12. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, 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**(5):991-1004.
13. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ: **Nidovirales: evolving the largest RNA virus genome**. *Virus Res* 2006, **117**(1):17-37.
14. Stenglein MD, Jacobson ER, Wozniak EJ, Wellehan JF, Kincaid A, Gordon M, Porter BF, Baumgartner W, Stahl S, Kelley K *et al*: **Ball python nidovirus: a candidate etiologic agent for severe respiratory disease in Python regius**. *MBio* 2014, **5**(5):e01484-01414.

15. den Boon JA, Snijder EJ, Chirnside ED, de Vries AA, Horzinek MC, Spaan WJ: **Equine arteritis virus is not a togavirus but belongs to the coronaviruslike superfamily**. *J Virol* 1991, **65**(6):2910-2920.

16. Cowley JA, Walker PJ: **The complete genome sequence of gill-associated virus of Penaeus monodon prawns indicates a gene organisation unique among nidoviruses**. *Arch Virol* 2002, **147**(10):1977-1987.

17. Lauber C, Ziebuhr J, Junglen S, Drosten C, Zirkel F, Nga PT, Morita K, Snijder EJ, Gorbalenya AE: **Mesoniviridae: a proposed new family in the order Nidovirales formed by a single species of mosquito-borne viruses**. *Arch Virol* 2012, **157**(8):1623-1628.

18. Adams MJ, Lefkowitz EJ, King AM, Carstens EB: **Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2014)**. *Arch Virol* 2014, **159**(10):2831-2841.

19. Pasternak AO, Spaan WJ, Snijder EJ: **Nidovirus transcription: how to make sense...?** *J Gen Virol* 2006, **87**(Pt 6):1403-1421.

20. Lauber C, Goeman JJ, Parquet MC, 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**(7):e1003500.

21. 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**(2):557-565.

22. Sagripanti JL, Zandomeni RO, Weinmann R: **The cap structure of simian hemorrhagic fever virion RNA**. *Virology* 1986, **151**(1):146-150.

23. van Vliet AL, Smits SL, Rottier PJ, de Groot RJ: **Discontinuous and non-discontinuous subgenomic RNA transcription in a nidovirus**. *EMBO J* 2002, **21**(23):6571-6580.

24. Lai MM, Brayton PR, Armen RC, Patton CD, Pugh C, Stohlman SA: **Mouse hepatitis virus A59: mRNA structure and genetic localization of the sequence divergence from hepatotropic strain MHV-3**. *J Virol* 1981, **39**(3):823-834.

25. Vanberlo MF, Horzinek MC, Vanderzeijst BAM: **Equine Arteritis Virus-Infected Cells Contain 6 Polyadenylated Virus-Specific Rnas**. *Virology* 1982, **118**(2):345-352.

26. Brierley I, Boursnell ME, Binns MM, Bilimoria B, Blok VC, Brown TD, Inglis SC: **An efficient ribosomal frame-shifting signal in the polymerase-encoding region of the coronavirus IBV**. *EMBO J* 1987, **6**(12):3779-3785.

27. Brierley I, Digard P, Inglis SC: **Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot**. *Cell* 1989, **57**(4):537-547.

28. Hussain S, Pan J, Chen Y, Yang Y, Xu J, Peng Y, Wu Y, Li Z, Zhu Y, Tien P *et al*: **Identification of novel subgenomic RNAs and noncanonical transcription initiation signals of severe acute respiratory syndrome coronavirus**. *J Virol* 2005, **79**(9):5288-5295.

29. Serrano P, Johnson MA, Almeida MS, Horst R, Herrmann T, Joseph JS, Neuman BW, Subramanian V, Saikatendu KS, Buchmeier MJ *et al*: **Nuclear magnetic resonance structure of the N-terminal domain of nonstructural protein 3 from the severe acute respiratory syndrome coronavirus**. *J Virol* 2007, **81**(21):12049-12060.

30. Ziebuhr J, Thiel V, Gorbalenya AE: **The autocatalytic release of a putative RNA virus transcription factor from its polyprotein precursor involves two paralogous papain-like proteases that cleave the same peptide bond**. *J Biol Chem* 2001, **276**(35):33220-33232.

31. Ziebuhr J, Snijder EJ, Gorbalenya AE: **Virus-encoded proteinases and proteolytic processing in the Nidovirales**. *J Gen Virol* 2000, **81**(Pt 4):853-879.

32. Plant EP, Dinman JD: **The role of programmed-1 ribosomal frameshifting in coronavirus propagation**. *Front Biosci* 2008, **13**:4873-4881.

33. van Hemert MJ, van den Worm SH, Knoops K, Mommaas AM, Gorbalenya AE, Snijder EJ: **SARS-coronavirus replication/transcription complexes are membrane-protected and need a host factor for activity in vitro**. *PLoS Pathog* 2008, **4**(5):e1000054.

34. Denison MR: **Seeking membranes: positive-strand RNA virus replication complexes**. *PLoS Biol* 2008, **6**(10):e270.

35. Cowley JA, Dimmock CM, Walker PJ: **Gill-associated nidovirus of Penaeus monodon prawns transcribes 3'-coterminal subgenomic mRNAs that do not possess 5'-leader sequences**. *J Gen Virol* 2002, **83**(Pt 4):927-935.

36. Jendrach M, Thiel V, Siddell S: **Characterization of an internal ribosome entry site within mRNA 5 of murine hepatitis virus**. *Arch Virol* 1999, **144**(5):921-933.

37. Schaecher SR, Mackenzie JM, Pekosz A: **The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles**. *J Virol* 2007, **81**(2):718-731.

38. Baric RS, Yount B: **Subgenomic negative-strand RNA function during mouse hepatitis virus infection**. *J Virol* 2000, **74**(9):4039-4046.

39. 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**(13):5108-5113.

40. Nedialkova DD, Gorbalenya AE, Snijder EJ: **Arterivirus Nsp1 modulates the accumulation of minus-strand templates to control the relative abundance of viral mRNAs**. *PLoS Pathog* 2010, **6**(2):e1000772.

41. Hogue BG, Machamer CE: **Coronavirus Structural Proteins and Virus Assembly**. In: *Nidoviruses.* Edited by Perlman S, Gallagher T, Snijder EJ. Washington, DC: ASM Press; 2008: 179-200.

42. Neuman BW, Chamberlain P, Bowden F, Joseph J: **Atlas of coronavirus replicase structure**. *Virus Res* 2014, **194**:49-66.

43. Thiel V, Weber F: **Interferon and cytokine responses to SARS-coronavirus infection**. *Cytokine Growth Factor Rev* 2008, **19**(2):121-132.

44. Perlman S, Netland J: **Coronaviruses post-SARS: update on replication and pathogenesis**. *Nat Rev Microbiol* 2009, **7**(6):439-450.

45. Mielech AM, Chen Y, Mesecar AD, Baker SC: **Nidovirus papain-like proteases: multifunctional enzymes with protease, deubiquitinating and deISGylating activities**. *Virus Res* 2014, **194**:184-190.

46. Butler JE, Lager KM, Golde W, Faaberg KS, Sinkora M, Loving C, Zhang YI: **Porcine reproductive and respiratory syndrome (PRRS): an immune dysregulatory pandemic**. *Immunol Res* 2014, **59**(1-3):81-108.

47. Brockway SM, Clay CT, Lu XT, Denison MR: **Characterization of the expression, intracellular localization, and replication complex association of the putative mouse hepatitis virus RNA-dependent RNA polymerase**. *J Virol* 2003, **77**(19):10515-10527.

48. Posthuma CC, Pedersen KW, Lu Z, Joosten RG, Roos N, Zevenhoven-Dobbe JC, Snijder EJ: **Formation of the arterivirus replication/transcription complex: a key role for nonstructural protein 3 in the remodeling of intracellular membranes**. *J Virol* 2008, **82**(9):4480-4491.

49. Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ: **Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles**. *MBio* 2013, **4**(4).

50. Neuman BW, Angelini MM, Buchmeier MJ: **Does form meet function in the coronavirus replicative organelle?** *Trends Microbiol* 2014, **22**(11):642-647.

51. Draker R, Roper RL, Petric M, Tellier R: **The complete sequence of the bovine torovirus genome**. *Virus Res* 2006, **115**(1):56-68.

52. Snijder EJ, Wassenaar AL, Spaan WJ: **The 5' end of the equine arteritis virus replicase gene encodes a papainlike cysteine protease**. *J Virol* 1992, **66**(12):7040-7048.

53. Nedialkova DD, Gorbalenya AE, Snijder EJ: **Arterivirus Papain-like Proteinase 1a**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2199-2204.

54. Nedialkova DD, Gorbalenya AE, Snijder EJ: **Arterivirus Papain-like Proteinase 1ß**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2205-2210.

55. Kikkert M, Snijder EJ, Gorbalenya AE: **Arterivirus nsp2 Cysteine Proteinase**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2210-2215.

56. Vatter HA, Di H, Donaldson EF, Radu GU, Maines TR, Brinton MA: **Functional analyses of the three simian hemorrhagic fever virus nonstructural protein 1 papain-like proteases**. *J Virol* 2014, **88**(16):9129-9140.

57. Sun Y, Xue F, Guo Y, Ma M, Hao N, Zhang XC, Lou Z, Li X, Rao Z: **Crystal structure of porcine reproductive and respiratory syndrome virus leader protease Nsp1alpha**. *J Virol* 2009, **83**(21):10931-10940.

58. Xue F, Sun Y, Yan L, Zhao C, Chen J, Bartlam M, Li X, Lou Z, Rao Z: **The crystal structure of porcine reproductive and respiratory syndrome virus nonstructural protein Nsp1beta reveals a novel metal-dependent nuclease**. *J Virol* 2010, **84**(13):6461-6471.

59. Ratia K, Mesecar A, O'Brien A, Baker SC: **Coronavirus Papain-like Peptidases**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2195-2199.

60. Gorbalenya AE, Lieutaud P, Harris MR, Coutard B, Canard B, Kleywegt GJ, Kravchenko AA, Samborskiy DV, Sidorov IA, Leontovich AM *et al*: **Practical application of bioinformatics by the multidisciplinary VIZIER consortium**. *Antiviral Res* 2010, **87**(2):95-110.

61. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. *Mol Biol Evol* 2015, **32**(1):268-274.

62. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences**. *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:175-182.

63. van Kasteren PB, Bailey-Elkin BA, James TW, Ninaber DK, Beugeling C, Khajehpour M, Snijder EJ, Mark BL, Kikkert M: **Deubiquitinase function of arterivirus papain-like protease 2 suppresses the innate immune response in infected host cells**. *Proc Natl Acad Sci U S A* 2013, **110**(9):E838-E847.

64. Bailey-Elkin BA, van Kasteren PB, Snijder EJ, Kikkert M, Mark BL: **Viral OTU deubiquitinases: a structural and functional comparison**. *PLoS Pathog* 2014, **10**(3):e1003894.

65. Herold J, Siddell SG, Gorbalenya AE: **A human RNA viral cysteine proteinase that depends upon a unique Zn2+-binding finger connecting the two domains of a papain-like fold**. *J Biol Chem* 1999, **274**(21):14918-14925.

66. Ratia K, Saikatendu KS, Santarsiero BD, Barretto N, Baker SC, Stevens RC, Mesecar AD: **Severe acute respiratory syndrome coronavirus papain-like protease: structure of a viral deubiquitinating enzyme**. *Proc Natl Acad Sci U S A* 2006, **103**(15):5717-5722.

67. Wojdyla JA, Manolaridis I, van Kasteren PB, Kikkert M, Snijder EJ, Gorbalenya AE, Tucker PA: **Papain-like protease 1 from transmissible gastroenteritis virus: crystal structure and enzymatic activity toward viral and cellular substrates**. *J Virol* 2010, **84**(19):10063-10073.

68. Tijms MA, van Dinten LC, Gorbalenya AE, Snijder EJ: **A zinc finger-containing papain-like protease couples subgenomic mRNA synthesis to genome translation in a positive-stranded RNA virus**. *Proc Natl Acad Sci U S A* 2001, **98**(4):1889-1894.

69. Tijms MA, Nedialkova DD, Zevenhoven-Dobbe JC, Gorbalenya AE, Snijder EJ: **Arterivirus subgenomic mRNA synthesis and virion biogenesis depend on the multifunctional nsp1 autoprotease**. *J Virol* 2007, **81**(19):10496-10505.

70. Kroese MV, Zevenhoven-Dobbe JC, Bos-de Ruijter JN, Peeters BP, Meulenberg JJ, Cornelissen LA, Snijder EJ: **The nsp1alpha and nsp1 papain-like autoproteinases are essential for porcine reproductive and respiratory syndrome virus RNA synthesis**. *J Gen Virol* 2008, **89**(Pt 2):494-499.

71. Lindner HA, Fotouhi-Ardakani N, Lytvyn V, Lachance P, Sulea T, Menard R: **The papain-like protease from the severe acute respiratory syndrome coronavirus is a deubiquitinating enzyme**. *J Virol* 2005, **79**(24):15199-15208.

72. Chen Z, Wang Y, Ratia K, Mesecar AD, Wilkinson KD, Baker SC: **Proteolytic processing and deubiquitinating activity of papain-like proteases of human coronavirus NL63**. *J Virol* 2007, **81**(11):6007-6018.

73. Frias-Staheli N, Giannakopoulos NV, Kikkert M, Taylor SL, Bridgen A, Paragas J, Richt JA, Rowland RR, Schmaljohn CS, Lenschow DJ *et al*: **Ovarian tumor domain-containing viral proteases evade ubiquitin- and ISG15-dependent innate immune responses**. *Cell Host Microbe* 2007, **2**(6):404-416.

74. Neuman BW, Joseph JS, Saikatendu KS, Serrano P, Chatterjee A, Johnson MA, Liao L, Klaus JP, Yates JR, 3rd, Wuthrich K *et al*: **Proteomics analysis unravels the functional repertoire of coronavirus nonstructural protein 3**. *J Virol* 2008, **82**(11):5279-5294.

75. Gorbalenya AE, Koonin EV, Lai MM: **Putative papain-related thiol proteases of positive-strand RNA viruses. Identification of rubi- and aphthovirus proteases and delineation of a novel conserved domain associated with proteases of rubi-, alpha- and coronaviruses**. *FEBS Lett* 1991, **288**(1-2):201-205.

76. Saikatendu KS, Joseph JS, Subramanian V, Clayton T, Griffith M, Moy K, Velasquez J, Neuman BW, Buchmeier MJ, Stevens RC *et al*: **Structural basis of severe acute respiratory syndrome coronavirus ADP-ribose-1''-phosphate dephosphorylation by a conserved domain of nsP3**. *Structure* 2005, **13**(11):1665-1675.

77. Putics A, Filipowicz W, Hall J, Gorbalenya AE, Ziebuhr J: **ADP-ribose-1"-monophosphatase: a conserved coronavirus enzyme that is dispensable for viral replication in tissue culture**. *J Virol* 2005, **79**(20):12721-12731.

78. Egloff MP, Malet H, Putics A, Heinonen M, Dutartre H, Frangeul A, Gruez A, Campanacci V, Cambillau C, Ziebuhr J *et al*: **Structural and functional basis for ADP-ribose and poly(ADP-ribose) binding by viral macro domains**. *J Virol* 2006, **80**(17):8493-8502.

79. Malet H, Coutard B, Jamal S, Dutartre H, Papageorgiou N, Neuvonen M, Ahola T, Forrester N, Gould EA, Lafitte D *et al*: **The crystal structures of Chikungunya and Venezuelan equine encephalitis virus nsP3 macro domains define a conserved adenosine binding pocket**. *J Virol* 2009, **83**(13):6534-6545.

80. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM: **A biochemical genomics approach for identifying genes by the activity of their products**. *Science* 1999, **286**(5442):1153-1155.

81. Nasr F, Filipowicz W: **Characterization of the Saccharomyces cerevisiae cyclic nucleotide phosphodiesterase involved in the metabolism of ADP-ribose 1",2"-cyclic phosphate**. *Nucleic Acids Res* 2000, **28**(8):1676-1683.

82. Chatterjee A, Johnson MA, Serrano P, Pedrini B, Joseph JS, Neuman BW, Saikatendu K, Buchmeier MJ, Kuhn P, Wuthrich K: **Nuclear magnetic resonance structure shows that the severe acute respiratory syndrome coronavirus-unique domain contains a macrodomain fold**. *J Virol* 2009, **83**(4):1823-1836.

83. Tan J, Vonrhein C, Smart OS, Bricogne G, Bollati M, Kusov Y, Hansen G, Mesters JR, Schmidt CL, Hilgenfeld R: **The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes**. *PLoS Pathog* 2009, **5**(5):e1000428.

84. Fang Y, Treffers EE, Li Y, Tas A, Sun Z, van der Meer Y, de Ru AH, van Veelen PA, Atkins JF, Snijder EJ *et al*: **Efficient -2 frameshifting by mammalian ribosomes to synthesize an additional arterivirus protein**. *Proc Natl Acad Sci U S A* 2012, **109**(43):E2920-E2928.

85. Gorbalenya AE, Donchenko AP, Blinov VM, Koonin EV: **Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases. A distinct protein superfamily with a common structural fold**. *FEBS Lett* 1989, **243**(2):103-114.

86. Ziebuhr J, Bayer S, Cowley JA, Gorbalenya AE: **The 3C-like proteinase of an invertebrate nidovirus links coronavirus and potyvirus homologs**. *J Virol* 2003, **77**(2):1415-1426.

87. Smits SL, Snijder EJ, de Groot RJ: **Characterization of a torovirus main proteinase**. *J Virol* 2006, **80**(8):4157-4167.

88. Ulferts R, Mettenleiter TC, Ziebuhr J: **Characterization of Bafinivirus main protease autoprocessing activities**. *J Virol* 2011, **85**(3):1348-1359.

89. Blanck S, Stinn A, Tsiklauri L, Zirkel F, Junglen S, Ziebuhr J: **Characterization of an alphamesonivirus 3C-like protease defines a special group of nidovirus main proteases**. *J Virol* 2014, **88**(23):13747-13758.

90. Barrette-Ng IH, Ng KK, Mark BL, Van Aken D, Cherney MM, Garen C, Kolodenko Y, Gorbalenya AE, Snijder EJ, James MN: **Structure of arterivirus nsp4. The smallest chymotrypsin-like proteinase with an alpha/beta C-terminal extension and alternate conformations of the oxyanion hole**. *J Biol Chem* 2002, **277**(42):39960-39966.

91. Anand K, Palm GJ, Mesters JR, Siddell SG, Ziebuhr J, Hilgenfeld R: **Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain**. *EMBO J* 2002, **21**(13):3213-3224.

92. Zirkel F, Kurth A, Quan PL, Briese T, Ellerbrok H, Pauli G, Leendertz FH, Lipkin WI, Ziebuhr J, Drosten C *et al*: **An insect nidovirus emerging from a primary tropical rainforest**. *MBio* 2011, **2**(3):e00077-00011.

93. Nga PT, Parquet MC, Lauber C, Parida M, Nabeshima T, Yu F, Thuy NT, Inoue S, Ito T, Okamoto K *et al*: **Discovery of the first insect nidovirus, a missing evolutionary link in the emergence of the largest RNA virus genomes**. *PLoS Pathog* 2011, **7**(9):e1002215.

94. Xiao Y, Ma Q, Restle T, Shang W, Svergun DI, Ponnusamy R, Sczakiel G, Hilgenfeld R: **Nonstructural proteins 7 and 8 of feline coronavirus form a 2:1 heterotrimer that exhibits primer-independent RNA polymerase activity**. *J Virol* 2012, **86**(8):4444-4454.

95. Zhai Y, Sun F, Li X, Pang H, Xu X, Bartlam M, Rao Z: **Insights into SARS-CoV transcription and replication from the structure of the nsp7-nsp8 hexadecamer**. *Nat Struct Mol Biol* 2005, **12**(11):980-986.

96. Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, Snijder EJ, Canard B, Imbert I: **One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities**. *Proc Natl Acad Sci U S A* 2014, **111**(37):E3900-3909.

97. Imbert I, Guillemot JC, Bourhis JM, Bussetta C, Coutard B, Egloff MP, Ferron F, Gorbalenya AE, Canard B: **A second, non-canonical RNA-dependent RNA polymerase in SARS coronavirus**. *EMBO J* 2006, **25**(20):4933-4942.

98. te Velthuis AJ, van den Worm SH, Snijder EJ: **The SARS-coronavirus nsp7+nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension**. *Nucleic Acids Res* 2012, **40**(4):1737-1747.

99. 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**(10):e1002294.

100. 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**(24):9372-9377.

101. Snijder EJ, den Boon JA, Horzinek MC, Spaan WJ: **Comparison of the genome organization of toro- and coronaviruses: evidence for two nonhomologous RNA recombination events during Berne virus evolution**. *Virology* 1991, **180**(1):448-452.

102. Mazumder R, Iyer LM, Vasudevan S, Aravind L: **Detection of novel members, structure-function analysis and evolutionary classification of the 2H phosphoesterase superfamily**. *Nucleic Acids Res* 2002, **30**(23):5229-5243.

103. Zhao L, Jha BK, Wu A, Elliott R, Ziebuhr J, Gorbalenya AE, Silverman RH, Weiss SR: **Antagonism of the interferon-induced OAS-RNase L pathway by murine coronavirus ns2 protein is required for virus replication and liver pathology**. *Cell Host Microbe* 2012, **11**(6):607-616.

104. Ahn DG, Choi JK, Taylor DR, Oh JW: **Biochemical characterization of a recombinant SARS coronavirus nsp12 RNA-dependent RNA polymerase capable of copying viral RNA templates**. *Arch Virol* 2012, **157**(11):2095-2104.

105. Gorbalenya AE, Pringle FM, Zeddam JL, Luke BT, Cameron CE, Kalmakoff J, Hanzlik TN, Gordon KH, Ward VK: **The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage**. *J Mol Biol* 2002, **324**(1):47-62.

106. Poch O, Sauvaget I, Delarue M, Tordo N: **Identification of four conserved motifs among the RNA-dependent polymerase encoding elements**. *EMBO J* 1989, **8**(12):3867-3874.

107. Deng Z, Lehmann KC, Li X, Feng C, Wang G, Zhang Q, Qi X, Yu L, Zhang X, Feng W *et al*: **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* 2014, **42**(5):3464-3477.

108. Seybert A, Hegyi A, Siddell SG, Ziebuhr J: **The human coronavirus 229E superfamily 1 helicase has RNA and DNA duplex-unwinding activities with 5'-to-3' polarity**. *RNA* 2000, **6**(7):1056-1068.

109. 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**(14):7833-7838.

110. Ivanov KA, Thiel V, Dobbe JC, van der Meer Y, Snijder EJ, Ziebuhr J: **Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase**. *J Virol* 2004, **78**(11):5619-5632.

111. Gorbalenya AE, Koonin EV: **Helicases - Amino-Acid-Sequence Comparisons and Structure-Function-Relationships**. *Curr Opin Struc Biol* 1993, **3**(3):419-429.

112. Kadlec J, Guilligay D, Ravelli RB, Cusack S: **Crystal structure of the UPF2-interacting domain of nonsense-mediated mRNA decay factor UPF1**. *RNA* 2006, **12**(10):1817-1824.

113. von Grotthuss M, Wyrwicz LS, Rychlewski L: **mRNA cap-1 methyltransferase in the SARS genome**. *Cell* 2003, **113**(6):701-702.

114. 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**(16):8071-8084.

115. Chen Y, Cai H, Pan J, Xiang N, Tien P, Ahola T, Guo D: **Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase**. *Proc Natl Acad Sci U S A* 2009, **106**(9):3484-3489.

116. Bouvet M, Debarnot C, Imbert I, Selisko B, Snijder EJ, Canard B, Decroly E: **In vitro reconstitution of SARS-coronavirus mRNA cap methylation**. *PLoS Pathog* 2010, **6**(4):e1000863.

117. Decroly E, Ferron F, Lescar J, Canard B: **Conventional and unconventional mechanisms for capping viral mRNA**. *Nat Rev Microbiol* 2011, **10**(1):51-65.

118. 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**(22):12135-12144.

119. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR: **Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics**. *PLoS Pathog* 2013, **9**(8):e1003565.

120. Bhardwaj K, Guarino L, Kao CC: **The severe acute respiratory syndrome coronavirus Nsp15 protein is an endoribonuclease that prefers manganese as a cofactor**. *J Virol* 2004, **78**(22):12218-12224.

121. Ivanov KA, Hertzig T, Rozanov M, Bayer S, Thiel V, Gorbalenya AE, Ziebuhr J: **Major genetic marker of nidoviruses encodes a replicative endoribonuclease**. *Proc Natl Acad Sci U S A* 2004, **101**(34):12694-12699.

122. Posthuma CC, Nedialkova DD, Zevenhoven-Dobbe JC, Blokhuis JH, Gorbalenya AE, Snijder EJ: **Site-directed mutagenesis of the Nidovirus replicative endoribonuclease NendoU exerts pleiotropic effects on the arterivirus life cycle**. *J Virol* 2006, **80**(4):1653-1661.

123. Nedialkova DD, Ulferts R, van den Born E, Lauber C, Gorbalenya AE, Ziebuhr J, Snijder EJ: **Biochemical characterization of arterivirus nonstructural protein 11 reveals the nidovirus-wide conservation of a replicative endoribonuclease**. *J Virol* 2009, **83**(11):5671-5682.

124. Frieman M, Ratia K, Johnston RE, Mesecar AD, Baric RS: **Severe acute respiratory syndrome coronavirus papain-like protease ubiquitin-like domain and catalytic domain regulate antagonism of IRF3 and NF-kappaB signaling**. *J Virol* 2009, **83**(13):6689-6705.

125. Laneve P, Altieri F, Fiori ME, Scaloni A, Bozzoni I, Caffarelli E: **Purification, cloning, and characterization of XendoU, a novel endoribonuclease involved in**

processing of intron-encoded small nucleolar RNAs in Xenopus laevis. *J Biol Chem* 2003, **278**(15):13026-13032.

126. Schwarz DS, Blower MD: **The calcium-dependent ribonuclease XendoU promotes ER network formation through local RNA degradation**. *J Cell Biol* 2014, **207**(1):41-57.

127. de Souza RF, Aravind L: **Identification of novel components of NAD-utilizing metabolic pathways and prediction of their biochemical functions**. *Mol Biosyst* 2012, **8**(6):1661-1677.

128. Schelle B, Karl N, Ludewig B, Siddell SG, Thiel V: **Selective replication of coronavirus genomes that express nucleocapsid protein**. *J Virol* 2005, **79**(11):6620-6630.

129. Bos EC, Luytjes W, van der Meulen HV, Koerten HK, Spaan WJ: **The production of recombinant infectious DI-particles of a murine coronavirus in the absence of helper virus**. *Virology* 1996, **218**(1):52-60.

130. Wilson L, McKinlay C, Gage P, Ewart G: **SARS coronavirus E protein forms cation-selective ion channels**. *Virology* 2004, **330**(1):322-331.

131. Kuo L, Masters PS: **The small envelope protein E is not essential for murine coronavirus replication**. *J Virol* 2003, **77**(8):4597-4608.

132. Yu IM, Oldham ML, Zhang J, Chen J: **Crystal structure of the severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein dimerization domain reveals evolutionary linkage between corona- and arteriviridae**. *J Biol Chem* 2006, **281**(25):17134-17139.

133. Zirkel F, Roth H, Kurth A, Drosten C, Ziebuhr J, Junglen S: **Identification and characterization of genetically divergent members of the newly established family Mesoniviridae**. *J Virol* 2013, **87**(11):6346-6358.

134. Veit M, Matczuk AK, Sinhadri BC, Krause E, Thaa B: **Membrane proteins of arterivirus particles: structure, topology, processing and function**. *Virus Res* 2014, **194**:16-36.

135. Cowley JA, Cadogan LC, Spann KM, Sittidilokratna N, Walker PJ: **The gene encoding the nucleocapsid protein of Gill-associated nidovirus of Penaeus monodon prawns is located upstream of the glycoprotein gene**. *J Virol* 2004, **78**(16):8935-8941.

136. Narayanan K, Huang C, Makino S: **SARS coronavirus accessory proteins**. *Virus Res* 2008, **133**(1):113-121.

137. Ito N, Mossel EC, Narayanan K, Popov VL, Huang C, Inoue T, Peters CJ, Makino S: **Severe acute respiratory syndrome coronavirus 3a protein is a viral structural protein**. *J Virol* 2005, **79**(5):3182-3186.

138. Liu DX, Fung TS, Chong KK, Shukla A, Hilgenfeld R: **Accessory proteins of SARS-CoV and other coronaviruses**. *Antiviral Res* 2014, **109**:97-109.

139. Luytjes W, Bredenbeek PJ, Noten AF, Horzinek MC, Spaan WJ: **Sequence of mouse hepatitis virus A59 mRNA 2: indications for RNA recombination between coronaviruses and influenza C virus**. *Virology* 1988, **166**(2):415-422.

140. de Groot RJ: **Structure, function and evolution of the hemagglutinin-esterase proteins of corona- and toroviruses**. *Glycoconj J* 2006, **23**(1-2):59-72.

141. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S, Graham RL, Baric RS, Stockwell TB *et al*: **Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing**. *PLoS Pathog* 2010, **6**(5):e1000896.

142. Kirkegaard K, Baltimore D: **The mechanism of RNA recombination in poliovirus**. *Cell* 1986, **47**(3):433-443.

143. Lai MM: **RNA recombination in animal and plant viruses**. *Microbiol Rev* 1992, **56**(1):61-79.

144. Simon-Loriere E, Holmes EC: **Why do RNA viruses recombine?** *Nat Rev Microbiol* 2011, **9**(8):617-626.

145. Romanova LI, Blinov VM, Tolskaya EA, Viktorova EG, Kolesnikova MS, Guseva EA, Agol VI: **The primary structure of crossover regions of intertypic poliovirus recombinants: a model of recombination between RNA genomes**. *Virology* 1986, **155**(1):202-213.

146. Baird HA, Galetto R, Gao Y, Simon-Loriere E, Abreha M, Archer J, Fan J, Robertson DL, Arts EJ, Negroni M: **Sequence determinants of breakpoint location during HIV-1 intersubtype recombination**. *Nucleic Acids Res* 2006, **34**(18):5203-5216.

147. Lowry K, Woodman A, Cook J, Evans DJ: **Recombination in enteroviruses is a biphasic replicative process involving the generation of greater-than genome length 'imprecise' intermediates**. *PLoS Pathog* 2014, **10**(6):e1004191.

148. Gmyl AP, Belousov EV, Maslova SV, Khitrina EV, Chetverin AB, Agol VI: **Nonreplicative RNA recombination in poliovirus**. *J Virol* 1999, **73**(11):8958-8965.

149. Lee HJ, Shieh CK, Gorbalenya AE, Koonin EV, La Monica N, Tuler J, Bagdzhadzhyan A, Lai MM: **The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative proteases and RNA polymerase**. *Virology* 1991, **180**(2):567-582.

150. Woo PC, Lau SK, Chu CM, Chan KH, Tsoi HW, Huang Y, Wong BH, Poon RW, Cai JJ, Luk WK *et al*: **Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia**. *J Virol* 2005, **79**(2):884-895.

151. Woo PC, Lau SK, Yip CC, Huang Y, Tsoi HW, Chan KH, Yuen KY: **Comparative analysis of 22 coronavirus HKU1 genomes reveals a novel genotype and evidence of natural recombination in coronavirus HKU1**. *J Virol* 2006, **80**(14):7136-7145.

152. Sutton G, Fry E, Carter L, Sainsbury S, Walter T, Nettleship J, Berrow N, Owens R, Gilbert R, Davidson A *et al*: **The nsp9 replicase protein of SARS-coronavirus, structure and functional insights**. *Structure* 2004, **12**(2):341-353.

153. Godeny EK, de Vries AA, Wang XC, Smith SL, de Groot RJ: **Identification of the leader-body junctions for the viral subgenomic mRNAs and organization of the simian hemorrhagic fever virus genome: evidence for gene duplication during arterivirus evolution**. *J Virol* 1998, **72**(1):862-867.

154. Boto L: **Horizontal gene transfer in evolution: facts and challenges**. *Proc Biol Sci* 2010, **277**(1683):819-827.

155. Zeng Q, Langereis MA, van Vliet AL, Huizinga EG, de Groot RJ: **Structure of coronavirus hemagglutinin-esterase offers insight into corona and influenza virus evolution**. *Proc Natl Acad Sci U S A* 2008, **105**(26):9065-9069.

156. Tengs T, Kristoffersen AB, Bachvaroff TR, Jonassen CM: **A mobile genetic element with unknown function found in distantly related viruses**. *Virol J* 2013, **10**:132.

157. Mihindukulasuriya KA, Wu G, St Leger J, Nordhausen RW, Wang D: **Identification of a novel coronavirus from a beluga whale by using a panviral microarray**. *J Virol* 2008, **82**(10):5084-5088.

158. Gorbalenya AE, Koonin EV: **Comparative analysis of amino-acid sequences of key enzymes of replication and expression of positive-strand RNA viruses: validity of approach and functional and evolutionary implications**. *Sov Sci Rev D Physicochem Biol* 1993, **11**:1-84.

159. Zust R, Cervantes-Barragan L, Habjan M, Maier R, Neuman BW, Ziebuhr J, Szretter KJ, Baker SC, Barchet W, Diamond MS *et al*: **Ribose 2'-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5**. *Nat Immunol* 2011, **12**(2):137-143.

160. Lehmann KC, Snijder EJ, Posthuma CC, Gorbalenya AE: **What we know but do not understand about nidovirus helicases**. *Virus Res* 2015, **202**:12-32.

161. Sniegowski PD, Gerrish PJ, Johnson T, Shaver A: **The evolution of mutation rates: separating causes from consequences**. *Bioessays* 2000, **22**(12):1057-1066.

162. Lynch M: **Evolution of the mutation rate**. *Trends Genet* 2010, **26**(8):345-352.

163. Eigen M: **Selforganization of matter and the evolution of biological macromolecules**. *Naturwissenschaften* 1971, **58**(10):465-523.

164. Eigen M: **Error catastrophe and antiviral strategy**. *Proc Natl Acad Sci U S A* 2002, **99**(21):13374-13376.

165. Holmes EC: **The Evolution and Emergence of RNA Viruses.** New York: Oxford University Press; 2009.

166. Boursnell ME, Brown TD, Foulds IJ, Green PF, Tomley FM, Binns MM: **Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus**. *J Gen Virol* 1987, **68 (Pt 1)**:57-77.

167. Hodgman TC: **A new superfamily of replicative proteins**. *Nature* 1988, **333**(6168):22-23.

168. Gorbalenya AE, Koonin EV: **Birnavirus RNA polymerase is related to polymerases of positive strand RNA viruses**. *Nucleic Acids Res* 1988, **16**(15):7735.

169. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM: **A novel superfamily of nucleoside triphosphate-binding motif containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination**. *FEBS Lett* 1988, **235**(1-2):16-24.

170. Gorbalenya AE: **Big nidovirus genome. When count and order of domains matter**. *Adv Exp Med Biol* 2001, **494**:1-17.

171. den Boon JA, Faaberg KS, Meulenberg JJ, Wassenaar AL, Plagemann PG, Gorbalenya AE, Snijder EJ: **Processing and evolution of the N-terminal region of the arterivirus replicase ORF1a protein: identification of two papainlike cysteine proteases**. *J Virol* 1995, **69**(7):4500-4505.

172. Snijder EJ, Wassenaar AL, Spaan WJ, Gorbalenya AE: **The arterivirus Nsp2 protease. An unusual cysteine protease with primary structure similarities to both papain-like and chymotrypsin-like proteases**. *J Biol Chem* 1995, **270**(28):16671-16676.

173. Sulea T, Lindner HA, Purisima EO, Menard R: **Deubiquitination, a new function of the severe acute respiratory syndrome coronavirus papain-like protease?** *J Virol* 2005, **79**(7):4550-4551.

174. Rost B: **Twilight zone of protein sequence alignments**. *Protein Eng* 1999, **12**(2):85-94.

175. Habermann BH: **Oh Brother, Where Art Thou? Finding Orthologs in the Twilight and Midnight Zones of Sequence Similarity**. In: *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods.* Edited by Pontarotti P. Cham: Springer International Publishing; 2016: 393-419.

176. Grigoriev A: **Mutational patterns correlate with genome organization in SARS and other coronaviruses**. *Trends Genet* 2004, **20**(3):131-135.

177. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat Methods* 2011, **8**(10):785-786.

178. Duckert P, Brunak S, Blom N: **Prediction of proprotein convertase cleavage sites**. *Protein Eng Des Sel* 2004, **17**(1):107-112.

179. Theis C, Reeder J, Giegerich R: **KnotInFrame: prediction of -1 ribosomal frameshift events**. *Nucleic Acids Res* 2008, **36**(18):6013-6020.

180. Higgins D, Lemey P: **Multiple sequence alignment**. In: *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.* Edited by Lemey P, Salemi M, Vandamme AM, 2 edn. Cambridge , UK: Cambridge University Press; 2009: 68-108.

181. Koonin EV, Galperin MY: **Sequence - Evolution - Function: Computational Approaches in Comparative Genomics**. In. Boston, MA: Springer; 2003.

182. Bottu G, Van Ranst M, Lemey P: **Sequence databases and database searching**. In: *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.* Edited by Lemey P, Salemi M, Vandamme AM, 2 edn. Cambridge , UK: Cambridge University Press; 2009: 33-67.

183. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

184. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

185. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.

186. Söding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**(7):951-960.

187. Söding J, Remmert M, Hauser A: **User Guide 2.0.15: HH-suite for sensitive protein sequence searching based on HMM-HMM alignment**. In: *HH-suite.* 2012.

188. Yang Z, Rannala B: **Molecular phylogenetics: principles and practice**. *Nat Rev Genet* 2012, **13**(5):303-314.

189. Schmidt HA, von Haeseler A: **Phylogenetic inference using maximum likelihood methods**. In: *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.* Edited by Lemey P, Salemi M, Vandamme AM, 2 edn. Cambridge , UK: Cambridge University Press; 2009: 181-209.

190. Ronquist F, van der Mark P, Huelsenbeck JP: **Bayesian phylogenetic analysis using MrBayes**. In: *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.* Edited by Lemey P, Salemi M, Vandamme AM, 2 edn. Cambridge , UK: Cambridge University Press; 2009: 210-266.

191. Pagel M, Meade A, Barker D: **Bayesian estimation of ancestral character states on phylogenies**. *Syst Biol* 2004, **53**(5):673-684.

192. Bao Y, Chetvernin V, Tatusova T: **PAirwise Sequence Comparison (PASC) and its application in the classification of filoviruses**. *Viruses* 2012, **4**(8):1318-1327.

193. Lauber C, Gorbalenya AE: **Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses**. *J Virol* 2012, **86**(7):3890-3904.

194. Muhire BM, Varsani A, Martin DP: **SDT: a virus classification tool based on pairwise sequence alignment and identity calculation**. *PLoS One* 2014, **9**(9):e108277.

CHAPTER 2

Domain organization and evolution of the highly divergent 5' coding region of genomes of arteriviruses, including the novel possum nidovirus

Anastasia A. Gulyaeva[#]
Magdalena Dunowska[#]
Erik Hoogendoorn
Julia Giles
Dmitry V. Samborskiy
Alexander E. Gorbalenya

[#]equal contribution

Journal of Virology (2017) 91(6):e02096-16
DOI: 10.1128/JVI.02096-16

ABSTRACT

In five experimentally characterized arterivirus species, the 5'-end genome coding region encodes the most divergent nonstructural proteins (nsp's), nsp1 and nsp2, which include papain-like proteases (PLPs) and other poorly characterized domains. These are involved in regulation of transcription, polyprotein processing, and virus-host interaction. Here we present results of a bioinformatics analysis of this region of 14 arterivirus species, including that of the most distantly related virus, wobbly possum disease virus (WPDV), determined by a modified 5' rapid amplification of cDNA ends (RACE) protocol. By combining profile-profile comparisons and phylogeny reconstruction, we identified an association of the four distinct domain layouts of nsp1-nsp2 with major phylogenetic lineages, implicating domain gain, including duplication, and loss in the early nsp1 evolution. Specifically, WPDV encodes highly divergent homologs of PLP1a, PLP1b, PLP1c, and PLP2, with PLP1a lacking the catalytic Cys residue, but does not encode nsp1 Zn finger (ZnF) and "nuclease" domains, which are conserved in other arteriviruses. Unexpectedly, our analysis revealed that the only catalytically active nsp1 PLP of equine arteritis virus (EAV), known as PLP1b, is most similar to PLP1c and thus is likely to be a PLP1b paralog. In all non-WPDV arteriviruses, PLP1b/c and PLP1a show contrasting patterns of conservation, with the N- and C-terminal subdomains, respectively, being enriched with conserved residues, which is indicative of different functional specializations. The least conserved domain of nsp2, the hypervariable region (HVR), has its size varied 5-fold and includes up to four copies of a novel PxPxPR motif that is potentially recognized by SH3 domain-containing proteins. Apparently, only EAV lacks the signal that directs −2 ribosomal frameshifting in the nsp2 coding region.

IMPORTANCE

Arteriviruses comprise a family of mammalian enveloped positive-strand RNA viruses that include some of the most economically important pathogens of swine. Most of our knowledge about this family has been obtained through characterization of viruses from five species: *Equine arteritis virus*, *Simian hemorrhagic fever virus*, *Lactate dehydrogenase-elevating virus*, *Porcine respiratory and reproductive syndrome virus 1*, and *Porcine respiratory and reproductive syndrome virus 2*. Here we present the results of comparative genomics analyses of viruses from all known 14 arterivirus species, including the most distantly related virus, WPDV, whose genome sequence was completed in this study. Our analysis focused on the multifunctional 5'-end genome coding region that encodes multidomain nonstructural proteins 1 and 2. Using diverse bioinformatics techniques, we identified many patterns of evolutionary conservation that are specific to members of distinct arterivirus species, both characterized and novel, or their groups. They are likely associated with structural and functional determinants important for virus replication and virus-host interaction.

INTRODUCTION

Arteriviruses are a family of enveloped nonsegmented positive-strand RNA viruses of mammals that belongs to the order *Nidovirales* [1, 2]. The arterivirus genetic diversity was recently classified into 14 species [3], five of which include relatively well-characterized viruses, i.e., equine arteritis virus (EAV) [4, 5], lactate dehydrogenase-elevating virus (LDV) [6, 7], simian hemorrhagic fever virus (SHFV) [8], porcine respiratory and reproductive syndrome virus 1 (PRRSV-1) [9], and PRRSV-2 [10]. Among the newly identified viruses is wobbly possum disease virus (WPDV), a marsupial virus that is most distantly related to the current members of the family *Arteriviridae* [11, 12]. Infection with WPDV has been linked to a fatal neurological disease of the Australian brushtail possum (*Trichosurus vulpecula*) [13]. The disease has been identified in both captive [14] and free-living [15] possum populations in New Zealand. It is currently unknown if the virus is present in other parts of the world. Experimental research on arteriviruses was driven by the need to develop robust control measures against PRRSV infection, which causes considerable losses to the swine industry [16], and aimed to reveal molecular mechanisms of replication and virus-host interactions of this family, which are often characterized using the EAV model. Comparative sequence analyses involving arteriviruses contributed to these goals by informing experimental research about natural constraints imposed on the structure and function of different genome regions and encoded products [5, 7, 17-25].

The arterivirus genome includes multiple open reading frames (ORFs), most of which overlap and are flanked by short noncoding regions at the 5′ and 3′ termini. Protein machineries controlling genome expression and replication are encoded in the first two ORFs, ORF1a and ORF1b, while capsid proteins controlling virus dissemination are encoded in the downstream ORFs, whose number varies among arteriviruses. ORF1a directs the synthesis of replicase polyprotein 1a (pp1a) and, together with ORF1b, also pp1ab. The latter involves a −1 programmed ribosomal frameshift (PRF) at the ORF1a/b overlap. pp1a and pp1ab are co- and posttranslationally processed by viral proteases to mature products (and their intermediates) that are designated nonstructural proteins (nsp's) 1 to 12 [1]. The release of the nsp1-to-nsp3 and nsp3-to-nsp12 proteins from pp1a/ab is mediated by papain-like proteases (PLPs) and chymotrypsin-like protease (3CLpro), respectively (nsp3 release is controlled by two proteases) [20]. No domain variation was reported for the nsp3-to-nsp12 region, with the size of ORF1b-encoded nsp9 to nsp12 being found to be under particularly strong constraint [21]. According to the characterization of mostly EAV and PRRSVs, these proteins include four RNA processing enzymes (residing in nsp9 to nsp11), diverse enigmatic cofactors (nsp6 to -8 and nsp12) of replication, 3CLpro (nsp4), and two transmembrane domains, TM2 and TM3, anchoring the replication-transcription complex (RTC) (nsp3 and nsp5) [1, 22, 23, 26].

In contrast, the domain organization and size of the nsp1-nsp2 region vary considerably among the five characterized arterivirus species. For other arteriviruses with fully sequenced genomes, this region has not been studied, while for the most distantly related virus, WPDV, the respective region of the genome was not available [11]. nsp2 is one of the two largest arterivirus proteins. Its size varies >2-fold, from 572 amino acids (aa) (EAV) to 1,232 aa (PRRSV-2). It invariably includes a multifunctional Zn-binding PLP2 domain at the N terminus and adjacent transmembrane (TM1) and Cys-rich (CR) domains at the C terminus [1, 27]. These conserved domains are separated by a poorly conserved domain known as the hypervariable region (HVR) [1], which is notable for its high content of proline in PRRSV [28, 29], while another poorly conserved domain of unknown function (hinge) is found upstream of PLP2 in some arteriviruses [27]. PLP2 mediates the processing of the nsp2-nsp3 junction and removes ubiquitin and ubiquitin-like moieties from target proteins, thus regulating both genome expression and virus-host interaction [30]. The HVR contains antigenic sites [31], and the TM1 domain seems to contribute to anchoring the RTC, but the function of the CR domain remains totally obscure [1]. In addition, EAV nsp2 is a cofactor essential for cleavage of the nsp4-nsp5 junction by 3CLpro [32]. Recently, two truncated and modified derivatives of nsp2 of unknown function, and possibly different localizations, were identified in PRRSVs [33, 34]. Their generation involves −1 and −2 PRFs in the nsp2 genome region encoding the HVR-TM1 junction and is directed by a slippery sequence and a downstream C-rich element conserved in several arteriviruses but not EAV [33, 34]. PRFs are transactivated by a complex of viral nsp1b and host poly(C) binding protein (PCBP) that binds to the downstream C-rich element [34, 35].

The scale of variation in the nsp1 coding region is even larger, since it may encode either one (nsp1; EAV), two (nsp1a and nsp1b; LDV and PRRSV-1 and -2), or three (nsp1a, nsp1b, and nsp1c; SHFV) proteins. Each of these nsp1 variants includes an enzymatically active PLP domain that liberates the respective nsp from pp1a/ab by cleavage at the C terminus and, for SHFV PLP1c, possibly also the N terminus [17, 36-41]. The name of each PLP includes a suffix matching that in the name of the nsp in which it resides. The only exception is EAV nsp1, which uniquely includes an enzymatically silent PLP (PLP1a) upstream of the active PLP, accordingly named PLP1b [36, 37]. The variable number of adjacent PLPs present in the nsp1 region is likely to have emerged by duplication, implying that they are paralogs. Yet similarity between paralogous PLPs is (extremely) low at both the primary and tertiary structure (available for PLP1a and PLP1b of PRRSV-2 [18, 42]) levels, which suggests that this region has been under diversifying positive selection and/or that considerable time has passed since the duplication. nsp1/nsp1a of the characterized arteriviruses also includes an N-terminal Zn finger (ZnF) domain [37]. All three domains of nsp1 are important for template-specific complex regulation of subgenomic mRNA production (transcription) and virion biogenesis in EAV [19, 43, 44];

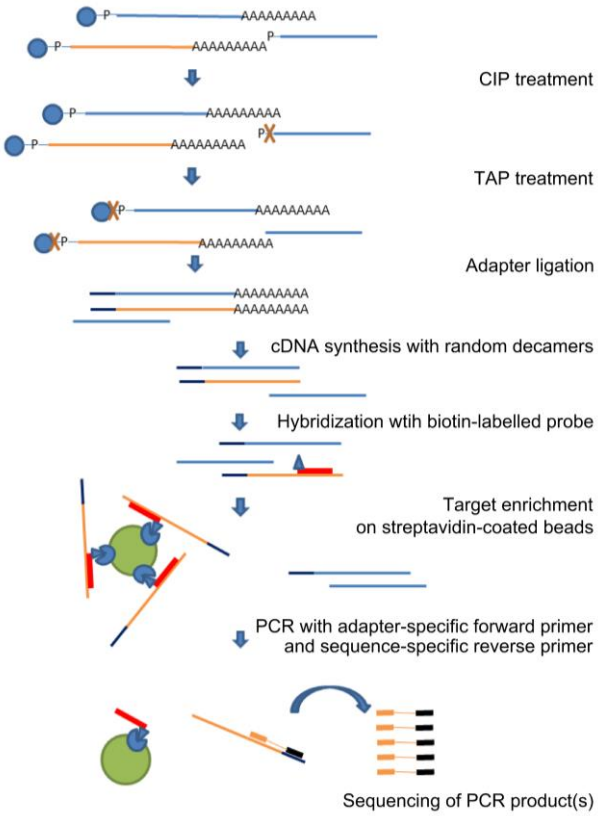


Figure 1 | Overview of the target-enriched 5′ RLM RACE protocol. Total RNA extracted from WPD-affected tissues was treated with calf intestine alkaline phosphatase (CIP) to remove free 5′ phosphates from all noncapped nucleic acids, followed by treatment with tobacco acid pyrophosphatase (TAP) to remove the cap structure from full-length mRNA (including capped positive-sense viral RNA), ligation of the RACE adapter to decapped mRNA containing 5′ phosphates, and reverse transcription of the ligated mRNA to cDNA by use of random decamers. These steps were performed according to the manufacturer's instructions (RLM RACE; Invitrogen). The ligated cDNA was then hybridized to biotinylated virus-specific probes. The viral sequences captured on streptavidin-coated magnetic beads were used in the PCR step of the 5′ RLM RACE protocol. The unknown 5′ end was amplified with a selection of virus-specific reverse primers and adapter-specific RACE primers. The target (WPDV) and nontarget (host) nucleic acids are depicted in orange and blue, respectively.

nsp1a of PRRSV-1 seems to play a similar role in transcriptional regulation [45]. Analysis of tertiary structure and biochemical characterization of nsp1b revealed a small domain with weak nuclease activity residing upstream of PLP1b in PRRSV-2 [42] (hereinafter called nuclease domain). This protein was also shown to facilitate −1 and −2 PRFs in the nsp2 genome region of PRRSV-1 and -2 [34]. All individual nsp1 subunits of PRRSV-2, LDV, SHFV, and EAV were found to have anti-innate-immunity activities [46, 47].

In the present study, we sought to gain insight into the structure and function of the nsp1-nsp2 genomic region by characterizing the evolution of this region in all known

arteriviruses, including seven newly identified arteriviruses of monkeys and one arterivirus of the forest giant pouched rat [12, 48-52]. To increase the resolution of this analysis, we also extended it to the most distantly related arterivirus, WPDV, whose sequence in this region is reported for the first time, thus completing its full genome sequence. We were interested in mapping the already identified domains to the nsp1-nsp2 region for arteriviruses that have not been characterized in this respect, reconstructing the evolutionary history of PLP duplications and other nsp1 domains, identifying molecular markers of PLP paralogs, and searching for new sites under constraint. Below we describe the obtained results along with the challenges of analysis of distant relations, summarize our conclusions, and outline directions for future studies.

RESULTS

Completion of genome sequencing of WPDV by a modified RACE method

The available WPDV sequence, obtained using a classic 5' rapid amplification of cDNA ends (RACE) protocol, comprised 10,087 nucleotides (nt), which included the published partial sequence of 9,509 nt [excluding the poly(A) tail] [11]. However, analysis of this sequence suggested that it may still lack the 5'-terminal region which encodes nsp1-nsp2 in other arteriviruses. Several attempts to further extend this sequence by use of a commercially available kit utilizing a modification of the classic approach to 5' RACE (5' RNA ligase-mediated [RLM] RACE) according to the manufacturer's instructions were unsuccessful. To address this challenge, we modified the 5' RLM RACE protocol by addition of a target enrichment step (Figure 1). Three rounds of RACE reactions (RACE 1–3; see Materials and Methods) were performed using different capture probes and different target-specific primers (Table S1). Using the modified protocol, three bands of different sizes (~0.4 kbp, 1.5 kbp, and 2.5 kbp) were obtained by primary PCR with the RACE.outer/WPD.S5.R primer pair (RACE 1) (Figure 2). Each band was reamplified separately with the RACE.inner primer in combination with either the WPD.S5.R, WPD.S7.R, or WPD.S8.R primer. Sequencing of the nested bands indicated that the 0.4-kbp band represented nonspecific amplification, while the 1.5-kbp and 2.5-kbp bands assembled into one sequence, which extended the available sequence of WPDV by another 2,006 nt. The longest PCR product obtained in RACE 2 extended the RACE 1 assembly by 808 nt. Small bands of about 250 bp were amplified following seminested PCRs with the RACE.outer/inner and WPD.S16.R primer pairs and either HOT FIREPol or Kappa LongRange enzyme mix (RACE 3). Both bands contained the same sequence, which aligned with the existing WPDV sequence but did not extend it. The RACE adapter was ligated at nt 343 of the WPDV sequence obtained in RACE 2.

Based on our inability to amplify any further sequences in the 5′-end direction by using a variety of primers and amplification conditions, combined with the bioinformatics analysis of the newly identified sequence (see below), we concluded that we most likely amplified the full genomic sequence of WPDV. If so, then the length of the WPDV genome is 12,901 nt, excluding the poly(A) tail, and the length of predicted polyprotein 1ab (pp1ab) of WPDV is 3,402 aa, whose coding sequence is flanked by a 245-nt 5′ untranslated region (5′-UTR) and a 97-nt 3′-UTR (Table 1).

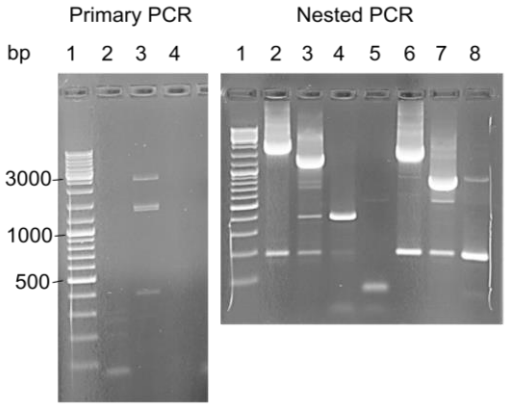


Figure 2 | Example of the results obtained using modified 5′ RLM RACE as described in the text. (Left) One of the primary PCRs (lane 3) using the RACE.outer/WPD.S5.R primer pair (see Table S1 in the supplemental material) and target-enriched cDNA captured on streptavidin-coated magnetic beads as the template produced three bands, with approximate sizes of 2,500 bp (band 1), 1,500 bp (band 2), and 400 bp (band 3), with no bands in the no-template control (lane 4). Lane 2 represents an unsuccessful 5′ RLM RACE reaction with a different source of starting material. (Right) Nested PCR with the RACE.inner primer and either the WPD.S5.R (lanes 2 to 5) or WPD.S7.R (lanes 6 to 8) reverse primer. DNA extracted from primary band 1 (lanes 2 and 6), band 2 (lanes 3 and 7), band 3 (lanes 4 and 8), or water (lane 5) was used as the template. No bands were visible in the no-template control with the RLM-RACE inner/WPD.S7.R primer pair (not shown in the picture). A DNA ladder (GeneRuler DNA ladder mix; Fermentas) was included in lane 1 of both gels.

Table 1 | Predicted ORFs in the genome of WPDV and their corresponding protein products.

Name	Start position (nt)	Stop position (nt)	Length (nt)	Protein product	Protein size (aa)
ORF1a	246	6,242	5,997	Polyprotein 1a	1,998
ORF1b	6,218	10,453	4,236	Polyprotein 1ab[a]	3,402
ORF2	10,309	10,932	624	Glycoprotein 2 (GP2)	207
ORF2a	10,553	10,690	138	Envelope protein (E)	45
ORF3	10,794	11,447	654	Glycoprotein 3 (GP3)	217
ORF4	11,330	11,581	252	Glycoprotein 4 (GP4)	83
ORF5	11,578	12,114	537	Glycoprotein 5 (GP5)	178
ORF5a	11,603	11,839	237	Glycoprotein 5a (GP5a)	78
ORF6	12,051	12,593	543	Membrane protein (M)	180
ORF7	12,424	12,804	381	Nucleocapsid protein (N)	126

[a]Polyprotein 1ab is predicted to be expressed from ORF1a and -b via a −1 ribosomal frameshift.

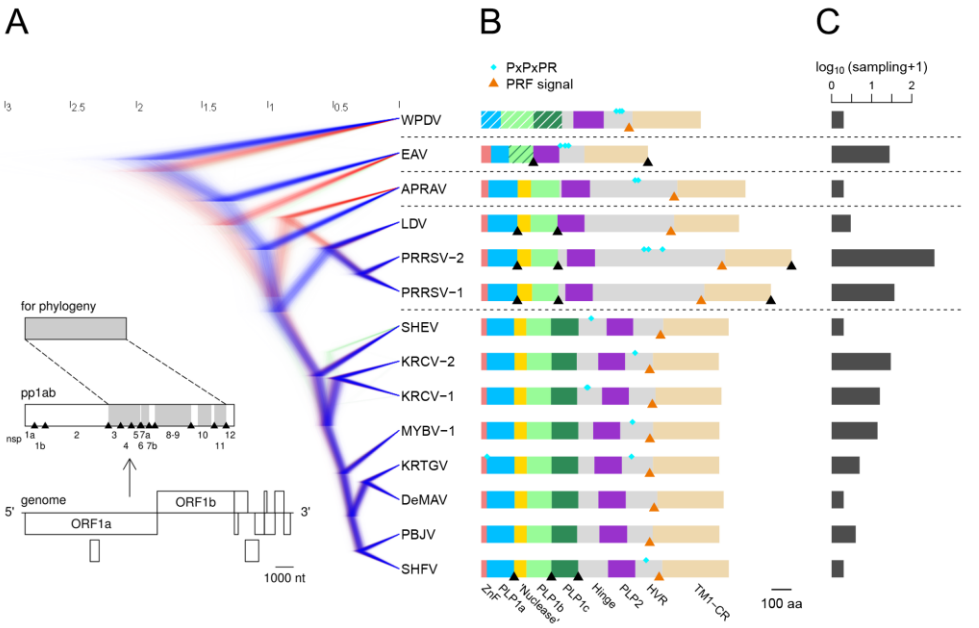


Figure 3 | Phylogeny and nsp1-nsp2 domain organization of arteriviruses. (**A**) The phylogeny is presented by a posterior sample of phylogenetic trees, reconstructed by BEAST software. The trees are colored blue, red, or green, in descending order of prevalent topology. The genome organization, polyprotein processing scheme, and polyprotein domains used for phylogeny reconstruction (shaded in gray) are detailed in the bottom left corner for PRRSV-2 (accession number NC_001961.1). (**B**) The domain organization of nsp1-nsp2 is shown for each arterivirus species. Protein domains are represented by colored bars. The bar representing PLP1b of EAV has dark green stripes to emphasize its affinity with PLP1c. Bars representing the PLP1 domains of WPDV have white stripes to show their weak sequence similarity with the PLP1 domains of other arteriviruses. The positions of nsp2 PRF-related motifs are indicated by orange triangles, those of experimentally established cleavage sites by black triangles, and those of PxPxPR motifs by cyan diamonds. (**C**) Number of genomes sequenced for each of the characterized species (with sampling size and bias).

Phylogeny of arteriviruses

To facilitate analysis of the nsp1-nsp2 genomic region, we reconstructed the phylogeny of arteriviruses by using multiple-sequence alignment (MSA) of seven conserved nsp domains (see Materials and Methods; see Table S2 in the supplemental material) of 14 viruses, including WPDV (Figure 3A; also see below). The total size of the analyzed MSA was 2,055 columns, which accounted for 43% of the pp1ab sites (bottom panel in Figure 3A). The viruses represented all species defined by DEmARC (Table 2) (see Materials and Methods), whose sampling varied by 2 orders of magnitude (Figure 3C). Based on the tree topology and the distance-based results of DEmARC, we recognized five clades (Table 2), including three represented by single virus species (WPDV, EAV, and African pouched rat arterivirus [APRAV]), one represented by eight virus species (named the simian clade), and one represented by three virus species (named the LDV-PRRSV clade). A large Bayesian sample of rooted trees revealed well-resolved branches for all viruses except APRAV and,

to a lesser extent, simian hemorrhagic encephalitis virus (SHEV) (Figure 3A). APRAV formed the basal branch to either the LDV-PRRSV (less favored; 33.66% of trees) or LDV-PRRSV and simian (most favored; 66.30% of trees) clades. Also, in a small fraction of trees (4.43%), SHEV formed a basal branch to the Kibale red colobus virus 1 (KRCV-1) and KRCV-2 clade, while in the majority of trees (95.50%) SHEV was basal to all simian arteriviruses other than SHEV.

Table 2 | Representative set of arteriviruses whose genome sequences were used in the present study.

Acronym[a]	Virus name	Accession no.	Cluster name[b]
PRRSV-2	Porcine reproductive and respiratory syndrome virus 2	EU624117.1	LDV-PRRSV
PRRSV-1	Porcine reproductive and respiratory syndrome virus 1	DQ489311.1	LDV-PRRSV
LDV	Lactate dehydrogenase-elevating virus	U15146.1	LDV-PRRSV
KRCV-2	Kibale red colobus virus 2	KC787658.1	Simian
PBJV	Pebjah virus	KR139839.1	Simian
SHFV	Simian hemorrhagic fever virus	AF180391.1	Simian
DeMAV	DeBrazza's monkey arterivirus	KP126831.1	Simian
KRTGV	Kibale red-tailed guenon virus 1	JX473849.1	Simian
KRCV-1	Kibale red colobus virus 1	KC787630.1	Simian
SHEV	Simian hemorrhagic encephalitis virus	KM677927.1	Simian
MYBV-1	Mikumi yellow baboon virus 1	KM110938.1	Simian
EAV	Equine arteritis virus	X53459.3	EAV
APRAV	African pouched rat arterivirus	KP026921.1	APRAV
WPDV	Wobbly possum disease virus	JN116253.3	WPDV

[a]Sequences of all full-length arterivirus genomes retrieved from GenBank and RefSeq, as well as that of the full-length WPDV genome, were grouped into 14 species by DEmARC. One sequence from each species was selected as a representative for further analysis.
[b]Deliniated species were further grouped into five clusters.

Domain organization of the nsp1-nsp2 genomic region of arteriviruses

We then analyzed the domain organization of the nsp1-nsp2 genomic region in different arteriviruses. Using arterivirus-wide MSA, we found that all poorly characterized viruses of the simian clade adopted the domain organization described for SHFV (Figure 3B; Table S3). The sequence affinity of viruses of the LDV-PRRSV clade in this region, except for the highly divergent hinge and HVR domains, was also previously documented and confirmed by our analysis. Quality MSAs of the nsp1-nsp2 region for both the simian and LDV-PRRSV clades and three other single-species clades were converted into the respective HMM profiles that were used for all-versus-all profile-profile comparisons by HHalign. The most informative comparisons were profile comparisons of the most diverse simian clade with itself and other clades, whose results are visualized as two-dimensional plots in Figure 4. Based on the number of high-scoring domains and the level of confidence (measured by both probability and E values), sequence affinity between the simian clade and other

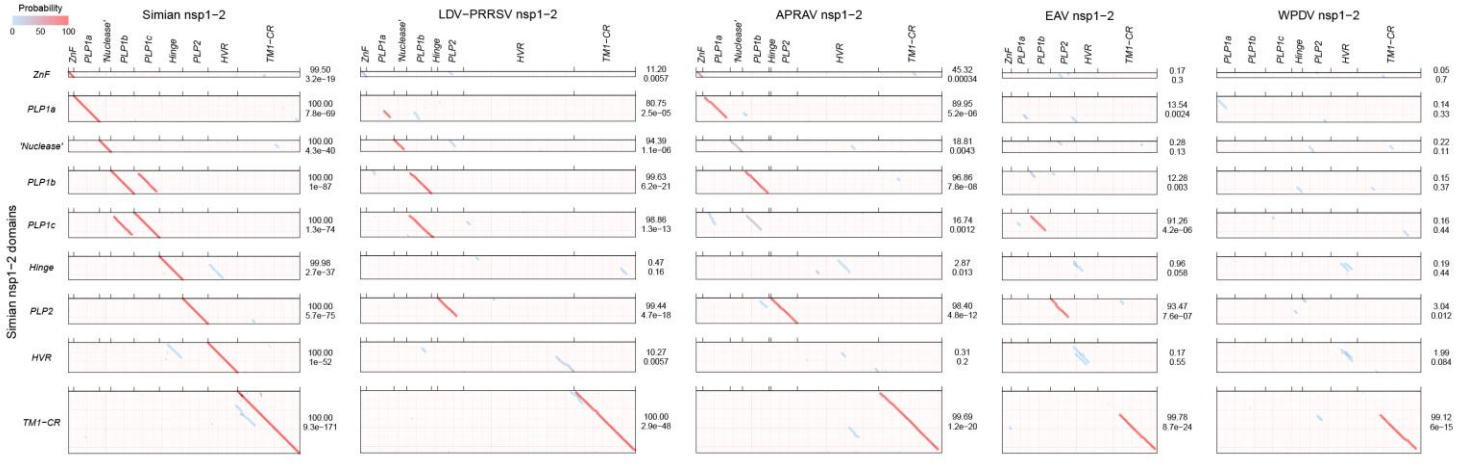


Figure 4 | Profile-profile comparisons of nsp1-nsp2 domains of the simian lineage and five other arterivirus lineages. The plots shown are HHalign dot plots, with domains and viruses indicated on the respective axes and alignment paths of the two top-scoring hits drawn with transparent lines. The color of each line indicates the probability of the hit. On the right side of each dot plot, the probability and E value of the top-scoring hit are depicted.

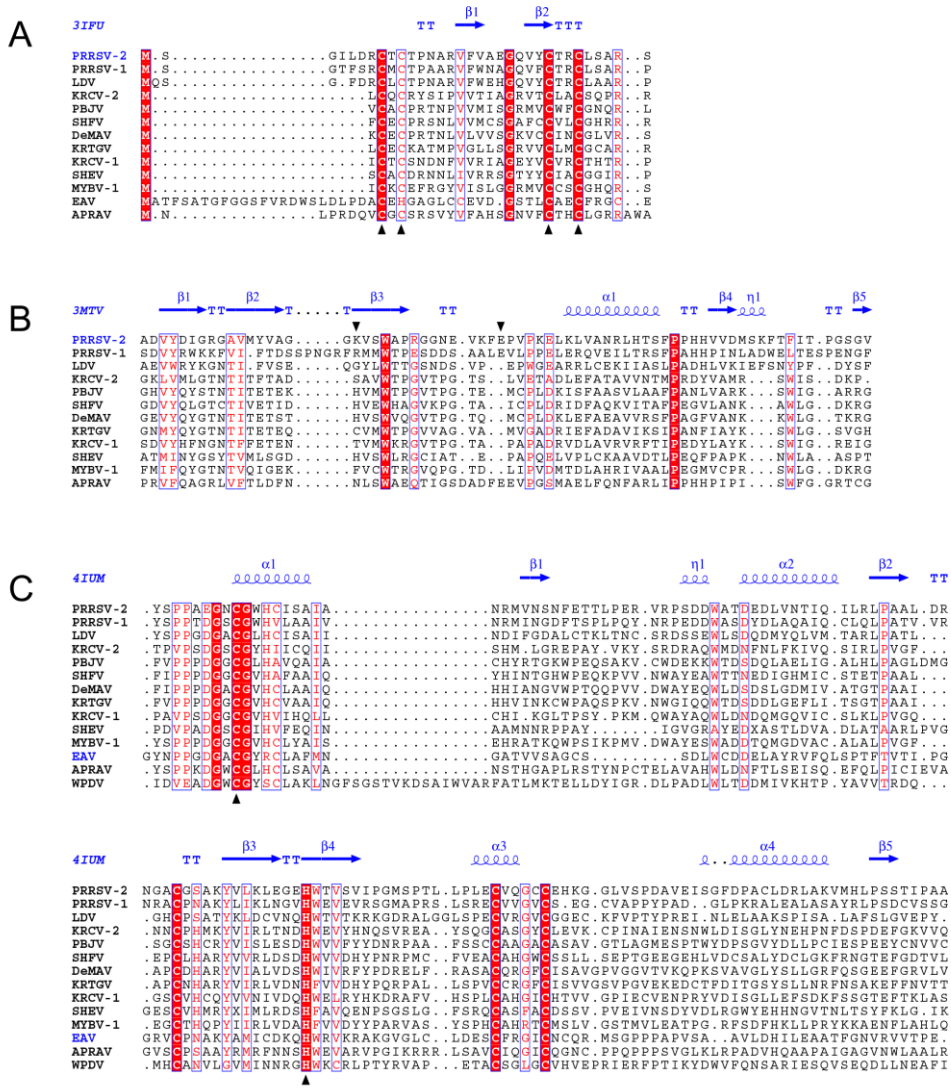


Figure 5 | Multiple-sequence alignments of selected nsp1-nsp2 domains of arteriviruses. (**A**) MSA of ZnF domains. Zinc-binding residues are marked with black triangles. (**B**) MSA of "nuclease" domains. Columns of the MSA that contain PRRSV-2 nsp1b residues whose mutation to alanine led to abolishment of PRRSV-2 nsp1b nuclease activity [42] are marked with black triangles. (**C**) MSA of PLP2 domains. Catalytic residues are marked with black triangles. MSAs were visualized with the help of Espript 2.1 [53]. Secondary structures were derived from PDB entries.

clades was ranked in the descending order LDV-PRRSV > APRAV > EAV > WPDV. Notably, this ranking was in agreement with the phylogenetic relationships of the clades (Figure 3A). This analysis also enabled nsp1-nsp2 domain assignment for APRAV (Figure 3B; Table S3; see below). At the domain level, support was obtained for the conservation of ZnF and

"nuclease" domains in all viruses except for EAV and WPDV, and for PLP2 of all non-WPDV arteriviruses (Figure 4 and 5). Since EAV also encodes the ZnF domain [37], this result showed that the conducted profile-profile comparisons were not sufficiently sensitive to reveal the most remote relationships. We then inspected residues conserved in the respective MSAs of the ZnF and nuclease domains (Figure 5A and B). As expected, four Zn-binding residues were conserved in the ZnF MSA of non-WPDV arteriviruses (Figure 5A). In contrast, the Lys and Glu residues implicated in the nuclease activity of PPRSV-2 [42] were found to be among the least conserved residues in the MSA of the nuclease domain (Figure 5B). Accordingly, the "nuclease" domain included only two residues (Trp and Pro) that were invariant in arteriviruses, further indicating that this domain is unlikely to have any enzymatic activity that can broadly be conserved in arteriviruses; its name is thus retained purely for historical reasons.

Relationships between paralogous and orthologous PLPs of all non-WPDV arteriviruses

Before proceeding to analyze WPDV further, we first clarified the relationships between paralogous and orthologous PLPs by using profile-profile plots. In agreement with prior observations, no significant similarity between PLP1a and either PLP1b or PLP1c was found for PLPs of any origin. In contrast, similarity between PLP1b and PLP1c variants of different origins was significant, although it varied considerably depending on the pair (Figure 6A). The most significant similarity was that between PLP1b variants of the simian and LDV-PRRSV clades, which was supported much more strongly than the next most significant hit between either of these PLP1b variants and simian PLP1c (3.9e−26 versus 1.1e−15). Likewise, PLP1b of APRAV showed much higher sequence similarity to PLP1b of the LDV-PRRSV or simian clade than to PLP1c of the simian cluster (8.2e−11 and 6.1e−09 versus 0.0001). In contrast, the enzymatically active PLP of EAV (known as PLP1b) was most similar to simian PLP1c rather than to PLP1b variants of different origins (7.28e−08 versus 0.00091 to 0.00013). In all comparisons of PLP1b and PLP1c, the result depended on the inclusion of EAV PLP1b: almost entire domains were similar without EAV PLP1b being involved, while the similarity was limited to the N-terminal half of the domain when EAV PLP1b was compared. To extend these observations further, MSA of combined PLP1b/PLP1c was used to infer the Bayesian phylogeny of these domains (Figure 6B). While the obtained sample of rooted trees lacked a prevalent topology, PLPs were clearly partitioned into two major PLP1b- and PLP1c-based clades according to the sequence affinities revealed in the profile analysis. For each clade, considerable uncertainty of the branching was observed for most viruses, likely due to the extremely large scale of divergence of the entire tree (more than four times that of the nsp-based tree of non-WPDV arteriviruses) (compare Figure 6B with Figure 3A), confounded by the small size of the PLP1 domains. The only notable exception to the domain-clade association was EAV

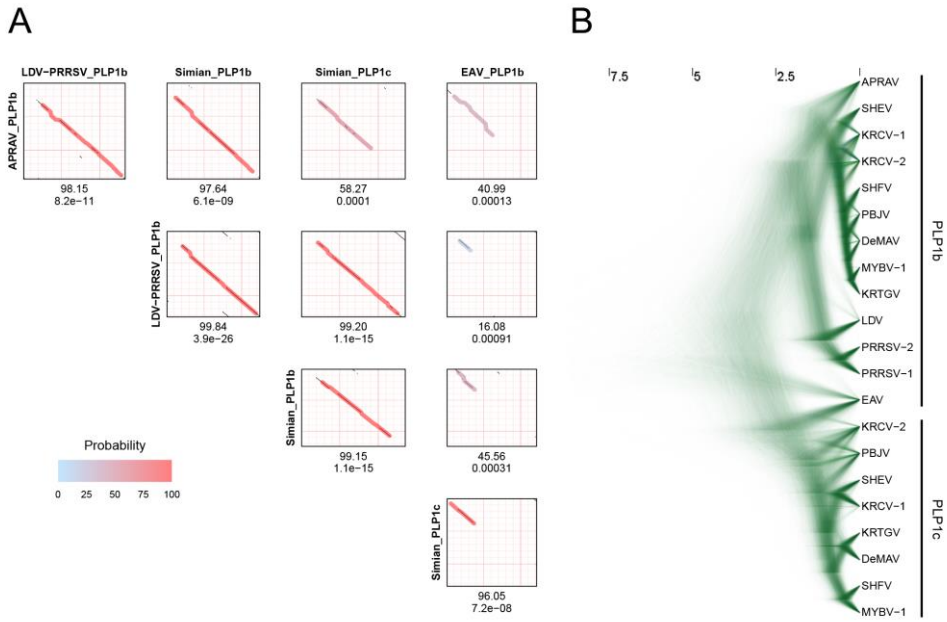


Figure 6 | Sequence similarity and evolutionary relationships of PLP1b and PLP1c. (**A**) HHalign comparisons between PLP1b and PLP1c domains of different arteriviruses. For each comparison, a dot plot is shown. On the dot plot, the alignment path of the top-scoring hit is drawn with a transparent line. The color of the line indicates the probability of the hit. Below the dot plot, the probability and E value of the top-scoring hit are given. (**B**) Posterior sample of phylogenetic trees generated by BEAST, based on MSA of PLP1b and PLP1c. For other designations, see the legend to Figure 3A.

PLP1b, which was basal to either PLP1c (70.26% of trees) or PLP1b and PLP1c (23.36% of trees), and this was also sustained in the comparable tree including WPDV (Figure S1; see below). These results combined strongly suggested an orthologous relationship between PLP1b variants of the simian, LDV-PRRSV, and APRAV clades but not that of EAV, which is most likely either an ortholog of PLP1c enzymes or a direct descendant of the ancestral enzyme for PLP1b and PLP1c (see Discussion).

Domain organization of the nsp1-nsp2 region in WPDV

To improve the limited resolution of the domains mapping in the nsp1-nsp2 region of WPDV by profile-profile comparison (Figure 4), we combined four clade-specific MSAs of domains of this region. These MSAs were then compared with the N-terminal 1,096 aa of WPDV pp1a/pp1ab in the profile-profile mode by using HHalign (Figure 7). Significant similarities were observed for the PLP2 (8.4e−06) and TM1-CR (1.7e−23) domains, which were much stronger than those observed using simian-based profiles only (Figure 4), facilitating mapping of these domains in the WPDV polyprotein. In line with the considerable divergence of WPDV, its PLP2 domain included a 16-aa insertion between

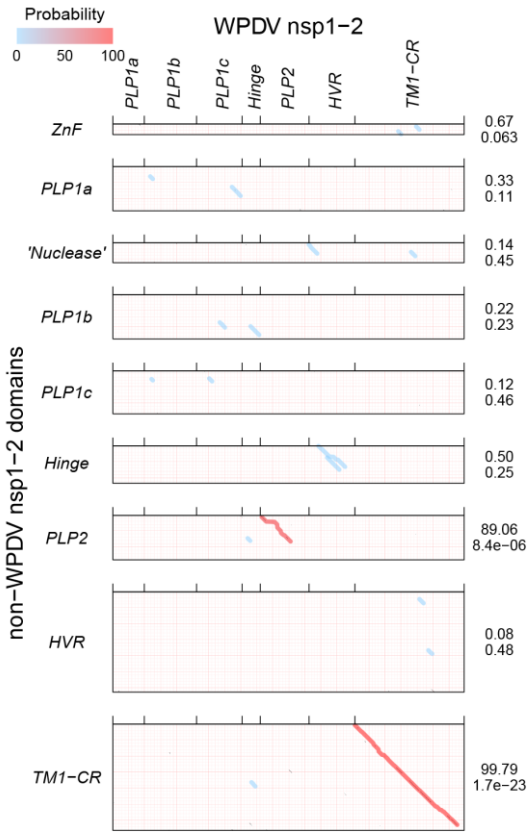


Figure 7 | HHalign profile-profile comparisons of nsp1-nsp2 domains of WPDV and non-WPDV arteriviruses. EAV PLP1b was regarded as PLP1c for this figure. For details, see the legend to Figure 4.

catalytic Cys and His residues in the otherwise uniformly compact PLP2 sequences of different origins (Figure 5C). Although no other domain showed statistically significant similarity, the sizes of regions upstream of PLP2 and between the TM1-CR and PLP2 domains in the WPDV pp1ab protein were sufficiently large to accommodate other canonical domains.

To learn whether WPDV could indeed encode highly divergent homologs of arterivirus PLP1, we scanned WPDV pp1ab with HMM profiles of short regions around the catalytic cysteine and histidine residues of nsp1 PLPs of other arteriviruses by using HHalign; since enzymatically silent PLP1a of EAV lacks the catalytic cysteine, it was not included in the corresponding HMM profile. Hit probability distributions (Figure 8) revealed that two top-scoring hits for the cysteine motif had considerably higher probabilities (1.62% and 0.47%, respectively) than those of other hits (≤0.04% and ≤0.06% for Cys and His motifs, respectively), indicating that they may be genuine. These top-scoring hits were mapped

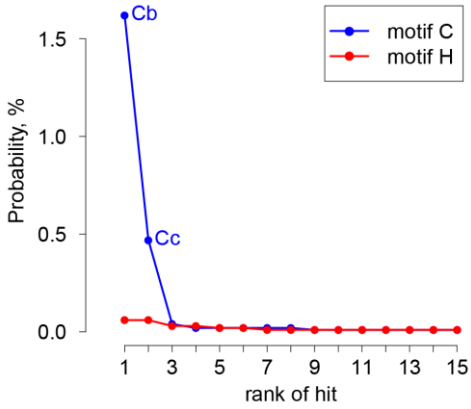


Figure 8 | Rank distribution of top HHalign hits between PLP1 active site motifs of arteriviruses and WPDV pp1ab. HMM profiles representing cysteine and histidine motifs of PLP1s of all non-WPDV arterivirus species, with the EAV PLP1a cysteine motif excluded, were compared with WPDV pp1ab. The 15 top hits were ranked in descending order of probability (indicated on the y axis). Hits potentially including the catalytic cysteines of WPDV PLP1b and PLP1c are designated Cb and Cc, respectively.

upstream of the putative PLP2 domain, to aa 121 to 125 and 301 to 311 of the WPDV polyprotein, positions compatible with belonging to two PLP1 varieties. Accordingly, these hits included CysTrp and CysTyr dipeptides, respectively, which either matched or closely resembled the CysTrp dipeptide with a catalytic Cys residue of PLP1b and PLP1c, besides conservation at other, less prominent positions (Figure 9 and 10A). These observations were used to guide MSAs between WPDV and arteriviruses for PLP1a, PLP1b, and PLP1c, including putative catalytic His residues (Figure 9), and to delineate the hinge domain in WPDV (Figure 3B; Table S3). Like its EAV counterpart, the delineated PLP1a domain of WPDV lacks the catalytic cysteine and is expected to be proteolytically silent. However, like PLP1a enzymes of all arteriviruses, it did include the most characteristic HXXXXXF motif (Figure 10A), which is the core of the Ha conservation peak in Figure 10B. Secondary structure predictions (Figure 9) and the modest impact of the WPDV inclusion in the respective MSAs on their mean conservation (Figure 10B) further supported the identification of these most divergent PLP1s (Figure S1). No ZnF or nuclease domains were evident in WPDV.

N- and C-terminal subdomains of PLP1 are enriched with sites that are conserved in paralogous PLP1b/c and PLP1a, respectively

Sequence similarity between PLP1a and PLP1b/PLP1c is limited to very few residues (Figure 9). This profound divergence was also evident upon comparison of the resolved crystal structures of PRRSV-2 nsp1a and nsp1b (Protein Data Bank [PDB] entries 3IFU [18] and 3MTV [42]) by use of DALI [54], which revealed the similarity between PLP1a and PLP1b to be below the Z-score cutoff (Z-scores of 4.2 and 3.6 and root mean square

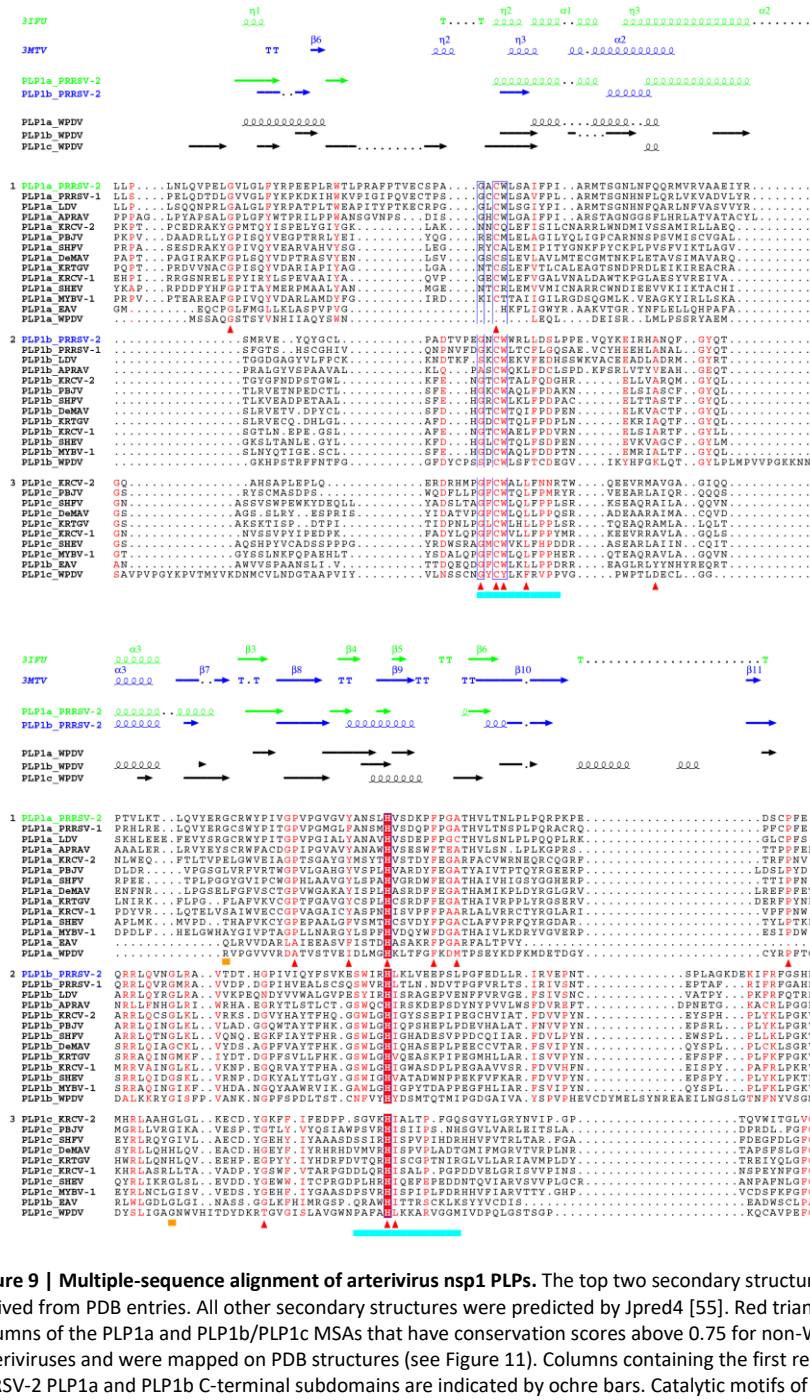


Figure 9 | Multiple-sequence alignment of arterivirus nsp1 PLPs. The top two secondary structures were derived from PDB entries. All other secondary structures were predicted by Jpred4 [55]. Red triangles indicate columns of the PLP1a and PLP1b/PLP1c MSAs that have conservation scores above 0.75 for non-WPDV arteriviruses and were mapped on PDB structures (see Figure 11). Columns containing the first residues of the PRRSV-2 PLP1a and PLP1b C-terminal subdomains are indicated by ochre bars. Catalytic motifs of nsp1 PLPs are underlined in cyan. The MSAs were visualized with Espript 2.1 [53].

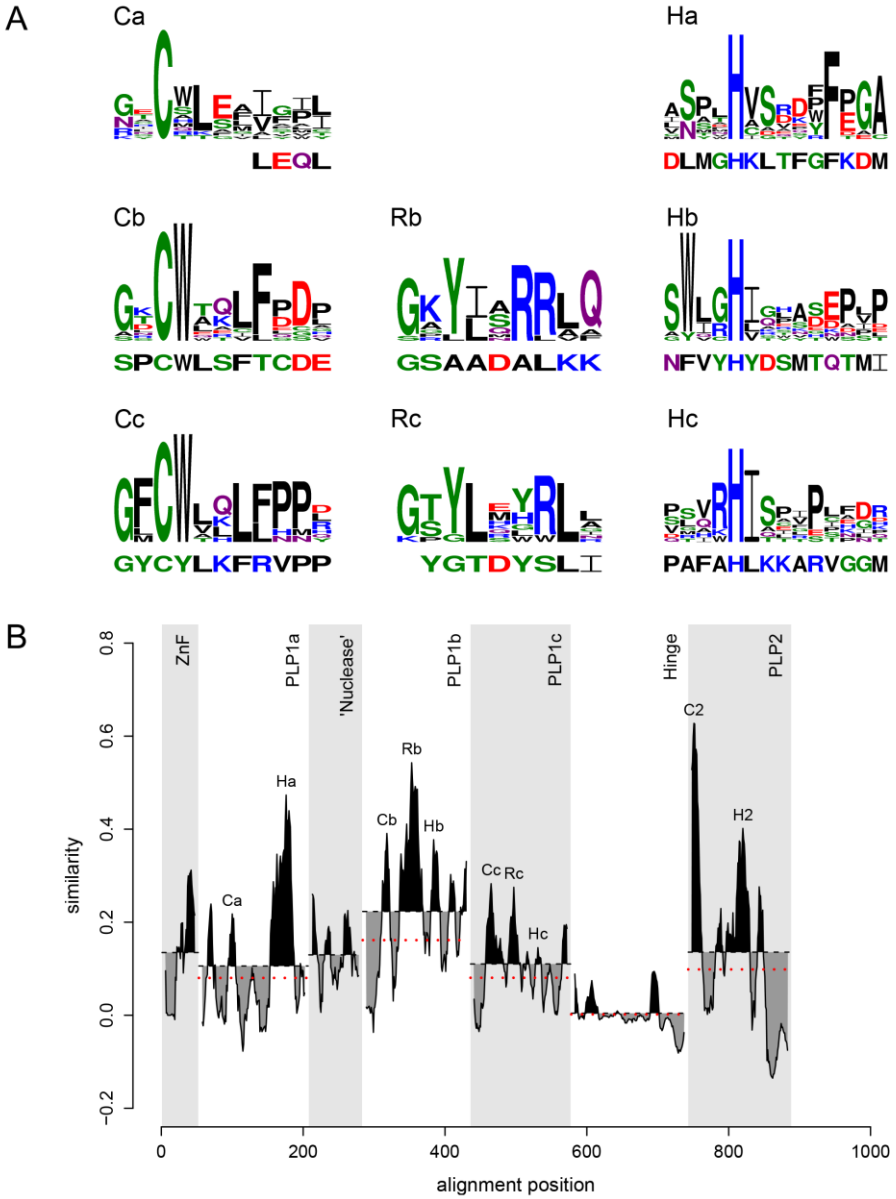


Figure 10 | Distribution of sequence conservation in the N-terminal region of pp1ab of arteriviruses. (**A**) MSAs of nsp1 PLP motifs of all non-WPDV arteriviruses are depicted as logos, with the homologous WPDV sequence specified below each logo. PLP motifs, including the catalytic residues Cys (C) and His (H) and putative RNA-binding residues (R), are labeled with domain-specific suffixes. Logos were prepared with the R package RWebLogo 1.0.3 [56]. (**B**) The conservation profile, calculated based on the MSA of sequences from non-WPDV clusters, is shown for each domain of nsp1 and the N-terminal domains of nsp2. Areas above and below the mean conservation lines are shaded in black and gray, respectively. Dotted red lines indicate the mean conservation of the domains after the addition of the WPDV sequence to the MSA. EAV PLP1b was regarded as PLP1c for this figure.

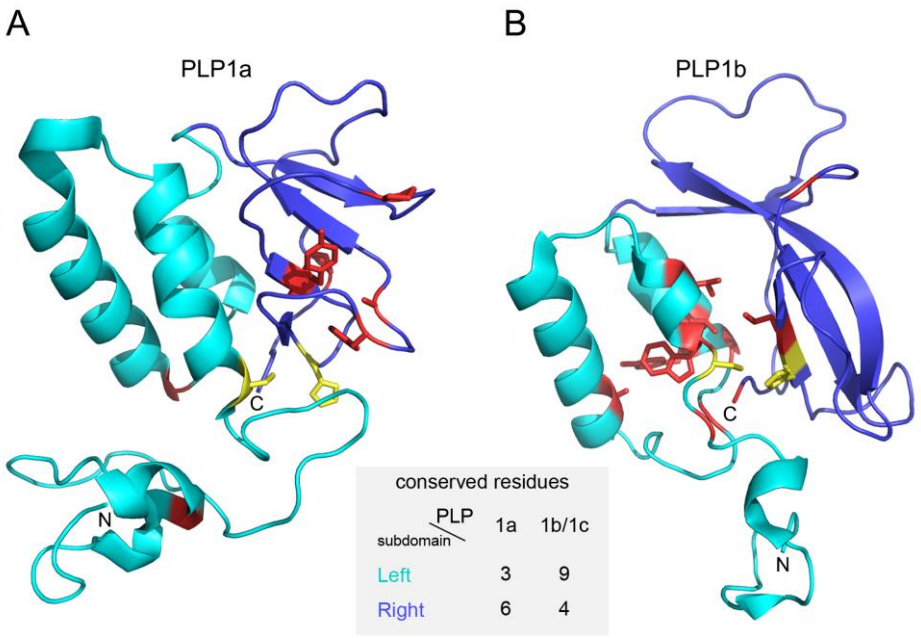


Figure 11 | Subdomain-specific distribution of residues conserved in PLP1a and PLP1b/c. The structures shown are tertiary structures of PRRSV-2 PLP1a (**A**) and PLP1b (**B**) with residues conserved in all non-WPDV arteriviral PLP1a and PLP1b/PLP1c domains, respectively. The N-terminal subdomain, formed by α-helices, is shown in cyan; and the C-terminal subdomain, consisting of antiparallel β-strands, is shown in blue. Conserved residues are shown in yellow (catalytic dyad) and red (all the rest). The following residues were conserved in the PLP1a alignment and mapped on PRRSV-2 (accession number EU624117.1) nsp1a: left subdomain, Gly45, Cys76, and Gly109; and right subdomain, Pro134, Tyr141, His146, Phe152, Ala155, and Pro175. The following residues were conserved in the PLP1b/c alignment and mapped on PRRSV-2 (accession number EU624117.1) nsp1b: left subdomain, Gly88, Cys90, Trp91, Leu94, Ala110, Gly120, Gly123, Tyr125, and Leu126; and right subdomain, Gly143, His159, Leu160, and Gly203. The figure was prepared with PyMOL [57].

deviations [RMSD] of 4.0 and 4.6 with PLP1b and PLP1a as queries, respectively) (see Materials and Methods). To gain insight into the selection that drove the divergence of these enzymes, we mapped residues conserved in PLP1a and PLP1b/c of non-WPDV arteriviruses on the structure of PRRSV-2 PLP1a (Figure 11A) and PLP1b (Figure 11B), respectively. Six of 9 residues conserved in PLP1a were found in the right subdomain of the papain fold, while 9 of 13 residues conserved in PLP1b/c were located in the left subdomain of the papain fold. This contrasting pattern suggests that the divergence of PLP1a and PLP1b/c has been constrained and/or promoted in a subdomain-specific fashion, which thus explains the exceptionally low similarity between these paralogs.

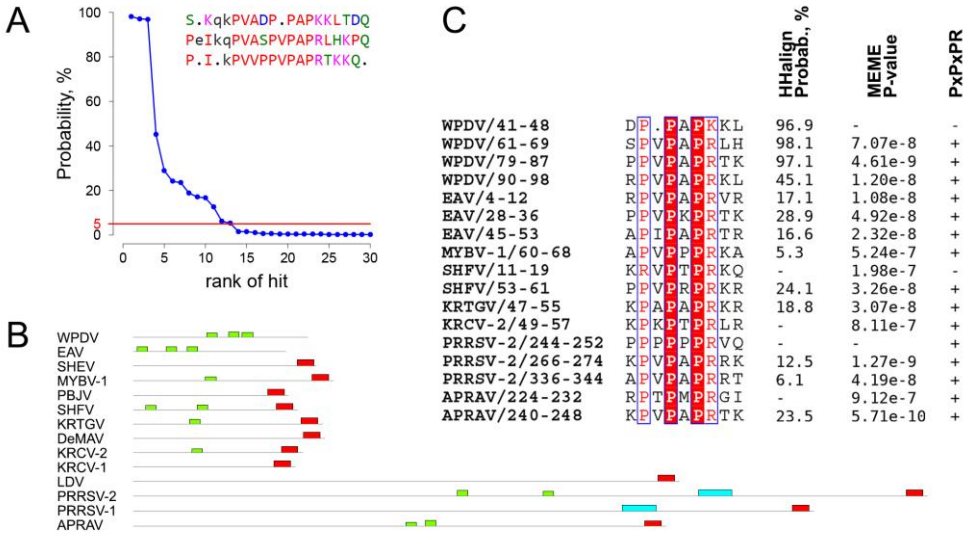


Figure 12 | Conservation of PxPxPR motifs in the HVR of arteriviruses. (**A**) Rank distribution of the top 30 hits obtained during HHalign comparison between WPDV HVR tandem repeats and individual HVR domain sequences of arteriviruses. The red line depicts the 5% probability threshold. WPDV HVR tandem repeats identified by RADAR are shown in the top right corner. (**B**) Locations of motifs identified by MEME in the HVR of arterivirus species. Extended PxPxPR motifs are shown in green, and conserved C-terminal motifs corresponding to the nsp2 PRF site are shown in red. (**C**) MSA of the PxPxPR motif and its derivatives in the HVR of viruses representing arterivirus species. Coordinates in the names of motifs refer to their domain position. Numbers to the right of the MSA show support for the identification of each motif by three methods. The first column shows probability values assigned to hits containing PxPxPR motifs by HHalign in analyses comparing HVR sequences of the respective arteriviruses to the MSA of tandem repeats of the WPDV HVR. The second column shows P values assigned to motifs by MEME. The third column shows matches (+) and mismatches (−) of the PxPxPR pattern.

Conservation of a novel proline-rich motif in the nsp2 HVR of WPDV and other arteriviruses

One of the two most divergent regions of nsp2 is the HVR, located between PLP2 and the TM1-CR domains (Figure 3B). We found that the size of this domain varied >5-fold, from 125 aa (EAV) to 716 aa (PRRSV-2). Since the size difference might have emerged as a result of duplications, we searched for repeats in this domain. The presence of tandem repeats was initially detected in the WPDV HVR when it was compared to itself by use of HHalign and was subsequently corroborated by RADAR [58]. The MSA of WPDV HVR tandem repeats was converted into an HMM profile and compared with the nsp2 HVRs of different arteriviruses by using HHalign, resulting in multiple significant hits that conformed to the pattern PxPxPR or a close derivative (Figure 12A and C). Similar results were obtained using MEME [59], which identified extended versions of this motif (E value = 9.3e−9) in representatives of eight species (Figure 12B and C). A subsequent search for strict matches to the PxPxPR motif in the pp1ab proteins of all sequenced arteriviruses

identified at least one copy of the motif in the HVR for most viruses of 10 arterivirus species, with the number of motif copies varying in some species (Table 3). Two of the remaining four species, SHEV and KRCV-1, were found to contain a PxPxPR motif(s) in the hinge domain, while none of the pp1ab domains of two other species, Pebjah virus (PBJV) and De Brazza's monkey arterivirus (DeMAV), contain this motif. Overall, PxPxPR motifs were found predominantly in the HVR and much less frequently in the hinge domain, with the only exception being two isolates of Kibale red-tailed guenon virus 1 (KRTGV) that contain one copy of the motif in the PLP1a domain.

Table 3 | Intraspecies variation in the number of PxPxPR motifs in the HVR and elsewhere in pp1ab.

		Motifs in HVR	
Species	**Total no. of genomes**	**No. of motifs per domain**	**No. of genomes**
WPDV	1	3	1
EAV	27	3	27
SHEV[a]	1	0	1
MYBV-1	13	1	13
PBJV	3	0	3
SHFV	1	1	1
KRTGV[b]	4	1	4
DeMAV	1	0	1
KRCV-2	29	2	1
		1	26
		0	2
KRCV-1[c]	15	0	15
LDV	2	1	1
		0	1
PRRSV-2	368	4	1
		3	310
		2	53
		1	4
PRRSV-1	36	1	34
		0	2
APRAV	1	2	1

[a]One motif is present in the hinge domain.
[b]One motif is present in the PLP1a domain of 2 out of 4 isolates.
[c]One or two motifs are present in the hinge domain of 11 or 4 out of 15 isolates, respectively.

EAV may be the only arterivirus that has no PRF motifs in the nsp2 region

The above-described MEME analysis also identified residue conservation in all arteriviruses except the most divergent ones, EAV and WPDV, at the very C terminus of the HVR (Figure 12B, red boxes), which is adjacent to the TM1-CR domains conserved in all arteriviruses (Figure 13). Upon conversion of the arterivirus-wide MSA of the HVR domain C terminus (Figure 14B) into the nucleotide MSA (Figure 14C), it became evident that

amino acid conservation identified by MEME corresponds to nucleotide PRF motif conservation, slippery sequence RG_GUU_UUU (R = G or A) and downstream element CCCANCUCC [33]. These motifs were shown to guide translation of the genome region encoding HVR/TM1-CR junction in two alternative open reading frames, −1TF and −2TF, in PRRSV-1 and -2 [33, 34]. These two ORFs are expressed via −1 and −2 PRF with the production of nsp2N and nsp2TF, respectively (Figure 14A). Previously, it was suggested that during nsp2 PRF in arteriviruses, complete codon-anticodon repairing is required at the closely monitored ribosomal A site, while mismatches are tolerated at the P site [33]. Accordingly, slippery sequences observed in our analysis conformed to the patterns NN_NUU_UUU, NN_NUU_UUC, and NN_NUC_UCU (with the exception of PRRSV-1 EU076704.1 slippery sequence GG_GUU_UGU), which allow the integrity of the A-site duplex to be maintained after the −1/−2 shift or, in the case of the latter pattern, only the −2 shift [60]. Deviations in the downstream element were rare and did not involve more than one nucleotide, while observed sizes of −2TF domains were comparable (with the exception of LDV L13298.1 47 aa -2TF domain) to the experimentally verified size of the −2TF domain of PRRSV (Table 4). These results suggest that the observed variations in PRF motifs in our large virus data set (Table 4) may be compatible with their function despite the detrimental effects of some of these variations artificially introduced into PRRSV-2 [33, 35].

To learn about WPDV in this respect, we compared HMM profiles of nsp2 PRF-related motifs of arteriviruses and the WPDV nsp2 nucleotide sequence by using NHMMER. The two motifs were found in close proximity and canonical order in the expected region of the WPDV nsp2 locus, with the third best hit to each of the two queries. Remarkably, the hit to the slippery sequence profile was the only one observed that allowed complete A-site duplex repairing in the −2 frame. While each of these hits was statistically insignificant (with E values of 4.6 and 2.3), the probability of observing their combination in this place by chance may be approximately 2 orders of magnitude smaller than that of observing each hit separately, given the size of the nsp2 locus. Importantly, no comparably located proximal hits were found upon scanning of the EAV nsp2 locus, which served as a negative control. Accordingly, WPDV compared to EAV deviated from PRRSV much less in both motifs, but these were separated by 18 rather than the canonical 10 nucleotides (Figure 14C). In WPDV, the −1 frameshift is expected to lead to immediate termination of translation (as observed in PRRSV) (Figure 15A), while translation in the −2 frame may result in the product being extended with a domain as in other arteriviruses, with the following caveats: the size of this domain is much smaller than those of arteriviruses (32 versus 169 to 230 aa) (but see Table 4), and it lacks a TM module (Figure 15B). The −1/−2 PRF is stimulated by a complex of PCBP and nsp1b in PRRSV [34, 35]. Its effector region, located in PLP1b, is most conserved in arteriviruses (peak and logo Rb in Figure 10) and,

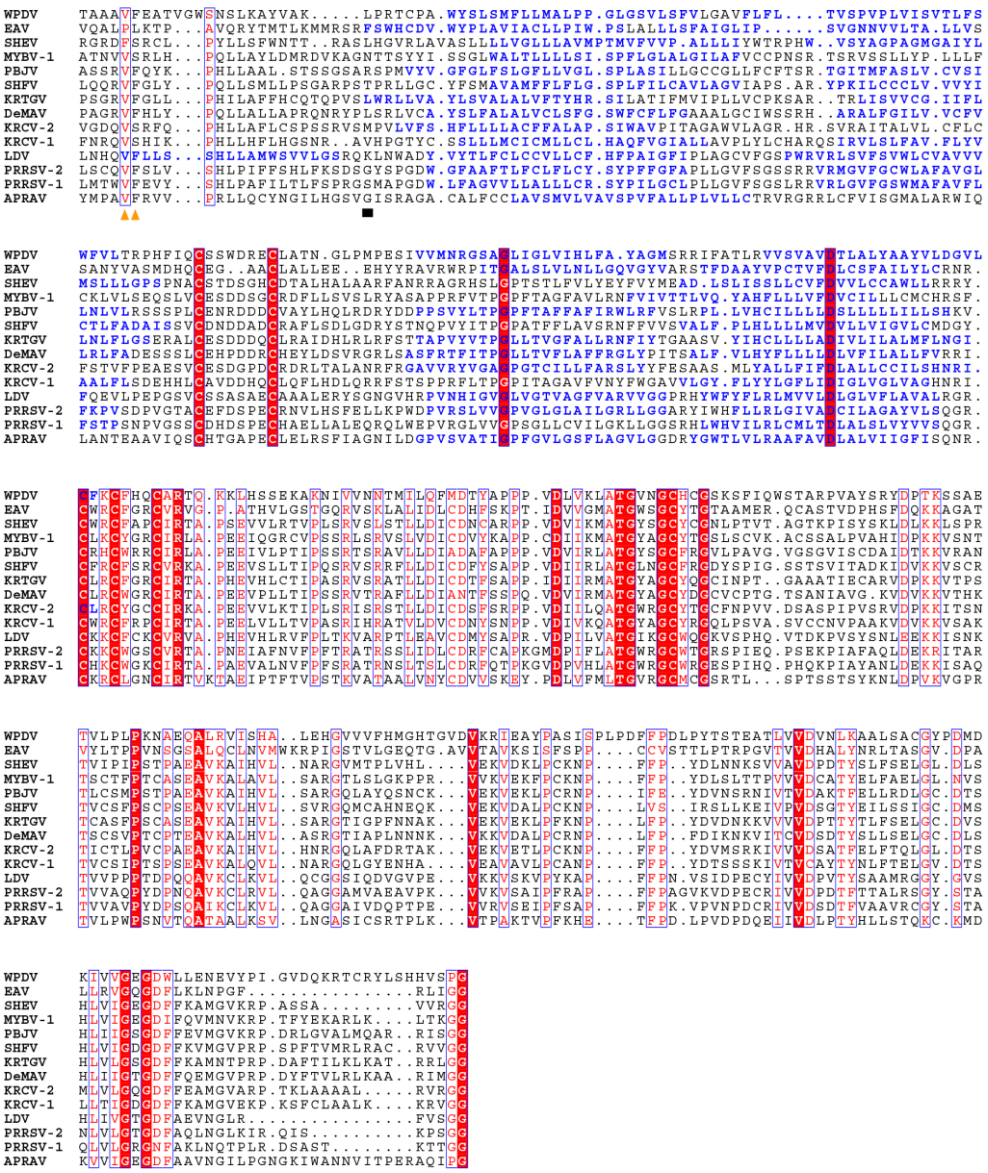


Figure 13 | Multiple-sequence alignment of the nsp2 C termini of arteriviruses. Columns containing amino acids whose tRNAs are expected to be present in the ribosomal P and A sites prior to −1/−2 frameshifting are marked with orange triangles. The first column of the TM1-CR domains is marked with a black box. Amino acid residues predicted by TMHMM 2.0 [61] to form transmembrane regions are colored blue. The MSA was visualized with Espript 2.1 [53].

further, has a conserved counterpart in PLP1c (Rc). WPDV deviates considerably from arteriviruses in this region, in both PLP1b and PLP1c, which may be due to either

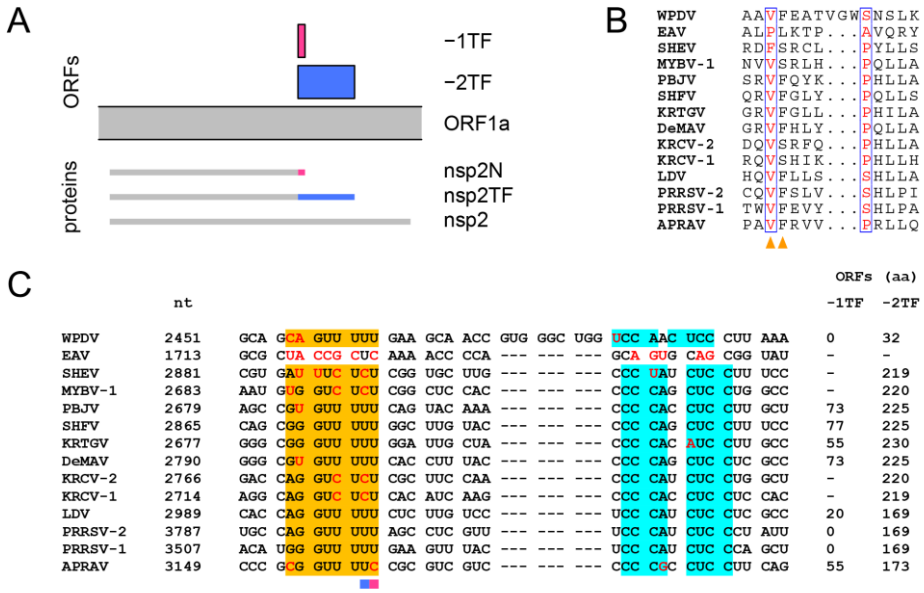


Figure 14 | Arteriviral nsp2 PRF. (**A**) Schematic representation of the expression of nsp2 moieties (based on LDV; accession number U15146.1). (**B**) Fragment of the pp1ab alignment corresponding to the site of nsp2 PRF. Columns containing amino acids whose tRNAs are present in the ribosomal P and A sites prior to frameshifting are highlighted with orange triangles. (**C**) Nucleotide alignment corresponding to the protein alignment presented in panel B. The slippery sequence is shown in orange and the C-rich element in cyan. Deviations from the canonical motifs, i.e., RG_GUU_UUU (R = G or A) and CCCANCUCC, are highlighted in red. For each sequence, the genome coordinate of the first nucleotide in the alignment is specified. If the frameshift site allows complete A-site duplex repairing in the −1 or −2 frame, then the length of the corresponding hypothetical protein product is specified. Otherwise, it is marked with a dash. Alignment columns containing the first nucleotides of −1TF and −2TF are highlighted with pink and blue bars, respectively.

coevolution with the PRF motifs or the lack of involvement of these domains in PRF regulation in WPDV.

DISCUSSION

In this report, we present the current state of the art for domain characterization of the nsp1-nsp2 genome region of arteriviruses by comparative sequence analysis. This work has confirmed and considerably extended the results of prior analyses of this region [5, 7, 17-20, 25, 39, 41]. Below, we briefly discuss the limitations and implications of the obtained results as well as the challenges of the conducted analyses.

We analyzed the genomes of all arteriviruses available on 11 June 2015 plus the genome sequence of WPDV, the most distantly related arterivirus, reported in full here for the first

Table 4 | Intraspecies variation of nsp2 PRF-related elements.

		Slippery sequence		C-rich region		-1TF		-2TF	
Species	Total no. of genomes	Sequence[a]	No. of genomes	Sequence[a]	No. of genomes	Length (aa)	No. of genomes	Length (aa)	No. of genomes
SHEV	1	A**u**_**u**U**c**_U**c**U	1	CC**u**ANCUCC	1	25	1	219	1
MYBV-1	13	**c**G_GU**c**_U**c**U	10	CCCANCUCC	13	10	4	220	13
		uG_GU**c**_U**c**U	3			0	4		
						21	3		
						15	1		
						13	1		
PBJV	3	G**u**_GUU_UUU	3	CCCANCUCC	3	73	3	225	3
SHFV	1	GG_GUU_UUU	1	CCCANCUCC	1	77	1	225	1
KRTGV	4	G**u**_GUU_UUU	2	CCCAN**a**UCC	4	60	2	230	4
		GG_GUU_UUU	2			55	2		
DeMAV	1	G**u**_GUU_UUU	1	CCCANCUCC	1	73	1	225	1
KRCV-2	29	AG_GU**c**_U**c**U	29	CCCANCUCC	29	24	29	220	29
KRCV-1	15	AG_GU**c**_U**c**U	15	CCCANCUCC	15	13	15	219	15
LDV	2	AG_GUU_UUU	2	CCCANCUCC	2	23	1	47	1
						20	1	169	1
PRRSV-2	368	AG_GUU_UUU	298	CCCANCUCC	366	0	314	169	365
		GG_GUU_UUU	40	CCC**g**NCUCC	2	23	33	168	1
		GG_GUU_UU**c**	17			16	16	128	1
		AG_GUU_UU**c**	6			18	5	115	1
		AG_**a**UU_UUU	5						
		uG_GUU_UUU	1						
		A**u**_GUU_UUU	1						
PRRSV-1	36	GG_GUU_UUU	35	CCCANCUCC	36	0	36	169	35
		GG_GUU_U**g**U	1					170	1
APRAV	1	**c**G_GUU_UU**c**	1	CCC**g**NCUCC	1	55	1	173	1

[a]Deviations from the canonical motifs, i.e., RG_GUU_UUU (R = G or A) and CCCANCUCC, are shown by lowercase bold letters.

Figure 15 | Multiple-sequence alignments of alternative nsp2 C termini. (**A**) C terminus of nsp2N, translated as a result of −1 PRF. (**B**) C terminus of nsp2TF, translated as a result of −2 PRF. MSAs were guided by the MSA presented in Figure 13. For other details, see the legend to Figure 13.

time. We extended the available WPDV genome sequence of 10,087 nt by 2,814 nt in the 5′ direction. This was accomplished with a modified 5′ RLM RACE protocol in two steps, RACE 1 (2,006-nt extension) and RACE 2 (808-nt extension), and benefited from bioinformatics analyses. No further extension was observed with an additional step, RACE 3, and no major genomic element was missing in this sequence according to our bioinformatics analysis. Collectively, these results attest to the completion of the genome sequence of WPDV, although we acknowledge that the exact terminal nucleotide(s) remains to be verified. Even with the target enrichment step, the full 5′ end was obtained in only one PCR, with several shorter PCR fragments amplified in various PCR runs. The difficulties encountered in amplification of the 5′ end of the WPDV sequence may be related to the presence of complex secondary structures at the 5′ end of the viral genomic

RNA. Such structures play a role in virus replication and have been described for the genomes of other nidoviruses [62].

The traditional 5' RACE protocol relies on homopolymer tailing of cDNA. The tail is then used to attach a linker sequence in the first rounds of PCR with a sequence-specific reverse primer and a linker-specific forward primer [63]. The main drawbacks of the technique are that it does not provide the ability to select full-length cDNA transcripts of interest and that it introduces bias for amplification of shorter sequences. As a result, a range of heterogeneous amplicons are produced, often including nonspecific products [64, 65]. To overcome these difficulties, 5' RLM RACE was used in the current study, which supports amplification of only capped, full-length mRNA. However, in all 5' RLM RACE reactions, only some of the amplified bands extended the sequence in the 5' direction despite the presence of a ligated adapter at the 5' ends of several other PCR products. This may have occurred due to one or more different factors of a biological and technical nature, including the presence of defective genomes and/or low efficiencies of the calf intestine alkaline phosphatase (CIP) and tobacco acid pyrophosphatase (TAP) treatments for preventing ligation of the adapter to noncapped RNA.

WPDV is the most distant of the arteriviruses based on conservation of the nonstructural proteins, and it infects the most distantly related mammalian host, a marsupial. In the arterivirus tree, its basal single-virus lineage could be contrasted with the sister lineage represented by a dozen arterivirus species which infect different placental hosts and which are separated by considerably shorter evolutionary distances. The divergence of WPDV from other arteriviruses is so profound that neither of the arterivirus-specific domains was delineated upstream of PLP2 in the putative nsp1 region by application of the most powerful profile-profile comparison techniques using conventional criteria. Only after the established homology of WPDV and arteriviruses in other nsp's was accounted for in the analysis did the profile-profile comparison identify three putative and highly divergent PLP1 domains in pp1ab. While this delineation will guide further experimental characterization of these domains, the mere presence of three paralogous PLP1 domains in WPDV is already significant for understanding the evolution of PLP1 domains in the sister lineage.

Given the presence of two or three PLP1 domains in all arteriviruses, the most recent common ancestor (MRCA) of all known arteriviruses most likely already encoded at least two PLP1 domains, one of which is expected to be the ancestor of the ubiquitous PLP1a domain (Figure 3). This consideration also implies that a duplication of PLP1 must have happened before the emergence of this MRCA; it remains to be established whether descendants of the ancestral arterivirus lineage with a single PLP1 domain have yet to be discovered or already went extinct. Gene duplication often results in subfunctionalization

or neofunctionalization, driven by positive selection to improve fitness that is facilitated by an increased evolvability of duplicates, each of which is less constrained than their ancestor [66, 67]. This framework may explain the observed subdomain-specific association of most conserved residues in PLP1a and PLP1b/c of all non-WPDV arteriviruses. Given the large evolutionary distance involved (Figure 3), these residues must be under the strong purifying selection that is commonly associated with a conserved function(s). Besides the Cys and His residues involved in catalysis, the functions of other residues have yet to be established.

Since PLP1b and PLP1c of the non-WPDV arteriviruses are much more closely related to each other than to PLP1a, they must have emerged through a second duplication and subsequent diversification. This duplication must have happened before the emergence of the MRCA of the simian group, all of whose members encode three PLP1 domains. The type- and lineage-specific evolutionary dynamics of PLP1 domains might have involved both divergent and convergent evolution and/or parallel duplication. These dynamics remain untested computationally due to poor sampling of three long-branch lineages that include just a single species each (WPDV, EAV, and APRAV), compounded by the observed variation in the number of PLP1 domains and the complexity of their similarities. For instance, similarity between PLP1b and PLP1c varies from very strong in viruses of the simian group to extremely weak in WPDV, while sequence affinity of the second PLP1 domain for PLP1b and PLP1c differs for EAV and LDV/PRRSV/APRAV, respectively. In the case of EAV, the observed affinity of the second PLP1 domain for PLP1c is both compatible with the published experimental research [33, 34] and incompatible with the current designation of this domain, PLP1b, which reflects its order in the pp1ab polyprotein. Consequently, our results predict EAV PLP1b to be closer functionally to PLP1c, which has not yet been characterized beyond its proteolytic activity in SHFV [39, 41, 46].

One of the recently identified functions of PLP1b is transactivation of nsp2 PRF, which was demonstrated for PRRSV-2 and shown to be lacking in EAV [33, 34], in line with PLP1b of the latter being similar to PLP1c (see above). Results of comparative sequence analyses by the discoverers of this phenomenon [33, 34] and those presented in this paper support the conservation of nsp2 PRF in all non-EAV arteriviruses. According to our analysis, the most divergent version of nsp2 PRF may be employed by WPDV, which deviates from other non-EAV arteriviruses in the sizes of the −2TF domain (smaller) and the region separating two PRF-related nucleotide motifs (larger). Both deviations may have functional implications. The −2TF domain of WPDV does not have hydrophobic regions predicted for other non-EAV arteriviruses, which may imply different localizations of nsp2TF proteins. The production of nsp2N and nsp2TF in PRRSV was highly sensitive to mutations that changed the size of the spacer between PRF motifs [35]. Consequently, during evolution, the unusually large size of this region in WPDV must have been

associated with changes elsewhere in the genome and/or been host specific. In this respect, PLP1b is a prime candidate to consider due to its role as the major domain of nsp1b in the transactivation of nsp2 PRF [34]. Compared to its orthologs, PLP1b of WPDV has a unique large insertion in the left subdomain and accepted many mutations to the putative equivalent of the positively charged α-helix that was implicated in the interaction with the PRF motifs in PRRSV. While further experimental research could address a possible connection between sequence specifics of the PLP1b and PRF motifs in WPDV, our results indicate that the −2 PRF in nsp2 may be a universal feature of non-EAV arteriviruses.

The apparent production of several molecular forms of nsp2 in arteriviruses may be linked to multifunctionality of this large nsp, which remains poorly characterized. We described here the large (5-fold) variation of the size of the most divergent domain of nsp2, the HVR, among arterivirus species, which is indicative of this domain being involved in arterivirus adaptation to hosts. Duplication was likely one of the mechanisms used to increase the size of this domain, as could be deduced from the presence of three tandem repeats with the formula PxPxPR in WPDV. These repeats may mediate a conserved function, since various numbers of their counterparts were identified in many but not all arteriviruses. Their interacting partners may be host proteins containing an SH3 domain(s), which have been shown to recognize PxPxPR motifs [68]. A similar suggestion was first made in a previous study [28], based on the presence of canonical SH3-binding PxxP motifs in nsp2 of PRRSV-1. However, PxxP motifs were not detected in our MEME analysis, suggesting that their presence in the HVR may be due to a disproportionally high Pro content in this domain.

In conclusion, our comparative genomic analysis of the most divergent region of replicative polyproteins revealed evolutionarily conserved patterns that are either specific to distinct species or common for different groups of arteriviruses. While the obtained insights were often the first ones for recently identified arteriviruses [12, 48-52], this analysis is also expected to promote further characterization of prototype arteriviruses, thus connecting the exploration of genetic diversity with experimental research on arteriviruses.

MATERIALS AND METHODS

Modified 5’ RLM RACE

The protocol supplied with a commercial kit (FirstChoice RLM RACE; Invitrogen) was modified by addition of a target enrichment step. Briefly, total RNA was isolated from a

standard inoculum (SI) that had been used in the previous WPD transmission studies [69] and from which the previously described partial viral sequence was obtained [11]. Extracted total RNA was used as a template for the initial steps of 5′ RML RACE, performed according to the manufacturer's instructions (Figure 1). The steps comprised treatment of total RNA with calf intestine alkaline phosphatase (CIP) to remove free 5′ phosphates from all noncapped nucleic acids, treatment with tobacco acid pyrophosphatase (TAP) to remove the cap structure from full-length mRNA (including capped positive-sense viral RNA), ligation of the provided RACE adapter to decapped mRNA containing 5′ phosphates, and reverse transcription of the ligated mRNA to cDNA by use of random decamers.

In the first round of 5′ RLM RACE (RACE 1), the cDNA was enriched for the target sequence before proceeding with the PCR step of the protocol. The enrichment step was performed using the magnetic bead, sequence capture, nested PCR method according to principles described by others [70-72]. Briefly, 0.24 pmol of a biotinylated capture probe that matched the available 5′ sequence of viral RNA (biotin-WPD.S5.F) (see Table S1 in the supplemental material) was added to a reaction mix that comprised 5 µl of cDNA, 1 µl of 10× buffer O (containing 50 mM Tris-HCl, 10 mM $MgCl_2$, 100 mM NaCl, and 0.1 mg/ml bovine serum albumin [BSA] at a 1× dilution; Fermentas), and water in a final volume of 10 µl. The nucleic acids were denatured at 95°C for 5 min and hybridized at 60°C for 23 h. An equal volume of 2× wash buffer (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 2 M NaCl) and 1 µl (5 µg) of streptavidin-coated magnetic beads (Dynabeads M280; Invitrogen) were then added to the hybridization reaction mixture, and the mixture was incubated for 3 h at 43°C with gentle shaking. The viral sequences captured on streptavidin-coated magnetic beads were then washed 3 times in 1× wash buffer and resuspended in 8 µl of water.

An aliquot (2 µl) of bead suspension was used in the PCR step of the 5′ RLM RACE protocol. Primary PCRs were performed using a 0.2 µM final concentration of each primer (RACE.outer forward primer and a virus-specific reverse primer) in 1× HOT FIREPol PCR master mix (Solis Biodyne) with 2 mM (final concentration) $MgCl_2$. The amplification conditions included 15 min of initial denaturation at 95°C followed by 35 cycles of denaturation (95°C for 10 s), annealing (60°C for 10 s), and elongation (72°C for 1 to 3 min), followed by a final extension step (72°C for 7 min). Nested PCRs were performed as primary reactions, but a nested adapter-specific primer (RACE.inner) was used in combination with each of several virus-specific reverse primers (Table S1). The template used for nested PCR was either the primary PCR product (1 µl) or a gel-purified band from the primary PCR (1 µl). Since the lengths of the expected PCR fragments were unknown, primary PCRs were also performed using an Expand long-range PCR kit (Roche) according to the manufacturer's instructions, with an initial elongation step of 4 min at 68°C.

In order to determine whether the longest PCR product represented the 5′ end of the full-length genomic RNA, the RLM RACE protocol was repeated using another capture probe (Biotyn_S12.F) targeting a region within the newly determined 5′ end of the sequence, in combination with virus-specific primers WPD.S10.R, WPD.S13.R, WPD.S14.R, and WPD.S15.R (RACE 2) (Table S1).

The final round of 5′ RLM RACE reactions (RACE 3) was performed using virus-specific primers (WPD.S16.R and WPD.S18.R) (Table S1) located close to the 5′ end identified in RACE 2. The RACE 3 reactions were performed according to the manufacturer's instructions, without the target enrichment step. Primary and nested PCR amplifications were performed with either HOT FIREPol PCR mix or Kappa LongRange HotStart ReadyMix (Kappa Biosystems). The long-range mix was used as recommended by the manufacturer to support amplification of fragments of up to 15 kbp. The non-TAP control was included in the RACE reaction mixtures to further assess whether any of the RACE-amplified bands originated from capped RNA sequences.

The final assembly of the newly identified 5′ end with the previously published sequence [11] was confirmed by amplification of a set of overlapping PCR fragments by use of virus-specific primers and SI cDNA as the template.

The previous GenBank record (accession number JN116253) was updated to include the 5′ end of the viral sequence.

Designation of nsp1 and PLP domains

In the literature, nsp1 PLPs and corresponding cleavage products are labeled with either the Latin letters a, b, and c or the Greek letters α, β, and γ (for example, PLP1a or PLP1α and nsp1a or nsp1α). In this report, we use Latin letters as labels.

Arterivirus genomes and classification

Full-length genomes of arteriviruses available on 11 June 2015 were retrieved from GenBank [73] and RefSeq [74] by using the homology-annotation hybrid retrieval of genetic sequences (HAYGENS) tool (http://veb.lumc.nl/HAYGENS). The sequence of the WPDV genome, including the newly sequenced 5′ terminus, whose annotation was updated accordingly (Table 1), was added to the set. With the help of DEmARC 1.3 ([75]; https://talk.ictvonline.org/files/ictv_official_taxonomy_updates_since_the_8th_report/m/animal-ssrna-viruses/5890), genomes of a total of 502 viruses were clustered into 14 species [3] that were grouped into five clusters. One virus representative was selected to represent each arterivirus species in further analyses (Table 2).

MSAs

Multiple-sequence alignments (MSAs) of pp1ab domains were generated using the Viralis platform [76] and assisted by use of the HMMER 3.1 [77], Muscle 3.8.31 [78], and ClustalW 2.012 [79] programs in default modes, with subsequent manual local refinement of MSAs of most divergent domains. Domain borders in nsp1-nsp2 proteins were tentatively identified (Table S3) through limited similarity with protein domains and cleavage cites that were studied experimentally [27, 36-39]. They may differ from the ones defined elsewhere. The MSA of nsp1 PLP paralogs was prepared using the profile mode of ClustalW in a stepwise manner: first, the PLP1b and PLP1c domain alignments were combined, and then the PLP1a MSA was added. MAFFT v7.123b [80] was used to align tandem repeats (see below). All presented protein MSAs were deposited at https://github.com/aag1/Arteriviridae_nsp1-2 in FASTA format.

Quantification of MSA conservation

To quantify residue conservation at each position of the MSA, we used the R package Bio3D 1.1.-5 [81], the “conserv” command, the “similarity” conservation assessment method, and the substitution matrix BLOSUM62 [82]. Individual columns of arteriviral PLP1a and PLP1b/PLP1c alignments (WPDV sequences excluded) were considered to be conserved if their conservation score exceeded 0.75. To transform conservation scores of individual columns in the arteriviral nsp1-nsp2 MSA into a conservation profile for plotting, a sliding window of 11 MSA columns was used to calculate mean conservation score values.

Secondary structure retrieval and prediction

Information about the PRRSV-2 nsp1a and nsp1b and EAV PLP2 secondary structures was retrieved from PDB structures 3IFU [18], 3MTV [42], and 4IUM [30], respectively, using the DSSP database [83] via the MRS system [84]. Secondary structure predictions were made for individual nsp1 PLP sequences of different origins by use of Jpred4 [55] in MSA mode.

Transmembrane region prediction

Transmembrane regions of proteins were predicted with the help of TMHMM 2.0 [61].

Profile-profile comparisons

We employed HHmake 2.0.16 to convert protein MSAs into HMM profiles and an in-house version of HHalign 2.0.16 [85] (deposited at https://github.com/dvs/hhsuite) to conduct profile-profile comparisons. The in-house version of HHalign enables the user control over the SMIN score threshold, otherwise hard coded to be 20. The SMIN score threshold is utilized by the HHalign algorithm to decide which hits will be reported, based on their raw Viterbi scores. By lowering the SMIN score threshold, the user can increase the number of

alternative hits reported, which may be informative for analyzing extremely remote relationships.

HHalign comparisons were performed with the following parameters: SMIN score threshold of 5, local alignment mode, and realignment by the MAC algorithm not applied. To visualize profile-profile comparisons in default mode, dot plots were generated.

Repeat and motif identification

We used a multistep procedure to characterize sequence repeats and associated motifs. First, the protein sequence of a virus was compared to itself by use of HHalign. In the produced diagonal plot, overlapping off-diagonal hits with high statistical support were indicative of tandem repeats. Subsequently, the protein sequence was submitted to the RADAR Web server [58] to verify the presence of tandem repeats and to delineate their exact positions. To study if an identified repeat motif was present in sequences representing other arterivirus species, the sequences were scanned with a RADAR-produced MSA of repeats by use of HHalign (probability threshold, 5%). At the next stage, the obtained results were verified and extended by use of MEME 4.11.2 [59], which was applied to the selected protein domain of representatives of all arterivirus species. In the MEME analysis, the number of unique motifs to be found was set to 10, the expected distribution of the unique motifs' occurrences in a sequence was defined as “any number of repetitions,” the lengths of motifs were allowed to range from 4 to 50 aa, and other parameters were set to their defaults.

Nucleotide sequence profile comparisons

We used NHMMER 3.1b1 [86] with the parameters rna-toponly-max-nonull2 to scan the EAV and WPDV genome regions encoding nsp2 for similarity to nucleotide MSAs of nsp2 PRF-related motifs from genomes representing 12 other arterivirus species.

Phylogeny reconstruction

The phylogeny of arteriviruses was reconstructed using a concatenated MSA of most conserved nsp domains (Table S2). To select a model of evolution that best fits the data, ProtTest 3.4 [87] was used. All models offered by ProtTest were tested. When a discrete gamma distribution was employed to model various rates of mutation among sites (+G), four rate categories were used. Maximum likelihood (ML) tree topology optimization strategy, employing a subtree pruning and regrafting (SPR) algorithm, was used. Two model selection criteria, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), were employed. According to both criteria, the LG+I+G+F model is the best. Subsequently, the phylogeny was reconstructed using the BEAST 1.8.2 package [88] and the LG+I+G4+F model. Two models, a strict clock and a relaxed clock

with an uncorrelated lognormal rate distribution, were tested. The latter was found to be superior (log10 Bayes factor of 5.48). Markov chain Monte Carlo (MCMC) chains were run for 10 million steps and sampled every 1,000 steps; the first 10% were discarded as burn-in. Mixing and convergence were verified with the help of Tracer (http://beast.bio.ed.ac.uk/Tracer).

A similar procedure was used to reconstruct the phylogeny of PLP1 domains by using MSA of PLP1b and PLP1c domains. Among the models available in BEAUti 1.8.2, ProtTest favored the LG+I+G4+F model, which was employed for BEAST phylogeny reconstruction. A relaxed clock with an uncorrelated lognormal rate distribution was favored over a strict clock (log10 Bayes factors of 4.88 and 3.80 for data sets with and without WPDV PLP1 domains, respectively). MCMC chains were run for 5 million steps and sampled every 500 steps; the first 10% were discarded as burn-in. The R package APE 3.5 was used to calculate the percentage of trees in the sample that differed in terms of the phylogenetic positions of major clades [89].

Tertiary protein structure comparison

We used the DALI server [54] for comparison of PLP tertiary structures. Conventionally, two folds are considered to be similar if their similarity Z-score is above 2. However, to be considered strongly supported, the similarity Z-score must be above the cutoff, defined as $n/10 - 4$, where n is the number of residues in the query structure [90]. For PRRSV-2 PLP1a and PLP1b queries, Z-score cutoffs were calculated to be 10.7 and 9.4, respectively.

Visualization of results of bioinformatics analyses

Protein MSAs with highlighted conservation and assigned secondary structure were visualized with Espript 2.1 [53], using the BLOSUM62 similarity coloring scheme and a similarity global score of 0.2. MSA conservation was also presented in the logo format with the help of the R package RWebLogo 1.0.3 [56]. To visualize the posterior sample of trees, DensiTree.v2.2.1 was used [91]. Protein tertiary structures were processed for presentation by use of PyMOL 1.7.6.6 [57]. R was used extensively for other data plotting [92].

ACKNOWLEDGMENTS

We thank Igor Sidorov for helpful discussions and for help with Viralis.

This study was partially funded by an Agreement about Cooperation in Bioinformatics between LUMC and MSU (MoBiLe Program) and by EU project EVAg 653316, the Leids

Universiteits Fonds, the Massey University Research Fund, and Lewis Fitch and McGeorge Veterinary Research grants.

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

SUPPLEMENTAL MATERIAL

Table S1 | Primer and probe sequences used to sequence the 5'-terminal region of WPDV genome.

Name	Sequence (5' to 3')	Used as	Position (nt)	Round
WPD.S5.R	TGGAGGTGGCGCGTAGGTGT	primer	3,028-3,047	RACE 1
Biotin-WPD.S5.F	Biotin-ATGCAGCTTATGTCCTTGATGGGGT	probe	2,893-2,917	RACE 1
WPD.S7.R	CAGGGCATGTGCGCGGTAGT	primer	2,510-2,530	RACE 1
WPD.S8.R	GCCCACGGTTGCTTCAAAAACTGCT	primer	2,062-2,086	RACE 1
WPD.S10.R	CCCACTCCAGTGCGTTTGTCAT	primer	1,288-1,309	RACE 2
WPD. S13.R	AGGCGCTGCAGTACCGTCGT	primer	1,096-1,115	RACE 2
WPD.S14.R	GATGAACGGCATCCCTGACA	primer	1,003-1,022	RACE 2
Biotin-WPD.S12.F	Biotin-CGGGGCGATCGTGGCTTACAG	probe	887-907	RACE 2
WPD.S15.R	CGTCTCCGGGTATCATGGTC	primer	869-888	RACE 3
WPD.S16.R	AAAATCGGGTGGACGGATGT	primer	545-564	RACE 3
WPD. S18.R	TTGTCGAATCGGGGGTAAGC	primer	150-169	RACE 3

Table S2 | Protein domains that are conserved in arteriviruses and were used for phylogeny reconstruction.

	Coordinates in NC_001961.1 genome (nt)[b]	
Domain[a]	**from**	**to**
nsp3	4,927	5,616
nsp4	5,617	6,228
nsp5	6,229	6,738
nsp7a	6,787	7,233
nsp8-9[c]	7,564	9,617
nsp10_HELcore	10,002	10,775
nsp11	10,941	11,609

[a]Domains conserved in all arteriviruses.
[b]Coordinates of conserved domains in NC_001961.1 genome of PRRSV-2, used to delineate domains in polyprotein MSA of selected arteriviruses (see Figure 3A).
[c]Translation involves -1 PRF.

Table S3 | Lengths (aa) of arteriviral nsp1-2 protein domains.

	Domain[b]								
Virus[a]	**ZnF**	**PLP1a**	**'Nuclease'**	**PLP1b**	**PLP1c**	**Hinge**	**PLP2**	**HVR**	**TM1-CR**
PRRSV-2	33	147	69	134	0	45	140	650	331
PRRSV-1	33	147	74	131	0	35	137	557	332
LDV	33	148	65	135	0	0	135	447	324
APRAV	36	146	66	142	0	11	142	436	340
KRCV-2	28	138	61	124	126	107	134	139	329
PBJV	28	137	62	124	128	111	139	127	333
SHFV	28	136	62	124	134	149	136	135	332
DeMAV	28	139	62	123	133	100	137	157	331
KRTGV	28	138	62	123	133	81	137	155	331
KRCV-1	28	136	62	122	131	124	134	133	329
SHEV	28	136	61	123	138	140	133	150	326
MYBV-1	28	137	62	124	134	74	135	164	331
EAV	49	90	0	121	0	0	130	125	317
WPDV	0	98	0	163	142	57	152	143	341

[a]One virus-representative from each of the fourteen arterivirus species, delineated by DEmARC, was analysed.
[b]Domains were delineated based on similarity with domains and cleavage sites of arteriviruses studied experimentally.

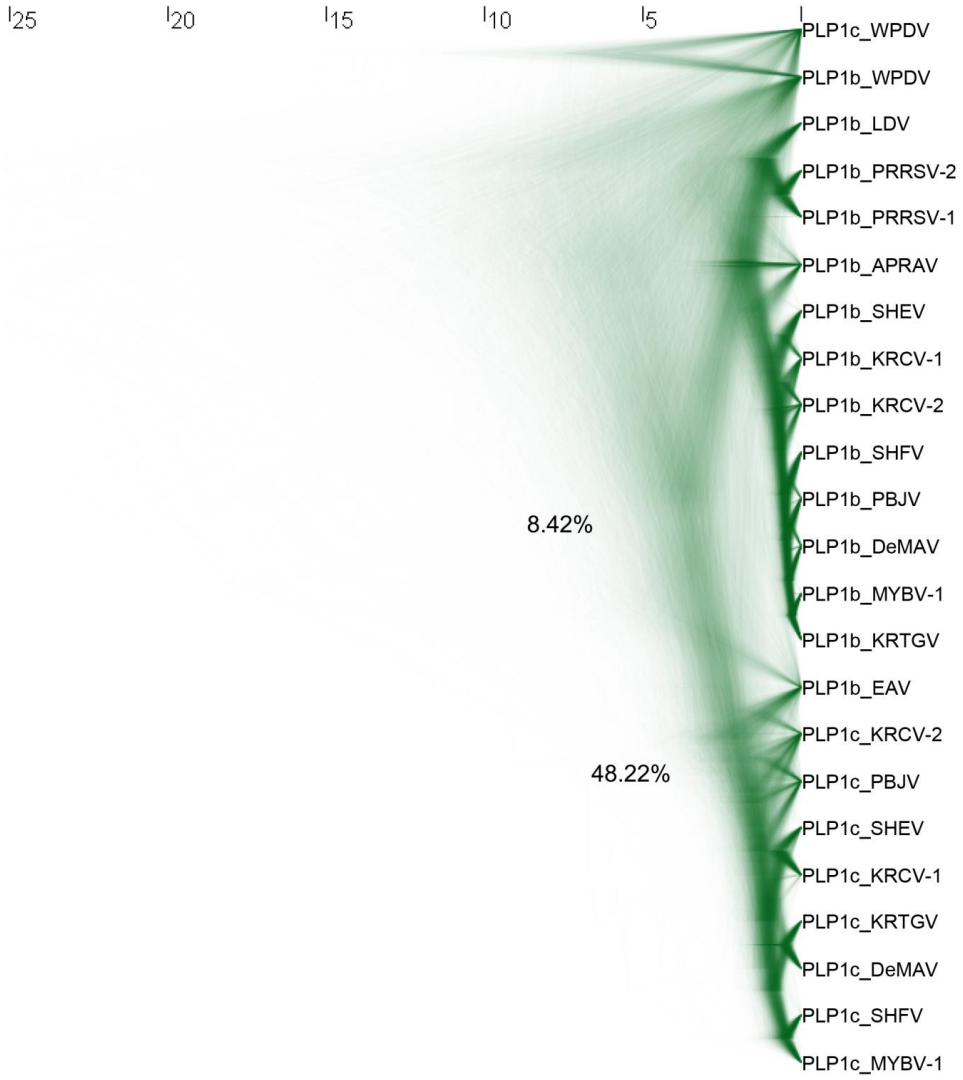


Figure S1 | Phylogeny of PLP1b and PLP1c of arteriviruses. Shown is a posterior sample of phylogenetic trees generated by BEAST using pan-arterivirus MSA of PLP1b and PLP1c. Percentages of trees in the sample, in which EAV PLP1b is basal to either non-WPDV PLP1c or non-WPDV PLP1bc clades are indicated near the MRCA of the corresponding clades. For other designations, see Figure 3A legend.

REFERENCES

1. Snijder EJ, Kikkert M, Fang Y: **Arterivirus molecular biology and pathogenesis**. *J Gen Virol* 2013, **94**(Pt 10):2141-2163.
2. Faaberg KS, Balasuriya UB, Brinton MA, Gorbalenya AE, Leung FC-C, Nauwynck H, Snijder EJ, Stadejek T, Yang H, Yoo D: **Family Arteriviridae.** In: *Virus Taxonomy, the 9th Report of the International Committee on Taxonomy of Viruses.* Edited by King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ: Eds. Academic Press; 2012: 796-805.
3. Adams MJ, Lefkowitz EJ, King AM, Harrach B, Harrison RL, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Mushegian AR *et al*: **Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2016)**. *Arch Virol* 2016, **161**:2921-2949.
4. Balasuriya UB, Snijder EJ, Heidner HW, Zhang J, Zevenhoven-Dobbe JC, Boone JD, McCollum WH, Timoney PJ, MacLachlan NJ: **Development and characterization of an infectious cDNA clone of the virulent Bucyrus strain of Equine arteritis virus**. *J Gen Virol* 2007, **88**(Pt 3):918-924.
5. den Boon JA, Snijder EJ, Chirnside ED, de Vries AA, Horzinek MC, Spaan WJ: **Equine arteritis virus is not a togavirus but belongs to the coronaviruslike superfamily**. *J Virol* 1991, **65**(6):2910-2920.
6. Palmer GA, Kuo L, Chen Z, Faaberg KS, Plagemann PG: **Sequence of the genome of lactate dehydrogenase-elevating virus: heterogenicity between strains P and C**. *Virology* 1995, **209**(2):637-642.
7. Godeny EK, Chen L, Kumar SN, Methven SL, Koonin EV, Brinton MA: **Complete genomic sequence and phylogenetic analysis of the lactate dehydrogenase-elevating virus (LDV)**. *Virology* 1993, **194**(2):585-596.
8. Zeng L, Godeny EK, Methven SL, Brinton MA: **Analysis of simian hemorrhagic fever virus (SHFV) subgenomic RNAs, junction sequences, and 5' leader**. *Virology* 1995, **207**(2):543-548.
9. Meulenberg JJ, Hulst MM, de Meijer EJ, Moonen PL, den Besten A, de Kluyver EP, Wensvoort G, Moormann RJ: **Lelystad virus, the causative agent of porcine epidemic abortion and respiratory syndrome (PEARS), is related to LDV and EAV**. *Virology* 1993, **192**(1):62-72.
10. Nelsen CJ, Murtaugh MP, Faaberg KS: **Porcine reproductive and respiratory syndrome virus comparison: divergent evolution on two continents**. *J Virol* 1999, **73**(1):270-280.
11. Dunowska M, Biggs PJ, Zheng T, Perrott MR: **Identification of a novel nidovirus associated with a neurological disease of the Australian brushtail possum (Trichosurus vulpecula)**. *Vet Microbiol* 2012, **156**(3-4):418-424.

12. Kuhn JH, Lauck M, Bailey AL, Shchetinin AM, Vishnevskaya TV, Bao Y, Ng TF, LeBreton M, Schneider BS, Gillis A *et al*: **Reorganization and expansion of the nidoviral family Arteriviridae**. *Arch Virol* 2016, **161**(3):755-768.

13. Giles J, Perrott M, Roe W, Dunowska M: **The aetiology of wobbly possum disease: Reproduction of the disease with purified nidovirus**. *Virology* 2016, **491**:20-26.

14. Mackintosh CG, Crawford JL, Thompson EG, McLeod BJ, Gill JM, O'Keefe JS: **A newly discovered disease of the brushtail possum: wobbly possum syndrome**. *N Z Vet J* 1995, **43**(3):126.

15. Perrott MR, Meers J, Cooke MM, Wilks CR: **A neurological syndrome in a free-living population of possums (Trichosurus vulpecula)**. *N Z Vet J* 2000, **48**(1):9-15.

16. Holtkamp DJ, Kliebenstein JB, Neumann EJ, Zimmerman JJ, Rotto HF, Yoder TK, Wang C, Yeske PE, Mowrer CL, Haley CA: **Assessment of the economic impact of porcine reproductive and respiratory syndrome virus on United States pork producers**. *Journal of Swine Health and Production* 2013, **21**(2):72-84.

17. Snijder EJ, Wassenaar AL, Spaan WJ: **The 5' end of the equine arteritis virus replicase gene encodes a papainlike cysteine protease**. *J Virol* 1992, **66**(12):7040-7048.

18. Sun Y, Xue F, Guo Y, Ma M, Hao N, Zhang XC, Lou Z, Li X, Rao Z: **Crystal structure of porcine reproductive and respiratory syndrome virus leader protease Nsp1alpha**. *J Virol* 2009, **83**(21):10931-10940.

19. Nedialkova DD, Gorbalenya AE, Snijder EJ: **Arterivirus Nsp1 modulates the accumulation of minus-strand templates to control the relative abundance of viral mRNAs**. *PLoS Pathog* 2010, **6**(2):e1000772.

20. Ziebuhr J, Snijder EJ, Gorbalenya AE: **Virus-encoded proteinases and proteolytic processing in the Nidovirales**. *J Gen Virol* 2000, **81**(Pt 4):853-879.

21. Lauber C, Goeman JJ, Parquet MC, 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**(7):e1003500.

22. Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, Janssen GM, 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**(17):8416-8434.

23. Lehmann KC, Hooghiemstra L, Gulyaeva A, Samborskiy DV, Zevenhoven-Dobbe JC, Snijder EJ, Gorbalenya AE, Posthuma CC: **Arterivirus nsp12 versus the coronavirus nsp16 2'-O-methyltransferase: comparison of the C-terminal cleavage products of two nidovirus pp1ab polyproteins**. *J Gen Virol* 2015, **96**(9):2643-2655.

24. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, 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**(5):991-1004.

25. Snijder EJ, Wassenaar AL, Spaan WJ, Gorbalenya AE: **The arterivirus Nsp2 protease. An unusual cysteine protease with primary structure similarities to both papain-like and chymotrypsin-like proteases**. *J Biol Chem* 1995, **270**(28):16671-16676.

26. Manolaridis I, Gaudin C, Posthuma CC, Zevenhoven-Dobbe JC, Imbert I, Canard B, Kelly G, Tucker PA, Conte MR, Snijder EJ: **Structure and genetic analysis of the arterivirus nonstructural protein 7alpha**. *J Virol* 2011, **85**(14):7449-7453.

27. Kikkert M, Snijder EJ, Gorbalenya AE: **Arterivirus nsp2 Cysteine Proteinase**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2210-2215.

28. Ropp SL, Wees CE, Fang Y, Nelson EA, Rossow KD, Bien M, Arndt B, Preszler S, Steen P, Christopher-Hennings J *et al*: **Characterization of emerging European-like porcine reproductive and respiratory syndrome virus isolates in the United States**. *J Virol* 2004, **78**(7):3684-3703.

29. Faaberg KS, Kehrli ME, Jr., Lager KM, Guo B, Han J: **In vivo growth of porcine reproductive and respiratory syndrome virus engineered nsp2 deletion mutants**. *Virus Res* 2010, **154**(1-2):77-85.

30. van Kasteren PB, Bailey-Elkin BA, James TW, Ninaber DK, Beugeling C, Khajehpour M, Snijder EJ, Mark BL, Kikkert M: **Deubiquitinase function of arterivirus papain-like protease 2 suppresses the innate immune response in infected host cells**. *Proc Natl Acad Sci U S A* 2013, **110**(9):E838-E847.

31. Fang Y, Snijder EJ: **The PRRSV replicase: exploring the multifunctionality of an intriguing set of nonstructural proteins**. *Virus Res* 2010, **154**(1-2):61-76.

32. Wassenaar AL, Spaan WJ, Gorbalenya AE, Snijder EJ: **Alternative proteolytic processing of the arterivirus replicase ORF1a polyprotein: evidence that NSP2 acts as a cofactor for the NSP4 serine protease**. *J Virol* 1997, **71**(12):9313-9322.

33. Fang Y, Treffers EE, Li Y, Tas A, Sun Z, van der Meer Y, de Ru AH, van Veelen PA, Atkins JF, Snijder EJ *et al*: **Efficient -2 frameshifting by mammalian ribosomes to synthesize an additional arterivirus protein**. *Proc Natl Acad Sci U S A* 2012, **109**(43):E2920-E2928.

34. Li Y, Treffers EE, Napthine S, Tas A, Zhu L, Sun Z, Bell S, Mark BL, van Veelen PA, van Hemert MJ *et al*: **Transactivation of programmed ribosomal frameshifting by a viral protein**. *Proc Natl Acad Sci U S A* 2014, **111**(21):E2172-E2181.

35. Napthine S, Treffers EE, Bell S, Goodfellow I, Fang Y, Firth AE, Snijder EJ, Brierley I: **A novel role for poly(C) binding proteins in programmed ribosomal frameshifting**. *Nucleic Acids Res* 2016, **44**(12):5491-5503.

36. Nedialkova DD, Gorbalenya AE, Snijder EJ: **Arterivirus Papain-like Proteinase 1ß**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2205-2210.

37. Nedialkova DD, Gorbalenya AE, Snijder EJ: **Arterivirus Papain-like Proteinase 1a**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2199-2204.

38. Li Y, Tas A, Sun Z, Snijder EJ, Fang Y: **Proteolytic processing of the porcine reproductive and respiratory syndrome virus replicase**. *Virus Res* 2015, **202**:48-59.

39. Vatter HA, Di H, Donaldson EF, Radu GU, Maines TR, Brinton MA: **Functional analyses of the three simian hemorrhagic fever virus nonstructural protein 1 papain-like proteases**. *J Virol* 2014, **88**(16):9129-9140.

40. den Boon JA, Faaberg KS, Meulenberg JJ, Wassenaar AL, Plagemann PG, Gorbalenya AE, Snijder EJ: **Processing and evolution of the N-terminal region of the arterivirus replicase ORF1a protein: identification of two papainlike cysteine proteases**. *J Virol* 1995, **69**(7):4500-4505.

41. Han M, Kim CY, Rowland RR, Fang Y, Kim D, Yoo D: **Biogenesis of non-structural protein 1 (nsp1) and nsp1-mediated type I interferon modulation in arteriviruses**. *Virology* 2014, **458-459**:136-150.

42. Xue F, Sun Y, Yan L, Zhao C, Chen J, Bartlam M, Li X, Lou Z, Rao Z: **The crystal structure of porcine reproductive and respiratory syndrome virus nonstructural protein Nsp1beta reveals a novel metal-dependent nuclease**. *J Virol* 2010, **84**(13):6461-6471.

43. Tijms MA, van Dinten LC, Gorbalenya AE, Snijder EJ: **A zinc finger-containing papain-like protease couples subgenomic mRNA synthesis to genome translation in a positive-stranded RNA virus**. *Proc Natl Acad Sci U S A* 2001, **98**(4):1889-1894.

44. Tijms MA, Nedialkova DD, Zevenhoven-Dobbe JC, Gorbalenya AE, Snijder EJ: **Arterivirus subgenomic mRNA synthesis and virion biogenesis depend on the multifunctional nsp1 autoprotease**. *J Virol* 2007, **81**(19):10496-10505.

45. Kroese MV, Zevenhoven-Dobbe JC, Bos-de Ruijter JN, Peeters BP, Meulenberg JJ, Cornelissen LA, Snijder EJ: **The nsp1alpha and nsp1 papain-like autoproteinases are essential for porcine reproductive and respiratory syndrome virus RNA synthesis**. *J Gen Virol* 2008, **89**(Pt 2):494-499.

46. Han M, Yoo D: **Modulation of innate immune signaling by nonstructural protein 1 (nsp1) in the family Arteriviridae**. *Virus Res* 2014, **194**:100-109.

47. Go YY, Li Y, Chen Z, Han M, Yoo D, Fang Y, Balasuriya UB: **Equine arteritis virus does not induce interferon production in equine endothelial cells: identification of nonstructural protein 1 as a main interferon antagonist**. *Biomed Res Int* 2014, **2014**:420658.

48. Lauck M, Hyeroba D, Tumukunde A, Weny G, Lank SM, Chapman CA, O'Connor DH, Friedrich TC, Goldberg TL: **Novel, divergent simian hemorrhagic fever viruses in a wild Ugandan red colobus monkey discovered using direct pyrosequencing**. *PLoS One* 2011, **6**(4):e19056.

49. Bailey AL, Lauck M, Weiler A, Sibley SD, Dinis JM, Bergman Z, Nelson CW, Correll M, Gleicher M, Hyeroba D *et al*: **High genetic diversity and adaptive potential of two simian hemorrhagic fever viruses in a wild primate population**. *PLoS One* 2014, **9**(3):e90714.

50. Lauck M, Sibley SD, Hyeroba D, Tumukunde A, Weny G, Chapman CA, Ting N, Switzer WM, Kuhn JH, Friedrich TC *et al*: **Exceptional simian hemorrhagic fever virus diversity in a wild African primate community**. *J Virol* 2013, **87**(1):688-691.

51. Bailey AL, Lauck M, Sibley SD, Pecotte J, Rice K, Weny G, Tumukunde A, Hyeroba D, Greene J, Correll M *et al*: **Two novel simian arteriviruses in captive and wild baboons (Papio spp.)**. *J Virol* 2014, **88**(22):13231-13239.

52. Lauck M, Alkhovsky SV, Bao Y, Bailey AL, Shevtsova ZV, Shchetinin AM, Vishnevskaya TV, Lackemeyer MG, Postnikova E, Mazur S *et al*: **Historical Outbreaks of Simian Hemorrhagic Fever in Captive Macaques Were Caused by Distinct Arteriviruses**. *J Virol* 2015, **89**(15):8082-8087.

53. 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**(13):3320-3323.

54. Holm L, Rosenstrom P: **Dali server: conservation mapping in 3D**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W545-549.

55. Drozdetskiy A, Cole C, Procter J, Barton GJ: **JPred4: a protein secondary structure prediction server**. *Nucleic Acids Res* 2015, **43**(W1):W389-394.

56. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res* 2004, **14**(6):1188-1190.

57. **The PyMOL Molecular Graphics System**. In., 1.7.6.6 edn: Schrödinger, LLC.

58. Heger A, Holm L: **Rapid automatic detection and alignment of repeats in protein sequences**. *Proteins* 2000, **41**(2):224-237.

59. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.

60. Percudani R: **Restricted wobble rules for eukaryotic genomes**. *Trends Genet* 2001, **17**(3):133-135.

61. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences**. *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:175-182.

62. van den Born E, Posthuma CC, Gultyaev AP, Snijder EJ: **Discontinuous subgenomic RNA synthesis in arteriviruses is guided by an RNA hairpin structure located in the genomic leader region**. *J Virol* 2005, **79**(10):6312-6324.

63. Frohman MA: **On beyond classic RACE (rapid amplification of cDNA ends)**. *PCR Methods Appl* 1994, **4**(1):S40-58.

64. Schaefer BC: **Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends**. *Anal Biochem* 1995, **227**(2):255-273.

65. Schramm G, Bruchhaus I, Roeder T: **A simple and reliable 5'-RACE approach**. *Nucleic Acids Res* 2000, **28**(22):E96.

66. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models**. *Nat Rev Genet* 2010, **11**(2):97-108.

67. Zhang JZ: **Evolution by gene duplication: an update**. *Trends in Ecology & Evolution* 2003, **18**(6):292-298.

68. Kowanetz K, Szymkiewicz I, Haglund K, Kowanetz M, Husnjak K, Taylor JD, Soubeyran P, Engstrom U, Ladbury JE, Dikic I: **Identification of a novel proline-arginine motif involved in CIN85-dependent clustering of Cbl and down-regulation of epidermal growth factor receptors**. *J Biol Chem* 2003, **278**(41):39735-39746.

69. Perrott MR, Wilks CR, Meers J: **Routes of transmission of wobbly possum disease**. *N Z Vet J* 2000, **48**(1):3-8.

70. Allen GP: **Antemortem detection of latent infection with neuropathogenic strains of equine herpesvirus-1 in horses**. *Am J Vet Res* 2006, **67**(8):1401-1405.

71. Tagle DA, Swaroop M, Lovett M, Collins FS: **Magnetic bead capture of expressed sequences encoded within large genomic segments**. *Nature* 1993, **361**(6414):751-753.

72. Morgan JG, Dolganov GM, Robbins SE, Hinton LM, Lovett M: **The selective isolation of novel cDNAs encoded by the regions surrounding the human interleukin 4 and 5 genes**. *Nucleic Acids Res* 1992, **20**(19):5173-5179.

73. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2016, **44**(D1):D67-D72.

74. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Res* 2016, **44**(D1):D733-745.

75. Lauber C, Gorbalenya AE: **Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses**. *J Virol* 2012, **86**(7):3890-3904.

76. Gorbalenya AE, Lieutaud P, Harris MR, Coutard B, Canard B, Kleywegt GJ, Kravchenko AA, Samborskiy DV, Sidorov IA, Leontovich AM *et al*: **Practical application of bioinformatics by the multidisciplinary VIZIER consortium**. *Antiviral Res* 2010, **87**(2):95-110.

77. Eddy SR: **A new generation of homology search tools based on probabilistic inference**. *Genome Inform* 2009, **23**(1):205-211.

78. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

79. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.

80. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Mol Biol Evol* 2013, **30**(4):772-780.

81. Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS: **Bio3d: an R package for the comparative analysis of protein structures**. *Bioinformatics* 2006, **22**(21):2695-2696.

82. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci U S A* 1992, **89**(22):10915-10919.

83. Touw WG, Baakman C, Black J, te Beek TA, Krieger E, Joosten RP, Vriend G: **A series of PDB-related databanks for everyday needs**. *Nucleic Acids Res* 2015, **43**(Database issue):D364-368.

84. Hekkelman ML, Vriend G: **MRS: a fast and compact retrieval system for biological data**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W766-769.

85. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment**. *Nat Methods* 2012, **9**(2):173-175.

86. Wheeler TJ, Eddy SR: **nhmmer: DNA homology search with profile HMMs**. *Bioinformatics* 2013, **29**(19):2487-2489.

87. Darriba D, Taboada GL, Doallo R, Posada D: **ProtTest 3: fast selection of best-fit models of protein evolution**. *Bioinformatics* 2011, **27**(8):1164-1165.

88. Drummond AJ, Suchard MA, Xie D, Rambaut A: **Bayesian phylogenetics with BEAUti and the BEAST 1.7**. *Mol Biol Evol* 2012, **29**(8):1969-1973.

89. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language**. *Bioinformatics* 2004, **20**(2):289-290.

90. Holm L, Kaariainen S, Rosenstrom P, Schenkel A: **Searching protein structure databases with DaliLite v.3**. *Bioinformatics* 2008, **24**(23):2780-2781.

91. Heled J, Bouckaert RR: **Looking for trees in the forest: summary tree from posterior samples**. *BMC Evol Biol* 2013, **13**:221.

92. R Core Team: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing; 2013.

CHAPTER 3

Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses

Kathleen C. Lehmann
Anastasia A. Gulyaeva
Jessika C. Zevenhoven-Dobbe
George M. C. Janssen
Mark Ruben
Hermen S. Overkleeft
Peter A. van Veelen
Dmitry V. Samborskiy
Alexander A. Kravchenko
Andrey M. Leontovich
Igor A. Sidorov
Eric J. Snijder
Clara C. Posthuma
Alexander E. Gorbalenya

Nucleic Acids Research (2015) 43(17):8416
DOI: 10.1093/nar/gkv838

ABSTRACT

RNA viruses encode an RNA-dependent RNA polymerase (RdRp) that catalyzes the synthesis of their RNA(s). In the case of positive-stranded RNA viruses belonging to the order *Nidovirales*, the RdRp resides in a replicase subunit that is unusually large. Bioinformatics analysis of this non-structural protein has now revealed a nidoviral signature domain (genetic marker) that is N-terminally adjacent to the RdRp and has no apparent homologs elsewhere. Based on its conservation profile, this domain is proposed to have nucleotidylation activity. We used recombinant non-structural protein 9 of the arterivirus equine arteritis virus (EAV) and different biochemical assays, including irreversible labeling with a GTP analog followed by a proteomics analysis, to demonstrate the manganese-dependent covalent binding of guanosine and uridine phosphates to a lysine/histidine residue. Most likely this was the invariant lysine of the newly identified domain, named nidovirus RdRp-associated nucleotidyltransferase (NiRAN), whose substitution with alanine severely diminished the described binding. Furthermore, this mutation crippled EAV and prevented the replication of severe acute respiratory syndrome coronavirus (SARS-CoV) in cell culture, indicating that NiRAN is essential for nidoviruses. Potential functions supported by NiRAN may include nucleic acid ligation, mRNA capping and protein-primed RNA synthesis, possibilities that remain to be explored in future studies.

INTRODUCTION

Positive-stranded (+) RNA viruses of the order Nidovirales can infect either vertebrate (families *Arteriviridae* and *Coronaviridae*) or invertebrate hosts (*Mesoniviridae* and *Roniviridae*) [1, 2]. Examples of nidoviruses with high economic and societal impact are the arterivirus porcine reproductive and respiratory syndrome virus (PRRSV) [3] and the zoonotic coronaviruses (CoVs) causing severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) in humans [4-6]. While nidoviruses constitute a monophyletic group, their genome size differences are striking, with genomes ranging from 13–16 kb for arteriviruses to 20–21 kb for mesoniviruses and 25–34 kb for roniviruses and coronaviruses, which may reflect different stages of the largest genome expansion known to have occurred in RNA viruses [7].

Nidoviruses are characterized by their distinct polycistronic genome organization, the conservation of key replicative enzymes, and a common genome expression and replication strategy [8] (Figure 1). Their distinctive transcription mechanism involves the synthesis of a variable set of subgenomic (sg) mRNAs, which are 3′ co-terminal with the viral genome (reviewed in [9, 10]). In most nidoviruses, sg mRNAs and genome also share a common 5′ leader sequence. The synthesis of sg mRNAs (transcription) and genome RNA (replication) is performed by a poorly characterized replication-transcription complex (RTC) that is comprised of multiple protein subunits (reviewed in [11-13]) encoded in two large open reading frames (ORFs), ORF1a and ORF1b, which are translated from the nidoviral genomic RNA (Figure 1A). The two polyproteins (pp), pp1a and pp1ab, the latter resulting from ribosomal frameshifting during genome translation, are auto-catalytically processed by multiple cognate proteases, one of which (the 3C-like ($3CL^{pro}$) or main (M^{pro}) protease) is responsible for the large majority of cleavages [14]. Downstream of ORF1b, nidovirus genomes contain multiple smaller ORFs, known as the 3′ ORFs [7], which are expressed from the sg mRNAs described above.

During evolution, most conserved proteins of nidoviruses have diverged more extensively than those of organisms of the Tree of Life. In line with the principal function of each region, genome conservation increases from 3′ ORFs to ORF1a to ORF1b [7]. Accordingly, the 3′ ORF region encodes virion proteins and, optionally, accessory proteins that are predominantly group- or family-specific and mediate virus–host interactions [15, 16]. ORF1a encodes a variable number of proteins that include co-factors of the RNA-dependent RNA polymerase (RdRp) and other ORF1b-encoded enzymes, three hydrophobic proteins mediating the association of the RTC with membranes and the viral proteases [13, 17-19]. The latter group includes the $3CL^{pro}$, which is the only ORF1a-encoded enzyme conserved in all nidoviruses. In contrast, ORF1b is highly conserved and encodes different RNA-processing enzymes that critically control viral RNA synthesis

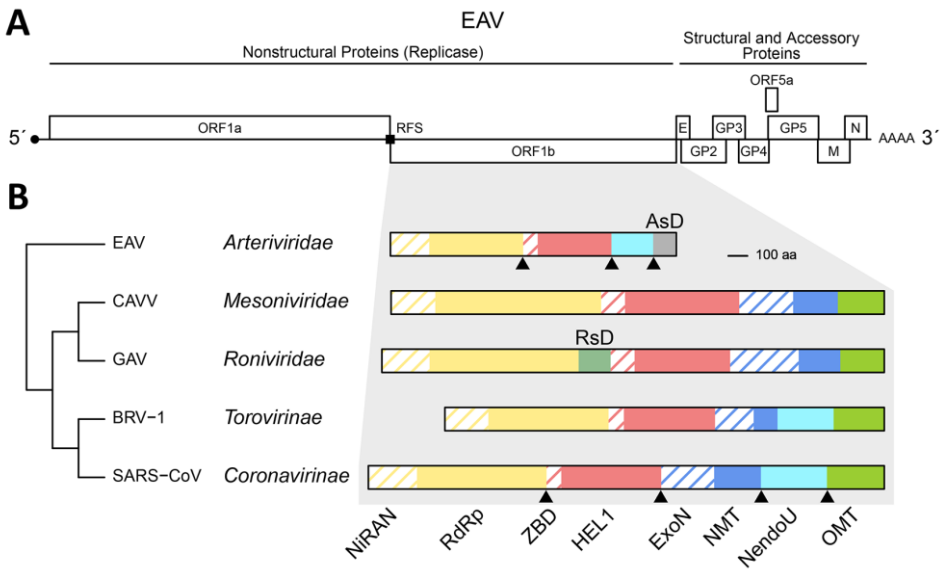


Figure 1 | Genome organization and ORF1b-encoded enzymes and domains of nidoviruses. (**A**) The genome organization of Equine arteritis virus (EAV), including replicase open reading frames (ORFs) 1a and 1b, and 3′ ORFs encoding structural proteins, is shown. Genomes of other nidoviruses employ similar organizations while they may vary in respect to size of different regions and number of 3′ ORFs. RFS, ribosomal frameshift site. (**B**) ORF1b size and domain comparison between the five nidovirus (sub)families is shown for EAV (*Arteriviridae*), Cavally virus (CAVV, *Mesoniviridae*), Gill-associated virus (GAV, *Roniviridae*), Breda virus (BRV-1, *Torovirinae*) and Severe acute respiratory syndrome coronavirus (SARS-CoV, *Coronavirinae*); see Supplementary Table 1 for details regarding these viruses. NiRAN, nidovirus RdRp-associated nucleotidyltransferase; RdRp, RNA-dependent RNA polymerase; ZBD, Zn-binding domain; HEL1, helicase superfamily 1 core domain; ExoN, exoribonuclease; N-MT, N7-methyltransferase; NendoU, nidovirus uridylate-specific endoribonuclease; O-MT, 2′-O-methyltransferase; AsD, arterivirus-specific domain; RsD, ronivirus-specific domain. Depicted is a simplified domain organization since most enzymes are part of multidomain proteins. Note that viruses of the *Torovirinae* subfamily encode a truncated version of N-MT. Triangles, established cleavage sites by $3CL^{pro}$ in two virus (sub)families; ORF1b-encoded proteins of other viruses may be proteolytically processed in a similar way. The order of emergence of different nidovirus (sub)families is presented by a simplified tree on the left.

(Figure 1B). These invariantly include the RdRp and a superfamily 1 helicase domain (HEL1), which is fused with a multinuclear Zn-binding domain (ZBD). RdRp and HEL1 are expressed as part of two different cleavage products residing next to each other in pp1ab [8]. The RdRp is believed to mediate the synthesis of all viral RNA molecules, while over the years the unwinding activity of the helicase was implicated in the control of replication, transcription, translation, virion biogenesis, and, most recently, post-transcriptional RNA quality control (reviewed in [20]). Among the lineage-specific proteins encoded in ORF1b are four enzymes. A 3′-5′ exoribonuclease (ExoN, in *Coronaviridae*, *Mesoniviridae* and *Roniviridae*) and an N7-methyltransferase (N-MT, in the *Coronavirinae* subfamily, *Mesoniviridae* and *Roniviridae*) constitute adjacent domains in the same pp1b cleavage product. They were implicated in RNA proofreading [19, 21, 22] and in 5′ end cap formation [23, 24], respectively. Downstream of this subunit, nidoviruses encode an

uridylate-specific endoribonuclease of unknown function (NendoU, in *Arteriviridae* and *Coronaviridae*) [25, 26] and/or a 2'-O-methyltransferase (O-MT) (in *Coronaviridae*, *Mesoniviridae* and *Roniviridae*), which was implicated in 5' end cap modification and immune evasion [23, 27-29]. All six ORF1b-encoded enzymes have distantly related viral and/or cellular homologs. Additionally, *Roniviridae* and *Arteriviridae* encode family-specific domains of unknown origin and function, RsD [30] and AsD [31, 32], respectively.

The protein subunit containing the RdRp domain is known as non-structural protein (nsp) 9 in *Arteriviridae* and nsp12 in *Coronaviridae* [8]. Its major ORF1b-encoded part varies in size from ~700 to ~900 amino acid residues and is N-terminally extended by a portion encoded in ORF1a. The RdRp-containing replicase subunit of nidoviruses thus seems to be larger than the characterized RdRps of other RNA viruses, which commonly comprise less than 500 amino acid residues [33]. RdRps adopt variations of an α/β fold (reviewed in [34]) and have characteristic conserved sequences (motifs). In nidoviruses, these motifs were mapped to the C-terminal one-third of the RdRp-containing protein [35, 36], whose tertiary structure is available only as a template-based model for SARS-CoV nsp12 [37, 38].

With one notable exception (N-MT; [24]), all ORF1b-encoded enzymes were initially identified by comparative genomic analysis involving viral and cellular proteins see [13, 31, 36, 39] and references there. These assignments were fully corroborated by their subsequent biochemical characterization [25, 26, 29, 40-45]. Furthermore, the (in)tolerance to replacement of active site residues as tested in reverse genetics studies of coronaviruses and arteriviruses in general correlated well with the observed enzyme conservation. Accordingly, the replacement of conserved residues of the nidovirus-wide conserved RdRp, ZBD and HEL1 were lethal [46-48], while virus mutants were crippled upon inactivation of ExoN, NendoU or O-MT enzymes [49-51], which are conserved in only some of the nidovirus families [30]. This correlation is noteworthy since it coherently links the results of the experimental characterization of a few nidoviruses in cell culture systems to evolutionary patterns that were shaped by natural selection in many hosts over an extremely large time frame. The fact that this correlation is evident for nidoviruses overall, rather than for separate families, indicates that nidovirus-wide comparative genomics provides sensible models to the functional characterization of the most conserved replicative proteins.

In the present study, we aimed to elucidate the domain organization, origin and function of the RdRp-containing proteins of nidoviruses by integrating bioinformatics, biochemistry and reverse genetics in a manner that was validated in many prior studies. Our extensive bioinformatics analysis revealed a novel domain, encoded upstream of the RdRp domain within the same cleavage product. It is conserved in all nidoviruses and has no apparent viral or cellular homologs, making it a second genetic marker for the order *Nidovirales*.

Based on results obtained using EAV and SARS-CoV, this domain was concluded to have an essential nucleotidylation activity and was named **ni**dovirus **R**dRp-**a**ssociated **n**ucleotidyltransferase (NiRAN). Its potential functions in nidovirus replication may include RNA ligation, protein-primed RNA synthesis, and the guanylyltransferase function that is necessary for mRNA capping.

MATERIALS AND METHODS

Virus genomes

Genomes of nidoviruses were retrieved from GenBank [52] and RefSeq [53] using Homology-Annotation hYbrid retrieval of GENetic Sequences (HAYGENS) tool http://veb.lumc.nl/HAYGENS. Genomes of all viruses were used to produce sequence alignments (see below), which were purged to retain only subsets of viruses representing the known diversity of each nidovirus family for downstream bioinformatics analyses. For the *Arteriviridae* and *Coronaviridae* families, one representative was drawn randomly from each evolutionary compact cluster corresponding to known and tentative species that were defined with the help of DEmARC 1.3 [54]. Twenty nine viruses of the family *Mesoniviridae* were clustered into six groups, whose intra- and inter-group evolutionary distance was below and above 0.075, respectively. One representative was chosen randomly from each of the six groups. For the *Roniviridae* family, two viruses, each prototyping a species, were used. To retrieve information about genomes, the SNAD program [55] was used. The final subsets include 30, 5, 10, 6 and 2 sequences representing all established and putative taxa of corona-, toro-, arteri-, mesoni- and roniviruses, respectively (Supplementary Table S1).

Multiple sequence alignments and secondary structure prediction

Multiple sequence alignments (MSAs) of proteins were generated using the Viralis platform [56] and assisted by HMMER 3.1 [57], Muscle 3.8.31 [58] and ClustalW 2.012 [59] programs in default modes. We have produced family-wide MSAs of nsp12 of coronaviruses, nsp9 of arteriviruses and their counterparts of mesoniviruses and roniviruses, whose borders have been tentatively mapped through limited similarity with known $3CL^{pro}$ cleavage sites of these viruses [60, 61]. They included NiRAN and RdRp domains delineated as described separately. For simplicity, we will refer to the proteins of mesoni- and roniviruses as nsp12t, with 't' standing for tentative, since the proteolytic cleavage of the replicase polyproteins of these viruses remains to be addressed in detail. Besides NiRAN and RdRp, we have also produced family-specific MSAs of three other nidovirus-wide conserved protein domains: $3CL^{pro}$, HEL1 and ZBD. Family-specific MSAs of the NiRAN domain were combined in a stepwise manner using the profile mode of

ClustalW with subsequent manual local refinement, which was limited and guided by results obtained using HHalign of the HH-suite 2.0.15 software [62, 63] when and if the two programs disagreed. The produced MSAs included one, two, three, four and five (sub)families, respectively, namely: *Coronavirinae* and *Torovirinae* (named CoTo), *Coronaviridae* and *Mesoniviridae* (CoToMe), *Coronaviridae*, *Mesoniviridae* and *Roniviridae* (CoToMeRo), *Coronaviridae*, *Mesoniviridae*, *Roniviridae* and *Arteriviridae* (CoToMeRoAr). The final MSA of NiRAN is presented in Supplementary Figure S1 in an annotated format and Supplementary Table S2 in FASTA format.

To reveal all local similarities between two MSAs, their profiles were compared using an align routine in HH-suite 2.0.15, whose results were visualized in a dot-plot fashion with the -dthr=0.25 and -dwin=10. Statistical significance of similarity was measured using % of confidence and expectation value (E). HH-suite calculates those for the best local hit in an MSA, regardless whether the latter was produced using the local or global mode of the program. Consequently, similarity of global MSAs may be underestimated. Based on family-specific MSAs of NiRAN and RdRp, the secondary structure of these domains was predicted using software Jpred 3 [64] and PSIPRED [65]. In both cases, the sequence with the least gaps was selected from the sequences forming the MSA. The prediction was made only for columns of the MSA in which the selected sequence does not contain gaps. The MSAs were converted into the final figure using ESPript [66].

Homology detection in protein databases

The obtained MSAs were converted into Hidden Markov Model (HMM) profiles or position-specific scoring matrices (PSSM) and used as queries to search for homologs in three different types of databases composed of: individual sequences (nr database, including GenBank CDS translations, RefSeq proteins, SwissProt, PIR and PRF [67]), profiles (PFAM A [68]), and protein 3D structures (PDB [69]). For GenBank scanning, HMMER 3.1 software [57] was used with the E-value cutoff of 10. To search for homologs among protein profiles and 3D structures, HHsearch of HH-suite 2.0.15 software [62, 63] and pGenTHREADER 8.9 software [70-72] were used, respectively.

In comparisons with the PDB (www.rcsb.org, [69]) using pGenTHREADER, RdRps of different viruses dominated the hit list for the best sampled nidoviruses, corona- and arteriviruses, and they were consistently present among the top hits for the two other families. Typically the similarity between a nidovirus query and a target encompassed the entire target and was limited to the C-terminal part of the query, with the N-terminal ~250 and ~350 amino acid residues remaining unmatched in arteriviruses and other nidoviruses, respectively (Figure 2A). Likewise, the C-terminal part of nsp9/nsp12/nsp12t

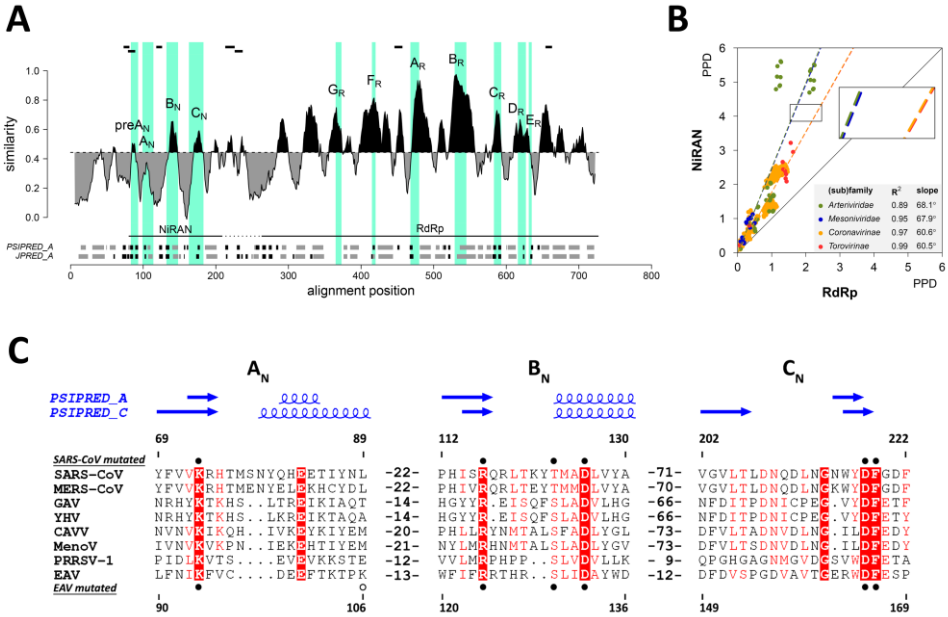


Figure 2 | Delineation and divergence of the NiRAN domain in the RdRp-containing proteins of nidoviruses. (**A**) Sequence variation, domain organization and secondary structure of the RdRp-containing protein of arteriviruses, and location of peptides identified by mass spectrometry after FSBG-labeling of arterivirus nsp9. Shown is the similarity density plot obtained for the multiple sequence alignment (MSA) of proteins including NiRAN and RdRp domains of arteriviruses. To highlight the regional deviation of conservation from that of the MSA average, areas above and below the mean similarity are shaded in black and grey, respectively. Uncertainty in respect to the domain boundary between NiRAN and RdRp is indicated by a dashed horizontal line. Positions of conserved sequence motifs of NiRAN and RdRp are indicated by vertical shading areas; motifs are labeled. Below the similarity density plot, secondary structure elements, predicted based on the arterivirus MSA using PSIPRED (PSIPRED_A) and Jpred 3 (JPRED_A), are presented in grey for α-helices, black for β-strands. (**B**) Relative scale of divergence of NiRAN versus RdRp in four different nidovirus (sub)families. Shown is scatter plot of PPDs of the NiRAN (y-axis) versus PPDs of RdRp (x-axis), which were calculated from the respective four PhyML trees. Dashed lines depict linear regressions fit in four differently highlighted PPD distributions, with its detail being magnified in the zoom-in; R^2 and slope values of the regressions are listed in the inset panel. The solid diagonal line corresponds to the matching rate of PPDs for the two domains and is provided for comparison. (**C**) MSA of the three conserved NiRAN motifs of eight representative nidoviruses and their predicted secondary structures. Absolutely conserved residues are in white font, while partially conserved residues are highlighted. Secondary structure predictions were made with PSIPRED [65] based on arterivirus (PSIPRED_A) or coronavirus (PSIPRED_C) MSAs. Residues mutated in recombinant SARS-CoV (*Coronaviridae*) non-structural protein (nsp) 12 and recombinant EAV (*Arteriviridae*) nsp9 are indicated by filled (conserved) and empty (control) circles, above and below the alignment respectively. Mutated residues D445A in EAV and K103A, D618A in SARS-CoV are not shown. Amino acid numbers above and below the alignment refer to SARS-CoV nsp12 and EAV nsp9, respectively. MERS-CoV, Middle East respiratory syndrome coronavirus (*Coronaviridae*); GAV (*Roniviridae*); YHV, yellow head virus (*Roniviridae*); CAVV (*Mesoniviridae*); MenoV, Meno virus (*Mesoniviridae*); PRRSV-1, porcine reproductive and respiratory syndrome virus, European genotype (*Arteriviridae*). For other abbreviations, see Figure 1.

matched the RdRp profiles of different virus families in PFAM [68] and an in-house database although this analysis was complicated by the presence of nidovirus sequences in the top-hit PFAM profile (see below). Based on these results we concluded that nsp9,

nsp12 and nsp12t contain N-terminal domains that are not part of canonical RdRps. This domain is referred to as NiRAN in this manuscript.

Evolutionary analyses

To estimate the divergence of NiRAN and RdRp, two analyses were conducted. Distribution of similarity density in MSAs of NiRAN and RdRp was plotted using R package Bio3D [73] under the conservation assessment method 'similarity', substitution matrix Blosum62 [74] and a sliding window of 11 MSA columns. Peaks of similarity were attributed to the known RdRp motifs G, F, A, B, C, D, E [35], or named and assigned to the newly recognized motifs of NiRAN, preA, A, B and C. Suffix R and N were added to motif labels of the RdRp and NiRAN domain, respectively. Reconstruction of phylogenetic trees of NiRAN and RdRp of different (sub)families was performed using PhyML 3.0, with the WAG amino acid substitution matrix, allowing substitution rate heterogeneity among sites (eight categories) and 1000 iterations of non-parametric bootstrapping [75]. Pairwise patristic distances (PPDs) between viruses were calculated from these trees using R package 'ape' [76]. They were used to assess relative rates of evolution of NiRAN and RdRp domains through the comparison of linear regressions, which were fit into the respective PPD distributions as implemented in R package 'stats' [77].

Protein expression and purification

Nucleotides 5256 to 7333 of the genome of the EAV Bucyrus strain were cloned into a pASK3 (IBA) vector essentially as described [47] to yield a construct that expresses nsp9 that is N-terminally fused to ubiquitin and tagged with hexahistidine at its C-terminus. Mutations were introduced according to the QuikChange protocol and verified by sequencing. Plasmids were transformed into *Escherichia coli* C2523/pCG1, which constitutively express the Ubp1 protease to remove the ubiquitin tag during expression and thereby generate the native nsp9 N-terminus. Cells were cultured in Luria Broth in the presence of ampicillin (100 µg/ml) and chloramphenicol (34 µg/ml) at 37°C until an OD_{600} >0.7. At this point protein expression was induced by the addition of anhydrotetracycline to a final concentration of 200 ng/ml and incubation was continued at 20°C overnight. Cell pellets were harvested by centrifugation and stored at −20°C until further use.

Proteins were batch purified by immobilized metal ion affinity chromatography using Co^{2+} Talon beads. In short, cell pellets were resuspended in lysis buffer (20 mM HEPES, pH 7.5, 10% glycerol (v/v), 10 mM imidazole, 5 mM β-mercaptoethanol) supplemented with 500 mM NaCl. Lysis was achieved by a 30-min incubation with 0.1 mg/ml lysozyme and five subsequent cycles of 10-s sonication to shear genomic DNA. Cellular debris was removed by centrifugation at 20 000 g for 20 min. The cleared supernatant was recovered and equilibrated Talon-beads were added. After 1 h of binding under agitation, beads were

washed four times for 15 min with a 25-times bigger volume of lysis buffer containing first 500 mM, than 250 mM, and finally twice 100 mM NaCl. In the end, proteins were eluted twice with lysis buffer containing 100 mM NaCl and 150 mM imidazole. Both fractions were pooled and dialyzed twice for 6 h or longer against an at least 100-fold bigger volume of 20 mM HEPES, pH 7.5, 50% glycerol (v/v), 100 mM NaCl, 2 mM DTT. All steps of the purification were performed at 4°C or on ice. All mutant proteins were expressed and purified in parallel with the wild-type protein used as reference in nucleotidylation assays. Protein concentrations were measured by absorbance at 280 nm using a calculated extinction coefficient of 93 170 $M^{-1}cm^{-1}$ and a molecular mass of 77 885 Da for wild-type nsp9. Typical protein yields were 5 mg/l culture and nucleotidylation activity was observed for at least 4 months if stored at −20°C at a concentration below 15 µM. Finally, the absence of the N-terminal ubiquitin tag was confirmed by mass spectrometry.

Nucleotidylation assay

Nucleotidylation assays were performed in a total volume of 10 µl containing, unless specified otherwise, 50 mM Tris, pH 8.5, 6 mM $MnCl_2$, 5 mM DTT, up to 2.5 µM nsp9 and 0.17 µM [α-^{32}P]NTP (Perkin Elmer, 3000 Ci/mmol). Furthermore, 12.5% glycerol (v/v), 25 mM NaCl, 5 mM HEPES, pH 7.5, and 0.5 mM DTT were carried over from the protein storage buffer. In preliminary experiments magnesium (1–20 mM) did not support nucleotidylation activity and was consequently not pursued further. Samples were incubated for 30 min at 30°C. Reactions were stopped by addition of 5 µl gel loading buffer (62.5 mM Tris, pH 6.8, 100 mM dithiothreitol (DTT), 2.5% sodium dodecyl sulphate (SDS), 10% glycerol, 0.005% bromophenol blue) and denaturing of the proteins by heating at 95◦C for 5 min. 12% sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) gels were run, stained with Coomassie G-250, and destained overnight. After drying, phosphorimager screens were exposed to gels for 5 h and scanned on a Typhoon variable mode scanner (GE healthcare), after which band intensities were analyzed with ImageQuant TL software (GE healthcare). The buffers used to find the pH optimum of the nucleotidylation reaction were MES (pH 5.5–6.5), MOPS (pH 7.0), Tris (pH 7.5–8.5) and CHES (pH 9.0–9.5) (20 mM).

To assess the chemical nature of the nucleotide-protein bond, the pH was temporarily shifted after product formation. To this end, 1 µl HCl or NaOH (both 1 M) was added before incubation at 95°C for 4 min. Afterward the original pH was restored by addition of the complementary base or acid, and samples were separated and analyzed as described.

FSBG labeling and mass spectrometry

Reaction mixtures were the same as described for the nucleotidylation assay with two modifications: radioactive nucleotides were replaced by up to 2 mM of the reactive

guanosine-5'-triphosphate (GTP) analog 5'-(4-fluorosulfonylbenzoyl)guanosine (FSBG) [78], of which the synthesis is described in supplementary Materials and Methods, and samples were incubated for 1 h at 30°C to increase the ratio between labeled and unlabeled protein. Subsequently, the protein (20 µg) was reduced by addition of 5 mM DTT and denatured in 1% SDS for 10 min at 70°C. Next, the samples were alkylated by addition of 15 mM iodoacetamide and incubation for 20 min at RT. Next, the protein was applied to a centrifugal filter (Millipore Microcon, MWCO 30 kDa) and washed three times with NH_4HCO_3 (25 mM) before a protease digestion was performed with 2 µg trypsin in 100 µl NH_4HCO_3 overnight at RT. Recovered peptides were treated with 50 mM NaOH for 25 min, desalted using Oasis spin columns (Waters) and finally analyzed by on-line nano-liquid chromatography tandem mass spectrometry on an LTQ-FT Ultra (Thermo, Bremen, Germany). Tandem mass spectra were searched against the Uniprot database, using mascot version 2.2.04, with a precursor accuracy of 2 ppm and product ion accuracy of 0.5 Da. Carbamidomethyl was set as a fixed modification, and oxidation, N-acetylation (protein N-terminus) and FSBG were set as variable modifications.

Label release

For analysis of the released nucleotides, 350 pmol of nsp9 were nucleotidylated with [α-^{32}P]nucleoside-5'-triphosphates ([α-^{32}P]NTPs) as described above for 1 h at 30°C. After the reaction free NTPs were removed by buffer exchange and extensive washing with the help of a centrifugal filter (Millipore ultrafree-0.5, MWCO 10 kDa). Protein was precipitated with a five times greater volume of acetone overnight at −20°C. The resulting pellet was resuspended in 20 mM Tris, pH 8.5, 100 mM NaCl. Equal amounts of the solutions were incubated at 95°C for 4 min after addition of HCl or NaOH (1 M). Samples were adjusted to their original pH and spotted onto polyethylenimine cellulose thin layer chromatography plates, which were developed in 80% acetic acid (1 M), 20% ethanol (v/v), 0.5 M LiCl. Plates were dried and phosphorimaging was performed as described above. Non-radioactive nucleotide standards were run on each plate and visualized by UV-shadowing to allow the identification of the radioactive products.

Reverse genetics of EAV

Alanine-coding mutations for conserved and control residues were introduced into full-length cDNA clone pEAV211 [79] using appropriate shuttle vectors and restriction enzymes. The presence of the mutations was confirmed by sequencing. pEAV plasmid DNA was *in vitro* transcribed with the mMessage-mMachine T7 kit (Ambion), and the synthesized RNA was transfected into BHK-21 cells after LiCl precipitation as described previously [80]. Virus replication was monitored by immunofluorescence microscopy until 72 h post transfection (p.t.) using antibodies directed against nsp3 and N protein as

described [81] and by plaque assays [80] using transfected cell culture supernatants, to monitor the production of viral progeny.

Sequence analysis of the nsp9-coding region was performed to either verify the presence of the introduced mutations or to monitor the presence of (second site) reversions. For this purpose, fresh BHK-21 cells were infected with virus-containing cell culture supernatants and total RNA was extracted with Tripure Isolation Reagent (Roche Applied Science) after appearance of cytopathic effect (CPE) (typically at 18 h post infection (p.i.)). EAV-specific primers were used to reverse transcribe RNA and PCR amplify the nsp9-coding region (nt 5256–7333). RT-PCR fragments of the EAV genome were sequenced after gel purification and sequences compared to those of the respective RNA used for transfection.

Reverse genetics of SARS-CoV

Mutations in the SARS-CoV nsp12-coding region were engineered in prSCV, a pBeloBac11 derivative containing a full-length cDNA copy of the SARS-CoV Frankfurt-1 sequence [82] by using 'en passant recombineering' as described in Tischer *et al.* [83]. The (mutated) BAC DNA was linearized with NotI, extracted with phenol–chloroform, and transcribed with T7 RNA Polymerase (mMessage-mMachine T7 kit; Ambion) using an input of 2 µg of BAC DNA per 20-µl reaction. Viral RNA transcripts were precipitated with LiCl according to the manufacturer's protocol. Subsequently, 6 µg of RNA were electroporated into 5 10^6 BHK-Tet-SARS-N cells, which expressed the SARS-CoV N protein following 4 h induction with 2 µM doxycycline as described previously [84]. Electroporated BHK-Tet-SARS-N cells were seeded in a 1:1 ratio with Vero-E6 cells. Viral protein expression and the production of viral progeny was followed until 72 h p.t. by immunofluorescence microscopy using antibodies directed against nsp4 and N protein and by plaque assays of cell culture supernatants, respectively (both methods were described previously in Subissi *et al.* [84]). All work with live SARS-CoV was performed inside biosafety cabinets in a biosafety level 3 facility at Leiden University Medical Center.

For sequence analysis of viral progeny, fresh Vero-E6 cells were infected with harvests from viable mutants taken at 72 h p.t., and SARS-CoV RNA was isolated 18 h p.i. using TriPure Isolation Reagent (Roche Applied Science) as described in the manufacturer's instructions. Random hexamers were used to prime the RT reaction, which was followed by amplification of the nsp12-coding region (nt 13398–16166) by using SARS-CoV-specific primers. RT-PCR products were sequenced to verify the presence of the introduced mutations.

RESULTS

Delineation of a novel, unique domain that is conserved upstream of the RdRp in polyproteins of all nidoviruses

Inspection of the intra-family sequence conservation for (sub)family-specific MSAs of nsp9, nsp12 and nsp12t (see 'Materials and Methods' section for technical details) using density similarity plots (Supplementary Figure S2) confirmed the association of characteristic RdRp motifs with some of the most prominent conservation peaks, located in the C-terminal half of nsp9 and nsp12 (RdRp domain). For nsp12t, similar conclusions could be drawn although the conservation profiles of these viruses, especially roniviruses, were of lesser resolution due to the overall higher similarity that was the result of the limited virus sampling and divergence. Importantly, also the N-terminal half of nsp9 and nsp12 (NiRAN domain) included a few above-average conservation peaks although the overall conservation was evidently highest around the established RdRp motifs (Figure 2A; Supplementary Figure S2). Likewise, NiRAN compared to RdRp accepted two-to-three times more substitutions in four nidovirus (sub)families (Figure 2B). In this comparison, slopes of the four PPD distributions were strikingly similar, particularly in the pairs of the *Coronavirinae* and *Torovirinae* (60.6 and 60.5, respectively) and the *Mesoniviridae* and *Arteriviridae* (67.9 and 68.1). Thus, NiRAN must have evolved under similar constraints in different lineages of nidoviruses, which is compatible with a common function of this domain.

Next, we investigated the relation of the NiRAN domains of the four different families by pairwise profile–profile comparisons using HHalign in local mode (see Supplementary Figure S3 and Figure 3 for all results and a selection of thereof, respectively). This analysis revealed strong support (~98% confidence and E = 7.7e-09–1.7e-08) for the similarity between NiRANs of coronavirus/torovirus nsp12 and mesonivirus nsp12t, and moderate support (~21–30% confidence and E = 0.00051–0.00091) for the similarity between the respective domains of mesoni- and roniviruses. Based on these observations, we have aligned the NiRAN domain of coronavirus nsp12 and mesonivirus nsp12t using the profile mode of ClustalW, with the MSA being slightly adjusted taking into account the HHalign-mediated results. This MSA of two families was superior compared to each of the two family-specific MSAs with respect to its similarity to the MSA of roniviruses (~54–75% confidence and E = 0.00011–0.00049). Consequently, the ronivirus MSA was added to the MSA of corona/toro- and mesoniviruses to generate an MSA of the NiRAN of the three families, hereafter called ExoN-encoding nidoviruses with reference to the domain that distinguishes this group from arteriviruses (Figure 1B). In the above HHsearch local alignments, almost the full-length NiRAN domains were aligned.

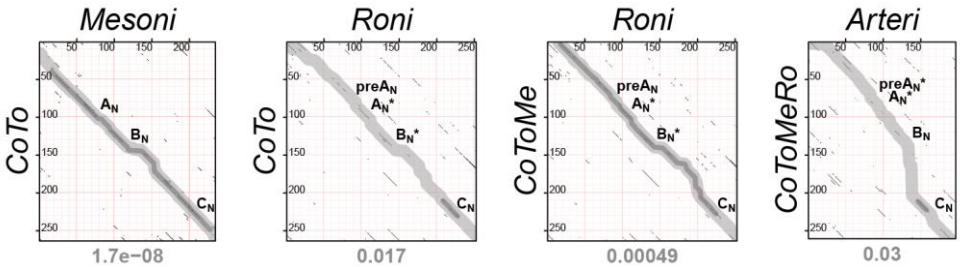


Figure 3 | Establishing sequence conservation between NiRAN domains of different (sub)families. Shown are four pairwise dot-plots that compare HMM profiles of NiRAN domains of different origins using HHalign. For the entire set of dot-plots generated, please see Supplementary Figure S3. First, third and fourth plots correspond to steps used to produce the nidovirus-wide NiRAN MSA (Supplementary Figure S1), while the second plot is shown for comparison. Coordinates of query and target HMMs are presented on y-axis and x-axis, respectively. All local similarities between two profiles are depicted as black dots. Transparent fat dark and light gray lines on the dot-plot show paths of HHalign alignments, obtained in local and global modes, respectively. The E-value of the top local alignment is specified below each dot-plot. In the profile–profile alignment produced in global mode, conserved amino acids of NiRAN motifs may have been properly aligned or not. If conserved residues of a motif were aligned, the corresponding region of the alignment path is labeled with the respective motif name without an asterisk. If the misalignment of conserved residues was limited to a shift of one or two residues (HMM–HMM alignment columns), the corresponding region of the alignment path is labeled with the respective motif name plus an asterisk.

In contrast to the above observations, the support for similarity between the NiRAN MSAs of arteriviruses and ExoN-encoding nidoviruses, separately or combined, in our HHalign-based analysis was relatively weak (E = 0.03–0.4), particularly with respect to confidence (1.5% or worse). This could be due to the similarity being recognized only in a small C-terminal region. This experience prompted us to compare conserved motifs and predicted secondary structures of the domains of these families (Supplementary Figures S1 and S2). Ten residues were found to be invariant in the conserved NiRAN of the ExoN-encoding nidoviruses. They map to three motifs designated A_N (with a K-x[6–9]-E pattern in ExoN-encoding nidoviruses), B_N (R-x[8–9]-D) and C_N (T-x-DN-x4-G-x[2,4]-DF), respectively, with motifs B_N and C_N representing the most prominent conservation peaks of this domain in coronaviruses (Supplementary Figure S2). Remarkably, similar conserved motifs are present in the NiRAN of arteriviruses (Figure 2A and C), where B_N and C_N again occupy the two most prominent peaks (Supplementary Figure S2). The three motifs are similarly positioned relative to the ORF1a/ORF1b frameshift signal in all nidoviruses, and, importantly, they were aligned in arteriviruses and the ExoN-encoding nidoviruses using HHalign in global mode (Figure 3, rightmost plot). Specifically, all four invariant residues of motifs A_N and B_N of ExoN-encoding nidoviruses are also conserved in arteriviruses although with slightly smaller distances separating the two residues of each pair (Supplementary Figure S1 and Figure 2C). In the most highly conserved motif C_N, the aspartate-phenylalanine dipeptide and likely glycine (the only deviating arginine at this position in the lactate dehydrogenase-elevating virus isolate U15146 may result from a

sequencing error) are absolutely conserved among nidoviruses while the other invariant residues of ExoN-encoding nidoviruses appear to have been replaced by similar residues in arteriviruses. Additionally, there is a good agreement between the predicted secondary structure for the domains of arteriviruses and ExoN-encoding nidoviruses, particularly in the area encompassing the three sequence motifs as well as regions immediately upstream of motif A_N (named preA_N motif) and downstream of motif C_N (Supplementary Figure S1). In ExoN-encoding nidoviruses, motifs B_N and C_N are separated by a variable region of 40–60 amino acid residues that does not include absolutely conserved residues, while in arteriviruses motifs B_N and C_N are adjacent. Based on these observations, we concluded that nsp9, nsp12 and nsp12t contain the NiRAN domain, which is conserved in all nidoviruses, although we acknowledge that the support for the conservation of different motifs between different nidovirus (sub)families is not equally strong. Also, we noted that, at this stage, it was not possible to precisely define the C-terminal border of the NiRAN domain. NiRAN and RdRp may thus, be adjacent or separated by another small domain of variable size in different nidoviruses (Supplementary Figure S2).

To gain insight into the origin and function of the NiRAN domain, we compared MSA-based profiles of this domain and its individual motifs of different nidovirus families and the entire order with the PFAM, GenBank, Viralis DB and PDB databases. As a control, we used the HMM profiles of four other domains that are conserved in all nidoviruses, $3CL^{pro}$, RdRp, ZBD and HEL1. We expected to find hits to either other nidovirus proteins, if NiRAN would have emerged by duplication or non-nidovirus proteins, if the NiRAN ancestor would have been acquired from an external source. None of the database scans involving the NiRAN retrieved a non-nidovirus hit whose E-value was better than 0.065 for HMMER and 1.3 for the HH-search program from HH-suite (Figure 4) and none of these hits had sequences similar to the motifs of the NiRAN. In contrast, statistically significant hits with virus and/or host proteins were identified for the nidoviral control proteins either in both or one of the scans; according to annotation, at least some of these hits were true positives in the functional and/or structural sense. Likewise, in scans of the PDB using pGenTHREADER, all top hits for the NiRAN of the four virus families had low support (P = 0.014 or worse) with no match of the conserved motifs. In contrast, top hits for four RdRp queries were supported with P-values of 0.0003 or better and targeted RdRps of other viruses, at least for arteri- and coronavirus queries.

EAV nsp9 has Mn^{2+}-dependent nucleotidylation activity with UTP/GTP preference

Since we could not identify any homologs of the NiRAN domain whose prior characterization would facilitate the formulation of a hypothesis about its function, we have reviewed the available information about nidovirus genome organization and

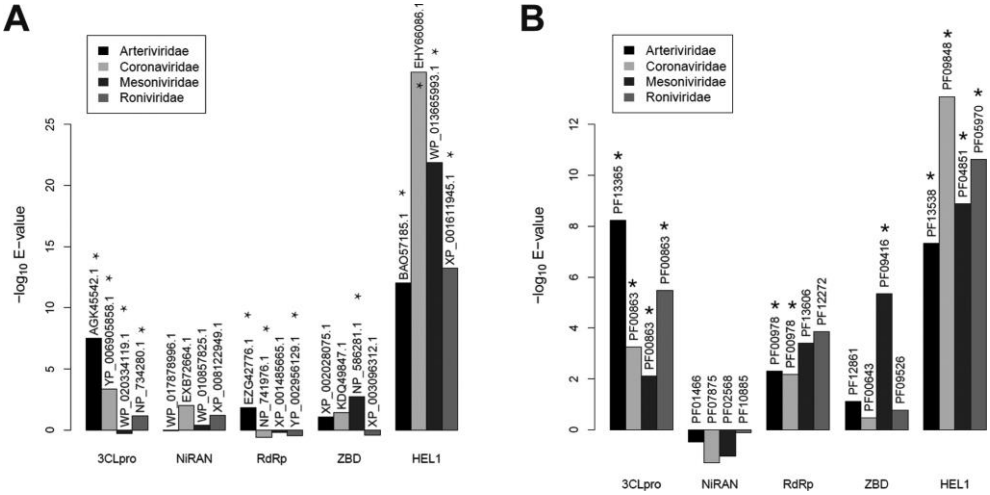


Figure 4 | Comparison of nidovirus-wide conserved domains with sequence databases. Shown are histograms depicting E-values of the best non-nidovirus hits obtained during HMMER-mediated profile-sequence (**A**) and HHsearch-mediated profile–profile (**B**) searches of the GenBank and PFAM A databases, respectively, using MSA profiles of five nidovirus-wide conserved domains encoded by four nidovirus families. The identity of the non-nidoviral top-hit in the respective databases is specified. Stars indicate hits whose homologous relationship with the respective query is also supported by the functional and/or structural annotation of the respective targets.

replicative enzymes, and the results described above. The data were most compatible with the hypothesis that this domain is an RNA processing enzyme, in view of (i) the abundance of RNA processing enzymes in the ORF1b-encoded polyprotein (Figure 1B); and (ii) the profile of invariant residues, composed of aspartate, glutamate, lysine, arginine and phenylalanine (and possibly glycine) (Figure 2C), the first four of which are among the most frequently employed catalytic residues [85]. Since the domain is uniquely conserved in nidoviruses, we hypothesized that its activity might work in concert with that of another, similarly unique RNA processing enzyme. At the time of this consideration, the NendoU endoribonuclease of nidoviruses was believed to be such an enzyme [25] (assessment revised in 2011, [30]). Consequently, we reasoned that a ligase function would be a natural counterpart for the endoribonuclease (NendoU), as observed in many biological processes, and would fit in the functional cooperation framework outlined in our previous analysis of the SARS-CoV proteome [39]. This hypothesis was also compatible with the predicted α/β structural organization of NiRAN (Supplementary Figure S1) and the lack of detectable similarity between NiRAN and the highly diverse nucleotidyltransferase superfamily, to which nucleic acid ligases belong. This superfamily is known to include groups that differ even in the most conserved sequence motifs, especially in proteins of viral origin [86, 87]. Based on mechanistic insights obtained with other ligases, we expected that the conserved lysine might be the principal catalytic residue of the NiRAN domain.

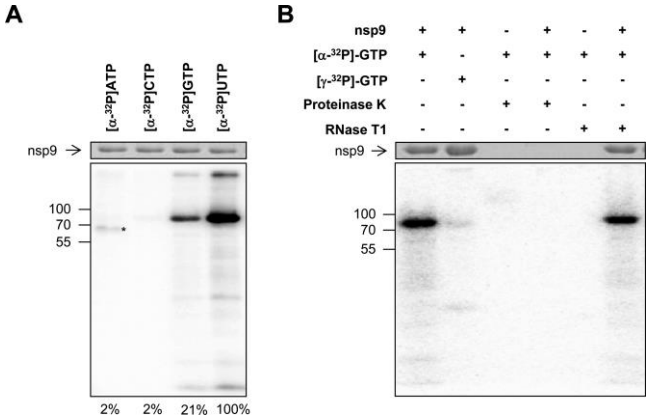


Figure 5 | EAV nsp9 has nucleotidylation activity. Purified recombinant EAV nsp9 (78 kDa) was incubated with the indicated ^{32}P-labeled NTP in the presence of $MnCl_2$. After denaturing SDS-PAGE, reaction products were visualized by Coomassie brilliant blue staining (top panels) and phosphor imaging (bottom panels). Positions of molecular weight markers are depicted on the left in kDa. (**A**) Uridylylation and guanylylation activity as revealed by covalent binding of the respective radioactive nucleotide to nsp9. Note that the protein indicated with an asterisk likely is an *Escherichia coli*-derived impurity reacting with ATP. Relative band intensities are shown at the bottom. (**B**) Guanylylation was distinguished from RNA polymerization by incubating the products generated during the nucleotidylation assay with proteinase K (1 mg/ml) or with RNase T1 (0.5 U), which cleaves single-stranded RNA after G residues, for 30 min at 37°C.

To detect this putative NTP-dependent RNA ligase activity, we took advantage of the universal ligase mechanism, which can be separated into three steps [88]. First, an NTP molecule, typically adenosine triphosphate (ATP), is bound to the enzyme's binding pocket, and a covalent bond is established between the nucleotide's α-phosphate, nucleoside-5′-monophosphate (NMP) and the side chain of either lysine or histidine, while pyrophosphate is released. Since this protein–NMP is a true, temporarily stable intermediate, it can be readily detected by biochemical methods. In contrast, demonstration of the following two steps, NMP transfer to the 5′ phosphate of an RNA substrate and subsequent ligation of a second RNA molecule under release of the NMP, depends on the availability of target RNA sequences whose identification is often not as straightforward. Thus, we first assessed our hypothesis by testing the covalent binding of a nucleotide, known as nucleotidylation.

To this end, recombinant EAV nsp9 was expressed in *E. coli*, purified, and incubated with each of the four NTPs, which were ^{32}P-labeled at the α-position. Samples were analyzed using denaturing SDS-PAGE to discriminate between covalent and affinity-based nucleotide binding. As can be seen in Figure 5A, we could indeed detect a radioactively labeled product with a mobility comparable to that of nsp9 in the presence of GTP and uridine-5′-triphosphate (UTP). To verify that this labeled band corresponded to a protein and did not result from 3′ end labeling of co-purified *E. coli* RNA or polyG synthesis by the RNA polymerase residing in the C-terminal domain of nsp9, guanylylation was followed by

the addition of either proteinase K or RNase T1, which cleaves single-stranded RNA after G residues. As expected, only protease treatment removed the band while incubation with RNase T1 had no effect on the product (Figure 5B). The same result was obtained after uridylylation using RNase A, which cleaves after pyrimidines in single-stranded RNA (data not shown). Furthermore, as the use of GTP labeled in the γ-position did not result in a radioactive product, we conclude that this phosphate, in agreement with the general nucleotidylation mechanism, is released during the reaction (Figure 5B).

Unexpectedly, we observed a marked substrate specificity of nsp9 for UTP, which resulted in the accumulation of five times more enzyme–nucleotide complex than observed with GTP. In contrast, we observed no covalent binding with ATP or cytidine-5'-triphosphate (CTP) as substrates (Figure 5A). The observed substrate preferences are remarkable for two reasons. First, since both UTP and GTP are present in significantly lower concentrations under physiological conditions than ATP [89] and are in general not used as primary energy source, it suggests that the identity of the base, rather than the energy stored within the phosphodiester bonds, may be critical for a subsequent step in the reaction pathway. This implies that reaction pathways other than RNA ligation, which predominantly utilizes ATP, must be considered. Second, the selective utilization of only one pyrimidine and one purine substrate raised questions about the nature and number of active sites involved, for instance, whether both nucleotides bind to separate binding sites or utilize different catalytic residues within the same binding site. Unfortunately, there are no crystal structures for any of the nidovirus nsp9/nsp12/nsp12t subunits available to date, which might have been used to resolve this matter in docking studies.

To characterize the NTP binding further, we compared the pH dependence of both activities. Interestingly, while the relative activities below pH 8.5 were identical with both substrates, the relative guanylylation activity was exceedingly higher than uridylylation at a pH above 8.5 (Figure 6A). To exclude that the observed pattern is due to a difference in the metal ion requirement, we determined the optimal manganese concentration for nucleotidylation with both substrates. As is apparent from Figure 6B, both activities share the same broad optimum between 6 and 10 mM $MnCl_2$. This result made it unlikely that manganese oxidation and a concomitant decrease of available Mn^{2+} ions, as we observed at a pH above 9.0, would selectively favour the utilization of one of the two substrates. The observed difference between guanylylation and uridylylation with regard to its pH optimum may thus be genuine. For instance, this slightly broadened or − more likely − shifted pH optimum of guanylylation may be the result of a GTP-induced spatial reorientation of amino acid side chains in the vicinity of the catalytic residue and a concomitant alteration of its pK_a. Alternatively, it may also be explained by the two substrates using different binding sites. These possibilities were partially addressed in the experiments described in the subsequent sections.

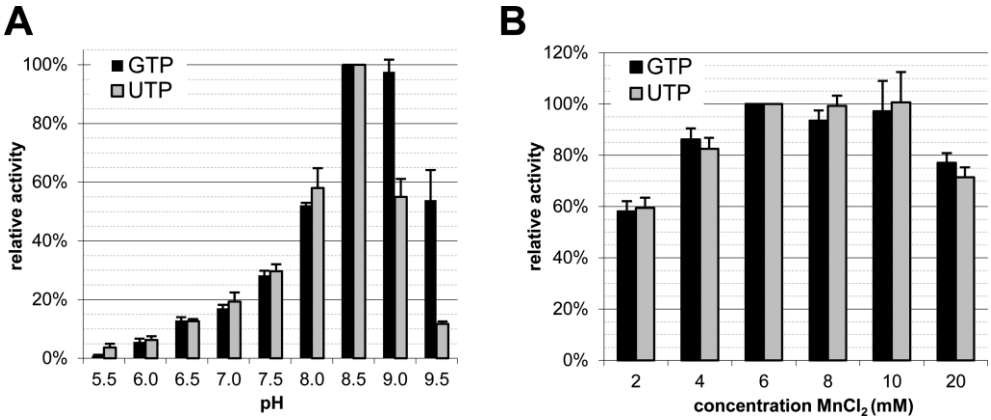


Figure 6 | EAV nsp9 guanylylation has a slightly broader or shifted pH optimum compared to uridylylation while the metal ion requirement is identical. (**A**) The pH optimum in the range from 5.5 to 9.5 was determined using the buffers listed in 'Materials and Methods' section. (**B**) Assessment of the optimal $MnCl_2$ concentration for nucleotidylation. Error bars represent the standard deviation of the mean based on three independent experiments.

FSBG labeling of nsp9 suggests the presence of a nucleotide binding site in the NiRAN domain

To verify that the newly discovered nucleotidylation activity is associated with the NiRAN domain, we first sought to establish the presence of the expected nucleotide binding site. To this end, we replaced the substrate in the nucleotidylation assay with the reactive guanosine analog 5'-(4-fluorosulfonylbenzoyl)guanosine (FSBG) (Supplementary Figure S4A) [78]. Depending on the exact shape of the nucleotide binding pocket this compound may be suitable for binding and reacting with any nucleophile within the pocket, leaving behind a stable sulfonylbenzoyl tag that can be readily detected by mass spectrometry. In this way, residues that are lining the binding site can be identified. However, because the points of attack of FSBG (sulfonyl group sulfur) and GTP (α-phosphorus) are spatially separated (~4A˚, Supplementary Figure S4A and B), these residues are not necessarily of biological relevance to nucleotidylation but rather are indicative of the local neighborhood of the nucleotidylation reaction.

After analysis of the nucleotidylation reaction mixture by mass spectrometry, seven modified peptides representing five distinct nsp9 regions could be assigned: three in (the vicinity of) the NiRAN domain and two in the RdRp domain (Figure 2A and Supplementary Figure S4C). In agreement with previously published results [78], we found only lysine and tyrosine residues to be modified, as these are thought to provide the chemically most stable bonds. The selectivity of the modification was evident from the fact that only seven lysine and tyrosine residues served as nucleophile for the reaction. Furthermore, we identified all these peptides in independent experiments using FSBG concentrations

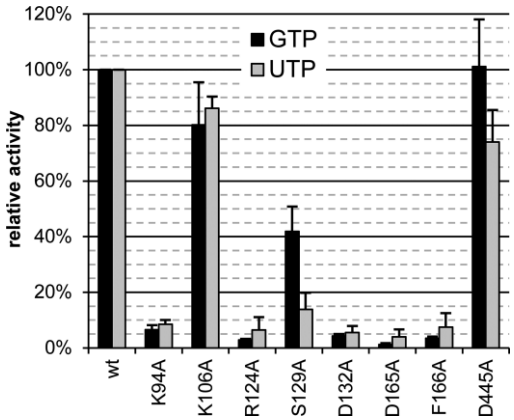


Figure 7 | Conserved NiRAN residues are essential for the nucleotidylation activity. Alanine substitution of conserved NiRAN residues dramatically decreased the nucleotidylation activity of nsp9. In contrast, mutation of the non-conserved K106 in the NiRAN domain or the conserved D445 in the RdRp domain had only a mild effect on activity. Error bars represent the standard deviation of the mean based on three independent experiments.

ranging from 25 µM to 2 mM. Within this range, a concentration of 100 µM was sufficient to detect all seven peptides. Together this strongly suggests that the reaction with FSBG only occurred after binding to a specific site(s) and did not originate from random collisions. Furthermore, the two modified residues in the EAV RdRp are located in either a predicted α-helix or in a loop not far upstream and downstream of the A_R and E_R motifs, respectively, which are involved in NTP binding in other, better characterized RdRps. The five modified residues in the EAV NiRAN domain are poorly conserved in related arteriviruses and are located in the vicinity of one of the three major motifs in either a predicted loop region (1 residue) or a β-strand (4 residues). These findings are compatible with the expected properties of the FSBG modification that may label any nucleophile within a 4 A˚ distance from the NTP-binding site(s). We therefore conclude that the peptides identified in this experiment reflect the presence of a nucleotide binding site within the RdRp required for RNA synthesis and a second binding site that is located in the NiRAN domain, which could serve for nucleotidylation.

Conserved residues of the NiRAN domain but not of the RdRp domain are required for nucleotidylation activity

In a next step, we examined the importance of conserved NiRAN residues for the guanylylation and uridylylation activities by characterization of alanine substitution mutants of several residues, including five invariant residues, in recombinant EAV nsp9. Notably, none of these mutations significantly reduced expression or stability (data not shown), indicating that they are most likely compatible with the protein's structure. Subsequent characterization demonstrated that all conserved NiRAN residues that were probed (Figure 2, Table 1) are important for nucleotidylation activity, as their replacement

with alanine led, with the exception of S129A, to a drop to below 10% of wild-type protein activity. In contrast, alanine substitution of a non-conserved N-terminal residue (K106A) as well as of a conserved residue in the RdRp domain (D445A of motif A_R), which is known to be essential for the polymerase activity in other RNA viruses [34], had only a mild effect, preserving at least 75% of the activity (Figure 7). Thus, we concluded that the identified sequence motifs in the EAV nsp9 NiRAN domain are functionally connected to the nucleotidylation activity. Whether the decrease in activity was due to a loss of affinity, impairment of catalysis or both remains to be established. In addition, as the level of remaining activity (again with exception of the S129A mutant) did not depend on the substrate used, both guanylylation and uridylylation are likely catalyzed by the same active site.

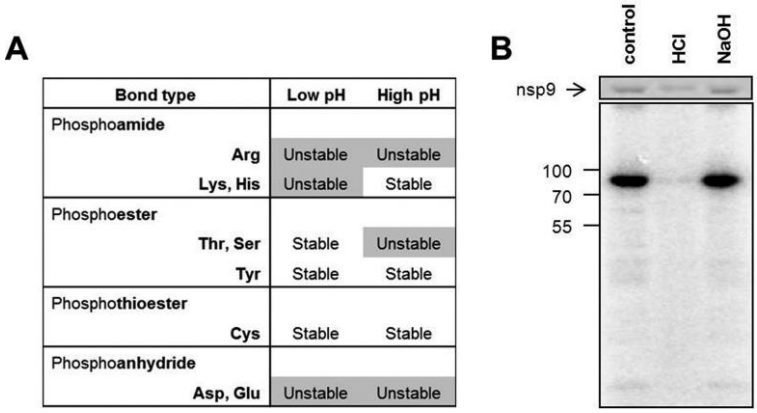
Table 1 | Reverse genetics analysis of EAV nsp9 and SARS-CoV nsp12 mutants.

	Motif	Mutant	Mutation	Virus titers (PFU/ml at 16-18 h p.t.)	nsp9/nsp12 sequence of P1 virus[a]
		wt		$1{\cdot}10^7$, $2{\cdot}10^8$	n.d.
	A_N	K94A	AAA→GCA	<20, <20	Reversion
	Non-conserved	K106A	AAA→GCA	$3{\cdot}10^5$, $2{\cdot}10^6$	GCA
	B_N	R124A	CGU→GCU	<20, <20	Reversion
EAV	B_N	S129A	UCG→GCG	$1{\cdot}10^4$, $5{\cdot}10^3$	Reversion
	B_N	D132A	GAU→GCU	$3{\cdot}10^4$, $6{\cdot}10^3$	Reversion
	C_N	D165A	GAU→GCU	$3{\cdot}10^3$, $1{\cdot}10^4$	Reversion
	C_N	F166A	UUU→GCU	<20, <20	n.a.
	A_R	D445A	GAC→GCC	<20, $1{\cdot}10^4$	Reversion
		wt		$4{\cdot}10^6$, $3{\cdot}10^5$	n.d.
	A_N	K73A	AAG→GCC	<20, <20	n.a.
	Non-conserved	K103A	AAG→GCA	<20, <20	GCA
	B_N	R116A	CGU→GCU	<20, <20	n.a.
SARS-CoV	B_N	T123A	ACA→GCU	$1{\cdot}10^5$, $4{\cdot}10^5$	GCU
	B_N	D126A	GAU→GCG	<20, <20	n.a.
	C_N	D218A	GAU→GCU	<20, <20	n.a.
	C_N	F219A	UUC→GCG	$2{\cdot}10^4$, $8{\cdot}10^2$	GCG
	A_R	D618A	GAU→GCG	<20, <20	n.a.

[a]Virus-containing supernatants were collected at 72 h p.t. and subsequently used for re-infection of fresh BHK-21 (EAV) or Vero-E6 (SARS-CoV) cells. Total RNA was isolated after appearance of CPE and nsp9/nsp12 coding regions were sequenced. All results were confirmed in a second independent experiment. n.d., not done; n.a., not applicable (non-viable phenotype).

In contrast to these results, the mutation at position S129, the only targeted residue that is fully conserved in arteriviruses but may be replaced by threonine in other nidoviruses,

exhibited a slightly different effect on guanylylation and uridylylation. Mutant S129A displayed an intermediate activity when using GTP but was almost as deficient as mutants of the nidovirus-wide conserved residues when UTP was used as substrate (Figure 7). This finding may indicate that S129 is specifically involved in the hydrogen bond network between protein and UTP. Alternatively, as the covalent binding of the nucleotide occurs via a nucleophilic attack on the α-phosphate, this serine may in principle be suitable to play this role. Although to our knowledge nucleic acid ligases typically employ lysine and rarely histidine as catalytic residues [88, 90], we cannot exclude that uridylylation occurs via this S129 while guanylylation utilizes another amino acid.



Bond type	Low pH	High pH
Phosphoamide		
Arg	Unstable	Unstable
Lys, His	Unstable	Stable
Phosphoester		
Thr, Ser	Stable	Unstable
Tyr	Stable	Stable
Phosphothioester		
Cys	Stable	Stable
Phosphoanhydride		
Asp, Glu	Unstable	Unstable

Figure 8 | A phosphoamide bond is formed between nsp9 and the guanosine phosphate. (**A**) Chemical stability of different phosphoamino acid bonds. Adapted from [91]. (**B**) The protein was labeled with [α-^{32}P]GTP and subsequently incubated at pH 8.5 (control) or under acidic or alkaline conditions. Reaction products were visualized after denaturing SDS-PAGE by Coomassie brilliant blue staining (top panel) and phosphor imaging (bottom panel). Size markers are depicted on the left in kDa.

Nucleotidylation occurs via the formation of a phosphoamide bond

In order to identify which type of amino acid is the catalytic residue involved in nucleotidylation, we probed the chemical stability of the bond formed between enzyme and nucleotide. To this end, we subjected the nucleotidylation product to either a higher or a lower pH for 4 min, while the protein was heat denatured. The loss of the radioactive label under acidic or alkaline conditions is an indicator for the type of bond that is formed (Figure 8A) [91]. As evident from Figure 8B, the bond between guanosine phosphate and nsp9 was acid-labile but stable under alkaline conditions, which is indicative of a phosphoamide bond originating from either a lysine or histidine. This result was also confirmed for uridylylation (data not shown), excluding a direct role for S129 in the attachment of the uridine phosphate. Since there is no conserved histidine present in the

NiRAN domain, K94 is the most likely candidate within this domain to fulfill the role of catalytic residue.

Guanosine and uridine phosphates may be attached via different phosphate groups

So far we have demonstrated that guanylylation and uridylylation are essentially equally sensitive to replacement of NiRAN residues, share the same metal ion requirements, and both rely on the formation of a phosphoamide bond. We therefore concluded that there is only one active site responsible for nucleotidylation, which allows utilization of both substrates. Interestingly, if this were true, discrimination of GTP and UTP against ATP and CTP would be solely based on the presence of an oxygen at C6 of GTP and C4 of UTP. However, given the pronounced size difference between UTP and GTP, the positions of both substrates within the binding site are unlikely to be equivalent. In principle, two binding scenarios are possible. First, the ribose and phosphate moieties of both nucleotides could occupy the same position within the binding site, for example by forming hydrogen bonds via the ribose's 2' and 3' hydroxyl groups and charge interactions between the protein and the phosphates. Yet, due to the size difference of the bases (pyrimidine vs. purine), any additional interactions between protein and bases would involve different hydrogen bond networks, potentially involving water molecules in the case of the smaller UTP. Alternatively, due to stacking interactions between an aromatic residue of the protein and the bases, uracil and the pyrimidine ring of guanine might occupy equivalent positions. As this would inevitably lead to the relative misplacement of the ribose and phosphates of UTP compared to GTP, the catalytic residue may compensate for the size difference by re-adjusting and attacking the β- instead of the α-phosphate of UTP.

To explore the above possibility, nsp9 was nucleotidylated as before and non-bound label was removed by extensive washing until no residual radioactivity was detected in the wash buffer. The nucleotide-protein bond was subsequently broken by lowering of the pH and the released nucleotide was analyzed by thin layer chromatography. While nsp9 incubated with GTP clearly released significantly more of the expected guanosine-5'-monophosphate (GMP) in an acidic environment than under alkaline conditions, the results after uridylylation were not as conclusive. Although also in this case the monophosphate was released after HCl treatment, the intensity did not match that of GMP and a second product was present in higher quantities (Figure 9A). This may indicate that uridine-5'-monophosphate (UMP) is either further hydrolyzed under these conditions or that in fact a UMP–protein adduct is only the minor product after uridylylation. Therefore, it remains unclear whether the binding of UTP indeed forces an attack of the β-phosphate. To exclude that the observed GMP release is caused by the treatment with

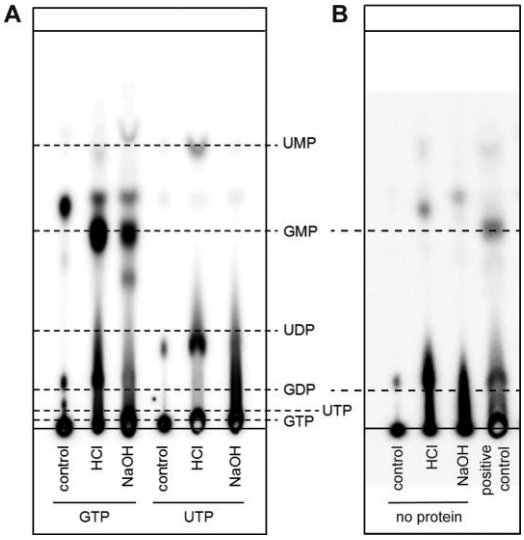


Figure 9 | GMP is released from labeled EAV nsp9 under acidic conditions. (**A**) nsp9 was labeled with [α-^{32}P]GTP or [α-^{32}P]UTP and was incubated at pH 8.5 (control) or under acidic or alkaline conditions after removal of non-incorporated nucleotides. Resulting products were separated with PEI-cellulose TLC. Solid lines represent the position where samples have been spotted (bottom) and the running front (top). Dashed lines represent the respective mobilities of the indicated nucleotides. (**B**) [α-^{32}P]GTP was incubated under the same conditions as in (A) but omitting nsp9. An nsp9-containing sample treated with HCl served as positive control.

HCl, control samples lacking nsp9 were also investigated. As expected this did not result in a product with equivalent mobility to GMP (Figure 9B).

NiRAN nucleotidylation is essential for EAV and SARS-CoV replication in cell culture

To establish the importance of the NiRAN domain for nidoviral replication, we used reverse genetics to engineer both EAV and SARS-CoV mutants in which conserved NiRAN residues were substituted with alanine. Following transfection of *in vitro*-transcribed full-length RNA into permissive cells, viral protein expression and progeny release were monitored (Table 1). As expected for such conserved residues, most alanine substitutions were either lethal for the virus or resulted in a severely crippled virus that reverted, thus confirming the essential role of the nucleotidylation activity during the viral replication cycle. Similarly, also replacement of a conserved aspartate in motif A of the downstream RdRp domain, which is known to be required for the activity of polymerases in other (+) RNA viruses [34], was tolerated in neither EAV nor SARS-CoV. Notable exceptions to this general pattern, in addition to the replacements of non-conserved lysine residues included as controls, were the T123A and F219A mutations in SARS-CoV nsp12. These mutations were stably maintained although they produced a mixed plaque phenotype comprising wild-type-sized and smaller plaques, with F219A also demonstrating a

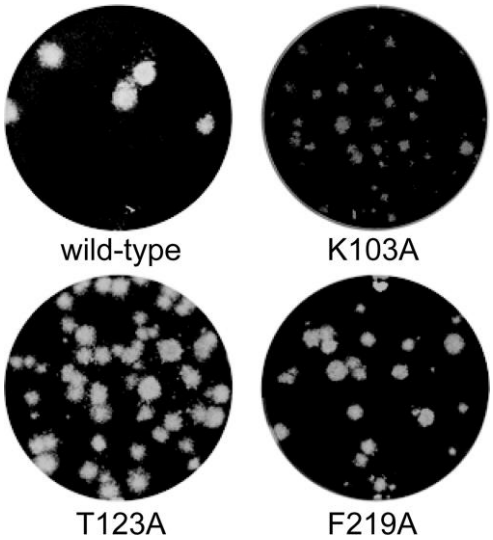


Figure 10 | Plaque phenotypes of viable SARS-CoV NiRAN mutants. Progeny virus harvested at 3 days post transfection was used for plaque assays (see 'Materials and Methods' section) on Vero-E6 cell monolayers, which were fixed and stained after 3 days to visualize virus-induced plaques.

markedly lower progeny titer (at least two logs reduced) than the wild-type control (Figure 10). The reason for this differential behavior of these two SARS-CoV mutants in comparison to those of EAV is unclear at the moment.

DISCUSSION

NiRAN is the first enzymatic genetic marker of the order *Nidovirales*

The NiRAN domain described in this study is the fourth ORF1b-encoded enzyme involved in RNA-dependent processes identified in arteriviruses and the seventh in coronaviruses. As in most prior studies of nidoviral replicative proteins, this identification was initiated by comparative genomics analysis. Unlike all other nidovirus enzymes, however, NiRAN was found to have no appreciable sequence similarity with proteins outside the order *Nidovirales*. Even the similarity between the arteriviral NiRAN and that of other nidoviruses was found to be marginal. These results suggested that NiRAN either is a unique enzyme specific to nidoviruses or has diverged from its paralogs beyond recognition, i.e. to an extent that cannot be ascertain by even the most powerful HMM-based tools currently available. The latter possibility is not merely hypothetical given that five out of the seven amino acid residues that are evolutionary invariant in the NiRAN domain belong to the most common residues found in proteins. We expect this uncertainty to be resolved in the future when the sampling of nidoviruses will be

expanded, sequence profile techniques will be further advanced, and tertiary structures of the proteins analyzed in this study may become available.

Besides technical challenges in the identification of NiRAN, this domain also stands out for its properties that are indicative of an unknown but critical role in nidovirus replication (see below). NiRAN is the only ORF1b-encoded domain that is located upstream of the RdRp and resides within the same non-structural protein. This implies that NiRAN may influence the folding of the downstream RdRp domain. It would be reasonable to expect cross-talk between these domains, potentially coupling the reactions and processes they catalyze. Thus, NiRAN is a prime candidate regulator and/or co-factor of the RdRp, a property that should be taken into account in future experiments aiming at the characterization of the RdRp or reconstitution of RTC activity *in vitro*.

The exclusive conservation of NiRAN in nidoviruses is indicative of its acquisition by a nidovirus ancestor before the currently known nidovirus families diverged. This makes the domain a genetic marker of this virus order, only the second after the previously identified ZBD and the first with enzymatic activity. It may not be a coincidence that each of these markers is associated with a key enzyme in (+) RNA virus replication, RdRp and HEL1, respectively. The HEL1-modulating role of the ZBD and its involvement in all major processes of the nidovirus replicative cycle have been documented (reviewed in [20]). Similar studies could be performed to probe the function(s) of NiRAN.

Possible functions of conserved NiRAN residues

We here demonstrated that NiRAN is essential for EAV and SARS-CoV replication in cell culture by testing mutants in which conserved residues had been replaced. The mutated viruses were either crippled (and in most cases reverted to wt) or dead, depending on the targeted residue and the virus studied. Importantly the magnitude of the observed effect paralleled that caused by replacement of an RdRp active site residue in the same virus. This parallel is most notable because of the much higher divergence of the NiRAN sequence compared to the RdRp. Also, the significance of NiRAN for virus replication must be different from that of NendoU, the only other ORF1b-encoded enzyme that has been probed extensively by mutagenesis in reverse genetics in both corona- and arteriviruses [25, 50, 92]. Two of those studies revealed that EAV and mouse hepatitis virus (MHV) NendoU mutants with replacements in the active site were stable and in the latter case even displayed similar plaque phenotypes as the wild-type virus while being only slightly delayed in growth [50, 92].

In our biochemical assays of the nidovirus RdRp subunit [40, 42, 93], we detected the new nucleotidylation activity that was associated with the NiRAN of EAV nsp9, as demonstrated by mass spectrometry analysis (Figure 2A and Supplementary Figure S4)

and the importance of conserved NiRAN residues for this activity (Figure 7). Nucleotidylation was most pronounced with UTP as substrate but was also observed with GTP (Figure 5A). Despite their size difference, both substrates appeared to be utilized by the same NiRAN binding site since uridylylation as well as guanylylation depended on the same conserved residues. To our knowledge such dual specificity has never been reported for a protein of an RNA virus and (likely) a host. Our results strongly suggested the nucleotidylated residue to be either a lysine or a histidine (Figure 8) located in the N-terminal part of nsp9. Since NiRAN lacks a conserved histidine, the conserved lysine of motif A (K94 in EAV nsp9) is the most likely target for nucleotidylation.

Given the non-radioactive endogenous NTP pool present in *E. coli*, these results imply that during its expression a part of the recombinant nsp9 may have already been converted to the described nucleoside adducts. Consequently, only the free nsp9 must have been available for nucleotidylation by its NiRAN domain using radioactive GTP/UTP. The nucleotidylated fraction of the total protein pool depends on many factors, including the adduct's stability, and remains unknown. However, this uncertainty does not undermine the validity of the established nucleotidylation activity of nsp9, given the specificity and selectivity documented here, which were determined using different techniques and various controls to arrive at a consistent set of properties of the enzyme. Combined, the results of our biochemical and bioinformatics analyses assigned nucleotidylation activity to the NiRAN domain beyond a reasonable doubt. To rationalize the protein's ability to bind nucleoside phosphates covalently, future studies may focus on the role of protein−nucleoside adducts as reaction intermediates for possible downstream processes, three of which are discussed below.

Next to K94 and/or conserved R124 of motif B_N, which may mediate NTP binding via interactions with the negatively charged phosphates, a third conserved residue which may contribute to NTP binding is the motif C_N phenylalanine (F166 in EAV). Since phenylalanine would most likely interact with the nucleotide substrate by base stacking, its contribution in terms of binding energy would be one order of magnitude lower than that of electrostatic interactions of lysine/arginine with the phosphates [94]. Based solely on this consideration, F166 could be expected to be of 'lesser' importance than the basic residues. However, this was apparently not the case since the replacement of the aromatic residue with alanine was lethal for EAV while substitution of either of the basic residues led to a low level of replication that eventually facilitated reversion (Table 1). All these substitutions require two nucleotide point mutations to revert back to wild-type, which should be an extremely rare event during a single round of replication. Consequently, the non-viable phenotype of the F166A mutant may hint at a lower tolerance of single-nucleotide partial revertants (F166V or F166S) in comparison to those originating from K94A (K94T or K94E) and R124A (R124P or R124G). Alternatively, the

observed non-viable F166A phenotype may be explained by a vital interaction between NiRAN and RdRp or other proteins involving F166. In contrast to EAV, the homologous residue in SARS-CoV nsp12, F219, appeared to be less essential since its replacement merely reduced progeny titers and altered the plaque phenotype, while the nucleotide changes were maintained. At present, the exact reason for this difference between EAV and SARS-CoV is unclear, but it suggests that the role and/or regulation of this conserved phenylalanine may have evolved in these distantly related nidoviruses, whose NiRAN domains are of strikingly different sizes; such evolution has parallels in other enzymes [95].

Since neither binding of phosphates nor base stacking would enable the enzyme to discriminate between the four bases, it is likely that some of the conserved residues are involved in the formation of a hydrogen bond network that is specific for GTP or UTP. The conserved serine/threonine of motif B_N could be a candidate as substitution of this serine in EAV nsp9 (S129) was the only mutation that had a differential effect on guanylylation and uridylylation (Figure 7). Finally, in agreement with observations for other nucleotidylate-forming enzymes [96-98], also nsp9 nucleotidylation is metal-dependent (Figure 5B), potentially due to an important role for metal ions in coordination of the triphosphate or charge neutralization of the pyrophosphate leaving group. In our *in vitro* system it was Mn^{2+} rather than the most common divalent cation Mg^{2+} that supported nucleotidylation activity when tested over a wide concentration range. We propose that at least one of the three acidic conserved residues (E100, D132 and D165 in EAV nsp9) is directly involved in the binding of this essential manganese ion(s). Since the concentration of this cation in cells is lower than that required to observe nucleotidylation *in vitro*, we cannot exclude the possibility that another co-factor or substrate modulates this property of the enzyme *in vivo*, and/or that another metal ion is used.

Possible roles of nucleotidylation in the context of viral replication

The identification of the nucleotidylation activity raises the question which role it may play in the nidovirus replicative cycle. In the discussion that follows, we will consider the pros and cons of the involvement of NiRAN's nucleotidylation activity in three previously described functions that are not involved in energy-dependent metabolic processes: nucleic acid ligation, mRNA capping and protein-primed RNA synthesis.

Ligase function

We initially considered NiRAN to be a non-canonical ATP-dependent RNA ligase. It was reasoned that, in the context of nidovirus replication, such an activity could be the functional complement of the NendoU endoribonuclease [7]. Moreover, at that time both enzymes were considered to have been conserved across all taxa during evolution of the

nidovirus lineage. However, it recently became clear that NendoU is conserved only in nidoviruses infecting vertebrate hosts. Consequently, our original hypothesis would not explain why this putative ligase would be conserved in roni- and mesoniviruses, which do not encode the endoribonuclease. Another complication regarding that original hypothesis has emerged from the present study, which identified NiRAN as being UTP/GTP-specific. Although the hydrolysis of all NTPs results in the release of the same amount of energy, ATP-dependent RNA ligases dominate the ligase family. It would therefore be surprising, if nidoviruses encoded a ligase that strongly discriminates against ATP. To our knowledge the GTP-specific tRNA-splicing ligase RtcB is the only currently known example of a protein involved in nucleic acid strand joining exhibiting this kind of substrate specificity [90]. Furthermore, thus far no substrates that would require a ligase function were identified in nidovirus replication, which however remains poorly characterized in general.

5' end cap guanylyltransferase function

Besides RNA ligases, also guanylyltransferases (GTases) employ a very similar mechanism of nucleotidylation and are used to permanently modify the 5' end of RNA with the bound GMP in a process called RNA capping (reviewed in [99]). Intriguingly, three of the four enzyme activities required for cap formation and modification, namely an RNA-triphosphatase and two methyltransferases, have been identified in coronaviruses [23, 44], with the missing activity being the GTase. Furthermore, recent characterization of EAV nsp10 in our lab (unpublished data) showed that it resembles its coronavirus homolog in terms of possessing RNA-triphosphatase activity, which is required prior to GTase activity in the conventional capping pathway. In line with these findings, experimental evidence supporting the presence of a cap structure on genomic RNA was reported for three very distantly related viruses of the *Nidovirales* order, namely for MHV (*Coronavirinae*) [100], Equine torovirus (*Torovirinae*) [101] and Simian hemorrhagic fever virus (*Arteriviridae*) [102]. Importantly, the known GTases of (+) RNA viruses, flavivirus NS5 [103], alphavirus nsP1 and orthologous proteins [97, 104], do neither share conserved features nor do they resemble host GTases. Thus, the possibility of NiRAN being a cap-synthesizing GTase could be reconciled with our current knowledge of the structural and sequence diversity of this class of enzymes. This cannot be said, however, about NiRAN's substrate preference for UTP over GTP, which has not been reported for GTases mediating cap formation.

Protein-priming function

If UTP binding by NiRAN reflects a genuine property of the enzyme, another mechanism that might utilize its nucleotidylation activity may be protein-primed RNA synthesis. This strategy is used by many viruses including the large group of picornavirus-like viruses,

which notably have evolutionary affinity to nidoviruses [35, 36]. In these viruses, a nucleotide is covalently attached to a protein that is commonly known as VPg (viral protein genome-linked), which may then be extended to a dinucleotide. This dinucleotide is subsequently base-paired to the 3' end of the viral RNA where it serves as primer for the synthesis of the complementary RNA strand [105]. Interestingly, the first nucleotide of the EAV genome is a G while the 3' end is equipped with a poly(A) tail. Thus, the dual specificity of nsp9 for GTP and UTP would be compatible with the different requirements for the initiation of the synthesis of genomic and subgenomic RNAs of positive and negative polarity, respectively. To which extent this property is conserved across nidoviruses remains to be established.

While considering this mechanism, it is instructive to take into account observations that distinguish nidoviruses from VPg-utilizing viruses. First, to our knowledge, all currently described nucleotide-VPg bonds are realized via the hydroxyl group of either a tyrosine or a serine/threonine [106-110], while NiRAN most likely uses the invariant lysine residue (Figure 8). This problem could be resolved if NiRAN assumes the role of the RdRp of VPg-encoding viruses and transfers the bound nucleotide to another protein that subsequently serves as VPg. Second, at least for coronaviruses, the VPg-based mechanism would not be compatible with the previously proposed primase-based mechanism [111] for the initiation of RNA synthesis. However, the latter mechanism remains tentative since it assigns primase activity to a protein complex that, according to a recent study [84], may merely be a processivity co-factor for the nsp12 RdRp. Finally, as mentioned before, the mRNAs of several nidoviruses were concluded to be capped at their 5' end, a modification that is not observed in known VPg-utilizing viruses. To use both VPg priming and capping, it would be necessary to actively or passively remove the attached protein in order to allow mRNA capping to commence. This sequence of events would constitute a novel, and perhaps unlikely, variant of the capping pathway, as the RNA's 5' end would not be di- or triphosphorylated after VPg removal, a requirement for entering any of the known viral capping pathways [99]. Thus, if NiRAN would be part of a VPg-utilizing mechanism, this might differ considerably from those currently described and could possibly also vary among nidoviruses.

In view of the considerations outlined for each of the three possible scenarios employing nucleotidylation activity, it is evident that presently none of these can be reconciled with the evolutionary, structural and functional characteristics of NiRAN described in this study without additional assumptions. This may reflect yet-to-be revealed specifics of the nidovirus RTC and its unparalleled complexity.

ACKNOWLEDGMENTS

The authors thank Bruno Canard, Etienne Decroly, Isabelle Imbert, Barbara Selisko, Lorenzo Subissi and Aartjan te Velthuis for helpful discussions; Chris Lauber and Erik Hoogendoorn for help with the DEmARC-based analysis, and Daniel Cupac and Linda Boomaars for technical assistance. A.E.G. is a member of the Netherlands Bioinformatics Center (NBIC) Faculty.

FUNDING

European Union Seventh Framework program through the EUVIRNA project (European Training Network on (+) RNA virus replication and Antiviral Drug Development [264286]; SILVER project [260644]; Netherlands Organization for Scientific Research (NWO) through TOP-GO [700.10.352]; Leiden University Fund; Collaborative Agreement in Bioinformatics between Leiden University Medical Center and Moscow State University (MoBiLe). Funding for open access charge: Leiden University Medical Center and Netherlands Organisation for Scientific Research.

Conflict of interest statement. None declared.

SUPPLEMENTARY DATA

Supplementary Materials and Methods

Synthesis of 5'-(4-fluorosulfonylbenzoyl)guanosine (FSBG)

Guanosine monohydrate (875 mg, 2.90 mmol) was co-evaporated twice with anhydrous DMF and subsequently dissolved in DMPU with gentle warming. The clear solution was cooled in an ice bath, and 4-(fluorosulfonyl)benzoyl chloride (812 mg, 3.65 mmol) was added. After 15 minutes the mixture was warmed to room temperature and stirred for another 4 hours. Petroleum ether 40/60 (50 mL) was added and a white precipitate formed. The organic layer was decanted and the residue triturated twice with a 1/1 mixture of ethyl acetate/diethyl ether (2 x 50 mL). The residue was re-crystallized from MeOH/water and further purified by C18-RP-HPLC (Phenomenex Gemini C18, pore size 110Å, particle size 5 µm, 150 x 21.2 mm, gradient 20 – 50% Acetonitrile in 0.1 % aqueous TFA, 20 mL/min) to yield the title compound as a white solid (232 mg, yield 17%) (Supplementary Figure 5).

Supplementary tables

Table S1 | Virus genome used for the bioinformatics analyses.

Virus name	Species	(Sub)family	Acronym	Accession number
Gill-associated virus	*Gill-associated virus*	*Roniviridae*	GAV	AF227196
Yellow head virus	to be established	*Roniviridae*	YHV	EU487200
Cavally virus	*Alphamesonivirus 1*	*Mesoniviridae*	CAVV	HM746600
Casuarina virus	to be established	*Mesoniviridae*	CASV	NC_023986
Dak Nong virus	to be established	*Mesoniviridae*	DKNV	AB753015.2
Hana virus	to be established	*Mesoniviridae*	HanaV	JQ957872
Nse virus	to be established	*Mesoniviridae*	NseV	JQ957874
Meno virus	to be established	*Mesoniviridae*	MenoV	JQ957873
SARS coronavirus Frankfurt 1	*Severe acute respiratory syndrome-related coronavirus*	*Coronavirinae*	SARS-CoV	AY291315
Rabbit coronavirus HKU14	*Betacoronavirus 1*	*Coronavirinae*	RbCoV_HKU14	JN874560
Murine hepatitis virus strain 2	*Murine coronavirus*	*Coronavirinae*	MHV-2	AF201929
Human coronavirus HKU1	*Human coronavirus HKU1*	*Coronavirinae*	HCoV_HKU1	AY884001
Betacoronavirus	to be established	*Coronavirinae*	EriCoV	KC545383
Bat coronavirus (BtCoV/133/2005)	*Tylonycteris bat coronavirus HKU4*	*Coronavirinae*	BtCoV/133/2005	DQ648794
Bat coronavirus HKU5-1	*Pipistrellus bat coronavirus HKU5*	*Coronavirinae*	BtCoV_HKU5	EF065509
MERS coronavirus EMC/2012	to be established	*Coronavirinae*	MERS-CoV	JX869059.2
Bat coronavirus HKU9-10-2	*Rousettus bat coronavirus HKU9*	*Coronavirinae*	BtCoV_HKU9	HM211101
Bat coronavirus CDPHE15/USA/2006	to be established	*Coronavirinae*	BtCoV_CDPHE15	KF430219
Human coronavirus NL63	*Human coronavirus NL63*	*Coronavirinae*	HCoV-NL63	AY567487
Miniopterus bat coronavirus HKU8	*Miniopterus bat coronavirus HKU8*	*Coronavirinae*	BtCoV_HKU8	EU420139
Rhinolophus bat coronavirus HKU2	*Rhinolophus bat coronavirus HKU2*	*Coronavirinae*	BtCoV_HKU2	EF203064
Bat coronavirus 1A	*Miniopterus bat coronavirus 1*	*Coronavirinae*	BtCoV_1A	EU420138
Alpaca respiratory coronavirus	*Human coronavirus 229E*	*Coronavirinae*	ACoV	JQ410000
Bat coronavirus (BtCoV/512/2005)	*Scotophilus bat coronavirus 512*	*Coronavirinae*	BtCoV/512/2005	DQ648858
Porcine epidemic diarrhea virus	*Porcine epidemic diarrhea virus*	*Coronavirinae*	PEDV	KC140102
Rousettus bat coronavirus HKU10	to be established	*Coronavirinae*	BtCoV_HKU10	JQ989271
Mink coronavirus strain WD1127	to be established	*Coronavirinae*	MCoV	HM245925
Feline coronavirus UU2	*Alphacoronavirus 1*	*Coronavirinae*	FCoV_UU2	FJ938060

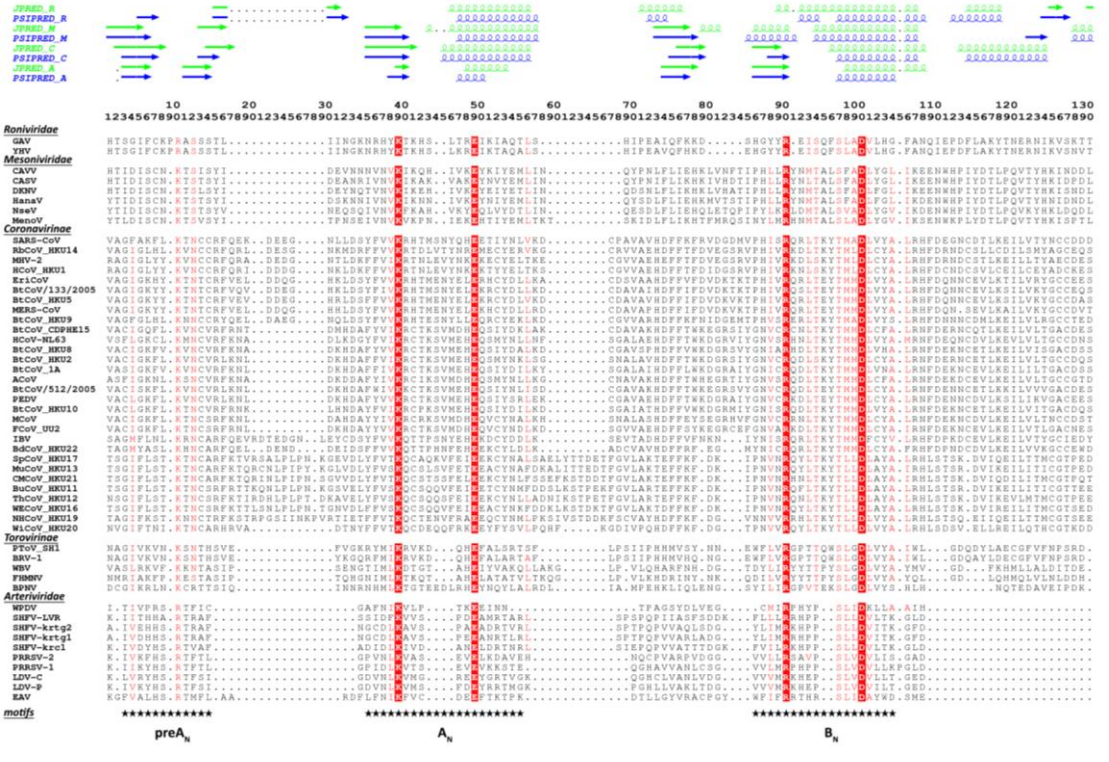
Table S1 (continued)

Virus name	Species	(Sub)family	Acronym	Accession number
Infectious bronchitis virus	*Avian coronavirus*	*Coronavirinae*	IBV	KC008600
Bottlenose dolphin coronavirus HKU22	*Beluga whale coronavirus SW1*	*Coronavirinae*	BdCoV_HKU22	KF793824
Sparrow coronavirus HKU17	to be established	*Coronavirinae*	SpCoV_HKU17	JQ065045
Munia coronavirus HKU13-3514	*Munia coronavirus HKU13*	*Coronavirinae*	MuCoV_HKU13	FJ376622
Common-moorhen coronavirus HKU21	to be established	*Coronavirinae*	CMCoV_HKU21	JQ065049
Bulbul coronavirus HKU11-934	*Bulbul coronavirus HKU11*	*Coronavirinae*	BuCoV_HKU11	FJ376619.2
Thrush coronavirus HKU12-600	*Thrush coronavirus HKU12*	*Coronavirinae*	ThCoV_HKU12	FJ376621
White-eye coronavirus HKU16	to be established	*Coronavirinae*	WECoV_HKU16	JQ065044
Night-heron coronavirus HKU19	to be established	*Coronavirinae*	NHCoV_HKU19	JQ065047
Wigeon coronavirus HKU20	to be established	*Coronavirinae*	WiCoV_HKU20	JQ065048
Porcine torovirus	*Porcine torovirus*	*Torovirinae*	PToV_SH1	NC_022787
Breda virus	*Bovine torovirus*	*Torovirinae*	BRV-1	AY427798
White bream virus	*White bream virus*	*Torovirinae*	WBV	DQ898157
Fathead minnow nidovirus	to be established	*Torovirinae*	FHMNV	GU002364.2
Ball python nidovirus	to be established	*Torovirinae*	BPNV	NC_024709
Possum nidovirus	to be established	*Arteriviridae*	WPDV	JN116253
Simian hemorrhagic fever virus	*Simian hemorrhagic fever virus*	*Arteriviridae*	SHFV-LVR	AF180391
Simian hemorrhagic fever virus	to be established	*Arteriviridae*	SHFV-krtg2	JX473847
Simian hemorrhagic fever virus	to be established	*Arteriviridae*	SHFV-krtg1	JX473848
Simian hemorrhagic fever virus	to be established	*Arteriviridae*	SHFV-krc1	HQ845737
Porcine reproductive and respiratory	*Porcine reproductive and respiratory syndrome virus*	*Arteriviridae*	PRRSV-2	JX138233
Porcine reproductive and respiratory	*Porcine reproductive and respiratory syndrome virus*	*Arteriviridae*	PRRSV-1	GU737264.2
Lactate dehydrogenase-elevating virus	*Lactate dehydrogenase-elevating virus*	*Arteriviridae*	LDV-C	L13298
Lactate dehydrogenase-elevating virus	*Lactate dehydrogenase-elevating virus*	*Arteriviridae*	LDV-P	U15146
Equine arteritis virus	*Equine arteritis virus*	*Arteriviridae*	EAV	DQ846750

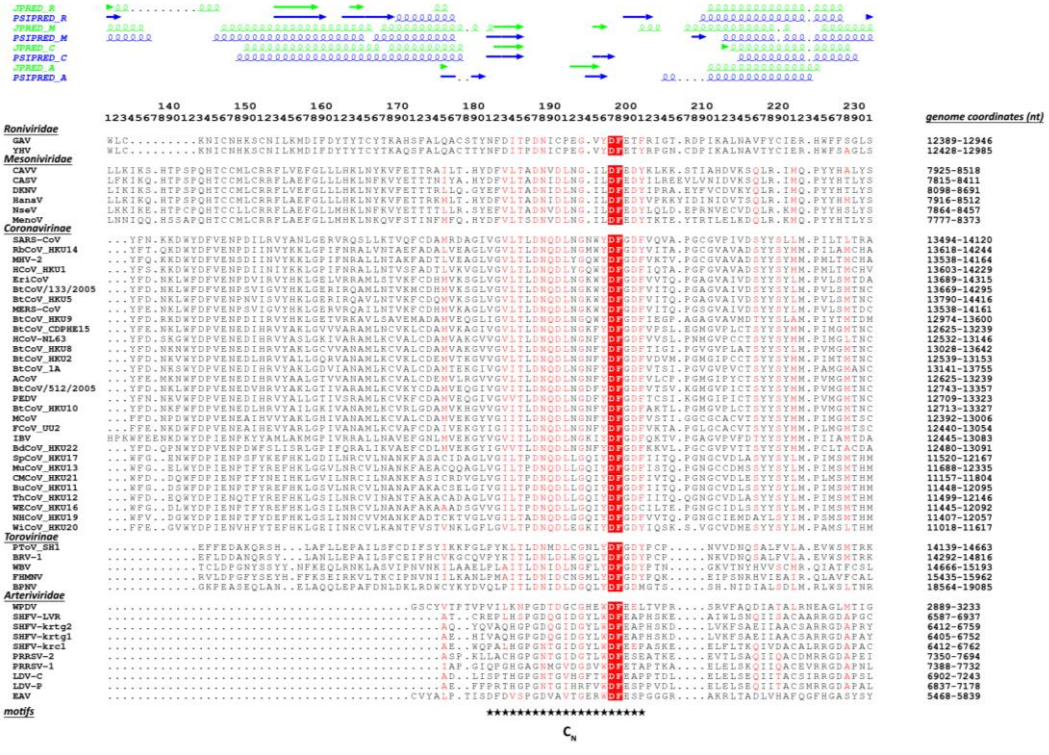
Table S2 | Multiple sequence alignment of NiRAN of nidoviruses.

FASTA file is available from https://doi.org/10.1093/nar/gkv838

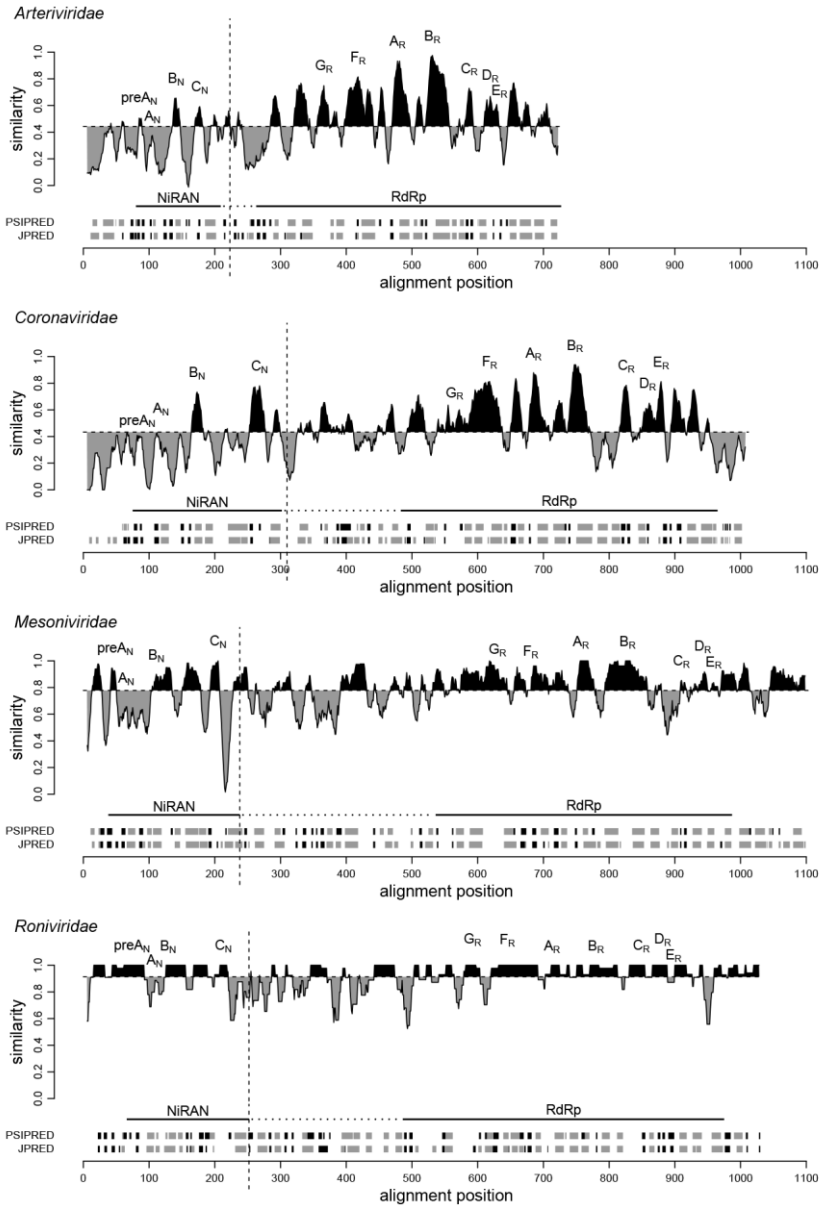
Supplementary figures



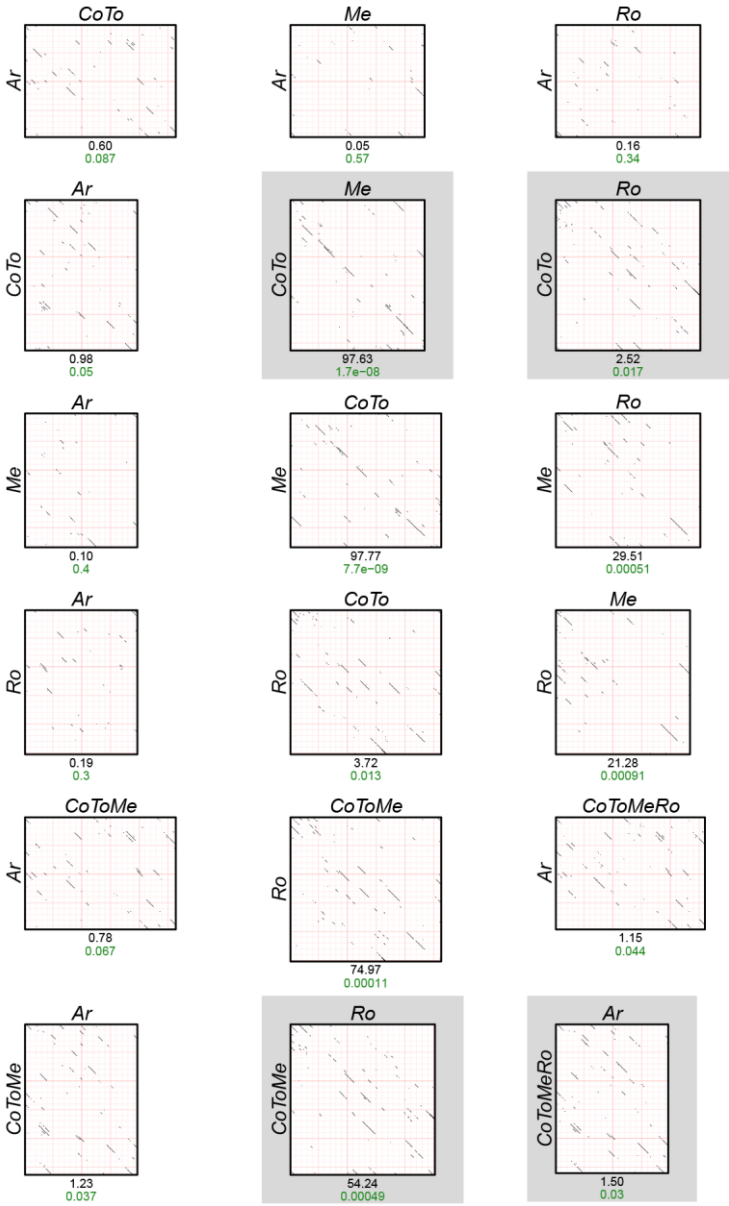
Supplementary Figure 1 | The nidovirus-wide NiRAN MSA encompassing conserved motifs. Virus names and accession numbers are listed in Table S1. Fully and partially conserved residues are depicted in white font with red background and red font, respectively. Sequence motifs are indicated by stars. Secondary structure predictions are shown on the top of the MSA. The name of each prediction indicates software used (Jpred 3 or PSIPRED) and which family-specific NiRAN MSA (R, *Roniviridae*; M, *Mesoniviridae*; C, *Coronaviridae*; A, *Arteriviridae*) were used to produce it. The plot was generated with ESPript.



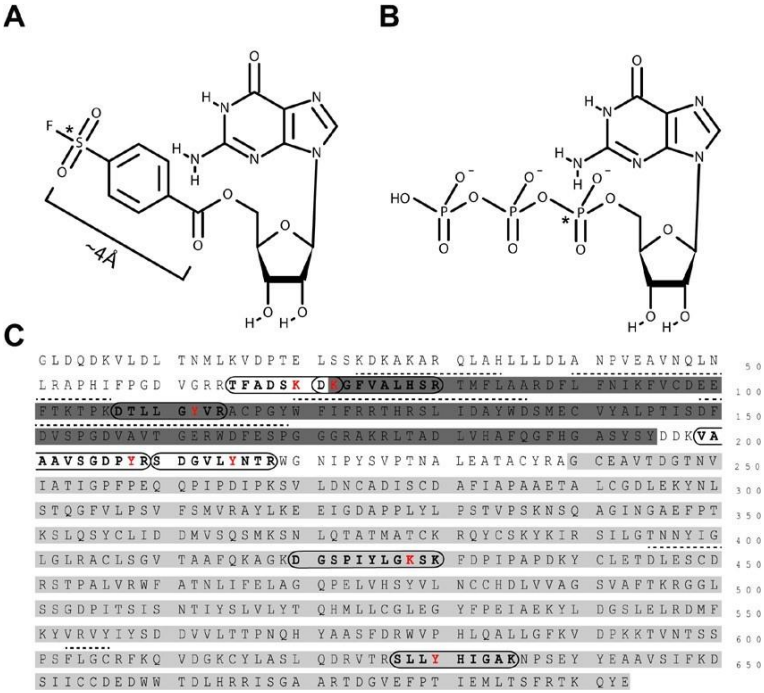
Supplementary Figure 1 (continued)



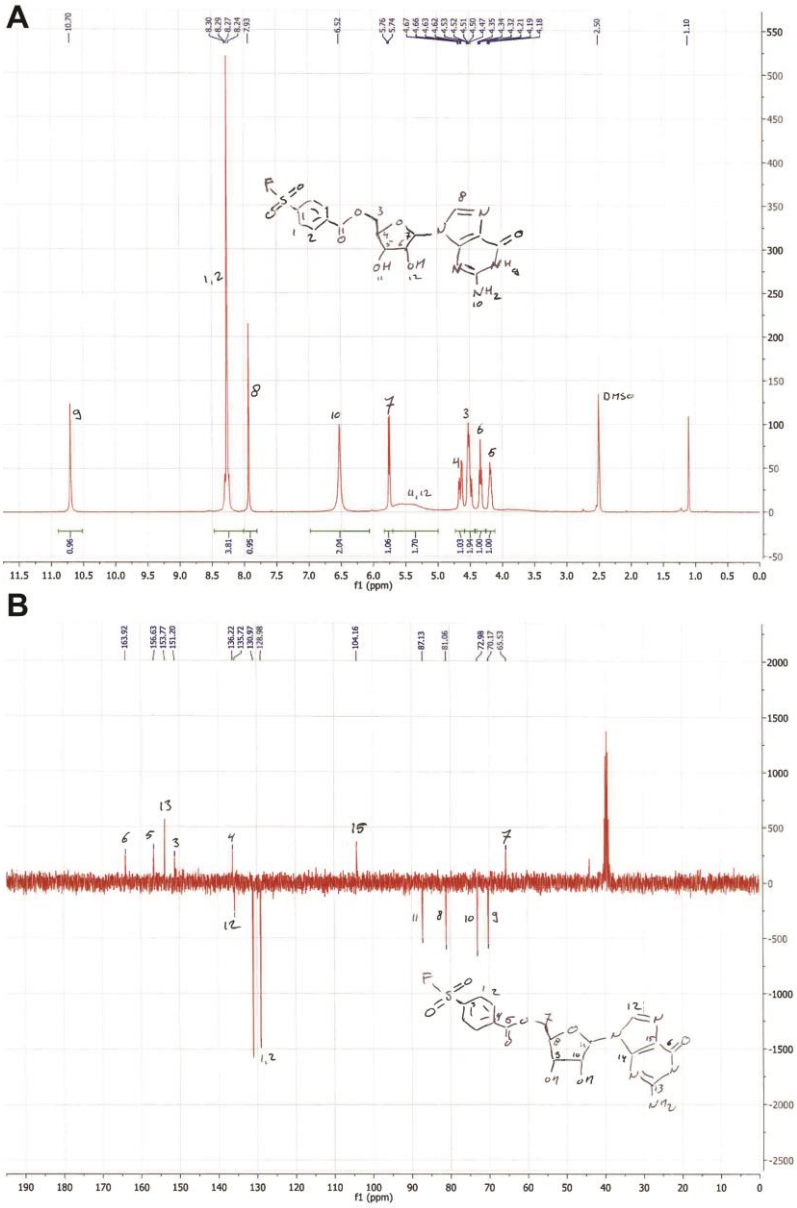
Supplementary Figure 2 | Sequence variation, domain organization, and secondary structure of NiRAN-RdRp-containing proteins of nidovirus families. For each family, the similarity density plot obtained for the MSA of proteins including the NiRAN and RdRp domains is shown. To highlight the regional deviation of conservation from that of the MSA average, areas above and below the mean similarity are shaded in black and gray, respectively. Sequence motifs of NiRAN and RdRp are labelled. Uncertainty in respect to the domain boundary between NiRAN and RdRp is indicated by dashed horizontal lines. Domain boundaries used for all bioinformatics analyses are indicated by dashed vertical lines. Below each similarity density plot predicted secondary structure elements are presented in gray for α-helices and black for β-strands.



Supplementary Figure 3 | Pairwise MSA-based HMM-HMM comparison of NiRANs of different origins. Each MSA of NiRAN was converted into an HMM profile, all possible pairs of different HMMs were aligned using HH-align. The label at the left and top of each plot specifies the group of viruses used as query and target in HMM-HMM comparison, respectively. Below each dot-plot the confidence (%) of the target being homologous to the query and the E value of the top local hit are shown in black and green, respectively. The four plots highlighted with grey background are also presented in Fig. 3.



Supplementary Figure 4 | (**A**) FSBG and (**B**) GTP structures indicating the spatial separation of the points of attack in FSBG and GTP. Asterisks mark the positions of the nucleophilic attack. (**C**) Mass spectrometry analysis of FSBG-linked EAV nsp9 identified seven unique, modified peptides (outlined) located either in vicinity of the NiRAN (dark gray background) or within the C-terminal RdRp domain (light gray background). Residues carrying the sulfonylbenzoyl modification are colored in red. Sequence or structural motifs are indicated by dashed lines above the sequence in the order preA_N, A_N, B_N, C_N, A_R, and E_R. See also Fig. 2A.



Supplementary Figure 5 | NMR analysis of 5'-(4-fluorosulfonylbenzoyl)guanosine. (**A**) ^{1}H NMR (300 MHz, DMSO-d_6) δ 10.70 (s, 1H), 8.38 – 8.12 (m, 4H), 7.93 (s, 1H), 6.52 (broad s, 2H), 5.75 (d, J = 4.8 Hz, 1H), 5.75 (broad s, 2H), 4.65 (dd, J = 11.9, 3.6 Hz, 1H), 4.59 – 4.42 (m, 2H), 4.34 (t, J = 5.1 Hz, 1H), 4.25 – 4.12 (m, 1H). (**B**) ^{13}C NMR (75 MHz, DMSO-d_6) δ 163.92, 156.63, 153.77, 151.20, 136.22, 135.72, 130.97, 128.98, 104.16, 87.13, 81.06, 72.98, 70.17, 65.53. Corresponding peaks and atoms are indicated by numbers.

REFERENCES

1. de Groot RJ, Cowley JA, Enjuanes L, Faaberg KS, Perlman S, Rottier PJM, Snijder EJ, Ziebuhr J, Gorbalenya AE: **Order *Nidovirales*.** In: *Virus Taxonomy, the 9th Report of the International Committee on Taxonomy of Viruses.* Edited by King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ: Elsevier; 2012: 785-795.
2. Lauber C, Ziebuhr J, Junglen S, Drosten C, Zirkel F, Nga PT, Morita K, Snijder EJ, Gorbalenya AE: **Mesoniviridae: a proposed new family in the order Nidovirales formed by a single species of mosquito-borne viruses**. *Arch Virol* 2012, **157**(8):1623-1628.
3. Neumann EJ, Kliebenstein JB, Johnson CD, Mabry JW, Bush EJ, Seitzinger AH, Green AL, Zimmerman JJ: **Assessment of the economic impact of porcine reproductive and respiratory syndrome on swine production in the United States**. *J Am Vet Med Assoc* 2005, **227**(3):385-392.
4. Coleman CM, Frieman MB: **Coronaviruses: important emerging human pathogens**. *J Virol* 2014, **88**(10):5209-5212.
5. Hilgenfeld R, Peiris M: **From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses**. *Antiviral Res* 2013, **100**(1):286-295.
6. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA: **Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia**. *N Engl J Med* 2012, **367**(19):1814-1820.
7. Lauber C, Goeman JJ, Parquet MC, 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**(7):e1003500.
8. Snijder EJ, Siddell SG, Gorbalenya AE: **The Order Nidovirales**. In: *Topley & Wilson's Microbiology and Microbial Infections: Virology Volume.* Edited by Mahy BW, ter Meulen V. London: Hodder Arnold; 2005: 390-404.
9. Pasternak AO, Spaan WJ, Snijder EJ: **Nidovirus transcription: how to make sense...?** *J Gen Virol* 2006, **87**(Pt 6):1403-1421.
10. Sawicki SG, Sawicki DL, Siddell SG: **A contemporary view of coronavirus transcription**. *J Virol* 2007, **81**(1):20-29.
11. Brian DA, Baric RS: **Coronavirus genome structure and replication**. *Curr Top Microbiol Immunol* 2005, **287**:1-30.
12. Masters PS: **The molecular biology of coronaviruses**. *Adv Virus Res* 2006, **66**:193-292.
13. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ: **Nidovirales: evolving the largest RNA virus genome**. *Virus Res* 2006, **117**(1):17-37.
14. Ziebuhr J, Snijder EJ, Gorbalenya AE: **Virus-encoded proteinases and proteolytic processing in the Nidovirales**. *J Gen Virol* 2000, **81**(Pt 4):853-879.

15. Weiss SR, Leibowitz JL: **Coronavirus pathogenesis**. *Adv Virus Res* 2011, **81**:85-164.

16. Perlman S, Netland J: **Coronaviruses post-SARS: update on replication and pathogenesis**. *Nat Rev Microbiol* 2009, **7**(6):439-450.

17. Subissi L, Imbert I, Ferron F, Collet A, Coutard B, Decroly E, Canard B: **SARS-CoV ORF1b-encoded nonstructural proteins 12-16: replicative enzymes as antiviral targets**. *Antiviral Res* 2014, **101**:122-130.

18. Neuman BW, Angelini MM, Buchmeier MJ: **Does form meet function in the coronavirus replicative organelle?** *Trends Microbiol* 2014, **22**(11):642-647.

19. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS: **Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity**. *RNA Biol* 2011, **8**(2):270-279.

20. Lehmann KC, Snijder EJ, Posthuma CC, Gorbalenya AE: **What we know but do not understand about nidovirus helicases**. *Virus Res* 2015, **202**:12-32.

21. 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**(24):9372-9377.

22. 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**(13):5108-5113.

23. Bouvet M, Debarnot C, Imbert I, Selisko B, Snijder EJ, Canard B, Decroly E: **In vitro reconstitution of SARS-coronavirus mRNA cap methylation**. *PLoS Pathog* 2010, **6**(4):e1000863.

24. Chen Y, Cai H, Pan J, Xiang N, Tien P, Ahola T, Guo D: **Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase**. *Proc Natl Acad Sci U S A* 2009, **106**(9):3484-3489.

25. Ivanov KA, Hertzig T, Rozanov M, Bayer S, Thiel V, Gorbalenya AE, Ziebuhr J: **Major genetic marker of nidoviruses encodes a replicative endoribonuclease**. *Proc Natl Acad Sci U S A* 2004, **101**(34):12694-12699.

26. Nedialkova DD, Ulferts R, van den Born E, Lauber C, Gorbalenya AE, Ziebuhr J, Snijder EJ: **Biochemical characterization of arterivirus nonstructural protein 11 reveals the nidovirus-wide conservation of a replicative endoribonuclease**. *J Virol* 2009, **83**(11):5671-5682.

27. 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**(10):e1002294.

28. Daffis S, Szretter KJ, Schriewer J, Li J, Youn S, Errett J, Lin TY, Schneller S, Zust R, Dong H *et al*: **2'-O methylation of the viral mRNA cap evades host restriction by IFIT family members**. *Nature* 2010, **468**(7322):452-456.

29. 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**(16):8071-8084.

30. Nga PT, Parquet MC, Lauber C, Parida M, Nabeshima T, Yu F, Thuy NT, Inoue S, Ito T, Okamoto K *et al*: **Discovery of the first insect nidovirus, a missing evolutionary link in the emergence of the largest RNA virus genomes**. *PLoS Pathog* 2011, **7**(9):e1002215.

31. Gorbalenya AE: **Big nidovirus genome. When count and order of domains matter**. *Adv Exp Med Biol* 2001, **494**:1-17.

32. Lehmann KC, Hooghiemstra L, Gulyaeva A, Samborskiy DV, Zevenhoven-Dobbe JC, Snijder EJ, Gorbalenya AE, Posthuma CC: **Arterivirus nsp12 versus the coronavirus nsp16 2'-O-methyltransferase: comparison of the C-terminal cleavage products of two nidovirus pp1ab polyproteins**. *J Gen Virol* 2015, **96**(9):2643-2655.

33. Lang DM, Zemla AT, Zhou CL: **Highly similar structural frames link the template tunnel and NTP entry tunnel to the exterior surface in RNA-dependent RNA polymerases**. *Nucleic Acids Res* 2013, **41**(3):1464-1482.

34. Ng KK, Arnold JJ, Cameron CE: **Structure-function relationships among RNA-dependent RNA polymerases**. *Curr Top Microbiol Immunol* 2008, **320**:137-156.

35. Gorbalenya AE, Pringle FM, Zeddam JL, Luke BT, Cameron CE, Kalmakoff J, Hanzlik TN, Gordon KH, Ward VK: **The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage**. *J Mol Biol* 2002, **324**(1):47-62.

36. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM: **Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis**. *Nucleic Acids Res* 1989, **17**(12):4847-4861.

37. Azzi A, Lin SX: **Human SARS-coronavirus RNA-dependent RNA polymerase: activity determinants and nucleoside analogue inhibitors**. *Proteins* 2004, **57**(1):12-14.

38. Xu X, Liu Y, Weiss S, Arnold E, Sarafianos SG, Ding J: **Molecular model of SARS coronavirus polymerase: implications for biochemical functions and drug design**. *Nucleic Acids Res* 2003, **31**(24):7117-7130.

39. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, 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**(5):991-1004.

40. Ahn DG, Choi JK, Taylor DR, Oh JW: **Biochemical characterization of a recombinant SARS coronavirus nsp12 RNA-dependent RNA polymerase capable of copying viral RNA templates**. *Arch Virol* 2012, **157**(11):2095-2104.

41. Bautista EM, Faaberg KS, Mickelson D, McGruder ED: **Functional properties of the predicted helicase of porcine reproductive and respiratory syndrome virus**. *Virology* 2002, **298**(2):258-270.

42. Beerens N, Selisko B, Ricagno S, Imbert I, van der Zanden L, Snijder EJ, Canard B: **De novo initiation of RNA synthesis by the arterivirus RNA-dependent RNA polymerase**. *J Virol* 2007, **81**(16):8384-8395.

43. Ivanov KA, Thiel V, Dobbe JC, van der Meer Y, Snijder EJ, Ziebuhr J: **Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase**. *J Virol* 2004, **78**(11):5619-5632.

44. 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**(14):7833-7838.

45. Seybert A, van Dinten LC, Snijder EJ, Ziebuhr J: **Biochemical characterization of the equine arteritis virus helicase suggests a close functional relationship between arterivirus and coronavirus helicases**. *J Virol* 2000, **74**(20):9586-9593.

46. Seybert A, Posthuma CC, van Dinten LC, Snijder EJ, Gorbalenya AE, Ziebuhr J: **A complex zinc finger controls the enzymatic activities of nidovirus helicases**. *J Virol* 2005, **79**(2):696-704.

47. te Velthuis AJ, van den Worm SH, Sims AC, Baric RS, Snijder EJ, van Hemert MJ: **Zn(2+) inhibits coronavirus and arterivirus RNA polymerase activity in vitro and zinc ionophores block the replication of these viruses in cell culture**. *PLoS Pathog* 2010, **6**(11):e1001176.

48. van Dinten LC, van Tol H, Gorbalenya AE, Snijder EJ: **The predicted metal-binding region of the arterivirus helicase protein is involved in subgenomic mRNA synthesis, genome replication, and virion biogenesis**. *J Virol* 2000, **74**(11):5213-5223.

49. 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**(22):12135-12144.

50. Posthuma CC, Nedialkova DD, Zevenhoven-Dobbe JC, Blokhuis JH, Gorbalenya AE, Snijder EJ: **Site-directed mutagenesis of the Nidovirus replicative endoribonuclease NendoU exerts pleiotropic effects on the arterivirus life cycle**. *J Virol* 2006, **80**(4):1653-1661.

51. Zust R, Cervantes-Barragan L, Habjan M, Maier R, Neuman BW, Ziebuhr J, Szretter KJ, Baker SC, Barchet W, Diamond MS *et al*: **Ribose 2'-O-methylation provides a**

molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. *Nat Immunol* 2011, **12**(2):137-143.

52. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2013, **41**(Database issue):D36-D42.

53. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM *et al*: **RefSeq: an update on mammalian reference sequences**. *Nucleic Acids Res* 2014, **42**(Database issue):D756-D763.

54. Lauber C, Gorbalenya AE: **Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses**. *J Virol* 2012, **86**(7):3890-3904.

55. Sidorov IA, Reshetov DA, Gorbalenya AE: **SNAD: Sequence Name Annotation-based Designer**. *BMC Bioinformatics* 2009, **10**:251.

56. Gorbalenya AE, Lieutaud P, Harris MR, Coutard B, Canard B, Kleywegt GJ, Kravchenko AA, Samborskiy DV, Sidorov IA, Leontovich AM *et al*: **Practical application of bioinformatics by the multidisciplinary VIZIER consortium**. *Antiviral Res* 2010, **87**(2):95-110.

57. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W29-37.

58. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

59. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.

60. Ziebuhr J, Bayer S, Cowley JA, Gorbalenya AE: **The 3C-like proteinase of an invertebrate nidovirus links coronavirus and potyvirus homologs**. *J Virol* 2003, **77**(2):1415-1426.

61. Blanck S, Stinn A, Tsiklauri L, Zirkel F, Junglen S, Ziebuhr J: **Characterization of an alphamesonivirus 3C-like protease defines a special group of nidovirus main proteases**. *J Virol* 2014, **88**(23):13747-13758.

62. Söding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**(7):951-960.

63. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment**. *Nat Methods* 2012, **9**(2):173-175.

64. Cole C, Barber JD, Barton GJ: **The Jpred 3 secondary structure prediction server**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W197-201.

65. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices**. *J Mol Biol* 1999, **292**(2):195-202.

66. Robert X, Gouet P: **Deciphering key features in protein structures with the new ENDscript server**. *Nucleic Acids Res* 2014, **42**(Web Server issue):W320-324.

67. Coordinators NR: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2015, **43**(Database issue):D6-17.

68. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic Acids Res* 2014, **42**(Database issue):D222-D230.

69. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

70. Jones DT: **GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences**. *J Mol Biol* 1999, **287**(4):797-815.

71. McGuffin LJ, Jones DT: **Improvement of the GenTHREADER method for genomic fold recognition**. *Bioinformatics* 2003, **19**(7):874-881.

72. Lobley A, Sadowski MI, Jones DT: **pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination**. *Bioinformatics* 2009, **25**(14):1761-1767.

73. Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS: **Bio3d: an R package for the comparative analysis of protein structures**. *Bioinformatics* 2006, **22**(21):2695-2696.

74. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci U S A* 1992, **89**(22):10915-10919.

75. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0**. *Syst Biol* 2010, **59**(3):307-321.

76. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language**. *Bioinformatics* 2004, **20**(2):289-290.

77. R Core Team: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing; 2011.

78. Hanoulle X, Van Damme J, Staes A, Martens L, Goethals M, Vandekerckhove J, Gevaert K: **A new functional, chemical proteomics technology to identify purine nucleotide binding sites in complex proteomes**. *J Proteome Res* 2006, **5**(12):3438-3445.

79. van den Born E, Gultyaev AP, Snijder EJ: **Secondary structure and function of the 5'-proximal region of the equine arteritis virus RNA genome**. *RNA* 2004, **10**(3):424-437.

80. Nedialkova DD, Gorbalenya AE, Snijder EJ: **Arterivirus Nsp1 modulates the accumulation of minus-strand templates to control the relative abundance of viral mRNAs**. *PLoS Pathog* 2010, **6**(2):e1000772.

81. van der Meer Y, van Tol H, Locker JK, Snijder EJ: **ORF1a-encoded replicase subunits are involved in the membrane association of the arterivirus replication complex**. *J Virol* 1998, **72**(8):6689-6698.

82. Pfefferle S, Krahling V, Ditt V, Grywna K, Muhlberger E, Drosten C: **Reverse genetic characterization of the natural genomic deletion in SARS-Coronavirus strain Frankfurt-1 open reading frame 7b reveals an attenuating function of the 7b protein in-vitro and in-vivo**. *Virol J* 2009, **6**:131.

83. Tischer BK, Smith GA, Osterrieder N: **En passant mutagenesis: a two step markerless red recombination system**. *Methods Mol Biol* 2010, **634**:421-430.

84. Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, Snijder EJ, Canard B, Imbert I: **One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities**. *Proc Natl Acad Sci U S A* 2014, **111**(37):E3900-3909.

85. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM: **Analysis of catalytic residues in enzyme active sites**. *J Mol Biol* 2002, **324**(1):105-121.

86. Henderson BR, Saeedi BJ, Campagnola G, Geiss BJ: **Analysis of RNA binding by the dengue virus NS5 RNA capping enzyme**. *PLoS One* 2011, **6**(10):e25795.

87. Shuman S, Schwer B: **RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases**. *Mol Microbiol* 1995, **17**(3):405-410.

88. Shuman S, Lima CD: **The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases**. *Curr Opin Struct Biol* 2004, **14**(6):757-764.

89. Traut TW: **Physiological concentrations of purines and pyrimidines**. *Mol Cell Biochem* 1994, **140**(1):1-22.

90. Chakravarty AK, Subbotin R, Chait BT, Shuman S: **RNA ligase RtcB splices 3'-phosphate and 5'-OH ends via covalent RtcB-(histidinyl)-GMP and polynucleotide-(3')pp(5')G intermediates**. *Proc Natl Acad Sci U S A* 2012, **109**(16):6072-6077.

91. Duclos B, Marcandier S, Cozzone AJ: **Chemical properties and separation of phosphoamino acids by thin-layer chromatography and/or electrophoresis**. *Methods Enzymol* 1991, **201**:10-21.

92. Kang H, Bhardwaj K, Li Y, Palaninathan S, Sacchettini J, Guarino L, Leibowitz JL, Kao CC: **Biochemical and genetic analyses of murine hepatitis virus Nsp15 endoribonuclease**. *J Virol* 2007, **81**(24):13587-13597.

93. te Velthuis AJ, Arnold JJ, Cameron CE, van den Worm SH, Snijder EJ: **The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent**. *Nucleic Acids Res* 2010, **38**(1):203-214.

94. Kumar NV, Govil G: **Theoretical studies on protein-nucleic acid interactions. III. Stacking of aromatic amino acids with bases and base pairs of nucleic acids**. *Biopolymers* 1984, **23**(10):2009-2024.

95. Bartlett GJ, Borkakoti N, Thornton JM: **Catalysing new reactions during evolution: economy of residues and mechanism**. *J Mol Biol* 2003, **331**(4):829-860.

96. Schmelz S, Naismith JH: **Adenylate-forming enzymes**. *Curr Opin Struct Biol* 2009, **19**(6):666-671.

97. Ahola T, Laakkonen P, Vihinen H, Kaariainen L: **Critical residues of Semliki Forest virus RNA capping enzyme involved in methyltransferase and guanylyltransferase-like activities**. *J Virol* 1997, **71**(1):392-397.

98. Nandakumar J, Shuman S, Lima CD: **RNA ligase structures reveal the basis for RNA specificity and conformational changes that drive ligation forward**. *Cell* 2006, **127**(1):71-84.

99. Decroly E, Ferron F, Lescar J, Canard B: **Conventional and unconventional mechanisms for capping viral mRNA**. *Nat Rev Microbiol* 2011, **10**(1):51-65.

100. 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**(2):557-565.

101. van Vliet AL, Smits SL, Rottier PJ, de Groot RJ: **Discontinuous and non-discontinuous subgenomic RNA transcription in a nidovirus**. *EMBO J* 2002, **21**(23):6571-6580.

102. Sagripanti JL, Zandomeni RO, Weinmann R: **The cap structure of simian hemorrhagic fever virion RNA**. *Virology* 1986, **151**(1):146-150.

103. Issur M, Geiss BJ, Bougie I, Picard-Jean F, Despins S, Mayette J, Hobdey SE, Bisaillon M: **The flavivirus NS5 protein is a true RNA guanylyltransferase that catalyzes a two-step reaction to form the RNA cap structure**. *RNA* 2009, **15**(12):2340-2350.

104. Ahola T, Ahlquist P: **Putative RNA capping activities encoded by brome mosaic virus: methylation and covalent binding of guanylate by replicase protein 1a**. *J Virol* 1999, **73**(12):10061-10069.

105. Paul AV, Rieder E, Kim DW, van Boom JH, Wimmer E: **Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the in vitro uridylylation of VPg**. *J Virol* 2000, **74**(22):10359-10370.

106. Ambros V, Baltimore D: **Protein is linked to the 5' end of poliovirus RNA by a phosphodiester linkage to tyrosine**. *J Biol Chem* 1978, **253**(15):5263-5266.

107. Pan J, Lin L, Tao YJ: **Self-guanylylation of birnavirus VP1 does not require an intact polymerase activity site**. *Virology* 2009, **395**(1):87-96.

108. Mitra T, Sosnovtsev SV, Green KY: **Mutagenesis of tyrosine 24 in the VPg protein is lethal for feline calicivirus**. *J Virol* 2004, **78**(9):4931-4935.

109. Jiang J, Laliberte JF: **The genome-linked protein VPg of plant viruses-a protein with many partners**. *Curr Opin Virol* 2011, **1**(5):347-354.

110. Zeddam JL, Gordon KH, Lauber C, Alves CA, Luke BT, Hanzlik TN, Ward VK, Gorbalenya AE: **Euprosterna elaeasa virus genome sequence and evolution of the Tetraviridae family: emergence of bipartite genomes and conservation of the VPg signal with the dsRNA Birnaviridae family**. *Virology* 2010, **397**(1):145-154.

111. Imbert I, Guillemot JC, Bourhis JM, Bussetta C, Coutard B, Egloff MP, Ferron F, Gorbalenya AE, Canard B: **A second, non-canonical RNA-dependent RNA polymerase in SARS coronavirus**. *EMBO J* 2006, **25**(20):4933-4942.

CHAPTER 4

A planarian nidovirus expands the limits of RNA genome size

Amir Saberi[#]
Anastasia A. Gulyaeva[#]
John L. Brubacher
Phillip A. Newmark
Alexander E. Gorbalenya

[#]equal contribution

PLoS Pathogens (2018) 14(11):e1007314
DOI: 10.1371/journal.ppat.1007314

ABSTRACT

RNA viruses are the only known RNA-protein (RNP) entities capable of autonomous replication (albeit within a permissive host environment). A 33.5 kilobase (kb) nidovirus has been considered close to the upper size limit for such entities; conversely, the minimal cellular DNA genome is in the 100–300 kb range. This large difference presents a daunting gap for the transition from primordial RNP to contemporary DNA-RNP-based life. Whether or not RNA viruses represent transitional steps towards DNA-based life, studies of larger RNA viruses advance our understanding of the size constraints on RNP entities and the role of genome size in virus adaptation. For example, emergence of the largest previously known RNA genomes (20–34 kb in positive-stranded nidoviruses, including coronaviruses) is associated with the acquisition of a proofreading exoribonuclease (ExoN) encoded in the open reading frame 1b (ORF1b) in a monophyletic subset of nidoviruses. However, apparent constraints on the size of ORF1b, which encodes this and other key replicative enzymes, have been hypothesized to limit further expansion of these viral RNA genomes. Here, we characterize a novel nidovirus (planarian secretory cell nidovirus; PSCNV) whose disproportionately large ORF1b-like region including unannotated domains, and overall 41.1-kb genome, substantially extend the presumed limits on RNA genome size. This genome encodes a predicted 13,556-aa polyprotein in an unconventional single ORF, yet retains canonical nidoviral genome organization and expression, as well as key replicative domains. These domains may include functionally relevant substitutions rarely or never before observed in highly conserved sites of RdRp, NiRAN, ExoN and 3CLpro. Our evolutionary analysis suggests that PSCNV diverged early from multi-ORF nidoviruses, and acquired additional genes, including those typical of large DNA viruses or hosts, which might modulate virus-host interactions. PSCNV's greatly expanded genome, proteomic complexity, and unique features – impressive in themselves – attest to the likelihood of still-larger RNA genomes awaiting discovery.

AUTHOR SUMMARY

RNA viruses are the only known RNA-protein (RNP) entities capable of autonomous replication. The upper genome size for such entities was assumed to be <35 kb; conversely, the minimal cellular DNA genome is in the 100–300 kilobase (kb) range. This large difference presents a daunting gap for the proposed evolution of contemporary DNA-RNP-based life from primordial RNP entities. Here, we describe a nidovirus from planarians, named planarian secretory cell nidovirus (PSCNV), whose 41.1 kb genome is 23% larger than any riboviral genome yet discovered. This increase is nearly equivalent in size to the entire poliovirus genome, and it equips PSCNV with an unprecedented extra coding capacity to adapt. The PSCNV has broken apparent constraints on the size of the genomic subregion that encodes core replication machinery in other nidoviruses, including coronaviruses, and has acquired genes not previously observed in RNA viruses. This virus challenges and advances our understanding of the limits to RNA genome size.

INTRODUCTION

Radiation of primitive life as it took hold on earth was likely accompanied by genome expansion, which was associated with increased complexity and a proposed progression from RNA-based through RNA-protein to DNA-based life [1]. The feasibility of an autonomous ancient RNA genome, and the mechanisms underlying such fateful transitions are challenging to reconstruct. It is especially unclear whether RNA entities ever evolved genomes close to the 100–300 kilobase (kb) range [2, 3] of the "minimal" reconstructed cellular DNA genome [4]. This range overlaps with the upper size limit of nuclear pre-mRNAs [5], which is likely the upper size limit for functional RNAs due to the relative chemical lability of RNA compared to DNA. However, pre-mRNAs are incapable of self-replication, the defining property of primordial genomic RNAs.

RNA viruses may uniquely illuminate the evolutionary constraints on RNA genome size [6-9], whether or not they descended directly from primitive RNA-based entities [10-13]. The same constraints may also inform research on biology and pathogenesis of RNA virus infections, because they shape the diversity of viral proteomes and RNA elements. The causes and consequences of changes in genome size can be understood in the context of a relationship that locks replication fidelity, genome size, and complexity within a unidirectional triangle [14]. RNA viruses appear to be trapped in the low state of this relationship (Eigen trap) [15], which is characterized by low fidelity (high mutation rate), small genome size (10 kb average), and low complexity (few protein/RNA elements). Specifically, low-fidelity replication without proofreading constrains genome expansion [16], since accumulation of mutations [17] would lead to the meltdown of larger genomes during replication (error catastrophe hypothesis) [18, 19].

This constraining relationship is supported by evidence from nidoviruses (order *Nidovirales*): enveloped viruses with positive-stranded RNA genomes in the range of 12.7 to 33.5 kb – the largest known RNA genomes [20-23] (Figure 1A,B, Table S1). The *Nidovirales* is composed of two vertebrate families, *Arteriviridae* and *Coronaviridae* (subfamilies *Coronavirinae* and *Torovirinae*), and two invertebrate families, *Mesoniviridae* and *Roniviridae* [24, 25], and includes important pathogens of humans (Severe acute respiratory syndrome coronavirus, SARS-CoV; Middle eastern respiratory syndrome coronavirus, MERS-CoV) and livestock (different arteriviruses, coronaviruses and roniviruses) [26-30]. All known nidoviruses with genomes larger than 20 kb also encode a proofreading exoribonuclease (ExoN) [14, 31-34] (Figure 1B), which, once acquired by an ancestral nidovirus, may have relieved the constraints on all three elements of the triangular relationship *simultaneously,* providing a solution to the Eigen trap [14].

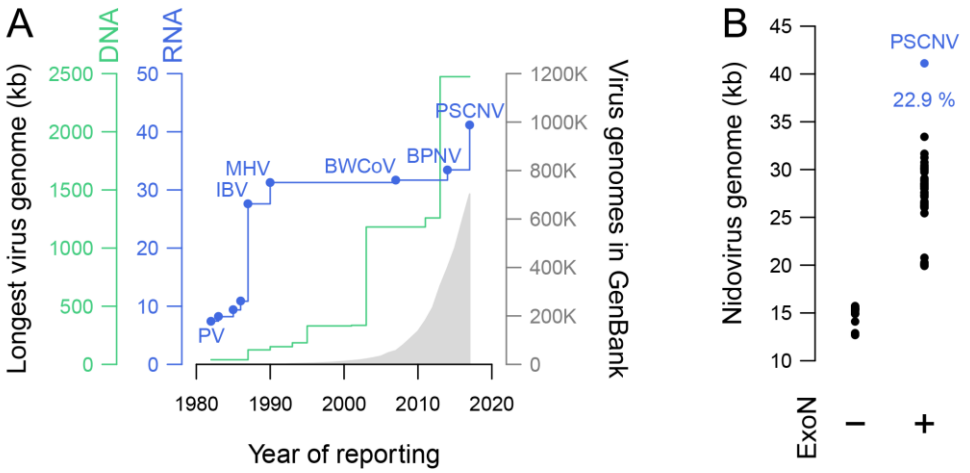


Figure 1 | Genome sizes of nidoviruses. (**A**) Timeline of discovery of largest RNA and DNA virus genomes versus accumulation of virus genome sequences in GenBank (1982–2017). PV, poliovirus; and nidoviruses: IBV, avian bronchitis virus, MHV, mouse hepatitis virus, BWCoV, beluga whale coronavirus SW1, BPNV, ball python nidovirus and PSCNV, planarian secretory cell nidovirus. (**B**) Comparison of genome sizes between nidoviruses that do not encode an ExoN domain, and those that do. Percentage indicates the difference between sizes of PSCNV and the next-largest entity.

In the last 20 years of virus discovery, however, despite the application of unbiased metagenomics to RNA virus discovery [35, 36], the largest-known RNA viral genome has only increased ~10% in size – a mere fraction of the nearly ten-fold increase observed for DNA viruses [37-39] (Figure 1A). Thus, other constraints have apparently limited genome size, even in RNA viruses equipped with proofreading capability. Further characterization of nidovirus molecular biology, variation, and evolution may provide insight into these other factors.

Nidovirus genomes are typically organized into many open reading frames (ORFs), which occupy >90% of genome and can be divided into three regions: overlapping ORF1a and ORF1b, and multiple ORFs at the 3'-end (3'ORFs) [14] (Figure 2). The products of these regions predominantly control genome expression/replication, and virus assembly/dissemination, respectively.

ORF1a and ORF1b are expressed by translation of the genomic RNA that involves a -1 programmed ribosomal frameshifting (PRF) at the ORF1a/ORF1b overlap [40, 41]. The two polyproteins produced without or with frameshifting, pp1a (ORF1a-encoded) and pp1ab (ORF1a/ORF1b-encoded), vary in size from 1,727 to 8,108 aa. They are processed to a dozen or more proteins by the virus' main protease (3CLpro, encoded in ORF1a; Figure 2) with possible involvement of other protease(s) [42]. These and other proteins form a

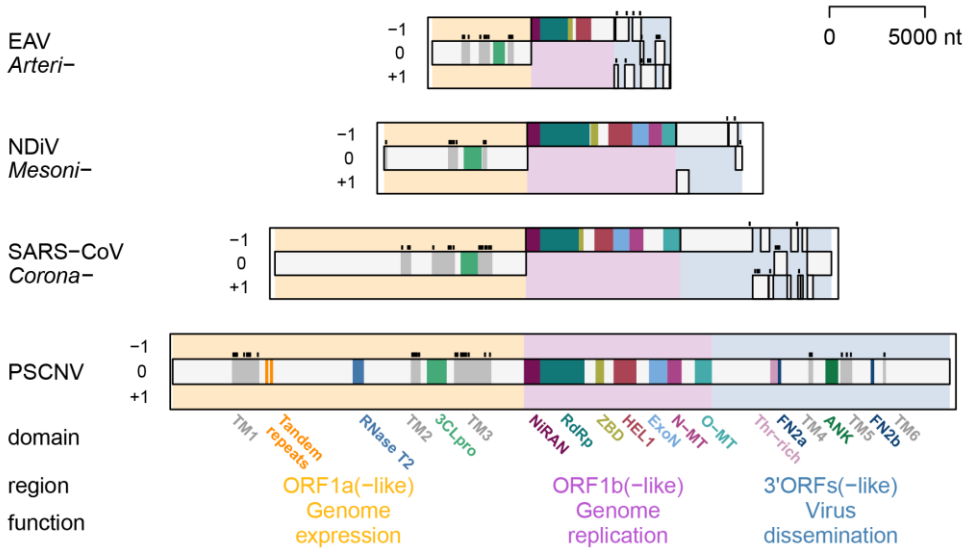


Figure 2 | Genomes and proteomes of nidoviruses. ORFs and encoded protein domains in genomes of viruses representing three nidovirus families and PSCNV. The protein-encoded part of the genomes is split in three adjacent regions, which are colored and labelled accordingly. EAV, equine arteritis virus; NDiV, Nam Dinh virus; SARS-CoV (see Table S1 for details on these viruses). ORF1a frame is set as zero. Protein domains conserved between these nidoviruses and PSCNV, and those specific to PSCNV are shown. TM, transmembrane domain (TM helices are shown by black bars above TM domains); Tandem repeats, two adjacent homologous regions of unknown function; RNase T2, ribonuclease T2 homolog; 3CLpro, 3C-like protease; NiRAN, nidovirus RdRp-associated nucleotidyltransferase; RdRp, RNA-dependent RNA polymerase; HEL1, superfamily 1 helicase with upstream Zn-binding domain (ZBD); ExoN, DEDDh subfamily exoribonuclease; N-MT and O-MT, SAM dependent N7- and 2'-O-methyltransferases, respectively; Thr-rich, region enriched with Thr residue; FN2a/b, fibronectin type 2 domains; ANK, ankyrin domain.

membrane-bound replication-transcription complex (RTC) [43, 44] that invariably includes two key ORF1b-encoded subunits: the Nidovirus RdRp-Associated Nucleotidyltransferase (NiRAN) fused to an RNA-dependent RNA polymerase (RdRp) [45, 46], and a zinc-binding domain (ZBD) fused to a superfamily 1 helicase (HEL1), respectively [47-50]. The RTC catalyzes the synthesis of genomic and 3'-coterminal subgenomic RNAs, the latter via discontinuous transcription that is regulated by leader and body transcription-regulating sequences (lTRS and bTRS) [51-53]. Subgenomic RNAs are translated to express virion and, in ExoN-positive viruses, accessory proteins encoded in the 3'ORFs [23, 54-59]. Most nidovirus proteins are multifunctional, but some released from the N-terminus of pp1a/pp1ab and/or encoded in the 3'ORFs are specialized in the modulation of virus-host interaction [26, 60-65].

Intriguingly, despite the large variation in genome size among extant nidoviruses, the size of ORF1b varies extremely little within either the ExoN-negative (12.7-15.7 kb genome range) or ExoN-positive (19.9-33.5 kb genome range) nidoviruses [66]. There is no overlap

between these two groups of viruses in the size range of ORF1b: the smallest ORF1b of an ExoN-positive nidovirus is almost double the length of the largest ExoN-negative ORF1b. In contrast, the ORF1a and 3'ORFs regions exhibit considerable size variation, and their sizes overlap between the ExoN-positive and ExoN-negative clades.

A current theoretical model of nidoviral genome dynamics, the three-wave model, proposes that genome expansion cycle is initiated by a rare increase of ORF1b (the first wave) in a common ancestor of ExoN-positive nidoviruses, which then permits parallel expansion of ORF1a and, often, 3'ORFs in subsequent overlapping waves in separate lineages [66]. Extant nidovirus genomes of different sizes have reached particular points on this trajectory of genome size, apparently due to the lineage-specific interplay of poorly understood genetic and host-specific factors. A single cycle of this process can account for genome expansion from the lower end of genome sizes (12.7 kb) to the upper end (31.7 kb); expansion of genomes far beyond that size range has been hypothesized to require a second cycle, beginning with a new wave of ORF1b expansion [66]. In the absence of newly discovered RNA viruses with significantly larger genomes since the time of that analysis, and due to the unknown nature of the ORF1b size constraint(s), however, the feasibility of a second cycle has remained uncertain, and the notion that ~34 kb is close to the actual limit of RNA virus genome size [35] has seemed plausible.

To examine whether this limit applies beyond the currently recognized ~3000 RNA virus species (isolated from only a few hundred host species), further sampling of virus diversity is required, particularly from host species in which viruses have thus far remained virtually unknown. To this end, we analyzed *de novo* transcriptomes from both major reproductive biotypes (strains) of the planarian *Schmidtea mediterranea* [67]: a hermaphroditic sexual strain, and an asexual strain whose members reproduce via transverse fission [68]. We report the discovery and characterization of the first known planarian RNA virus, dubbed the planarian secretory cell nidovirus. PSCNV has the largest RNA genome by a considerable margin – a feat made more remarkable by the fact that its genome is organized as a single ORF. Concomitantly, it has adapted the nidoviral regulatory toolkit in novel ways, and acquired many features that revise the known limits of viral genomic and proteomic variation – some of these features being unique among nidoviruses, others among RNA viruses, and still others among all known viruses. Our results imply that viruses with the nidoviral genetic plan have potential to expand RNA genomes further along the trajectory envisioned by the multi-cycle three-wave model.

RESULTS

Identification and genomic assembly of a large RNA virus from planarians

To identify potential nidovirus-like sequences in the planarian transcriptome, we queried two in-house *de novo*-assembled *Schmidtea mediterranea* transcriptomes [67] for sequences that significantly resembled a reference coronavirus genome. Two nearly identical (99.97%) nested transcripts, txv3.2-contig_1447 (originating from the sexual strain) and txv3.1-contig_12746 (from the asexual strain), showed a statistically significant similarity to known nidoviruses as reciprocal BLAST top hits. We hypothesized that these transcripts are genomic fragments of a new nidovirus species. We further identified several overlapping EST clones with >99% nucleotide identity to the transcriptome contigs, and assembled these into a putative partial genome (Figure S1). Finally, with additional transcriptome search iterations and Sanger sequencing of the transcript 5'-end, we assembled a 41,103-nt transcript (excluding the polyA tail). Based on several criteria (see below), we assigned this RNA sequence to the genome of a virus we dubbed Planarian Secretory Cell Nidovirus (PSCNV) (Figure S1) This sequence was the reference genome used for further analyses (see Materials and Methods for more detail).

The complete PSCNV genome encodes a single 40,671-nt ORF that is flanked by a 128-nt 5'-UTR and a 304-nt 3'-UTR (Figures 1B,2). In addition, we detected multiple small ORFs in the genome region of the main ORF whose lengths exceeded 150 nt: 8 ORFs in the same strand as the large ORF (plus-strand), length ranging from 156 to 267 nt, 5 of which mapped to the 3'-terminal quarter of the genome; and 24 ORFs in the reverse complement strand (minus-strand), distributed throughout the genome, with lengths ranging from 153 to 681 nt. To further verify the presence of the viral genome *in vivo,* we amplified large overlapping genomic subregions by RT-PCR (Table S2, Figure S1) [69]. These sequences could not be amplified from *S. mediterranea* genomic DNA, nor could they be found in the reference planarian genome [70]; thus, they appear to derive from an exogenous source.

PSCNV variants in worldwide planarian laboratories imply recent virus transmission

A survey of 14 *S. mediterranea* RNA-seq datasets from nine laboratories worldwide uncovered PSCNV reads in five datasets from three American locations. Of the positive datasets, three originated from the sexual strain, and two from the asexual strain. Overall, viral sequences were much more abundant in transcriptomes obtained from sexual strains (Table S3).

The PSCNV sequences detected in these studies vary little from one another. The three most complete sequences (tentatively reconstructed from PRJNA319973, PRJNA79031, and PRJNA421285) are characterized by >99.9% identity across a nearly 13 kb span of the genome, where all three are based on reference genome coverage by reads of at least 2x (and at least 10x for >95% of positions). Indeed, sequences from PRJNA319973 and PRJNA79031 – the two datasets from the Newmark laboratory – exhibit only a single mutation relative to the reference genome, and the sequence from PRJNA421285 – from the Sanchez Alvarado laboratory – differs at only 9 positions (Table S4). This low variation is notable, as two of the datasets analyzed (PRJNA79031 and PRJNA421285) are derived from sexual *S. mediterranea*, and the other one (PRJNA319973) from an asexual *S. mediterranea* lab strain. The source populations of these two strains are separated from each other by about 500 km of the Mediterranean Sea: the asexual laboratory strain was established from a population in Barcelona [71], and the sexual strain originates from a Sardinian population. A recent study of the evolutionary history of *S. mediterranea* suggests that these populations diverged from each other at least 4 million years ago [72].

Given the long-separate history of these two planarian strains prior to becoming subject of research, and the relatively high mutation rate in characterized nidoviruses, the detection of nearly identical viral transcripts in both is strong evidence that the virus is transmissible. The absence of viral sequences from asexual strains in most labs, and their presence in all labs that have reported RNA-seq data from the sexual strain, strongly suggest that the virus first infected (or was endemic to) the sexual strain, and has subsequently spread to asexual stocks.

PSCNV infects the secretory cells of planarians

We examined PSCNV infection in planarian tissues by whole-mount in situ hybridization (ISH). PSCNV RNA was detected abundantly in cells of the secretory system in both sexuals and asexuals (Figure 3A). Fluorescent ISH revealed viral RNA in gland cell projections that form secretory canals (Figure 3B). Notably, viral RNA was detected largely in ventral cells (Figure 3C) whose localization corresponds to mucus-secreting cells that produce the slime planarians use for gliding locomotion, and to immobilize prey [73].

We then analyzed planarians by electron microscopy (EM) for the presence of viral structures. In one specimen, membrane-bound compartments containing 90–150 nm spherical-to-oblong particles resembling nidoviral nucleocapsids [74, 75] were found in the cytoplasm of mucus-secreting cells. These sub-epidermal gland cells are notable for their abundant rough endoplasmic reticulum and long projections into the ventral epithelium, through which they secrete mucus (Figure S2). These cells provide an ideal environment for nidoviral replication, which co-opts host membranes to produce viral replication complexes [76, 77]. Putative viral particles were found both in deep regions of

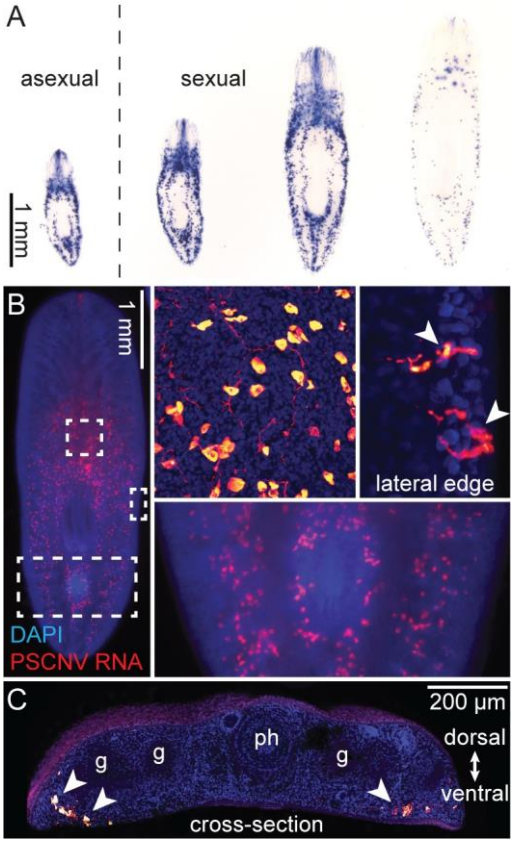


Figure 3 | Expression of PSCNV RNA in planarians. (**A**) PSCNV RNA (blue) detected in asexual (left) and sexual *S. mediterranea* by whole-mount ISH. (**B**) Fluorescent ISH showing PSCNV expression in a sexual planarian. Insets show higher magnification of areas indicated by boxes. Top two insets are confocal projections. Secretory cell projections to lateral body edges are indicated by arrowheads. (**C**) Tiled confocal projections of PSCNV expression in a cross-section. Cells expressing PSCNV are ventrally located (arrowheads). Gut ("g") and pharynx ("ph") are indicated. DAPI (blue) labels nuclei.

these cells, and in their trans-epidermal projections (Figure 4A–C). The latter location suggests a route for viral transmission. Notably, particles in sub-epidermal layers have a "hazy" appearance and are embedded in a relatively electron-dense matrix (Figure 4D). In contrast, particles closer to the apical surface of the epidermis appear as relatively discrete structures, standing out against electron-lucent surrounding material (Figure 4E). The size, ultrastructure, and host-cell locations are all consistent with these structures being nidoviral nucleocapsids [74, 75].

In 280 images from the positive specimen, all other ultrastructural features were normal. Importantly, typical mucus vesicles were evident in this specimen, often immediately adjacent to vesicles containing putative virions (Figure 4C, see also Figure S2). As such, we

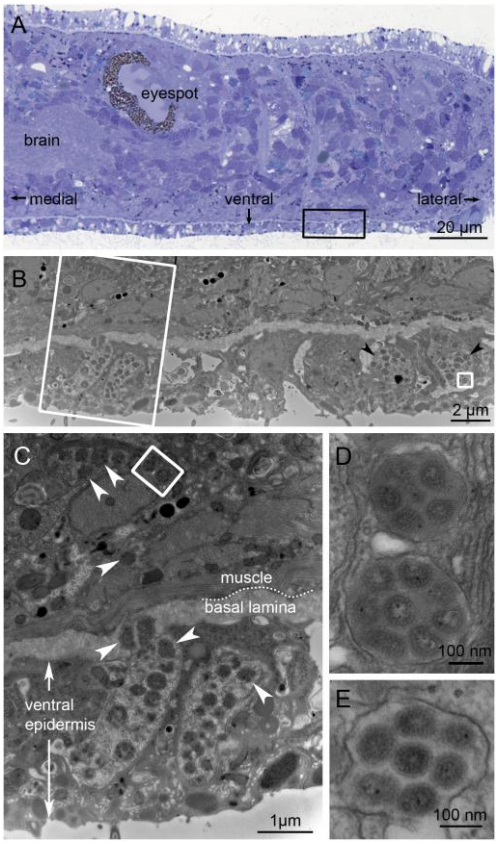


Figure 4 | Putative PSCNV particles revealed by electron microscopy. (**A**) Adjacent histological transverse section, to orient EM images. Black rectangle corresponds to location of (**B**), a low magnification EM view to provide context. White rectangle corresponds to location of (**C**), in which putative viral particles enclosed within membrane sacs are indicated by arrowheads. White rectangle in (C) and square in (B) indicate positions of higher magnification views shown in (**D**) and (**E**), respectively, each illustrating several viral particles within a membrane sac. In top-left of (C), note the mucus granules adjacent to virus laden sacs (see also Fig. S2). Scale bars as indicated.

determined that these structures do not represent artefacts caused by atypical fixation of this specimen.

Overview of the PSCNV proteome reveals a unique nidovirus

The genome and proteome of PSCNV are by far the largest yet reported for an RNA virus. Its RNA genome is ~25% larger than that of the next-largest known RNA virus (BPNV, [21]), which is separated by a comparable margin from the first nidovirus genome sequenced 30 years ago (IBV, [78]) (Figure 1A). The size of the predicted PSCNV polyprotein (13,556 amino acids, aa) is 58–67% larger than the largest known RNA virus proteins produced

from a single ORF (8,572 aa; Gamboa mosquito virus, [79]) or multiple ORFs through frameshifting (8,108 aa; BPNV, [21]) (Figure 5).

Functional annotation of the PSCNV polyprotein by comparative genomics [14, 31, 80, 81] presented a distinct bioinformatics challenge, due to its weak similarity to other proteins and its extremely large size, which exceeds the average size of protein domains by approximately 75-fold. We delineated at least twenty domains in the PSCNV polyprotein, including twelve domains conserved in nidoviruses or other entities, using a multistage computational procedure that combined different analyses within a probabilistic framework (Figure 2; Figure S3-S16; Table S5; see Materials and Methods). We initially identified six regions highly enriched in hydrophobic residues characteristic of transmembrane domains, named TM1 to TM6 accordingly (Figure 2). The number and relative location of the TM domains resemble those found in the proteomes of nidoviruses, which commonly have five or more TM domains in non-structural and structural proteins [82-85]. We then identified fourteen regions enriched in individual amino acid residues (Figure S4), with the strongest signal observed for Thr-rich region (residues 10429–10559, 44.3% Thr residues, up to 13.4 SD above the mean). Notably, the Thr-rich region overlaps with a Ser-rich region (10461–10501 aa, 19.5% Ser residues, up to 5.5 SD above the mean). Subsequently, two tandem repeats were identified toward the N-terminus of the polyprotein (residues 1616–1682 and 1686–1751, Probability 96.6%, Figure S5), which showed no significant similarity to other proteins in the databases using HHsearch.

We used the domains described above to split the polyprotein into nine regions, which were analyzed by an iterative HHsearch-based procedure (outlined in Figure S3 and SI Materials and Methods). Our approach identified eight domains that, together with TM2 and TM3, form a canonical synteny of replicative domains in the central part of the polyprotein (genome), which is characteristic of known invertebrate nidoviruses (Figure 2): 3CLpro, NiRAN, RdRp, ZBD, HEL1, ExoN, and S-adenosylmethionine (SAM)-dependent N7- and 2'-O-methyltransferases (N-MT and O-MT, respectively). Five of these domains (3CLpro, NiRAN, RdRp, HEL1, and O-MT) were identified by hits exceeding the 95% Probability threshold, while three others were based on weaker hits: 35.0% for ZBD, 39.1% for ExoN, and 80.8% for N-MT. Despite the lower Probability values obtained for the latter three domains, synteny and conservation of essential functional residues strongly suggest that they encode true homologs of canonical nidoviral proteins. Overall, the analysis demonstrates the existence of the three definitive nidoviral genomic subregions in the PSCNV single-ORF genome: ORF1a-, ORF1b-, and 3'ORFs-like. Within these regions, TM2, 3CLpro, and TM3 map to the ORF1a-like region, while NiRAN, RdRp, ZBD, HEL1, ExoN, N-MT, and O-MT map to the ORF1b-like region.

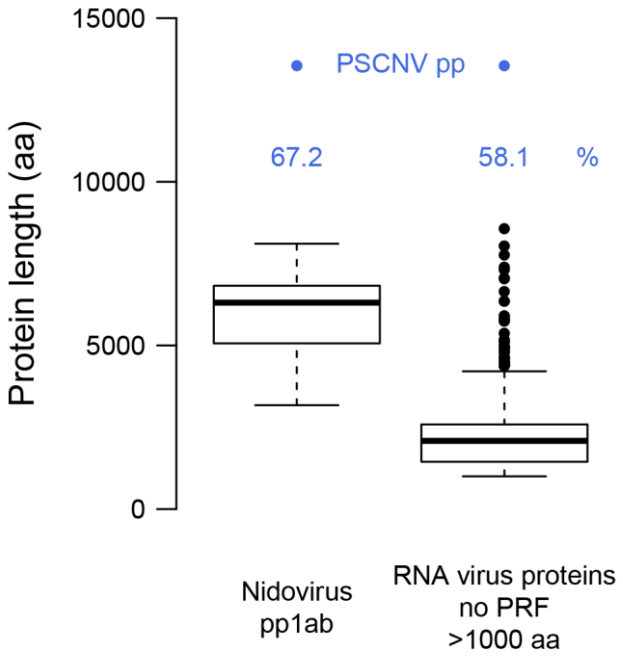


Figure 5 | Largest proteins of nidoviruses and other RNA viruses in comparison with PSCNV polyprotein. Percentage indicates the difference between sizes of the PSCNV polyprotein (pp) and that of the next-largest entity. For details, see SI Materials and Methods.

In addition to the canonical replicative domains present in the canonical order and location, we found four domains that are novel for nidoviruses: one upstream and three downstream of the array of the conserved replicative domains (Table S5). These include a homolog of ribonuclease T2 (RNase T2, Probability 80.0%) upstream of the TM2, two fibronectin type 2 domains (FN2a and FN2b, 91.3% and 78.5%, respectively), and an ankyrin repeats domain (ANK, 98.9%) downstream of the O-MT. For the three domains identified with the under-threshold hits, additional support came from conservation of functionally important residues (see below).

We subsequently generated multiple sequence alignments (MSAs) of these domains for a representative set of established nidovirus species, followed by phylogenetic reconstruction to characterize PSCNV by revealing common and unique features of its conserved domains. The next three sections summarize the salient features of the replicative, novel, and structural domains of the polyprotein.

Conserved and distinctive features in PSCNV's replicative and regulatory proteins

3CL protease (main protease of polyprotein processing)

Nidoviruses employ an ORF1a-encoded protease, 3CLpro, with a narrow substrate specificity that controls expression of ORF1a and ORF1b by releasing itself and downstream domains comprising replicative machinery, up to and including the most C-terminal domain encoded by ORF1b [42]. This protease includes a catalytic domain composed of a two-barrel chymotrypsin-like fold and a C-terminal accessory domain whose fold varies among nidoviruses [86, 87]. It is flanked by two TM domains in the polyprotein (TM2 and TM3), which anchor the RTC to the membrane [43] (Figure 2). The catalytic domain of PSCNV 3CLpro was identified in the canonical position between TM2 and TM3 (Figure S3) through hits to hidden Markov model (HMM) profiles of cellular serine proteases with chymotrypsin-like folds, while its similarity to the HMM profile of the nidovirus 3CLpro was extremely low (Probability 2.8%; see Table S5), indicating unique properties. The long distance (~250 aa) between the C-terminus of the putative catalytic domain of PSCNV 3CLpro and the N-terminus of TM3, suggests that PSCNV 3CLpro possesses a highly divergent C-terminal domain. Unlike other characterized invertebrate nidoviruses, which all employ cysteine as the catalytic nucleophile [88, 89], PSCNV 3CLpro appears to use the Ser-His-Asp catalytic triad typical of cellular chymotrypsin-like proteases (Figure S7). PSCNV 3CLpro was also found to have a residue variation that has never been observed in 3CLpro-encoding viruses before: it encodes a Val residue in the position commonly occupied by a His residue in the putative substrate-binding pocket (GX**V** vs G/YX**H**, highlighted in bold) [42, 88-91].

NiRAN, RdRp, ZBD, HEL1 (RNA replicative enzyme domains)

Consistent with the essential enzymatic activities of RdRp (the catalytic domain of RNA polymerase) and HEL1 (helicase), the PSCNV polyprotein hits to HMM profiles of these domains were ranked as the top two by two measures of statistical significance (Table S5). Mutiple sequence alignments confirm the high conservation of canonical motifs and residues in these domains (Figures S9 and S11). The only exception concerns the RdRp C motif: a Ser residue of the nidovirus-specific SDD signature [23] is replaced by Gly in PSCNV. As in previously described nidoviruses, PSCNV's HEL1-associated ZBD includes 12 Cys or His residues that are homologous to putative Zn-binding residues (Figure S10). The PSCNV RdRp-associated NiRAN retains six out of the seven invariant residues observed in all known nidoviruses [45] (Figure S8). The outlier is in motif B_N, in which Thr takes the place of an invariant Asp as the distal residue. In addition, the B_N motif in PSCNV also contains an Asn at a highly conserved Ser/Thr position. These substitutions might represent the "swapping" of the two residues, assuming that the chemically similar Asp

and Asn residues play an equivalent role in the respective proteins. This hypothesis is plausible, given that the two affected residues are expected to be in close proximity to each other, separated only by an incomplete turn of the putative alpha-helix of the motif B_N (Figure S8). Another notable feature of the PSCNV NiRAN is the large distance between invariant Lys and Glu residues of the motif A_N: 20 aa in PSCNV compared to 5–9 aa in other nidoviruses. The conservation of NiRAN and ZBD in PSCNV is significant for assignment of this virus to nidoviruses, since both domains are the only known genetic markers of the order *Nidovirales*.

ExoN, N-MT, O-MT (proofreading and RNA-modifying enzyme domains)

ExoN is a 3'-5' exoribonuclease that improves the fidelity of replication and transcription by excision of a 3' mismatched nucleotide in characterized nidoviruses [31-34, 92-94]. Like its orthologs, the PSCNV ExoN contains the characteristic D-E-D-H-D pentad, which includes counterparts of catalytic and other active site residues. The H-D subset is embedded within a highly conserved domain, whose structure is maintained by two Cys and two His residues coordinating a Zn^{2+} in characterized nidoviruses. However, these residues are substituted in PSCNV (H-C-H-C by E-S-Q-Q), which may therefore lack this Zn-finger (Figure S12). In this respect, PSCNV ExoN is more similar to its cellular homologs than to those of nidoviruses (Table S5). In contrast, the ExoNs of all ExoN-positive nidoviruses, including PSCNV, include another (upstream) Zn-finger, which distinguishes them from related enzymes of other origins. The N-MT and O-MT are implicated in viral RNA capping machinery [31, 92, 95-100]. In both transferases, a number of residues crucial for substrate and ligand binding are conserved in PSCNV homologs, including Zn-binding residues of N-MT (Figure S13), and the catalytic K-D-K-E tetrad of O-MT (Figure S14). Notably, like ExoN, O-MT is conserved in all nidoviruses with genomes >20 kb.

PSCNV encodes protein domains that are novel to nidoviruses

RNase T2. The PSCNV RNase T2 homolog was identified upstream of the TM2 domain. It conserves both active-site motifs typical of such RNases, CASI and CASII, including catalytic His, Glu, and Lys residues, (Figure S6) suggesting an enzymatically active protein [101].

Fibronectin type II (FN2) domains

We identified two FN2 domains, FN2a and FN2b, with only 21.7% pairwise identity to each other, including few residues aside from the most conserved Cys and aromatic residues (Figure S15). According to the *Schmidtea mediterranea* genome database (SmedGD; [102]), several proteins of *S. mediterranea* include putative FN2 domains, but neither these nor FN2 domains of other origins show particular sequence affinity to those of PSCNV. Thus, the historical acquisition and subsequent evolution of these domains is unclear at this time.

Ankyrins

We identified three divergent ankyrin repeats in a PSCNV polyprotein region of ~100 aa (Figure S16). In searches of Uniprot and the host proteome (Smed Unigene) using BLAST, the PSCNV ANK domain yielded highly significant hits (E-values ranging from 3E-23 to 8E-14, Figure 6) to proteins from *S. mediterranea* and another free-living planarian, *Dendrocoelum lacteum* [103]. The cellular domains clustered together in a phylogenetic reconstruction of the evolutionary relationship between these proteins and the PSCNV ANK using BEAST software (LG+G4 model, relaxed clock with uncorrelated log-normal rate distribution) (Figure 6). The topology of this tree implies that an ancestor of PSCNV acquired a host ANK domain prior to the divergence of the *S. mediterranea* and *D. lacteum* lineages, but we cannot exclude an alternative explanation in case if viral ANK repeats experienced accelerated evolution compared to host sequences.

Putative structural proteins of PSCNV

The 3'ORFs region of nidoviruses encodes components of the enveloped virion [23, 54], which define receptor specificity [55-57] and typically include the nucleocapsid protein (N), characterized by biased amino acid composition and structurally disordered region(s) [104, 105], spike glycoprotein(s) (S protein in corona- and toroviruses) and transmembrane matrix protein (M in corona- and toroviruses) enriched with TM regions [58, 59, 106]. As expected from the weak sequence conservation of this region in other nidoviruses [14, 107] and its weak similarity with other viruses [108], we were unable to find statistically significant similarity between the PSCNV polyprotein and structural proteins of the known nidoviruses. Nevertheless, important nidoviral themes are evident.

First we noted that the genome distribution of the TM-encoding regions in PSCNV conformed to that observed in other nidoviruses, with TM1 and TM2 located upstream of 3CLpro, TM3 C-terminal to 3CLpro, and TM4–TM6 downstream, in the 3'ORFs-like region (Figure 2). In nidoviruses, the TM domains encoded in the 3'-genome region are known to be part of the S and M proteins or their equivalents, and occasionally additional accessory proteins [14, 58, 59, 106, 109]. The extracellular portion of the S protein is supported by multiple disulfide bridges between conserved Cys residues [56]. In PSCNV, a Cys-rich region was observed downstream of TM5 (Figure S4). In an approximately 650 aa region surrounding the TM6 domain (4.7% of the polyprotein length), we identified six areas enriched in Pro, Leu, Gly, Gln, Asn, or Arg, in close proximity to each other (Figure S4). This region accounted for 43% of all residue-enriched areas in the polyprotein; such an exceptionally high concentration of sequences enriched with specific amino acids is indicative of unusual properties. Accordingly, this area was predicted to include the longest stretch of disordered regions. In nidoviruses, disordered hydrophilic-rich areas are characteristic of N proteins.

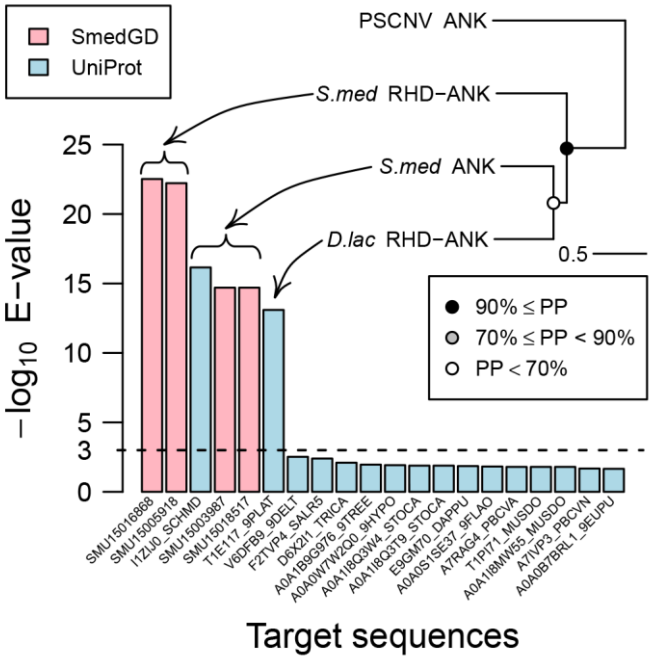


Figure 6 | ANK domain of PSCNV and its homologs. The closest cellular homologs of PSCNV ANK are ranked by similarity (left, above the broken baseline) and depicted through phylogeny (right; reconstructed and rooted by BEAST, summarized as maximum clade credibility tree; PP, posterior probability of clades) along with protein domain architecture: *S. med*, *Schmidtea mediterranea*; *D. lac*, *Dendrocoelum lacteum*; RHD, Rel homology DNA binding domain.

In PSCNV, the polyprotein region downstream of O-MT is ~4000 aa, more than twice as large as the largest known structural protein of nidoviruses [106]. We reasoned that this part of its polyprotein might be processed by cellular signal peptidase (SPase) and/or furin to produce several proteins, as documented for maturation of the structural proteins of many RNA viruses, including nidoviruses [110-114]. Indeed, our analysis of potential cleavage sites of these proteases revealed highly uneven distributions (Figure S4), with sites predicted only in the N- and C-terminal parts of the polyprotein: 1400–3100 aa (one SPase and four furin sites) and 10200–13200 aa (three SPase and five furin sites). All of these are outside of the region that must be processed by 3CLpro. With the exception of the most C-terminal furin site, all predicted sites are in close vicinity to provisional borders of the domains described above, as would be expected if these domains function as distinct proteins. Specifically, if the predicted SPase and furin sites are cleaved, TM1, TM4, TM5, and TM6 would end up in separate proteins, with one protein including the TM4 and ANK domains. With predicted cleavage sites flanking it from both sides, TM5 may be released as a separate protein, most similar to M proteins in size and hydrophobicity. We also note that two putative proteins may combine a FN2 module with a disordered region:

FN2a with a Thr/Ser-rich region and FN2b with the Pro/Leu/Gly/Gln/Asn/Arg-rich region, respectively. Based on the reasoning outlined above, the latter combination may constitute a region of the N protein.

Overall, our analysis of the predicted PSCNV proteins suggests that its genome is functionally organized in much the same manner as in the multi-ORF nidoviruses: with the non-structural and structural proteins encoded in the 5'- and 3'- regions, respectively.

PSCNV clusters with invertebrate nidoviruses in phylogenetic analyses

Next we sought to determine when PSCNV emerged, relative to other nidoviruses. The proteome analysis described above indicates that PSCNV shares the main features characteristic of invertebrate nidoviruses, although it also exhibits distinctive properties indicative of a distant relationship with previously characterized nidoviruses. To resolve very deep branching, we used an outgroup in our analysis, and selected astroviruses for this purpose [23]. Astroviruses [115] and nidoviruses share multi-ORF genome organization, a central role for 3CLpro in polyprotein processing, and similarities in the RdRp domain. Conversely, astroviruses do not encode a HEL1, NiRAN or ZBD, and their 3CLpro is highly divergent. Given the divergent 3CLpro of PSCNV, RdRp remained as the only domain most suitable for phylogeny reconstruction; this domain has been used in many studies on macroevolution of nidoviruses [21, 23, 35, 116].

We performed phylogenetic analysis of the RdRp core region by Bayesian inference (BEAST software, LG+I+G4 model, relaxed clock with uncorrelated log-normal rate distribution). Nidoviruses including PSCNV formed a monophyletic group in >90% of the trees in the analyzed Bayesian sample, with PSCNV being one of the basal branches in the cluster of invertebrate nidoviruses in 88.7% of the trees, basal to either mesoni- and roniviruses (54.7% of the trees), or roniviruses (20.6%), or mesoniviruses (13.4%) (Figure 7 and Figure S17).

In addition, we built a nidovirus phylogeny without an outgroup (BEAST software, LG+I+G4 model, relaxed clock with uncorrelated log-normal rate distribution), based on a concatenated alignment of five domains conserved in all nidoviruses (3CLpro, NiRAN, RdRp, ZDB, HEL1). Again, PSCNV belonged to the cluster of invertebrate nidoviruses in the majority of trees and was basal to either mesoni- and roniviruses (11.8% of the trees), or roniviruses (83.0%), or mesoniviruses (3.6%).

Origin of single-ORF genome organization

Is the unique single-ORF genomic organization of PSCNV an ancestral characteristic of nidoviruses, or has it evolved from an ancestral multi-ORF organization? To choose between these alternative scenarios, we need to reconstruct a genomic ORF organization

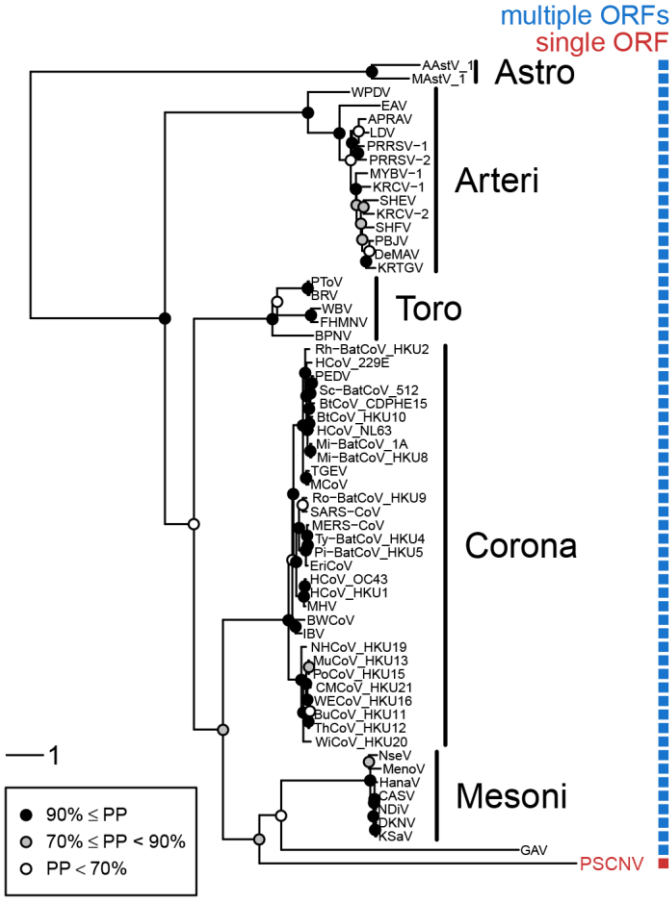


Figure 7 | Phylogeny of PSCNV. RdRp-based Bayesian maximum clade credibility tree and the genomic ORF organization (character state) for PSCNV, a representative set of nidoviruses, and astroviruses (outgroup). PP, posterior probability of clades. For virus names, see Table S1.

of the most recent common ancestor (MRCA) of nidoviruses. Such reconstruction by orthology, which was used for RdRp-based phylogeny, is not feasible with the current dataset, as none of the open reading frames or their overlaps (with the exception of the ORF1a/ORF1b junction) are conserved in all known multi-ORF nidoviruses.

To address this challenge, we noted that nidoviruses with multi-ORF organization, unlike PSCNV, recurrently use initiation and termination codons to delimit ORF-specific proteins in the 3'ORFs region, indicative of pervasive selection forces that operate in all nidoviruses except PSCNV. Therefore, we reasoned that multi- and single-ORF organizations in nidoviruses could be treated as two alternative discrete states of a single trait (ORF organization), regardless of the complexity of their actual evolutionary relations in the 3'ORFs region and assuming the rate of transition between any two multi-ORF

organizations to be extremely high compared to that between single- and multi-ORF organizations. This reasoning allows us to reformulate the question in the framework of ancestral state reconstruction analysis: if each extant nidovirus is characterized by one of the two states of a trait (ORF organization), which state of the trait was inherent for their MRCA?

To conduct this analysis, we applied the BayesTraits [117] program to the RdRp-based Bayesian sample of phylogenetic trees including the outgroup, which accounts for uncertainty in the phylogeny inference of nidoviruses. The results strongly favored multi-ORF organization of the ancestral nidovirus (Log Bayes Factor (BF) 6.06 and 6.16, when multi-ORF genome organization, or no information about genome organization, were specified as states of the trait for astroviruses, respectively) (Figure S17). Similarly, strong support (Log BF 4.79) for multi-ORF ancestral organization was obtained when the analysis was conducted based on a phylogeny without an outgroup, reconstructed using five nidovirus-wide conserved domains.

PSCNV expanded disproportionately in the ORF1b-like region

Each of the three main regions of the PSCNV genome is larger than its counterparts in all other nidoviruses (Figure 8A, Tables S1,S6). However, the size differences between PSCNV and the next largest nidovirus in each of these regions are smaller than those observed for complete genomes (Figure 8A: 5.7%, 20.6% and 15.6% for ORF1a, ORF1b and 3'ORFs, respectively, vs 22.9% for the genome). This paradoxical observation is due to profound differences in regional size variation among nidoviruses [66] such that different nidoviruses are the next largest to PSCNV for each of the three main regions (Table S1).

To account for these and other differences in sizes of the three regions while assessing the regional size increases of PSCNV, we employed two measures in addition to the percentage size increase between PSCNV and the next largest nidovirus (see Materials and Methods, formulas D_2 and D_3 versus formula D_1). First, for each genome region, we normalized the size difference between PSCNV and the next largest virus against the difference between the latter and the median-sized virus for that region (formula D_2). Second, we checked how much the deviation calculated with formula D_2 differs from that expected under a hypothesis that size changes are uniform across the three genome regions and therefore proportional to genome-wide changes (formula D_3). These measures show that, relative to the size variation among known ExoN-positive nidoviruses, the size increase in the ORF1b region was extraordinarily large (D_2=1270.5% and D_3=968.1%), while the corresponding increases in the two other regions were modest and smaller than could be expected (18.9% and 14.4% for ORF1a, and 44.3% and 33.7% for 3'ORFs) (Figure 8B, Table S6).

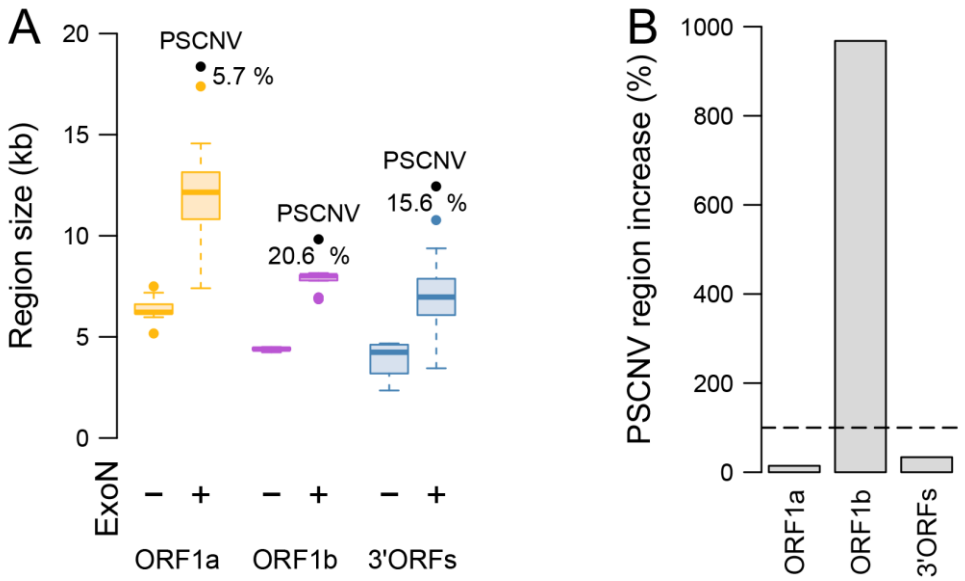


Figure 8 | Nidovirus genome and region size differences. (**A**) Sizes of three nidovirus ORF regions. Percentage indicates the difference between a genome region's size in PSCNV, and that of the next-largest entity. Color scheme as in Fig. 2. (**B**) Size increase of the three genome regions in PSCNV (grey bars) relative to the increase expected if all regions had expanded evenly (broken line); calculated using formula D3, see text and Table S6.

PSCNV genome features suggest mechanisms to regulate the stoichiometry of proteins encoded by a single-ORF genome

Virus reproduction requires different viral protein stoichiometries at distinct replicative cycle stages, a challenge for a single-ORF genome theoretically producing equimolar quantities of encoded polypeptides. To this end, all previously described nidoviruses employ -1 PRF during translation of ORF1a+ORF1b in addition to ORF1a alone from genomic template to produce two polyproteins: pp1ab and pp1a, respectively [40, 41]. The net result of this mechanism is relatively high expression of the ORF1a- compared to ORF1b-encoded proteins, since PRF occurs at the ORF1a/1b junction in 15–60% of ORF1a translation events. In contrast, proteins encoded in the 3'ORFs region are produced by translation of subgenomic (sg) mRNAs, synthesized on specific minus-strand templates [51-53], which are in turn produced by discontinuous RNA synthesis on genomic templates. Discontinuous minus-strand template synthesis relies on lTRS and bTRS, which are nearly identical, short repeats at sites where RNA synthesis pauses (upstream of 3'ORFs) and resumes (in the 5'-UTR), respectively. Templates of some sg mRNAs may be terminated at bTRS. Both transcription and translation of sg mRNAs provide a means to produce relatively large quantities of structural proteins, compared to non-structural

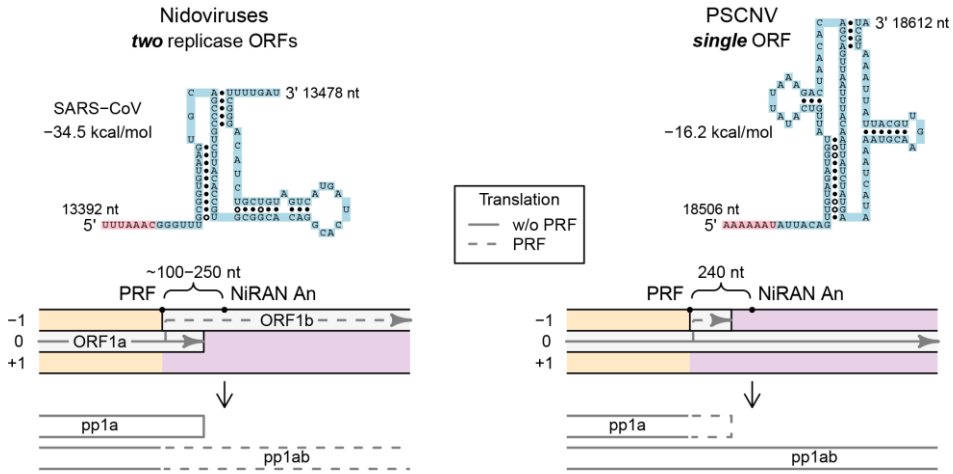


Figure 9 | Genome translation. Comparison of mechanisms by which ORFs 1a and 1b are translated in previously described nidoviruses (left) and PSCNV (right, hypothetical). On the top, RNA structure of the PRF sites, predicted by KnotInFrame, is presented: slippery sequence, pink; pseudoknot, blue.

(replicative) proteins, late in the replicative cycle, and to regulate production of accessory proteins. We analysed the PSCNV genome for evidence of such mechanisms.

Genome translation and frameshifting

ORF1a/1b -1 PRF in nidoviruses is facilitated by a pseudoknot preceded by a slippery sequence, which lies ~100–250 nt upstream of the region encoding the A_N motif of the NiRAN domain. To check if an analogous structure is present in the PSCNV genome, KnotInFrame was applied to the 1000-nt genome fragment immediately upstream of the region encoding the NiRAN A_N motif. The top prediction identified nucleotide 18512 as a putative PRF site. This nucleotide is positioned 240 nt upstream of the region encoding the NiRAN A_N motif, and the free energy of the downstream pseudoknot is -16.2 kcal/mol (Figure 9, right). Notably, when the identical procedure was applied to SARS-CoV, the top prediction (Figure 9, left) correctly identified the experimentally verified PRF site with only minor deviations between the predicted and experimentally verified structure of the downstream pseudoknot [118]. As a result of -1 PRF at the identified PSCNV site, translation would shift from the main PSCNV ORF to a small 39-nt ORF. If -1 PRF at this site indeed occurs in a fraction of ORF1a-like region translation events, translation of the ORF1b-like region (and also 3'ORFs-like region) will be attenuated, with a net result that should be similar to that of other nidoviruses: proteins encoded in the ORF1a-like region will be expressed in higher quantities than proteins encoded in the ORF1b-like region.

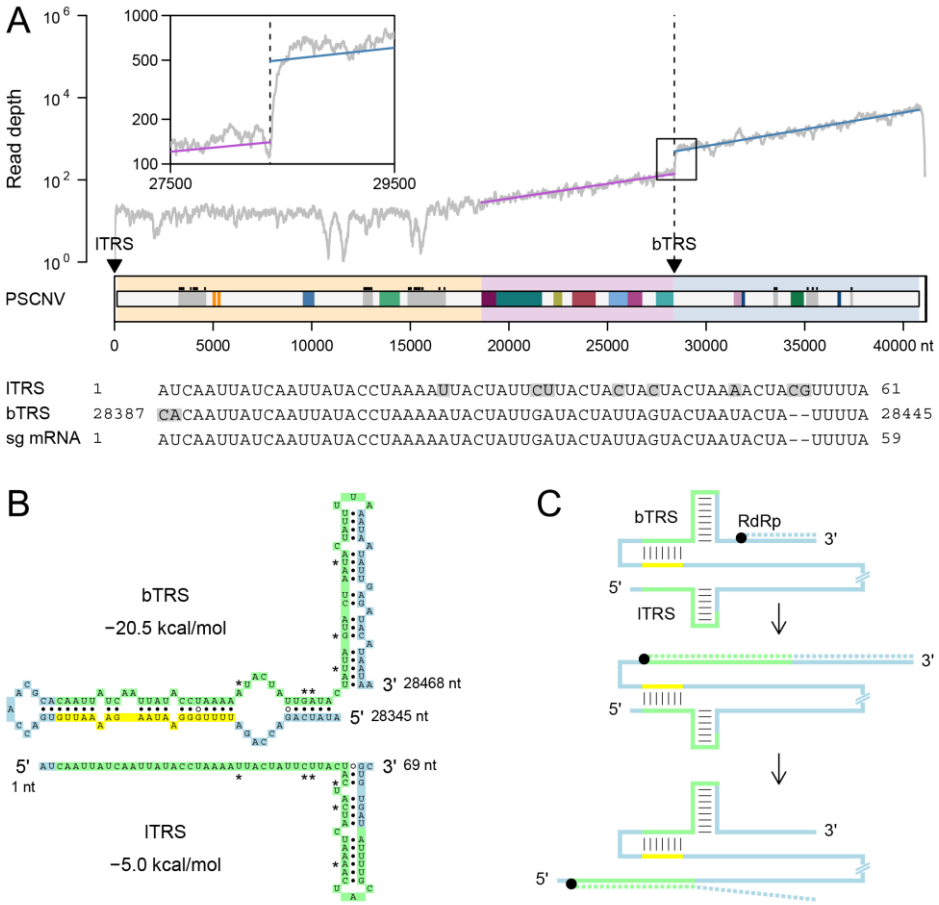


Figure 10 | Genome transcription. (**A**) Mean depth of RNA-seq coverage along the PSCNV genome (approximated by exponential regression in ORF1b-like and 3'ORFs-like regions) calculated based on five datasets used to assemble the transcriptomes in which PSCNV was found [67]. Indicated on the genome map (coloured as in Fig. 2) are the positions of oligonucleotide repeats (leader and body TRSs) in the genome, and below is their alignment with a sg mRNA 5' terminus identified by 5'-RACE (nucleotide mismatches between sg mRNA and TRSs are shown with grey backgrounds). (**B**) Predicted secondary structure of TRSs. TRSs are highlighted in green, region upstream of bTRS, interacting with its 5'-terminus in yellow, asterisks indicate mismatching nucleotides of TRSs. (**C**) Model of discontinuous RNA synthesis mediated by TRSs and their secondary structure. Genome is shown by solid line and nascent minus strand by dashed line. Color code matches that of panel B.

Discontinuous genome synthesis (transcription)

To search for TRSs in the PSCNV genome, its 5'-UTR was compared with the whole genome sequence using nucleotide BLAST. A pair of highly similar sequences (86% identity, E-value 2E-14) was identified in the 5'-UTR (3–61 nt) and immediately upstream of the 3'ORFs-like region (28389–28445 nt) (Figure 10A). If these repeats are indeed utilized as TRSs in discontinuous RNA synthesis, a template for a 12717 nt sg mRNA

(excluding the polyA tail) would be produced. Indeed, we observed a ~3x rise in transcriptomic read coverage beginning at the bTRS genome position, and confirmed the presence of the expected template-switching junction in a sg RNA by 5'-RACE conducted on infected planarians (Figure 10A). That sg mRNA contains a 12327-nt ORF identical to the 3'-terminus of the main PSCNV ORF (28473–40799 nt in genome coordinates), if its translation starts from the 5'-most Met codon of the sg mRNA.

To explore a mechanistic basis for RNA strand translocation during the postulated discontinuous transcription, we predicted RNA secondary structure for the PSCNV genome in the vicinity of the TRS signals (Figure 10B). According to the prediction, 3'-terminal nucleotides of both TRSs, starting from the 36th TRS nucleotide, form hairpins involving nucleotides of the downstream region. In contrast, 5'-terminal parts of the TRSs may be folded differently: the first 35 nucleotides of the lTRS remain unstructured, while the first 35 nucleotides of the bTRS form a hairpin involving the upstream sequence. Two parts, tip and basal, could be recognized in this hairpin. The tip part includes 22 nucleotides of bTRS that seems to form 17 canonical base pairs with a genome region just 11 nucleotides upstream (yellow in Figure 10B). Since these 22 nucleotides of bTRS are identical to those of the lTRS, the latter might alternatively form a stable secondary structure with the yellow region (upstream of bTRS; Figure 10C). The basal part of the hairpin is much smaller and may not be conserved in the possible interaction involving lTRS.

Identification of partial genome sequences of putative planarian viruses related to PSCNV

Finally, we used the PSCNV polyprotein as a query sequence to survey several flatworm species' transcriptomes in the PlanMine database [119] for the presence of other nidoviruses related to PSCNV. We identified six contig sequences with highly significant similarity to PSCNV, indicative of at least two nidoviruses (Figure S18). These contigs originate from transcriptomes of *S. mediterranea* (uc_Smed_v2 and ox_Smed_v2 assemblies, two and one contigs, respectively; the latter contig was excluded from consideration due to being almost identical to one of the former contigs) and another planarian species, *Planaria torva* (dd_Ptor_v3 assembly, three contigs). Translations of the two uc_Smed_v2 contigs of 814 nt and 1839 nt gave hits of >99% aa identity to the very C-terminus of PSCNV polyprotein, indicative of a variant of PSCNV circulating in the same host species (see section above). In contrast, the dd_Ptor_v3 transcriptome included two short contigs (283 nt and 289 nt) with hits to the PSCNV RdRp domain (38 and 48% aa identity) as well as an 8811-nt contig, whose translation in the +1 frame gave 3 discontinuous hits, one to the O-MT domain of the ORF1b-like region (37% aa identity) and two to the 3'ORFs-like region and its FN2b domain (25% and 37% aa identity). These domains are separated by different distances in PSCNV and the 8811-nt contig. It is

notable that all three hits from the *P. torva* contig correspond to its translation in the same frame, uninterrupted by stop-codons, suggesting that ORF1b-like and 3'ORFs-like regions of this putative and divergent virus could also be expressed from a single ORF.

DISCUSSION

The advent of metagenomics and transcriptomics has greatly accelerated the pace of virus discovery, leading to studies reporting genome sequences of dozens to thousands of new RNA viruses in poorly characterized hosts [35, 36, 79, 120-126]. These developments have substantially advanced our appreciation of RNA virus diversity, and improved our understanding of the mechanisms of its generation [127, 128]. Notwithstanding that sea change, the largest known RNA genomes continue to belong to nidoviruses, as has been the case for 30 years, since the first coronavirus genome of 27 kb was sequenced [14, 21, 78] (Figure 1A).

This study's transcriptomics-based discovery of PSCNV in planarians reinforces the status of nidoviruses as relative giants among RNA viruses, and also demonstrates that RNA genomes may be substantially larger than previously understood. The discovery of a virus with this large 41.1-kb RNA genome was unexpected in the context of accumulating genomic data on viruses and emerging concepts in the field. Below, we discuss the implications of PSCNV's distinctive features, and future directions of research.

PSCNV is distantly related to previously described nidoviruses

The PSCNV polyprotein includes distant homologs of all ten domains common to invertebrate nidoviruses, as well as the vertebrate *Coronavirinae* subfamily [14, 45]. These were identified with high statistical confidence, using an iterative bioinformatics procedure with profile searches at its core. These domains include the definitive nidovirus markers NiRAN and ZBD, and all ten are syntenic between PSCNV and other nidoviruses. Most are located in ORF1b-like (replicase) region, which also includes four subregions left unannotated (Figure 2). Of these unannotated subregions, one flanked by ZBD and HEL1 may correspond to the regulatory domain 1B, which is uniformly present but poorly conserved in helicases of nidoviruses [48, 49], while the other three may represent domains uniquely acquired by a PSCNV ancestor. Like all characterized invertebrate nidoviruses and unlike most vertebrate nidoviruses [14, 129], PSCNV does not encode a homolog of an uridylate-specific endonuclease (NendoU) [31]. Accordingly, our rooted RdRp-based phylogenetic analysis assigned PSCNV to a monophyletic clade of invertebrate nidoviruses. Another topologically similar tree was inferred using five nidovirus-wide conserved domains using a dataset that did not include an outgroup. The observed tree

topology is also broadly compatible with other observations of this study (see below), and with RdRp-based trees of known nidoviruses produced in other studies [14, 21, 35]. Given that PSCNV infects planarian hosts, consistent placement of this virus in the invertebrate nidovirus clade by different analyses makes biological sense. On the other hand, the precise position of PSCNV in the invertebrate nidovirus clade remains poorly resolved for several reasons, including the highly skewed host representation in the analyzed small sample of 57 nidoviruses, and the large divergence of invertebrate nidoviruses from each other.

The dominant trees topology placed PSCNV in a very long and deeply rooted branch, which have been recognized as a suborder in the pending taxonomic proposal [130]. This is further supported by the presence of the GDD tripeptide in the RdRp C motif (Figure S9), most common in ssRNA+ viruses other than nidoviruses, which typically (except for the arterivirus Wobbly possum disease virus, WPDV, [81]) have an SDD signature instead [131]. The pronounced divergence of PSCNV is also evident in other conserved protein domains, 3CLpro, NiRAN and ExoN, each of which carries substitutions not observed in other invertebrates or all nidoviruses.

Two prominent replacements in PSCNV 3CLpro are functionally meaningful (Figure S7). The replacement of the otherwise invariant His by Val in the putative substrate pocket is indicative of a modified P1 substrate specificity for this enzyme, which exhibits a strong preference for Glu or Gln residues in P1 position in most other ssRNA+ viruses, including vertebrate nidoviruses [42, 88-91]. Accordingly, we were unable to identify typical 3CLpro cleavage sites at the expected inter-domain borders in the portion of the PSCNV polyprotein that must be processed by 3CLpro. Furthermore, the nucleophilic catalytic residue of PSCNV's 3CLpro is Ser, while its counterpart in other characterized invertebrate nidoviruses is Cys. Similar variation of this residue has been described among vertebrate arteri- and toroviruses versus coronaviruses [42, 88-91], with distinct variants being associated with deeply separated virus lineages at the rank of (sub)family. Diversification of the nucleophile residue was also observed in other ssRNA+ viruses that employ 3C(L) proteases [132, 133]. This recurrent Ser-Cys toggling of the catalytic nucleophile in other well-established viral families argues against independent origins of 3CLpros in PSCNV and other nidoviruses, despite their weak sequence similarity.

Besides its exceptionally large genome size, the single-ORF organization of the PSCNV genome is unprecedented for nidoviruses. This single-ORF organization was unexpected, given that multi-ORF organization is conserved across the vast diversity of nidoviruses separated by large evolutionary distances, and infecting vertebrate or invertebrate hosts. In contrast, other large monophyletic groups of ssRNA+ viruses with comparable host

ranges (e.g., the order *Picornavirales* or Flavi-like viruses), include many viruses with either single- or multi-ORF organizations that intertwine phylogenetically [79, 132, 133].

The PSCNV single-ORF genome may be expressed in a manner similar to that of multi-ORF nidoviruses

The use of 3CLpro as the main protease responsible for the release of key RTC subunits from polyproteins would be anticipated to remain essential in the single-ORF PSCNV. In contrast, two other conserved mechanisms of genome expression, ORF1a/1b -1 PRF and discontinuous transcription, might not be expected to operate in this virus, since they are associated with the use of multiple ORFs in nidoviruses. We reasoned otherwise, however, on the grounds that these mechanisms allow differential expression of three functionally different regions of the nidovirus genome, which are also conserved in PSCNV. We located a potential -1 PRF signal in the PSCNV genome. This signal is located at the canonical position observed in other nidoviruses, and could potentially attenuate in-frame translation downstream of the ORF1a-like region in a manner different from a mechanism used by other characterized nidoviruses, but with similar end-products (Figure 9). Such a postulated mechanism is used by encephalomyocarditis virus to attenuate the expression of replicase components in favour of capsid proteins from its main long ORF [134].

Likewise, we obtained several lines of evidence for upregulated transcription of the 3'ORFs-like region as a subgenomic RNA (Figure 10). The products of this region may also be derived from the polyprotein, but are likely required in greater abundance toward the end of the viral replication cycle, and separate expression from sg mRNA would more efficiently address this need. Importantly, no evidence, either bioinformatic or experimental, was obtained for other sg mRNAs, although we cannot exclude their existence. PSCNV's putative TRSs are exceptionally long for nidoviruses (59 and 57 nt versus typically a dozen nt), perhaps because smaller repeats might emerge in its extraordinarily long genome by chance, interfering with the transcription accuracy. Other unknown factors may also contribute to this large TRS repeat size.

The putative leader TRS (lTRS) and body TRS (bTRS), along with their predicted RNA secondary structures, suggest a model for transcriptional regulation of the PSCNV genome. We postulate that during anti-genomic RNA synthesis, the virus RTC unwinds two bTRS hairpins (Figure 10C, top). As a result, the region immediately upstream of the bTRS (yellow in the figure) becomes available for base-pairing with the 5'-terminus of the lTRS (Figure 10C, middle). This interaction will bring the two distant regions of the genome in close proximity, facilitating translocation of the nascent minus-strand from body to leader TRS (Figure 10C, bottom). The latter step is considered routine in the current model of sg RNA synthesis in well-characterized arteriviruses and coronaviruses [51, 135]. However, its

mechanistic details are poorly understood and may operate differently among nidovirus families.

Although we cannot exclude the possibility that smaller ORFs may be expressed by PSCNV, it seems unlikely that they would contribute substantially to the virus proteome, in line with the apparent inverse relationship between genome size and gene overlap [136]. Rather, such ORFs could be used for regulatory purposes, as in the case of the very small ORF at the border of ORF1a- and ORF1b-like regions, through the PRF mechanism proposed above.

The combined genomic and proteomic characteristics of PSCNV defy central role of multiple ORFs in the life cycle and evolution of nidoviruses, despite their universal presence in all other nidoviruses [26, 60]. Contrary to conventional wisdom, single-ORF genome expression can involve the synthesis of subgenomic mRNAs. Rather than multi-ORF genome organization, functional constraints linked to the synteny of key replicative enzymes may be the hallmark characteristic of nidoviruses [137].

PSCNV has acquired novel proteins with potential functions in host-virus interactions

Most of the domains that we annotated in the PSCNV giant polyprotein are homologs of canonical nidovirus domains. However, we also mapped several unique domains. Below we discuss possible functions of five small domains, all of which plausibly modulate different aspects of virus-host interaction.

PSCNV encodes a ribonuclease T2 homolog upstream of the putative 3CLpro in the ORF1a-like region (Figure 2). Ribonucleases of the T2 family (RNase T2) are ubiquitous cellular enzymes that non-specifically cleave ssRNA in acidic environments [138]. DNA polydnaviruses and RNA pestiviruses are the only two other virus groups that are known to encode related enzymes [139, 140]. In pestiviruses, the RNase T2 homolog is a domain of secreted glycoprotein E^{rns} found in virions, but dispensable for virus entry [141]. The E^{rns} structure is supported by four disulfide bridges that are formed by eight conserved Cys residues [139]. None of these residues were found in the PSCNV RNase T2 homolog, consistent with its location in the polyprotein region that produces cytoplasmic proteins in other nidoviruses. In polydnaviruses and pestiviruses, the RNase T2 homolog modulates cell toxicity and immunity [139, 140], and a similar role could be considered for the PSCNV RNase T2 homolog. The origin of this domain in PSCNV remains uncertain due to the lack of close homologs in either its host, *S. mediterranea*, or other cellular and viral species.

Two other unique domains of PSCNV are fibronectin type II (FN2) homologs, protein modules of approximately 40 aa with two conserved disulfide bonds, which are ubiquitous

in extracellular proteins of both vertebrates and invertebrates [142, 143]. Because of the low similarity of FN2a and FN2b to each other and other homologs, it is not clear whether they emerged by duplication or were acquired independently. No other known virus encodes an FN2 homolog (although the putative nidovirus identified in *P. torva* may include ortholog of FN2b, Figure S18), suggesting that PSCNV's FN2 domains function in a unique aspect of its replication cycle. FN2 domains are known to possess collagen-binding activity, and are found in a variety of proteins that bind to and remodel the extracellular matrix [144, 145]. Thus, it is conceivable that these domains might play a role in the shedding or transmission of PSCNV virions. This hypothesis is compatible with the accumulation of PSCNV RNA and particles, presumably virions, in the planarian mucus-secreting cells. Besides FN2 domains, this process might also involve the Thr/Ser-rich region adjacent to FN2a in polyprotein, since Thr-rich and Thr/Ser-rich regions have been implicated in mediating adherence of fungal and bacterial extracellular (glyco) proteins to various substrates [146, 147].

The identification of the ankyrin repeats domain (ANK) in PSCNV is unprecedented and intriguing. In proteins of other origins, the ANK domain is a tandem array of ankyrin repeat motifs (~33 residues each) of variable number and divergence that fold together to form a protein-binding interface [148]. Ankyrin-containing proteins are involved in a wide range of functions in all three domains of cellular life. In viruses described to date, they have been identified exclusively in large DNA viruses with genome sizes ranging from ~100 kb to 2474 kb, the latter of *Pandoravirus salinus*, the largest viral genome described so far [38, 148-150]. Acquisition of this domain, likely from a planarian host, might have provided a PSCNV ancestor with a mechanism to evade host innate immunity. Notably, according to SmedGB [102] annotation, host proteins SMU15016868 and SMU15005918, whose C-terminal domains are the closest homologs of PSCNV ANK (Figure 6), contain a Rel homology domain (RHD) at their N-termini. This N-RHD-ANK-C domain architecture is typical of the NF-κB protein, a precursor of a cellular transcription factor that triggers inflammatory immune responses upon virus infection or other cell stimulation [151]. NF-κB is activated for translocation to the nucleus by degradation of its inhibitor, C-terminal ANK domain of NF-κB protein or its closely related paralog, IκB protein [148, 152, 153]. Several large DNA viruses have been shown to encode IκB-mimicking proteins that prevent NF-κB from entering the nucleus in response to the infection, and thus downregulate the host immune response [154, 155]. PSCNV ANK may represent the first example of an IκB-mimicking protein in RNA viruses, although RNA viruses including nidoviruses can target NF-κB protein using other mechanisms [156]. This striking parallel between PSCNV and large DNA viruses blurs the distinction between these viruses regarding to how they adapt to hosts [157]. It further highlights the exceptional coding capacity of PSCNV genome among RNA viruses.

Emergence and evolution of the PSCNV genome: implications for the viability of large RNA genomes

The single-ORF organization of PSCNV's exceptionally large genome is intriguing, but we cannot determine whether this association between genome size and organization is causal or coincidental from observation of a single species. In this respect, determining whether the putative nidovirus we identified in *P. torva* also employs a single-ORF organization could be illuminating. An evolutionary switch between multi- and single-ORF organizations, regardless of its direction, must be a multi-step process, since it affects many translation regulatory signals. In our study, we used a simple model of this process with two character states within Bayesian phylogenetic framework to obtain support for the single-ORF organization of PSCNV emerging from the multi-ORF organization. This approach is apparently not sensitive to choice of domains used for phylogeny reconstruction or inclusion of an outgroup. However, given the deep position of the PSCNV lineage in the nidovirus tree, the ambiguous rooting of PSCNV relative to other invertebrate nidovirus families, and PSCNV being the only single-ORF nidovirus known, further analysis of this transition using improved sampling of nidoviruses and their sister clades [35, 36], and more sophisticated models is warranted.

In the few experimentally characterized coronaviruses with genomes of 27–31 kb, the mutation rate is low by RNA virus standards, due to ExoN proofreading activity [34, 158, 159]. This observation is in line with the inverse relationship between genome size and mutation rate in viruses and prokaryotes [160, 161]. Accordingly, we may expect mutation rates to differ in ExoN-containing nidoviruses with different genome sizes, with PSCNV having a particularly low mutation rate. While characterization of mutation rates of PSCNV and other nidoviruses must await future studies, we already note a distinctive similarity between cellular proofreading exonucleases and ExoN of PSCNV that separates it from its orthologs in other ExoN-positive nidoviruses. Specifically, there is a correlation between the presence of the Zn-finger motif in the exonuclease active site [33, 92] and genome size of biological entity encoding exonuclease: non-PSCNV nidoviruses with genome sizes in the range of 20-34 kb include a Zn-finger embedding catalytic His, while PSCNV and DNA-based entities with genome sizes >41 kb do not (Figure S12) [162]. Based on these observations, it is plausible that this Zn-finger might limit ExoN's capacity to improve replication fidelity while providing other benefits, and its loss in the PSCNV lineage could have been a factor promoting genome expansion.

Besides the lack of the Zn-finger in ExoN, the reported size increase of the ORF1b-like region in PSCNV relative to other nidoviruses (about 10-fold greater than expected under an assumption of uniform expansion in all genome subregions) is particularly notable in the context of the theoretical framework presented in the introduction. Briefly, expansion

of RNA genomes requires escape from the so-called Eigen trap (or Eigen paradox): such genomes are confined to a low-size state, in which low replication fidelity prevents the evolution of larger genomes, which in turn prevents the evolution of greater complexity, which could introduce tools to increase replication fidelity [15]. The three-wave model of genome expansion in nidoviruses notes that the ORF1b region, which encodes the core replicative machinery, appears to play a central role in such constraints. It proposes that a common nidovirus-wide wave of expansion in the ORF1b region precedes and permits subsequent lineage-specific waves in the ORF1a and 3'ORFs subregions. In the order *Nidovirales*, a wave of expansion in ORF1b involved the acquisition of the ExoN proofreading exonuclease, which permitted further expansion of other subregions due to a reduced mutation rate. Until now, however, the genomes of large nidoviruses (the 20-to-34 kb size range) appeared to have reached a plateau at the low-30 kb range, associated with very little variability in the size of ORF1b among members of this group (6,9-to-8,2 kb). The three-wave model predicts that further genome expansion far beyond 34 kb would require a second cycle of waves, beginning again with ORF1b [66]. The disproportionate increase in PSCNV's ORF1b-like region is consistent with this prediction. The acquisition of additional, still-uncharacterized domains in this region of the PSCNV genome, as well as the distinctive features of its ExoN domain, may help to explain this "second escape" from the Eigen trap. Further characterization of novel ORF1b domains is required, to assess their contribution to replication fidelity.

Our discovery of PSCNV, and analysis of its genome, show that nidoviruses can overcome the ORF1b-size barrier and adopt divergent ORF organizations. If the multi-cycle three-wave model of genome expansion in RNA viruses holds, one would expect that a large expansion of ORF1b, as evident in PSCNV, would permit yet greater expansion of the ORF1a and 3'ORFs regions in other viruses of the PSCNV lineage. Thus, nidoviruses of yet-to-be-sampled hosts might prove to have evolved even larger RNA genomes than that reported here, further decreasing the gap between virus RNA and host DNA genome sizes.

MATERIALS AND METHODS

All Materials and Methods are described in S1 Materials and Methods in detail.

PSCNV genome and its variants in *S. mediterranea* RNA-seq data

The genome sequence of human coronavirus OC43 (GenBank KY014282.1) was used to query two in-house *de novo*-assembled *Schmidtea mediterranea* transcriptomes (transcripts assembled from multiple asexual and sexual planarian stocks, designated with txv3.1 and txv3.2 prefixes, respectively) [67] using tblastx (BLAST+ v2.2.29 [163]). With E-

value cut-off 10, 25 *S. mediterranea* transcripts were identified and used in reciprocal BLAST searches against the NCBI NR database. Two nested transcripts, txv3.2-contig_1447 (assembled from sexual planarians, GenBank BK010449) and txv3.1-contig_12746 (assembled from asexual planarians, GenBank BK010448), showed statistically significant similarity to other nidoviruses, which exceeded its similarity to other entries. Sequences of these two transcripts overlap by 23,529 nt with only 7 nt mismatches (0.03%). The larger transcript, txv3.1-contig_12746, was used to search in planarian EST clones [69, 164], which found the following overlapping clones showing >99% nucleotide identity: PL06016B2F06, PL06005B2C04. PL06007A2B12, PL06008B2B03 PL08002B1C07, and PL08001B2B04 (GenBank DN313906.1, DN309834.1, DN310382.1, DN310925.1, HO005314.1, and HO005110.1, respectively). Transcripts txv3.1-contig_12746 and txv3.2-contig_1447, and the six EST clones were assembled into an incomplete putative genome. Conflicts between overlapping sequences were always resolved in favor of the txv3.1-contig_12746 sequence. Fifteen 3'-terminal nt of the reverse complement of txv3.1-contig_12746 ("TATTATGTGATACAC") and two 3'-terminal nt of HO005314.1 and HO005110.1 ("TG") were discarded due to their likely technical origin. The assembled sequence contains a stop codon followed by a short untranslated region and a polyadenylated (polyA) tail. The planarian transcriptomes were surveyed again for transcripts with >50 nt overlap at the 5'-end of the incomplete genome by consecutive rounds of nucleotide BLAST. This identified txv3.1-contig_349344 (from asexual planarians; 11,647 nt; 100-nt overlap with txv3.1-contig_12746 with no mismatches; GenBank BK010447) upstream of the original transcripts, and no further extension was achieved with more BLAST iterations. The 5'-end of the genome was then extended using 5'-RACE followed by Sanger sequencing (primers in Table S2).

Reads from planarian RNA-seq datasets (used to assemble the two transcriptomes described above, and those available from EBI ENA [165]) were mapped to the PSCNV genome sequence by either CLC Genomics Workbench 7, or Bowtie2 version 2.1.0 [166]. Read counts and coverage were estimated using SAMtools 0.1.19 [167], and genome sequence variants were called by BCFtools 1.4 [168].

Reverse transcription, PCR, and 5'-RACE

Freshly prepared RNA from mature sexual planarians was used for cDNA synthesis (iScript, Bio-Rad) or 5'-RACE (RLM-RACE, Ambion) according to manufacturer instructions. Large overlapping amplicons across the PSCNV genome (primers in Table S2) were amplified by standard Phusion® High-Fidelity DNA polymerase reactions, with 65ºC primer annealing temperature and 10 min extension steps.

In situ hybridization

Colorimetric and fluorescent in situ hybridizations were done following published methods [169]. Digoxigenin (DIG)-labelled PSCNV probes were generated by antisense transcription of the planarian EST clone PL06016B2F06 (GenBank DN313906.1) [69]. Following color development, all samples were cleared in 80% (v/v) glycerol and imaged on a Leica M205A microscope (colorimetric) or a Carl Zeiss LSM710 confocal microscope (fluorescent).

Histology and Transmission Electron Microscopy

Sexual and asexual planarians originating from the Newmark laboratory were fixed and processed for epoxy (Epon-Araldite) embedding as previously described [170]. For light-microscopic histology, 0.5 µm sections were stained with 1% (w/v) toluidine blue O in 1% (w/v) borax for 30 s at 100°C, and imaged on a Zeiss Axio Observer. For transmission electron microscopy, 50–70 nm sections were collected on copper grids, stained with lead citrate [171] and imaged with a AMT 1600 M CCD camera on a Hitachi H-7000 STEM at 75 kV. Putative virions were seen by TEM in sections from a single worm, which led us to re-examine a collection of 1697 electron micrographs, drawn from 16 additional worms (12 sexuals, four asexuals) from cultures known to harbor PSCNV. All images that included some portion of a mucus cell were chosen for further examination (n=165); the total number of cells represented cannot be determined without three-dimensional reconstruction from serial sections, which is not practical for such large and irregularly shaped cells. No additional examples of putative viral structures were found among the specimens included in these samples.

Genome and Protein databases

For various analyses we used the following databases: PlanMine [119], Smed Unigene [102], scop70_1.75, pdb70_06Sep14 and pfamA_28.0 supplemented with profiles of conserved nidovirus domains [172-174], Uniprot [175], genome sequences representing the current 57 nidovirus species that were delineated by DEmARC [176] and recognized by ICTV on year 2016 [177], NCBI Viral Genomes Resource [178], GenBank [179] and RefSeq [180].

Computational RNA sequence analysis

To predict RNA secondary structure and PRF sites we used Mfold web server [181] and KnotInFrame [182], respectively. Blastn (BLAST+ v2.2.29) [163] was used to identify RNA repeats.

Computational protein analyses

Virus protein sequences were analyzed to predict disordered regions (DisEMBL 1.5 [183]), transmembrane regions (TMHMM v.2.0), secondary structure (Jpred4 [184]), signal

peptides (SignalP 4.1 [185]), N-glycosylation sites (NetNGlyc 1.0) and furin cleavage sites (ProP 1.0 [186]). Multiple sequence alignments of RNA virus proteins were generated by the Viralis platform [187]. Protein homology profile-based analyses were assisted with HMMER 3.1 [188], and HH-suite 2.0.16 [189]. To identify sites enriched with amino acid residue, distribution of each residue along polyprotein sequence was assessed using permutation test executed with a custom R script.

To establish homology for ZBD, ExoN, and N-MT, which top HHsearch hits were under the 95% Probability threshold, we considered several criteria about the source hits: 1) being among the top three for the respective query of a database; 2) being similar to several homologous profiles in two or three databases; 3) residing in the polyprotein position conserved in nidoviruses for the respective domain (Figure S3, Table S5); and 4) including most residues that are critical for function of the respective domain (see below). For ZBD, we also observed a statistically significant enrichment in cysteine (Cys) residues (Figure S4), in line with the coordination of three Zn^{2+} ions by characterized ZBDs, which involves predominantly Cys and His residues [48, 49].

Genome region size comparison between PSCNV and nidoviruses

Size differences between genome regions of PSCNV and nidoviruses (Table S1) were estimated using three measures, D_1, D_2, and D_3, that accounted for: 1) the region size, D_1(region)=(p-M)/M*100%; 2) the region size variation, D_2(region)=(p-M)/(M-m)*100%; and 3) the region size variation and genome size increase, D_3(region)=D_2(region)/D_2(genome)*100%, where m and M are median and maximum sizes of the region in ExoN-containing nidoviruses, respectively, and p is region size in PSCNV.

Evolutionary analyses

Phylogeny was reconstructed by Bayesian approach using a set of tools including BEAST 1.8.2 package [190] and ProtTest 3.4 [191] as described in [81]. BayesTraits V2 [117] was used to perform ancestral state reconstruction. Preference for a state at a node was considered statistically significant only if Log BF exceeded 2 [192].

Visualization of results

Protein alignments were visualized with the help of ESPript 2.1 [193]. To visualize Bayesian samples of trees, DensiTree.v2.2.1 was used [194]. R was used for visualization [195].

DATA AVAILABILITY STATEMENT

Contigs and 5'-RACE sequences used to assemble the PSCNV genome and subgenome were deposited to GenBank (accession nos. BK010447–BK010449, MH933723–MH933734). The complete PSCNV genome sequence is available on GenBank (accession no. MH933735).

ACKNOWLEDGMENTS

We thank Paul Ahlquist for helping initiate this collaboration and Johan den Boon, Andrey M. Leontovich, Dmitry V. Samborskiy, and Igor A. Sidorov for discussions and assistance.

FUNDING

This work was supported by NIH R01 HD043403 to PAN and by EU Horizon2020 EVAg 653316 project and LUMC MoBiLe program to AEG; PAN is an investigator of the Howard Hughes Medical Institute and AEG was Leiden University Fund Professor at the time of this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

AUTHOR CONTRIBUTION

Conceptualization: AS, AAG, PAN, AEG. Data curation: AS, JLB, AEG. Formal analysis: AS, AAG, AEG. Funding acquisition: PAN, AEG. Investigation: AS, AAG, JLB, AEG. Methodology: AS, AAG, JLB, AEG. Project administration: PAN, AEG. Resources: PAN, AEG. Software: AAG. Supervision: PAN, AEG. Validation: AS, AAG. Visualization: AS, AAG, JLB. Writing – original draft: AS, PAN, AEG. Writing – review & editing: AS, AAG, JLB, PAN, AEG.

SUPPORTING INFORMATION

S1 Materials and Methods

Search for nido-like viruses in transcriptomes of S. mediterranea

Two *de novo* transcriptomes of planarian *S. mediterranea* [67] were searched for sequences similar to human coronavirus OC43 (GenBank KY014282.1) by the tblastx application in BLAST+ v2.2.29 [163] using BLOSUM80 matrix, word size 2, and E-value cut-off 10. The resulting hits were translated in six frames by EMBOSS:6.6.0.0 transeq [196] and used to search for similar domains in the NCBI non-redundant protein database (NR) by deltablast (BLAST+ v2.2.29) [197] with the same parameters, except using an E-value cut-off of 1.

Assessment of PSCNV genome coverage by RNA-seq reads

Reads from five independent in-house *S. mediterranea* RNA-seq datasets, previously used to assemble the transcriptomes in which PSCNV was found [67], were mapped to the PSCNV genome sequence (1–41103 nt) using either CLC Genomics Workbench 7 (alignment criteria: mismatch cost 2, insertion/deletion cost 3, length fraction > 0.9, similarity fraction > 0.9), or Bowtie2 version 2.1.0 with default parameters [166]. PSCNV genome coverage by reads from each dataset was estimated using SAMtools 0.1.19 [167].

Search for viruses related to PSCNV in planarian RNA database

The PlanMine database [119] was downloaded from http://planmine.mpi-cbg.de/planmine/ on 2017.10.06, contigs were translated in six frames by EMBOSS:6.6.0.0 transeq [196], and compared with PSCNV polyprotein by blastp (BLAST+ v2.2.29) [163]. Only hits with E-value < 0.001 were considered with the exception of those that involved PSCNV HEL1 or ANK domains. For these domains, whose homologs are common in many proteomes, an additional condition for consideration was to have one or more extra hits between the particular contig translation and other regions of PSCNV polyprotein.

Identification of PSCNV variants in S. mediterranea RNA-seq data

RNA-seq data from fourteen *S. mediterranea* studies (Table S3) were downloaded from the EBI ENA [165] and aligned to PSCNV genome sequence (1–41103 nt) using Bowtie2 version 2.1.0 with default parameters [166]. Read counts and coverage were estimated using SAMtools 0.1.19 [167]. Genome sequence variants were called by BCFtools 1.4 [168] with the following parameters: maximum per-file depth 100000 (including for INDEL calling), the original variants calling method, *p*-value threshold 0.5, ploidy 1.

Nidoviral species and their genomes and proteomes

One representative genome sequence per nidovirus species [177] (in total 57 sequences) was selected for this study (Table S1). Their proteomes, including protein sizes (Fig. 2), were defined using respective entries in the RefSeq database [180] (where available), the literature, and comparative sequence analysis. Boundaries of genome regions were defined as follows: ORF1a region, from the first nucleotide (nt) of the ORF1a start codon to the last nt of the last in-frame codon translated before ORF1a/1b programmed ribosomal frameshifting (PRF); ORF1b region, from the first nt of the first ORF1b codon translated after ORF1a/1b PRF to the last nt of the ORF1b stop codon; 3'ORFs region, from the first nt following ORF1b stop codon to the last nt of the stop codon of the most 3'-terminal ORF.

The single-ORF genome organization of PSCNV presents a distinctive challenge for defining boundaries of three genome regions evident in the multi-ORF nidoviruses. We defined two boundaries, tentatively equivalent to the ORF1a/ORF1b and ORF1b/3'ORFs, in vicinity of the protein motifs universally conserved in all nidoviruses and PSCNV. As result, three regions were defined as follows: ORF1a-like, from the first nt of the start codon of the main ORF to the 18512 nt, the predicted -1PRF site 240 nt upstream of the codon encoding absolutely conserved lysine (Lys) residue of the NiRAN An motif; ORF1b-like, from the 18513 nt to the 28346 nt, which is 260 nt downstream of the codon encoding catalytic glutamate (Glu) residue of O-MT; 3'ORFs-like, from the 28347 nt to the last nt of the main ORF stop codon.

RNA virus polyproteins

For the purpose of this study (Fig. 5), we compiled a list of RNA virus polyproteins larger than 1000 amino acids (aa), based on the information available from the NCBI Viral Genomes Resource on 2017.04.13 [178] and RefSeq entries [180] specified there.

Virus discovery and genome sequencing timelines

The number of viral genomes that were sequenced each year, starting from 1982, was estimated using NCBI Entrez query [198], as the number of GenBank Nucleotide database (2018.01.02) entries belonging to the "Viral sequences" division and containing the phrase "complete cds" in the title, with publication dates within the year of interest [179]. To plot timelines of discovery of viruses with largest RNA and DNA genomes, those viruses were identified and associated information was retrieved for each year using NCBI Viral Genomes Resource on 2017.04.13 [178] and the relevant literature. We used poliovirus (PV), and nidoviruses avian bronchitis virus (IBV), mouse hepatitis virus (MHV), Beluga whale coronavirus SW1 (BWCoV), and ball python nidovirus (BPNV) to highlight the

longest RNA virus genome at 1981 and from 1987 onward, respectively, in Fig. 1A (see Table S1 for the genome sizes of the above nidoviruses).

Multiple sequence alignments of proteins

Multiple sequence alignments (MSAs) of 3CLpro, NiRAN, RdRp, ZBD, HEL1, ExoN, N-MT and O-MT protein domains were prepared for individual nidovirus families using the Viralis platform [187] and assisted by the HMMER 3.1 [188], Muscle 3.8.31 [199] and ClustalW 2.012 [200] programs in default modes. For each domain, MSAs of different nidovirus families and PSCNV were later combined using ClustalW in the profile mode, with subsequent manual local refinement. MSAs of RNase T2, FN2, and ANK domains and PSCNV tandem repeats were prepared using MAFFT v7.123b [201].

Host proteome

Proteome of *S. mediterranea*, Smed Unigene 2015.02.17 [102], was obtained from http://smedgd.stowers.org/.

Identification of ORFs

PSCNV genome was scanned for ORFs in six reading frames by ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) using the standard genetic code and minimal ORF length of 150 nt.

Protein secondary structure retrieval and prediction

Secondary structure was retrieved from PDB structures using the DSSP database [202] via the MRS system [203] for the following proteins: TGEV 3CLpro, 1LVO [87]; SARS-CoV ExoN and N-MT, 5C8T [92]; SARS-CoV O-MT, 3R24 [98]; POLG_BVDVC, 4DW3 [139]; RNT2_HUMAN, 3T0O [204]; MMP2_HUMAN, 1J7M [205]. In all other cases, secondary structure was predicted for individual sequences using Jpred4 [184] in the MSA mode.

Identification of PSCNV polyprotein sequence regions enriched in particular amino acid residues

To identify polyprotein regions enriched in a given amino acid residue, we calculated the distribution of that residue along the polyprotein and compared it to that of permuted sequences within a statistical framework that was applied to each residue type separately. Specifically, we calculated the cumulative count of a particular residue type within the ever expanding [1, *i*] window, where 1 is the first position and *i* is each position from the 1st to the last 13,556th in the polyprotein. The produced discrete data were approximated by R function “smooth.spline” with default parameters, and the first derivative of the approximation was obtained for each *i* value [195]. The procedure was then applied to 100 random permutations of the polyprotein sequence, and mean μ and standard deviation (SD) σ of the resulting derivative values were used to define significance threshold

$T=\mu+Z(1-0.05/L)*\sigma=\mu+4.5*\sigma$, where Z( ) is a quantile function of the standard normal distribution and L is the polyprotein sequence length. Protein sequence regions with derivative values larger than the threshold (4.5 SD above the mean) were considered enriched in the amino acid residue. To avoid artefacts of the approximation, we excluded data corresponding to the N- and C- terminal 100 amino acids of the polyprotein.

Prediction of disordered protein regions

Intrinsically disordered regions of the PSCNV polyprotein were predicted by DisEMBL 1.5 using Remark465 predictor with default parameters [183].

Prediction of transmembrane regions

Transmembrane (TM) regions of proteins were predicted using TMHMM Server v.2.0 (http://www.cbs.dtu.dk/services/TMHMM/) with default parameters. To conform to the input sequence length limitation (8000 aa), PSCNV polyprotein sequence was split into consecutive 8000 and 6556 aa fragments, with a 1000 aa overlap; predictions belonging to the overlap region were accepted even if supported only for one of the fragments.

Prediction of signal peptides

To predict signal peptides, SignalP 4.1 [185] was used. Prediction was made for all PSCNV polyprotein sequence fragments of length 70 aa with default parameters. A D-score threshold of 0.75 was applied to predictions; when predicted signal peptides overlapped, the one with the highest D-score was selected.

Prediction of N-glycosylation sites

N-glycosylation sites were predicted using NetNGlyc 1.0 Server (http://www.cbs.dtu.dk/services/NetNGlyc/) with default parameters. Only predictions with potential above 0.75, supported by all nine networks were accepted. Predictions where potentially glycosylated asparagine (Asn) is followed by proline (Pro), and predictions overlapping with TM helices were discarded. To conform to the input sequence length limitation (4000 aa), PSCNV polyprotein sequence was split into 4000 aa fragments, with 1000 aa overlaps starting from the N-terminus (the most C-terminal fragment was 1556 aa long; 5 fragments in total); predictions belonging to the overlaps were accepted even if supported only for one of the fragments.

Prediction of furin cleavage sites

Furin cleavage sites were predicted by ProP 1.0 Server [186] in default mode and with the PSCNV polyprotein sequence submitted as overlapping fragments as described for the N-glycosylation sites prediction.

Identification of protein sequence repeats

To search for repeats in PSCNV polyprotein, its sequence was compared to itself using an in-house version of HHalign 2.0.16 with the following parameters: SMIN score threshold 5, E-value threshold 10, local alignment mode, realignment by the MAC algorithm not applied, up to 1000 alternative alignments allowed to be shown [81].

Identification of protein domains conserved in PSCNV and other viruses or hosts

We used HHsearch 2.0.16 [189] to query databases scop70_1.75, pdb70_06Sep14 and modified pfamA_28.0 [172-174] with the PSCNV polyprotein fragments using iterative procedure. The modified pfamA_28.0 included original pfamA_28.0 and Hidden Markov Model (HMM) profiles of the most conserved nidovirus domains 3CLpro, NiRAN, RdRp, ZBD, HEL1, ExoN, N-MT, and O-MT, composed of sequences representing *Coronaviridae*, *Mesoniviridae* and *Roniviridae* species (Table S1). This modification facilitates statistical evaluation of similarity between the PSCNV polyprotein and the nidovirus conserved domains within a framework that is used for the pfamA domains. During the first iteration of the procedure, polyprotein was split into fragments by TM clusters (TM helices separated by less than 300 aa), tandem repeats and Thr-rich region. Overlapping hits characterized by Probability above 95% were clustered, clusters were used to split polyprotein into smaller regions that served as HHsearch queries on subsequent iteration. Procedure was repeated until iteration during which no hits satisfying the 95% Probability threshold were detected. Finally, regions of polyprotein without hits were split into successive fragments of 300 aa length starting from N- and C-termini (shorter regions were discarded), which were again scanned for hits by HHsearch. To evaluate the statistical significance of HHsearch hits, we used two measures, E-value and Probability (estimates probability of the query being homologous to the target). We considered homology to be established for PSCNV regions and a database entry that were connected by hits with Probability >95%, and made additional considerations when evaluating hits with Probability ≤95%, as advised in the HH-suite User Guide [189]. In this subsequent analysis, we considered rank, size, and E-value of hits, and conservation of key functionally important residues in the query.

Search for the closest homologs of PSCNV protein domains not previously described in nidoviruses

PSCNV protein domains that were *not* previously described in nidoviruses (RNase T2, FN2, ANK) were compared with Uniprot (2017.01.16) [175] and Smed Unigene (2015.02.17) [102] databases using blastp (BLAST+ v2.2.29) [163]. Domains were extended by 100 amino acids at N- and C-termini in order to capture homology extending beyond that identified by HHsearch. The FN2a domain was not extended at the N-terminus because of the low-complexity Thr-rich domain located immediately upstream. For searches in Smed

Unigene database, effective length of the search space was made equal to that of the search in Uniprot with the same query, in order to make E-values comparable. Domain composition of Smed Unigene hits was obtained from this database , while that of Uniprot hits – from InterPro database [206].

Identification of individual ankyrin repeats

Full alignments corresponding to Ank and Ank_3 families of Pfam 28.0 [174], each representing individual ankyrin repeat, were combined. The resulting alignment was converted to HMM profile by HHmake 2.0.16. The HMM profile had a consensus "xxxGxTpLHxAxxxxxxxxivxxLlxxGadxnxxd", with positions 6–9 and 20–25 corresponding to two conserved ankyrin repeat motifs: TPLH and V/I-V-x-L/V-L-L [148]. It was compared to the PSCNV Ankyrin domain (11360–11570 aa) using in-house version of HHalign 2.0.16 (parameters as detailed for comparison of PSCNV polyprotein sequence with itself). Hits to the PSCNV polyprotein were regarded as individual ankyrin repeats if the alignment included 6–25 positions of the HMM profile.

Phylogeny reconstruction

Phylogeny was reconstructed based on the MSA of the conserved core of RdRp domain (517 columns, 1958–2356 aa in the EAV pp1ab CAC42775.2 of X53459.3), including one representative of each nidovirus species (Table S1) and PSCNV, as well as an outgroup consisting of viruses of two species prototyping the astrovirus genera (*Avastrovirus 1*, Y15936.2; *Mamastrovirus 1*, L23513.1) [207]. Phylogeny was reconstructed using BEAST 1.8.2 package [190] with the model of amino acid replacement selected by ProtTest 3.4 [191] (Akaike information criterion and Bayesian information criterion employed for model selection; maximum likelihood (ML) tree topology optimization strategy utilizing subtree pruning and regrafting moves). Both strict clock and relaxed clock with uncorrelated log-normal rate distribution were tested, and a better-fitting model was selected based on Bayes factor estimate. Markov Chain Monte Carlo (MCMC) chains were run for 10 million iterations and sampled every 1000 iterations; the first 10% iterations were discarded as burn-in. Mixing and convergence were verified with the help of Tracer 1.5 (http://beast.bio.ed.ac.uk/Tracer). Results were summarized as maximum clade credibility (MCC) tree. R package APE 3.5 was used to calculate percentage of trees in the Bayesian sample, characterized by various phylogenetic positions of PSCNV [208]. The same procedure was used to reconstruct 1.) a phylogeny based on the MSA of five nidovirus-wide conserved domains (3CLpro, NiRAN, RdRp, ZBD, HEL1; 1569 columns, 1065-1227, 1740-1881, 1958-2356, 2373-2427, 2520-2774 aa in the EAV pp1ab CAC42775.2 of X53459.3) including one representative of each nidovirus species (Table S1) and PSCNV; 2.) a phylogeny based on the MSA of PSCNV ANK and its closest cellular homologs (Fig. S16, from first to last column without gaps).

Ancestral state reconstruction

BayesTraits V2, MCMC method was used to test support for one ancestral state over the other at a given node [117]. A sample of phylogenetic trees, reconstructed by BEAST as detailed above, was utilized. State "1", single ORF, was assigned to PSCNV, while state "0", multiple ORFs, was assigned to all other viruses in the phylogeny. We also run a version of the analysis where state "-", that is the lack of information about genome organization, was assigned to astroviruses. To derive prior distributions for the rate parameters of the model, we calculated a ML estimate of the rate parameters on each tree in our sample, and set mean and variance of the gamma priors to conform to those of the obtained distributions. MCMC chains (10 million iterations, first 1% iterations discarded as burn-in) were run with the node of interest fossilized in both states. The Harmonic Mean value was recorded at the final iteration of each chain. Log Bayes Factor (Log BF) was calculated as twice the difference between Harmonic Mean values of the better and the worse fitting models. The procedure was repeated three times and the smallest value of the Log BF was reported. Preference for a state at a node was considered statistically significant only if Log BF exceeded 2 [192].

Identification of putative transcription-regulating sequences (TRSs)

Nidoviruses utilize non-adjacent nucleotide repeats (conserved signals) in the 5'-UTR and the second half of the genome to regulate synthesis of subgenomic (sg) mRNAs (transcription). These repeats are known as leader and body transcription-regulating sequences, lTRS and bTRS, respectively. To search for potential TRSs, the 5'-UTR sequence was compared with the PSCNV genome using blastn (BLAST+ v2.2.29) [163].

RNA secondary structure prediction

RNA secondary structure prediction for PSCNV genome regions encompassing lTRS and bTRS (1–9000 nt and 20441–29440 nt, respectively) was assisted by the Mfold web server [181]. Only the top-ranking predictions with the lowest free energy were considered. Maximal distance between paired bases was set to 150 nt. Free energy for fragments of the prediction was calculated using http://unafold.rna.albany.edu/?q=mfold/Structure-display-and-free-energy-determination.

PRF site prediction

KnotInFrame [182] was applied to a 1000 nt region of PSCNV genome immediately upstream of the region encoding the NiRAN An motif. Only the top prediction was considered.

Visualization of the results

Protein alignments were visualized by ESPript 2.1 [193] using the Risler similarity matrix [209] and similarity global score 0.7. To visualize Bayesian samples of trees, DensiTree.v2.2.1 was used [194]. R was used extensively for visualization [195].

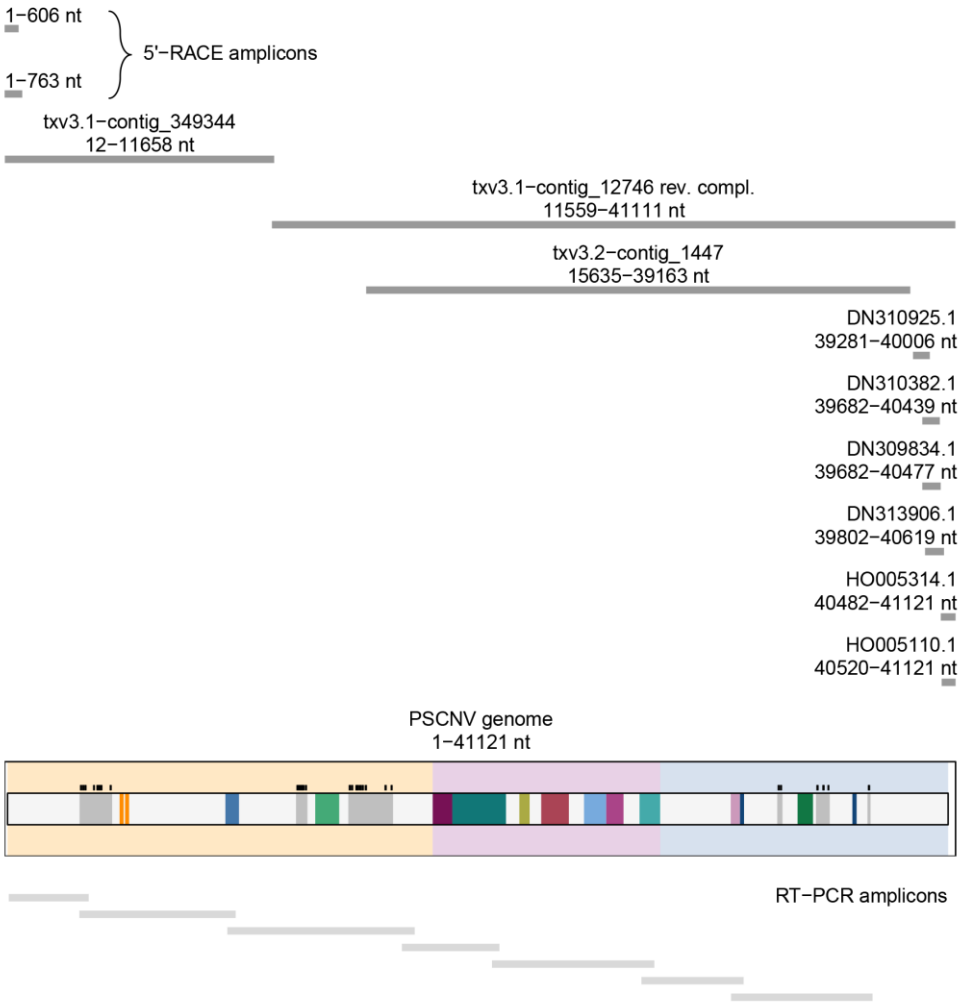


Figure S1 | PSCNV genome assembly and its verification. Contigs and 5'-RACE amplicons, used to assemble the PSCNV genome sequence are shown above the PSCNV genome map (see Fig. 2 for designations) by dark grey lines, with coordinates of the corresponding PSCNV genome regions specified on top of each line. The genome sequence was verified by obtaining products of expected sizes in seven RT-PCR reactions with pairs of primers that were designed to amplify large overlapping PSCNV genome regions (shown by light grey lines below the PSCNV genome map).

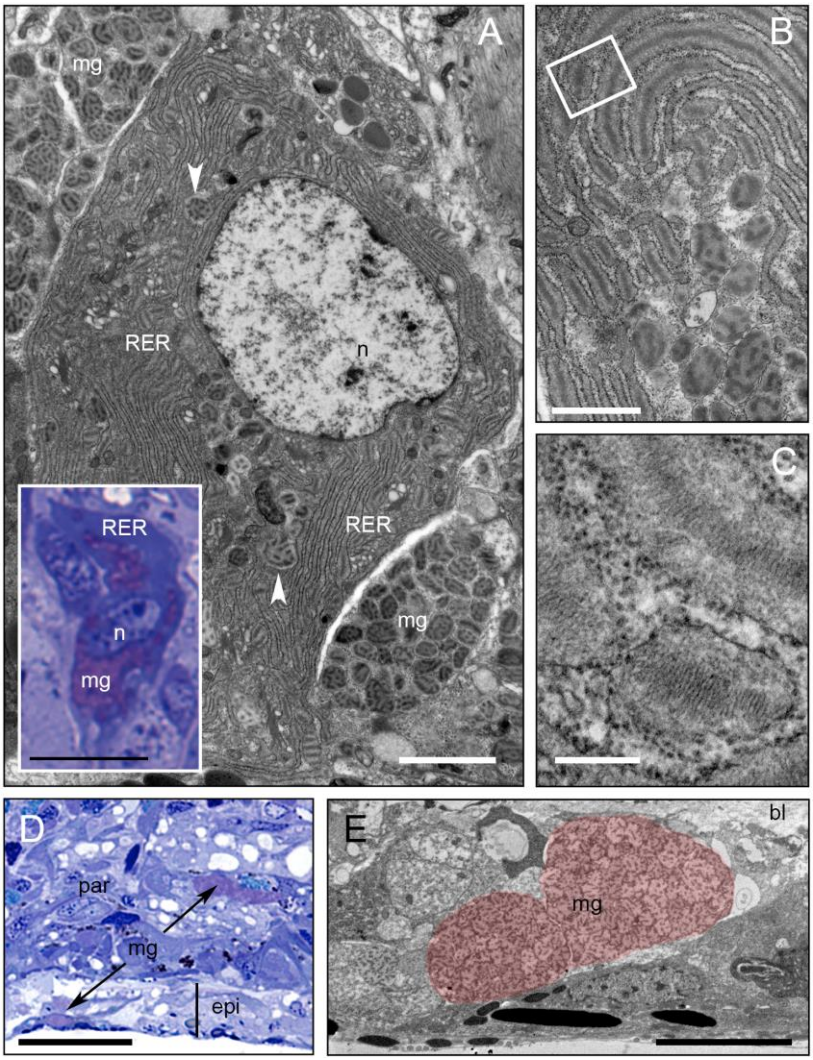


Figure S2 | Characteristics of mucus cells in *S. mediterranea*. (**A**) Transmission electron micrograph of typical mucus cell [210]; n = nucleus. Cell bodies of such cells are filled with rough endoplasmic reticulum (RER). Distinctive mottled structures indicated by arrowheads are mucus granules. Extensions of other cells filled with these granules are also visible (mg). Inset shows a light micrograph of such a cell, stained with toluidine blue O. Mucus-rich regions of cytoplasm stain metachromatically (reddish-purple), while RER is a more-uniform blue. (**B**) Region of RER from mucus-cell cytoplasm (different cell from panel *A*) showing dilated ER lumens, and nascent mucus granules. (**C**) Higher magnification of RER in boxed region from panel *B*. (**D**) Light micrograph of cross section through ventral parenchyma (par) and epidermis (epi) stained with toluidine blue O. Reddish-purple patches indicated by arrows are fields of mucus granules (mg). (**E**) Transmission electron micrograph of ventral epithelium, showing mucus granules (mg, tinted red) just under the external surface. Scale bars: *A*, 2 µm (inset, 10 µm); *B*, 1 µm; *C*, 200 nm; *D*, 20 µm; *E*, 5 µm.

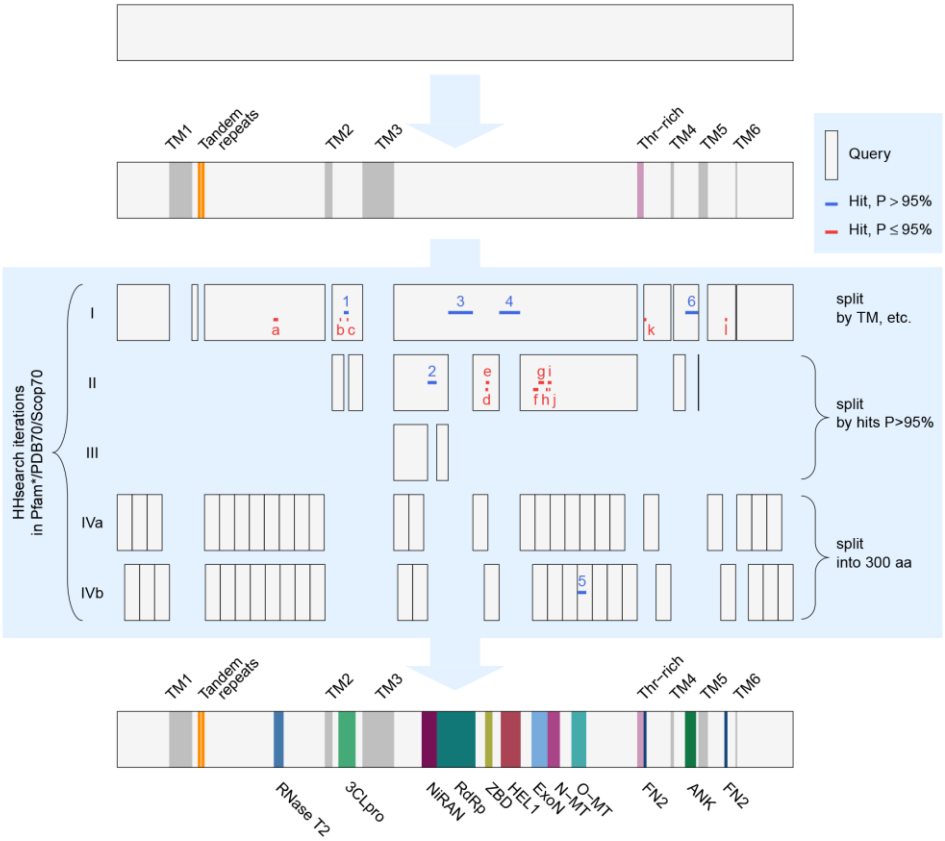


Figure S3 | Outline of iterative HHsearch-based procedure to annotate PSCNV polyprotein. Grey bars on the top and bottom represent PSCNV polyprotein with annotation available prior to the procedure and obtained as a result of the procedure (see Table S5), respectively. Outline of the procedure (see SI Text) is presented on blue background. Iterations of the procedure are designated by Latin numbers on the left. Grey bars represent regions of PSCNV polyprotein that served as HHsearch queries to three profile databases. Basis used to split polyprotein into regions during each iteration is indicated on the right. Locations of clusters of hits with Probability (P) >95% are depicted in dark blue, with numeric indices that reflect their relative position in polyprotein, from the N- to C-terminus. Locations of accepted hits with Probability ≤95% are depicted in red, with letter indices that reflect their relative position in polyprotein, from the N- to C-terminus.

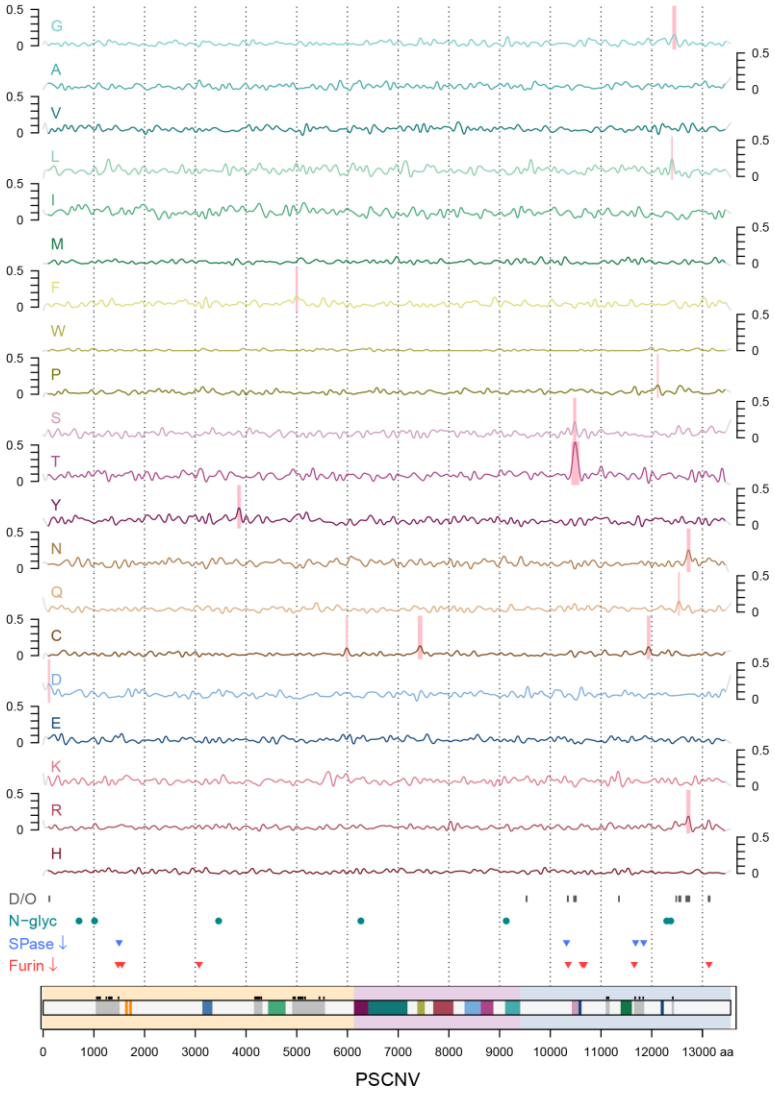


Figure S4 | Density distribution of twenty amino acid residues and predicted functional sites of PSCNV polyprotein. Top: first derivative of cumulative amino acid residue content is plotted for each of the 20 residues with residue-specific colors; values corresponding to the N- and C- terminal 100 residues were excluded from consideration to avoid artefacts and are shown in grey. Sites enriched with a particular residue at statistically significant level are highlighted by pink background. Bottom: polyprotein location of predicted intrinsically disordered regions (D/O), N-glycosylation sites (N-glyc), signal peptidase (SPase ↓) and furin (Furin ↓) cleavage sites are shown by grey boxes, green dots, blue and red triangles, respectively (see Fig. 2 for PSCNV genome map designations). Single-letter abbreviations for the amino acid residues are as follows: G, Gly; A, Ala; V, Val; L, Leu; I, Ile; M, Met; F, Phe; W, Trp; P, Pro; S, Ser; T, Thr; Y, Tyr; N, Asn; Q, Gln; C, Cys; D, Asp; E, Glu; K, Lys; R, Arg; H, His.

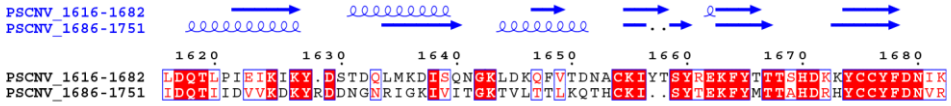


Figure S5 | Alignment of PSCNV tandem repeats. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to the first repeat.

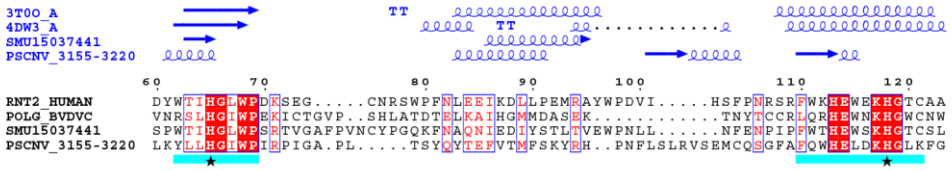


Figure S6 | MSA of RNase T2 domains of diverse origins, including PSCNV. CAS I and CAS II motifs are underlined in cyan, and catalytic histidine residues are denoted with black stars. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering above of the alignment refers to the top sequence.

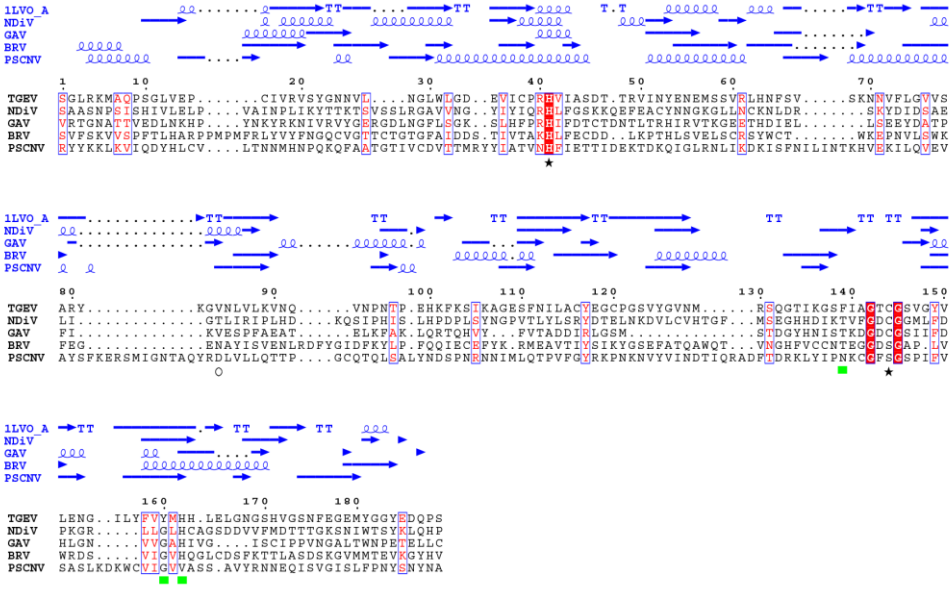


Figure S7 | The aligned proteases employ either catalytic Cys-His dyad or catalytic Ser-His-Asp triad. MSA of 3CLpro domains from four distantly related nidoviruses and PSCNV (4438–4664 aa). Columns containing TGEV 3CLpro catalytic dyad residues are marked by black stars. TGEV 3CLpro Val84 residue that is spatially equivalent to the catalytic acidic residue of serine proteases is marked with empty circle. Residues of the TGEV 3CLpro substrate-binding pocket are underlined with green bars [87]. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to TGEV nsp5.

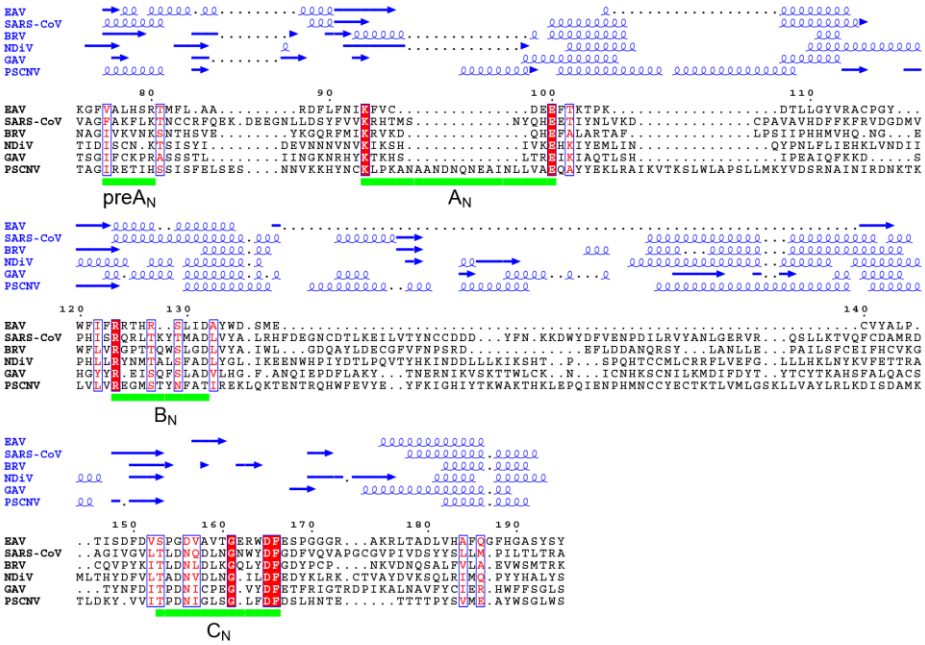


Figure S8 | MSA of NiRAN domains from five distantly related nidoviruses and PSCNV (6181–6410 aa). Conserved motifs are underlined in green. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to EAV nsp9.

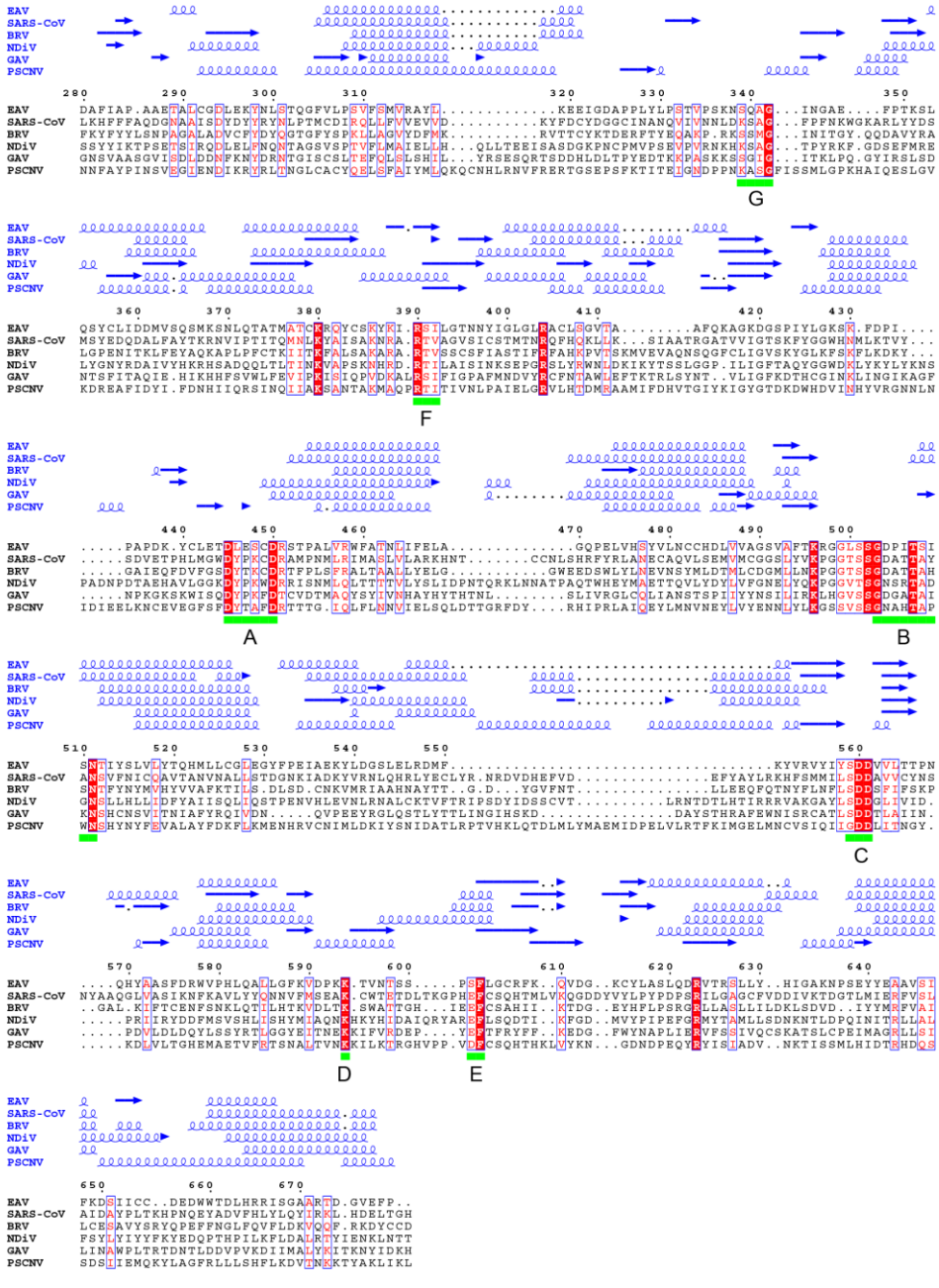


Figure S9 | MSA of RdRp domains from five distantly related nidoviruses and PSCNV (6632–7125 aa). Conserved motifs are underlined in green. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to EAV nsp9.

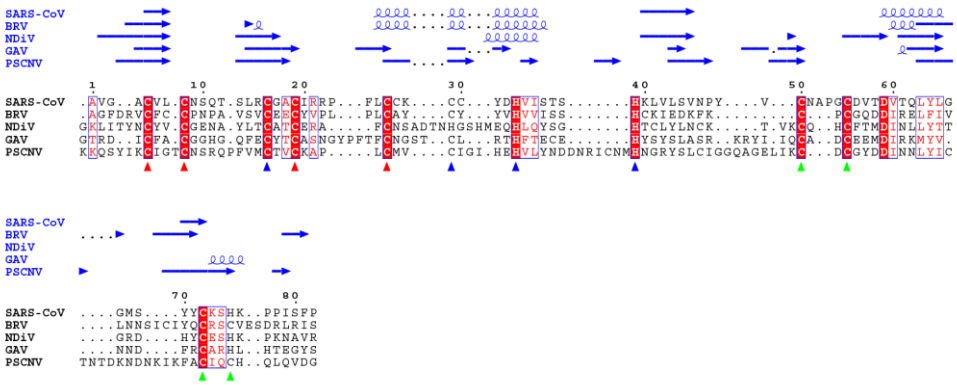


Figure S10 | MSA of ZBD domains from four distantly related nidoviruses and PSCNV (7379–7484 aa). Residues of three zinc fingers coordinating zinc ions (delineated according to the solved EAV ZBD structure [48]) are marked by red, blue and green triangles, respectively. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to SARS-CoV nsp13.

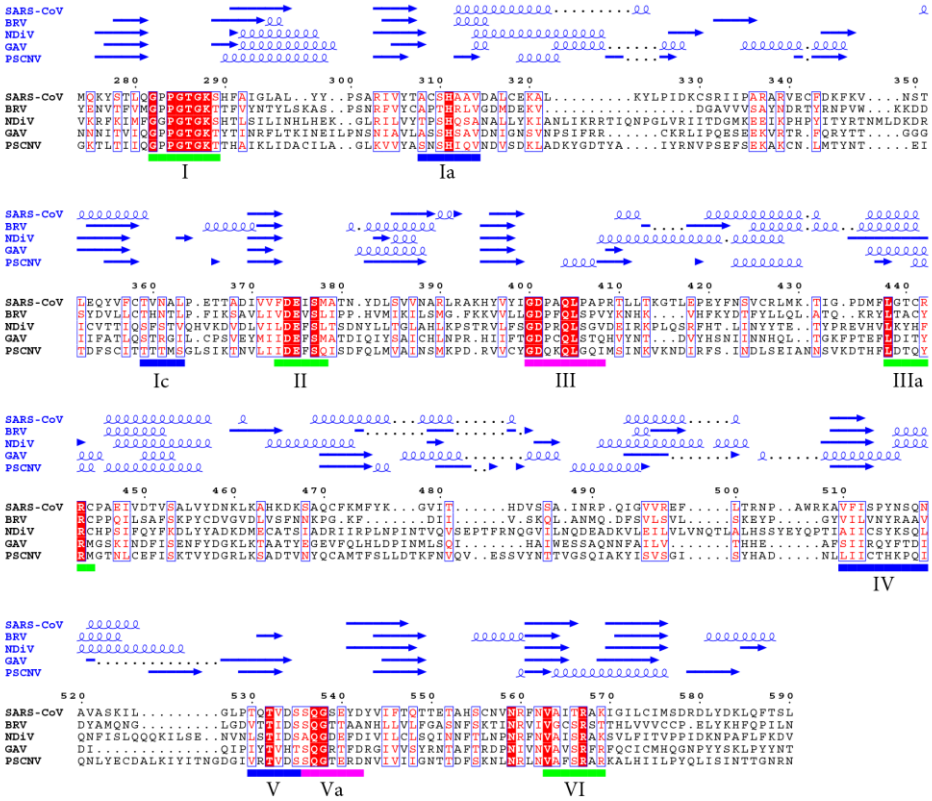


Figure S11 | MSA of HEL1 domains from four distantly related nidoviruses and PSCNV (7718–8056 aa). Conserved motifs are highlighted by color indicating their predominant function [47]: NTP binding and hydrolysis, green; nucleic acid binding, blue; coupling of NTP and nucleic acid binding, purple. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to SARS-CoV nsp13.

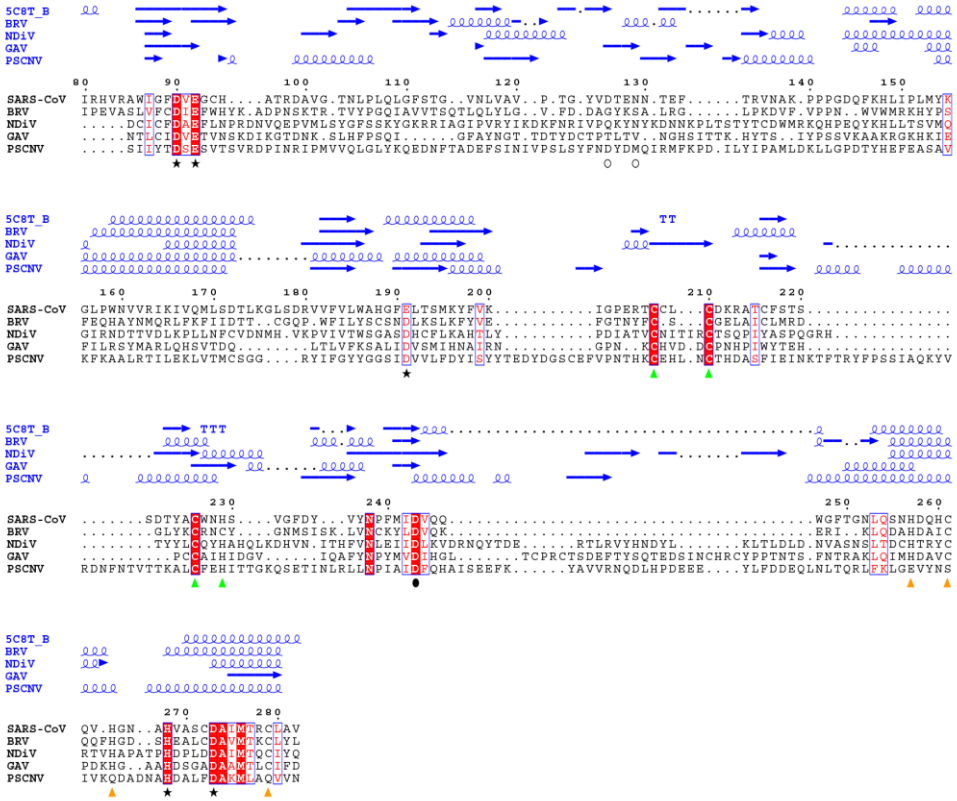


Figure S12 | MSA of ExoN domains from four distantly related nidoviruses and PSCNV (8342–8629 aa). Columns containing SARS-CoV ExoN catalytic residues and Asp243 residue, essential for nuclease activity, are marked by black stars and circle, respectively. Green and orange triangles mark columns that contain residues of two SARS-CoV ExoN zinc fingers; empty circles indicate columns that contain SARS-CoV ExoN residues interacting with nsp10 (the majority of such residues are not shown, as they belong to the N-terminal 1-76 aa region of SARS-CoV nsp14) [92]. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to SARS-CoV nsp14.

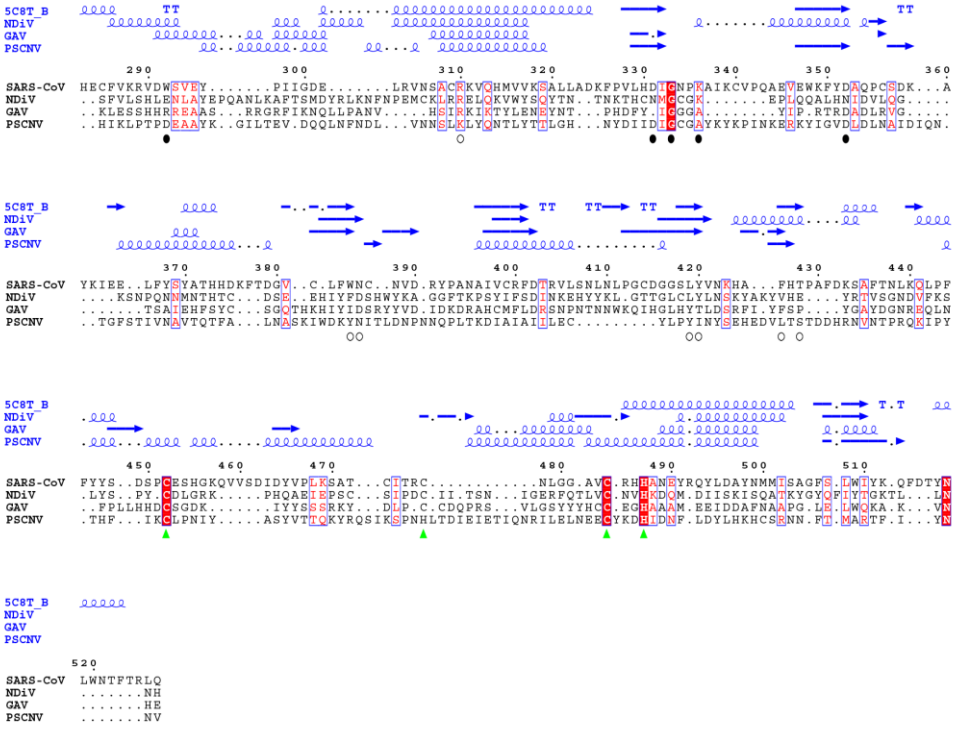


Figure S13 | MSA of N-MT domains from three distantly related nidoviruses and PSCNV (8632–8878 aa). Columns containing SARS-CoV SAH- and GpppA-binding residues, such that their mutation significantly reduced N7-MTase activity, are marked by black and empty circles, respectively. Residues of SARS-CoV N-MT involved in formation of zinc-finger are marked by green triangles [92]. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to SARS-CoV nsp14.

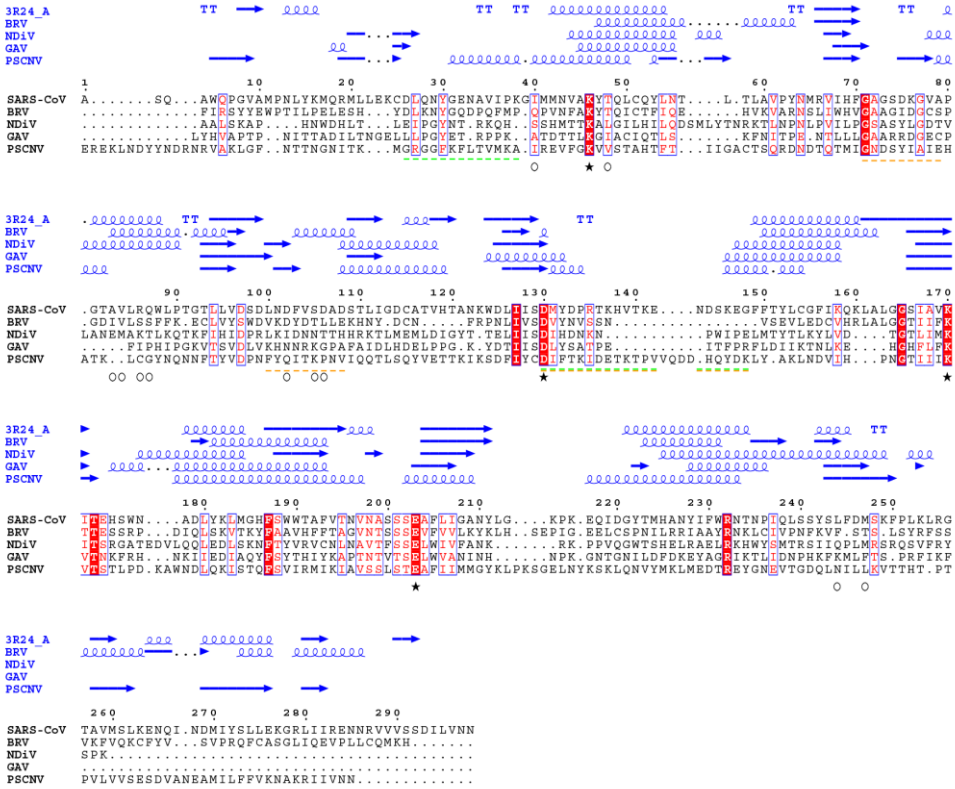


Figure S14 | MSA of O-MT domains from four distantly related nidoviruses and PSCNV (9110–9406 aa). Columns containing SARS-CoV O-MT catalytic tetrad residues are marked by black stars. SARS-CoV O-MT residues involved in interaction with nsp10 are marked by empty circles. Loops constituting SAM-binding cleft and cap-binding groove of SARS-CoV O-MT are underlined in orange and green, respectively [98]. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to SARS-CoV nsp16.

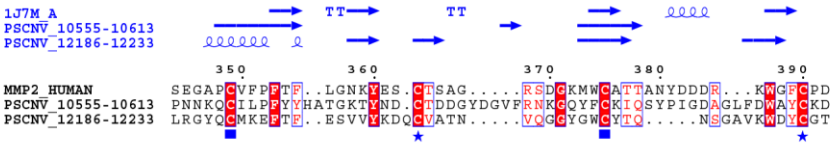


Figure S15 | Comparison of FN2 domains from human matrix metalloproteinase-2 and PSCNV. Shown is the MSA of the third FN2 domain of human matrix metalloproteinase-2 (MMP2) and FN2a (10555–10613 aa) and FN2b (12186–12233 aa) of PSCNV. Pairs of cysteine residues, predicted to form disulfide bridges, are designated by blue bars (first pair) and stars (second pair). Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structures, derived from MMP2 1J7M and predicted for PSCNV domains, is shown in blue. Residue numbering above the alignment refers to the top sequence.

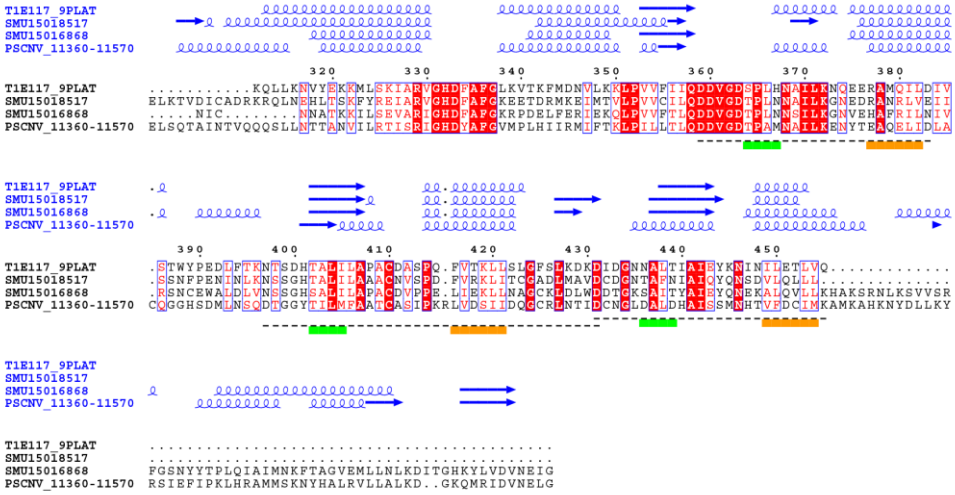


Figure S16 | Comparison of PSCNV ANK domain with most closely related flatworm proteins. Individual ankyrin repeats in PSCNV polyprotein are underlined by black dashed lines. Signature motifs of individual ankyrin repeats are highlighted in green and orange. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Predicted secondary structure is shown in blue. Residue numbering above the alignment refers to the top sequence.

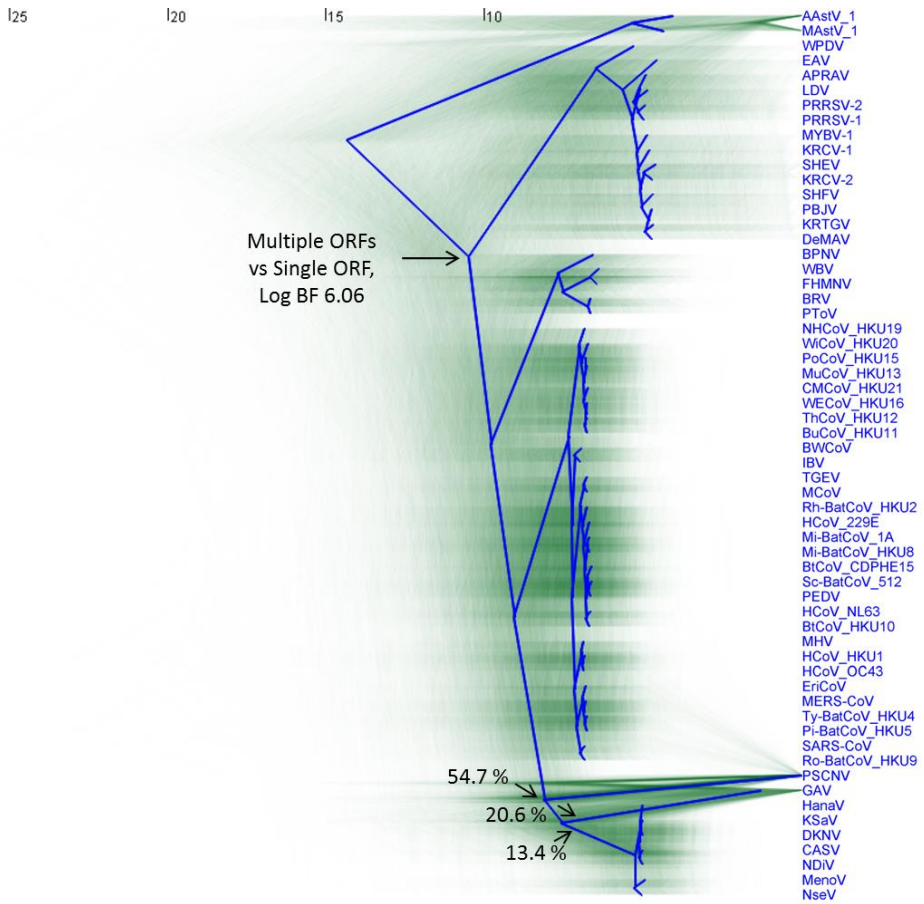


Figure S17 | Phylogeny reconstructed by BEAST based on the alignment of RdRp core of PSCNV, nidoviruses, and astroviruses. Bayesian sample of trees is shown in green, consensus tree with the highest clade support is shown in blue. Support for multiple ORFs vs single ORF in the genome of MRCA of nidoviruses as calculated using BayesTraits V2 is indicated. Short arrows show three most frequently observed (percentages of trees in the sample indicated) positions of the PSCNV branch, which collectively account for 88.7% of PSCNV topologies in the tree sample analyzed. Position of the PSCNV branch in the depicted consensus tree is the one that is most frequently observed (54.7% of trees in the sample).

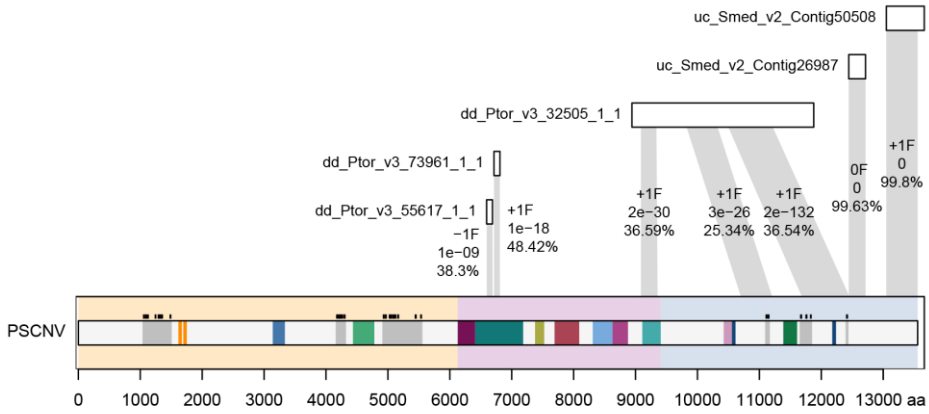


Figure S18 | Statistically significant BLAST hits between translated contigs of PlanMine database and PSCNV polyprotein. Contigs from two assemblies, dd_Ptor_v3 and uc_Smed_v2, are shown as white rectangles. For each hit, depicted as a grey band, a frame in which the contig was translated ("F" stands for forward), E-value, and percentage of amino acid identity are specified. Contig ox_Smed_v2_19364 was also identified but is not depicted due to being identical (with the exception of four 3'-terminal nt) to uc_Smed_v2_Contig50508. See Fig. 2 for PSCNV genome map designations.

Table S1 | Genome sequences and size characteristics of representatives of nidovirus species used in bioinformatics analyses.

(Sub)family	Species	Acronym	Accession number	Genome, nt	Genome region, nt		
					ORF1a	ORF1b	3'ORFs
Arteriviridae	*Equine arteritis virus*	EAV	X53459.3	12704	5181	4347	2894
Arteriviridae	*Lactate dehydrogenase-elevating virus*	LDV	U15146.1	14104	6615	4236	3018
Arteriviridae	*Porcine respiratory and reproductive syndrome virus 1*	PRRSV-1	M96262.2	15111	7185	4380	3199
Arteriviridae	*Porcine respiratory and reproductive syndrome virus 2*	PRRSV-2	U87392.3	15411	7506	4377	3189
Arteriviridae	*Simian hemorrhagic fever virus*	SHFV	AF180391.2	15717	6312	4476	4634
Arteriviridae	*Kibale red-tailed guenon virus 1*	KRTGV	JX473849.1	15264	6177	4476	4379
Arteriviridae	*Kibale red colobus virus 1*	KRCV-1	KC787630.1	15446	6141	4395	4678
Arteriviridae	*Kibale red colobus virus 2*	KRCV-2	KC787658.1	15596	6153	4458	4530
Arteriviridae	*Mikumi yellow baboon virus 1*	MYBV-1	KM110938.1	14927	6165	4461	4101
Arteriviridae	*Simian hemorrhagic encephalitis virus*	SHEV	KM677927.1	15370	6270	4401	4385
Arteriviridae	*DeBrazza's monkey arterivirus*	DeMAV	KP126831.1	15684	6249	4503	4622
Arteriviridae	*Pebjah virus*	PBJV	KR139839.1	15478	6183	4452	4615
Arteriviridae	*African pouched rat arterivirus*	APRAV	KP026921.1	14953	6717	4353	3400
Arteriviridae	*Wobbly possum disease virus*	WPDV	JN116253.3	12917	5973	4236	2351
Coronavirinae	*Alphacoronavirus 1*	TGEV	AJ271965.2	28586	12024	8031	7939
Coronavirinae	*Human coronavirus 229E*	HCoV_229E	AF304460.1	27317	12228	8049	6287
Coronavirinae	*Human coronavirus NL63*	HCoV_NL63	AY567487.1	27553	12153	8037	6791
Coronavirinae	*Miniopterus bat coronavirus 1*	Mi-BatCoV_1A	EU420138.1	28326	12777	8022	6970
Coronavirinae	*Miniopterus bat coronavirus HKU8*	Mi-BatCoV_HKU8	EU420139.1	28773	12666	8025	7575
Coronavirinae	*Porcine epidemic diarrhea virus*	PEDV	AF353511.1	28033	12324	8022	7169
Coronavirinae	*Rhinolophus bat coronavirus HKU2*	Rh-BatCoV_HKU2	EF203065.1	27164	12150	8034	6428
Coronavirinae	*Scotophilus bat coronavirus 512*	Sc-BatCoV_512	DQ648858.1	28203	12357	8025	7286
Coronavirinae	*Bat coronavirus HKU10*	BtCoV_HKU10	JQ989271.1	28489	12318	8028	7596
Coronavirinae	*Bat coronavirus CDPHE15*	BtCoV_CDPHE15	KF430219.1	28035	12453	8025	7109
Coronavirinae	*Mink coronavirus 1*	MCoV	HM245925.1	28941	12027	8022	8327
Coronavirinae	*Betacoronavirus 1*	HCoV_OC43	AY585228.1	30741	13131	8157	8929

Table S1 (continued)

(Sub)family	Species	Acronym	Accession number	Genome, nt	Genome region, nt		
					ORF1a	ORF1b	3'ORFs
Coronavirinae	*Human coronavirus HKU1*	HCoV_HKU1	AY597011.1	29942	13395	8154	7892
Coronavirinae	*Murine coronavirus*	MHV	AF201929.1	31276	13230	8145	9378
Coronavirinae	*Pipistrellus bat coronavirus HKU5*	Pi-BatCoV_HKU5	EF065509.1	30482	13425	8124	8353
Coronavirinae	*Rousettus bat coronavirus HKU9*	Ro-BatCoV_HKU9	EF065513.1	29114	12687	8070	7862
Coronavirinae	*Severe acute respiratory syndrome-related coronavirus*	SARS-CoV	AY274119.3[1]	29751	13134	8088	7903
Coronavirinae	*Tylonycteris bat coronavirus HKU4*	Ty-BatCoV_HKU4	EF065505.1	30286	13284	8076	8343
Coronavirinae	*Middle East respiratory syndrome-related coronavirus*	MERS-CoV	JX869059.2	30119	13155	8082	8293
Coronavirinae	*Hedgehog coronavirus 1*	EriCoV	KC545383.1	30148	13344	8109	8134
Coronavirinae	*Avian coronavirus*	IBV	M95169.1	27608	11826	8064	6685
Coronavirinae	*Beluga whale coronavirus SW1*	BWCoV	EU111742.1	31686	11865	8127	10771
Coronavirinae	*Bulbul coronavirus HKU11*	BuCoV_HKU11	FJ376619.2	26487	10746	8049	6867
Coronavirinae	*Thrush coronavirus HKU12*	ThCoV_HKU12	FJ376621.1	26396	10812	8049	6722
Coronavirinae	*Munia coronavirus HKU13*	MuCoV_HKU13	FJ376622.1	26552	10998	7926	6812
Coronavirinae	*Coronavirus HKU15*	PoCoV_HKU15	JQ065043.1	25432	10875	7929	5866
Coronavirinae	*White-eye coronavirus HKU16*	WECoV_HKU16	JQ065044.1	26041	10839	8049	6420
Coronavirinae	*Night heron coronavirus HKU19*	NHCoV_HKU19	JQ065047.1	26077	10830	8013	6553
Coronavirinae	*Wigeon coronavirus HKU20*	WiCoV_HKU20	JQ065048.1	26227	10704	7917	7114
Coronavirinae	*Common moorhen coronavirus HKU21*	CMCoV_HKU21	JQ065049.1	26223	10584	8043	6813
Torovirinae	*Bovine torovirus*	BRV	AY427798.1	28475	13332	6870	7219
Torovirinae	*Porcine torovirus*	PToV	JQ860350.1	28301	13248	6870	7199
Torovirinae	*White bream virus*	WBV	DQ898157.1	26660	13599	6969	4877
Torovirinae	*Fathead minnow nidovirus 1*	FHMNV	GU002364.2	27318	14565	6960	4813
Torovirinae	*Ball python nidovirus 1*	BPNV	KJ541759.1	33452	17394	6933	7170
Mesoniviridae	*Alphamesonivirus 1*	NDiV	DQ458789.2	20192	7491	7788	3466
Mesoniviridae	*Alphamesonivirus 2*	KSaV	KC807171.1	20795	8073	7788	3519

[1]To generate Fig. S12 and S13, GU553365.1 was used; to generate Fig. S14 – AY394850.2

Table S1 (continued)

(Sub)family	Species	Acronym	Accession number	Genome, nt	Genome region, nt		
					ORF1a	ORF1b	3'ORFs
Mesoniviridae	*Alphamesonivirus 3*	DKNV	AB753015.2	20307	7644	7782	3493
Mesoniviridae	*Alphamesonivirus 4*	CASV	KJ125489.1	19917	7416	7782	3448
Mesoniviridae	*Alphamesonivirus 5*	HanaV	JQ957872.1	20070	7488	7776	3447
Mesoniviridae	*Mesonivirus 1*	NseV	JQ957874.1	20074	7482	7791	3488
Mesoniviridae	*Mesonivirus 2*	MenoV	JQ957873.1	19979	7404	7791	3463
Roniviridae	*Gill-associated virus*	GAV	AF227196.2	26253	12153	7869	5508

Table S2 | Primers used for viral genome detection, 5'-RACE, and genome-wide overlapping amplification.

Primer name	Region	Sequence	Paired with	Amplicon size (bp)	Purpose
PSCNV-detect-fwd	36764..36782	AGGTGGTTATGGATGGTGT	PSCNV-detect-rev	1047	Genome detection
PSCNV-detect-rev	complement(37793..37810)	GGTGATTGATTGCGTGGT			
PSCNV-FPR-rev-606	complement(584..606)	AGACACCATCTCTTTCCATTTGT	RLM-RACE kit	606	Genomic 5'-RACE
PSCNV-FPR-rev-763	complement(744..763)	GCTATATCACCTTGGTCGCC	RLM-RACE kit	763	Genomic 5'-RACE
PSCNV-FPR-rev-28815	complement(28796..28815)	CCAAATCGGTCAAAATTCGT	RLM-RACE kit	429	Sg 5'-RACE
PSCNV-FPR-rev-29433	complement(29414..29433)	TGTCGCTTGGCATAAGTTCA	RLM-RACE kit	1047	Sg 5'-RACE
PSCNV-FPR-fwd-171	182..201	ACGAAAGGATGGCGTTCAAA	PSCNV-BlpI-rev	3456	Large amplicon 1
PSCNV-BlpI-rev	complement(3618..3637)	ACATGGGCATCTGTGAACAT			
PSCNV-BlpI-fwd	3234..3258	AGAATCCAATCATATCGACGAATTC	PSCNV-BglI-rev	6758	Large amplicon 2
PSCNV-BglI-rev	complement(9971..9991)	TCATCTGAACAACCTGTTGCT			
PSCNV-BglI-fwd	9633..9653	GGAGCACCGTTGACATCATAT	PSCNV-BstEII-rev	8101	Large amplicon 3
PSCNV-BstEII-rev	complement(17714..17733)	CGATAGCGGCAACAATCGAA			
PSCNV-BstEII-fwd	17182..17201	TAAACAGCCCACCACCAACA	PSCNV-MluI-rev	4194	Large amplicon 4
PSCNV-MluI-rev	complement(21375..21395)	AGAACTTTGGTCATGTCGTGT			
PSCNV-MluI-fwd	21076..21097	TGGGTGAGCTAATGAATTGTGT	PSCNV-AgeI-rev	7019	Large amplicon 5
PSCNV-AgeI-rev	complement(28072..28094)	AATAAAAGCCTCAGTGCTCAAAC			
PSCNV-AgeI-fwd	27539..27559	AAAGATGGGACGTGGTGGATT	PSCNV-StuI-rev	4416	Large amplicon 6
PSCNV-StuI-rev	complement(31935..31954)	GCCCAATCAAACAAGCCTGC			
PSCNV-StuI-fwd	31416..31436	CCAACAACACAACTTCGGACA	PSCNV-SacI-rev	6114	Large amplicon 7
PSCNV-SacI-rev	complement(37509..37529)	TCCACCACGGAAAAATACTCG			

Table S3 | *S. mediterranea* RNA-seq datasets screened for presence of PSCNV reads.

			Sequencing experiments		
Laboratory	**Strain**	**BioProject**	**All**	**With PSCNV reads**	**PSCNV reads, ppm[1]**
Aboobaker	Asexual	PRJNA79649	1[2]	0	0
Bartscherer	Asexual	PRJNA222859	8	0	0
Graveley	Asexual	PRJNA151483	3	2	10
Graveley	Sexual	PRJNA151483	6	6	69
Newmark	Asexual	PRJNA319973	15	15	19
Newmark	Sexual	PRJNA79031	4	4	1834
Pearson	Asexual	PRJNA205281	9	0	0
Pearson	Asexual	PRJNA415947	5	0	0
Rajewsky	Asexual	PRJNA79997	4	0	0
Reddien	Asexual	PRJNA320389	8	0	0
Rink	Asexual	PRJNA208294	8	0	0
Sanchez Alvarado	Asexual	PRJNA215411	1	0	0
Sanchez Alvarado	Sexual	PRJNA215411	1	0	0
Sanchez Alvarado	Sexual	PRJNA324545	40	0[3]	0
Sanchez Alvarado	Sexual	PRJNA421285	32	32	1258
Sanchez Alvarado	Sexual	PRJNA421831	15	0	0

[1]Number of reads mapped to the PSCNV reference genome sequence per million reads in the BioProject.

[2]Data obtained using ABI SOLiD sequencing platform (5 runs) were not analyzed.

[3]A single read from SRR3629921 run mapped to the PSCNV genome and was considered an artefact.

Table S4 | PSCNV genome sequence variants in the 28389–41000 nt region[1].

Reference			PRJNA319973			PRJNA79031			PRJNA421285		
genome coordinate	nt	aa	p-value	nt	aa	p-value	nt	aa	p-value	nt	aa
31585	U	I	3.20E-23	C	T	3.20E-23	C	T	3.20E-23	C	T
31828	A	H		*	*		*	*	4.00E-19	G	R
35506	G	R		*	*		*	*	3.20E-23	A	K
35714	G	Q		*	*		*	*	3.20E-23	A	*
37558	G	R		*	*		*	*	0.031	A	H
37648	A	Q		*	*		*	*	3.20E-23	C	P
39112	U	I		*	*		*	*	3.20E-23	C	T
39185	U	F		*	*		*	*	3.20E-23	C	*
40748	C	Y		*	*		*	*	1.30E-20	U	*

[1]Asterisks indicate nt/aa identical to the reference.

Table S5 | Domain identification in PSCNV polyprotein through comparison with various protein databases using HHsearch (see Fig. S3 for outline).

				Hit					
Domain	**Iteration[1]**	**Index[2]**	**Database[3]**	**Name[4]**	**Probability**	**E-value**	**PSCNV coo[5]**	**PSCNV len[6]**	**Template HMM[7]**
RNase T2	I	a	pfam*	PF00445, Ribonuclease_T2	80	0.18	3133–3226	94	6–107 (178)
3CLpro	I	b	pdb	3k6y_A, Serine_protease	73.3	39	4462–4491	30	55–79 (237)
	I	1	scop	d2o8la1, V8 protease	95.5	0.032	4545–4641	97	90–188 (216)
	I	c	pfam*	3CLproCore_CoToMeRo	2.8	420	4605–4636	32	132–158 (187)
NiRAN	II	2	pfam*	NiRAN_CoToMeRo	95.1	0.0073	6226–6406	181	34–198 (202)
RdRp	I	3	pfam*	RdRpCore_CoToMeRo	99.1	1.00E-09	6639–7133	495	7–450 (457)
ZBD	II	d	pfam*	PF14569, Zinc-binding RING-finger	35	2.6	7387–7438	52	17–64 (77)
	II	e	pfam*	ZBD_CoToMeRo	22.7	39	7395–7460	66	13–64 (80)
HEL1	I	4	pfam*	HEL1_CoToMeRo	99.9	7.50E-28	7719–8044	326	2–307 (319)
ExoN	II	f	scop	d1w0ha, human DEDDh 3'-5'-exoribonuclease	26.2	12	8342–8446	105	7–95 (200)
	II	g	pfam*	ExoN_CoToMeRo	4.2	240	8449–8560	112	98–168 (205)
	II	h	pdb	3mxm_B, TREX1 3' Exonuclease	39.1	14	8598–8631	34	178–211 (242)
N-MT	II	i	pfam*	PF07091, Ribosomal RNA methyltransferase	80.8	0.19	8636–8708	73	46–134 (243)
	II	j	pfam*	NMT_CoMeRo	0.8	1200	8659–8686	28	24–54 (238)
O-MT	IVb	5	pfam*	OMT_CoToMeRo	96.6	0.00033	9237–9407	171	122–280 (305)
FN2a	I	k	pfam*	PF00040, Fibronectin type II domain	91.3	0.026	10561–10611	51	2–42 (42)
ANK	I	6	pdb	2rfa_A, ankyrin repeat domain of TRPV6	98.9	3.30E-08	11394–11555	162	35–218 (232)
FN2b	I	l	pfam*	PF00040, Fibronectin type II domain	78.5	0.35	12191–12231	41	1–42 (42)

[1]Iteration of HHsearch-based procedure during which hit was obtained.
[2]Index of cluster of significant hits (numeric, black font) or individual sub-significant hit (letter, grey font). For each cluster of significant hits, only the top hit is presented in the table.
[3]Databases: pfam*, pfamA_28.0 extended to include eight nidovirus domains; pdb, pdb70_06Sep14; scop, scop70_1.75.
[4]Names of nidoviral domains that were added to pfamA_28.0 have suffixes _CoToMeRo or _CoMeRo (each syllable designates a (sub)family of nidoviruses, included in the profile).
[5]Coordinates of hit in residues of PSCNV polyprotein.
[6]Length of hit in residues of PSCNV polyprotein.
[7]Coordinates of hit in match states of HMM profile from database. Number of match states in HMM profile is shown in parentheses.

Table S6 | Genome region size increases in PSCNV compared to ExoN-containing nidoviruses.

Region	p, nt[1]	M, nt[2]	m, nt[3]	D_1, %[4]	D_2, %[5]	D_3, %[6]
genome	41121	33452	27608	22.9	131.2	100
ORF1a	18384	17394	12153	5.7	18.9	14.4
ORF1b	9834	8157	8025	20.6	1270.5	968.1
3'ORFs	12453	10771	6970	15.6	44.3	33.7

[1]Region size in PSCNV.
[2]Maximum region size in ExoN-containing nidoviruses.
[3]Median region size in ExoN-containing nidoviruses.
[4]D_1(region)=(p-M)/M*100%, PSCNV region size increase calculated accounting for the region size of ExoN-containing nidoviruses.
[5]D_2(region)=(p-M)/(M-m)*100%, PSCNV region size increase calculated accounting for the region size variation of ExoN-containing nidoviruses.
[6]D_3(region)=D_2(region)/D_2(genome)*100%, PSCNV region size increase calculated accounting for the region size variation of ExoN-containing nidoviruses and PSCNV genome size increase.

REFERENCES

1. Joyce GF: **The antiquity of RNA-based evolution**. *Nature* 2002, **418**(6894):214-221.
2. Leipe DD, Aravind L, Koonin EV: **Did DNA replication evolve twice independently?** *Nucleic Acids Res* 1999, **27**(17):3389-3401.
3. Poole AM, Logan DT: **Modern mRNA proofreading and repair: clues that the last universal common ancestor possessed an RNA genome?** *Mol Biol Evol* 2005, **22**(6):1444-1455.
4. Xavier JC, Patil KR, Rocha I: **Systems biology perspectives on minimal and simpler cells**. *Microbiol Mol Biol Rev* 2014, **78**(3):487-509.
5. Li S, Guo W, Dewey CN, Greaser ML: **Rbm20 regulates titin alternative splicing as a splicing repressor**. *Nucleic Acids Res* 2013, **41**(4):2659-2672.
6. Holmes EC: **Error thresholds and the constraints to RNA virus evolution**. *Trends Microbiol* 2003, **11**(12):543-546.
7. Lauber C, Gorbalenya AE: **Taxonomy Advancement and Genome Size Change: Two Perspectives on RNA Virus Genetic Diversity**. In: *Virus Evolution: Current Research and Future Directions.* Edited by Weaver SC, Denison M, Roossinck M, Vignuzzi M: Caister Academic Press; 2016: 215-232.
8. Belshaw R, Gardner A, Rambaut A, Pybus OG: **Pacing a small cage: mutation and RNA viruses**. *Trends Ecol Evol* 2008, **23**(4):188-193.
9. Chirico N, Vianelli A, Belshaw R: **Why genes overlap in viruses**. *Proc Biol Sci* 2010, **277**(1701):3809-3817.
10. Gorbalenya AE: **Host-related sequences in RNA viral genomes**. *Seminars in Virology* 1992, **3**:359-371.
11. Koonin EV, Dolja VV: **A virocentric perspective on the evolution of life**. *Curr Opin Virol* 2013, **3**(5):546-557.
12. Forterre P: **Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain**. *Proc Natl Acad Sci U S A* 2006, **103**(10):3669-3674.
13. Agol VI: **Which came first, the virus or the cell?** *Paleontological Journal* 2010, **44**(7):728-736.
14. Nga PT, Parquet MC, Lauber C, Parida M, Nabeshima T, Yu F, Thuy NT, Inoue S, Ito T, Okamoto K *et al*: **Discovery of the first insect nidovirus, a missing evolutionary link in the emergence of the largest RNA virus genomes**. *PLoS Pathog* 2011, **7**(9):e1002215.
15. Eigen M: **Selforganization of matter and the evolution of biological macromolecules**. *Naturwissenschaften* 1971, **58**(10):465-523.

16. Steinhauer DA, Domingo E, Holland JJ: **Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase**. *Gene* 1992, **122**(2):281-288.

17. Drake JW, Holland JJ: **Mutation rates among RNA viruses**. *Proc Natl Acad Sci U S A* 1999, **96**(24):13910-13913.

18. Bull JJ, Sanjuan R, Wilke CO: **Theory of lethal mutagenesis for viruses**. *J Virol* 2007, **81**(6):2930-2939.

19. Eigen M: **Error catastrophe and antiviral strategy**. *Proc Natl Acad Sci U S A* 2002, **99**(21):13374-13376.

20. den Boon JA, Snijder EJ, Chirnside ED, de Vries AA, Horzinek MC, Spaan WJ: **Equine arteritis virus is not a togavirus but belongs to the coronaviruslike superfamily**. *J Virol* 1991, **65**(6):2910-2920.

21. Stenglein MD, Jacobson ER, Wozniak EJ, Wellehan JF, Kincaid A, Gordon M, Porter BF, Baumgartner W, Stahl S, Kelley K *et al*: **Ball python nidovirus: a candidate etiologic agent for severe respiratory disease in Python regius**. *MBio* 2014, **5**(5):e01484-01414.

22. Bodewes R, Lempp C, Schurch AC, Habierski A, Hahn K, Lamers M, von Dornberg K, Wohlsein P, Drexler JF, Haagmans BL *et al*: **Novel divergent nidovirus in a python with pneumonia**. *J Gen Virol* 2014, **95**(Pt 11):2480-2485.

23. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ: **Nidovirales: evolving the largest RNA virus genome**. *Virus Res* 2006, **117**(1):17-37.

24. de Groot RJ, Cowley JA, Enjuanes L, Faaberg KS, Perlman S, Rottier PJM, Snijder EJ, Ziebuhr J, Gorbalenya AE: **Order *Nidovirales*.** In: *Virus Taxonomy, the 9th Report of the International Committee on Taxonomy of Viruses.* Edited by King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ: Elsevier; 2012: 785-795.

25. Lauber C, Ziebuhr J, Junglen S, Drosten C, Zirkel F, Nga PT, Morita K, Snijder EJ, Gorbalenya AE: **Mesoniviridae: a proposed new family in the order Nidovirales formed by a single species of mosquito-borne viruses**. *Arch Virol* 2012, **157**(8):1623-1628.

26. Snijder EJ, Kikkert M, Fang Y: **Arterivirus molecular biology and pathogenesis**. *J Gen Virol* 2013, **94**(Pt 10):2141-2163.

27. Masters PS, Perlman S: ***Coronaviridae***. In: *Fields Virology.* Edited by Knipe DM, Howley PM, vol. 1. Philadelphia: Wolters Kluwer Health/Lippincott Williams and Wilkins; 2013: 825-858.

28. Cowley JA, Dimmock CM, Wongteerasupaya C, Boonsaeng V, Panyim S, Walker PJ: **Yellow head virus from Thailand and gill-associated virus from Australia are closely related but distinct prawn viruses**. *Dis Aquat Organ* 1999, **36**(2):153-157.

29. Zhou P, Fan H, Lan T, Yang XL, Shi WF, Zhang W, Zhu Y, Zhang YW, Xie QM, Mani S *et al*: **Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin**. *Nature* 2018, **556**(7700):255-258.

30. Al-Tawfiq JA, Memish ZA: **Middle East respiratory syndrome coronavirus: transmission and phylogenetic evolution**. *Trends Microbiol* 2014, **22**(10):573-579.

31. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, 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**(5):991-1004.

32. 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**(13):5108-5113.

33. 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**(2):E162-E171.

34. 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**(22):12135-12144.

35. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS *et al*: **Redefining the invertebrate RNA virosphere**. *Nature* 2016, **540**:539-543.

36. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L *et al*: **The evolutionary history of vertebrate RNA viruses**. *Nature* 2018, **556**:197-202.

37. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM: **The 1.2-megabase genome sequence of Mimivirus**. *Science* 2004, **306**(5700):1344-1350.

38. Philippe N, Legendre M, Doutre G, Coute Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C *et al*: **Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes**. *Science* 2013, **341**(6143):281-286.

39. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ *et al*: **Giant viruses with an expanded complement of translation system components**. *Science* 2017, **356**(6333):82-85.

40. Plant EP, Dinman JD: **The role of programmed-1 ribosomal frameshifting in coronavirus propagation**. *Front Biosci* 2008, **13**:4873-4881.

41. Firth AE, Brierley I: **Non-canonical translation in RNA viruses**. *J Gen Virol* 2012, **93**(Pt 7):1385-1409.

42. Ziebuhr J, Snijder EJ, Gorbalenya AE: **Virus-encoded proteinases and proteolytic processing in the Nidovirales**. *J Gen Virol* 2000, **81**(Pt 4):853-879.

43. Neuman BW, Angelini MM, Buchmeier MJ: **Does form meet function in the coronavirus replicative organelle?** *Trends Microbiol* 2014, **22**(11):642-647.

44. Snijder EJ, Decroly E, Ziebuhr J: **The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing**. *Adv Virus Res* 2016, **96**:59-126.

45. Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, Janssen GM, 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**(17):8416-8434.

46. Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, Snijder EJ, Canard B, Imbert I: **One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities**. *Proc Natl Acad Sci U S A* 2014, **111**(37):E3900-3909.

47. Lehmann KC, Snijder EJ, Posthuma CC, Gorbalenya AE: **What we know but do not understand about nidovirus helicases**. *Virus Res* 2015, **202**:12-32.

48. Deng Z, Lehmann KC, Li X, Feng C, Wang G, Zhang Q, Qi X, Yu L, Zhang X, Feng W *et al*: **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* 2014, **42**(5):3464-3477.

49. Hao W, Wojdyla JA, Zhao R, Han R, Das R, Zlatev I, Manoharan M, Wang M, Cui S: **Crystal structure of Middle East respiratory syndrome coronavirus helicase**. *PLoS Pathog* 2017, **13**(6):e1006474.

50. Seybert A, Hegyi A, Siddell SG, Ziebuhr J: **The human coronavirus 229E superfamily 1 helicase has RNA and DNA duplex-unwinding activities with 5'-to-3' polarity**. *RNA* 2000, **6**(7):1056-1068.

51. Pasternak AO, Spaan WJ, Snijder EJ: **Nidovirus transcription: how to make sense...?** *J Gen Virol* 2006, **87**(Pt 6):1403-1421.

52. Sola I, Almazan F, Zuniga S, Enjuanes L: **Continuous and Discontinuous RNA Synthesis in Coronaviruses**. *Annu Rev Virol* 2015, **2**(1):265-288.

53. Di H, Madden JC, Jr., Morantz EK, Tang HY, Graham RL, Baric RS, Brinton MA: **Expanded subgenomic mRNA transcriptome and coding capacity of a nidovirus**. *Proc Natl Acad Sci U S A* 2017, **114**(42):E8895-E8904.

54. Perlman S, Netland J: **Coronaviruses post-SARS: update on replication and pathogenesis**. *Nat Rev Microbiol* 2009, **7**(6):439-450.

55. Tian D, Wei Z, Zevenhoven-Dobbe JC, Liu R, Tong G, Snijder EJ, Yuan S: **Arterivirus minor envelope proteins are a major determinant of viral tropism in cell culture**. *J Virol* 2012, **86**(7):3701-3712.

56. Li F: **Receptor recognition mechanisms of coronaviruses: a decade of structural studies**. *J Virol* 2015, **89**(4):1954-1964.

57. de Groot RJ: **Structure, function and evolution of the hemagglutinin-esterase proteins of corona- and toroviruses**. *Glycoconj J* 2006, **23**(1-2):59-72.

58. Veit M, Matczuk AK, Sinhadri BC, Krause E, Thaa B: **Membrane proteins of arterivirus particles: structure, topology, processing and function**. *Virus Res* 2014, **194**:16-36.

59. Ujike M, Taguchi F: **Incorporation of spike and membrane glycoproteins into coronavirus virions**. *Viruses* 2015, **7**(4):1700-1725.

60. Fehr AR, Perlman S: **Coronaviruses: an overview of their replication and pathogenesis**. *Methods Mol Biol* 2015, **1282**:1-23.

61. Kindler E, Thiel V, Weber F: **Interaction of SARS and MERS Coronaviruses with the Antiviral Interferon Response**. *Adv Virus Res* 2016, **96**:219-243.

62. Totura AL, Baric RS: **SARS coronavirus pathogenesis: host innate immune responses and viral antagonism of interferon**. *Curr Opin Virol* 2012, **2**(3):264-275.

63. Silverman RH, Weiss SR: **Viral phosphodiesterases that antagonize double-stranded RNA signaling to RNase L by degrading 2-5A**. *J Interferon Cytokine Res* 2014, **34**(6):455-463.

64. Deng X, Baker SC: **An "Old" protein with a new story: Coronavirus endoribonuclease is important for evading host antiviral defenses**. *Virology* 2018, **517**:157-163.

65. Menachery VD, Mitchell HD, Cockrell AS, Gralinski LE, Yount BL, Jr., Graham RL, McAnarney ET, Douglas MG, Scobey T, Beall A *et al*: **MERS-CoV Accessory ORFs Play Key Role for Infection and Pathogenesis**. *MBio* 2017, **8**(4).

66. Lauber C, Goeman JJ, Parquet MC, 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**(7):e1003500.

67. Saberi A, Jamal A, Beets I, Schoofs L, Newmark PA: **GPCRs Direct Germline Development and Somatic Gonad Function in Planarians**. *PLoS Biol* 2016, **14**(5):e1002457.

68. Newmark PA, Sanchez Alvarado A: **Not your father's planarian: a classic model enters the era of functional genomics**. *Nat Rev Genet* 2002, **3**(3):210-219.

69. Zayas RM, Hernandez A, Habermann B, Wang Y, Stary JM, Newmark PA: **The planarian Schmidtea mediterranea as a model for epigenetic germ cell specification: analysis of ESTs from the hermaphroditic strain**. *Proc Natl Acad Sci U S A* 2005, **102**(51):18491-18496.

70. Grohme MA, Schloissnig S, Rozanski A, Pippel M, Young GR, Winkler S, Brandl H, Henry I, Dahl A, Powell S *et al*: **The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms**. *Nature* 2018, **554**(7690):56-61.

71. Newmark PA, Sanchez Alvarado A: **Bromodeoxyuridine specifically labels the regenerative stem cells of planarians**. *Dev Biol* 2000, **220**(2):142-153.

72. Lazaro EM, Harrath AH, Stocchino GA, Pala M, Baguna J, Riutort M: **Schmidtea mediterranea phylogeography: an old species surviving on a few Mediterranean islands?** *BMC Evol Biol* 2011, **11**:274.

73. Zayas RM, Cebria F, Guo T, Feng J, Newmark PA: **The use of lectins as markers for differentiated secretory cells in planarians**. *Dev Dyn* 2010, **239**(11):2888-2897.

74. Ortego J, Ceriani JE, Patino C, Plana J, Enjuanes L: **Absence of E protein arrests transmissible gastroenteritis coronavirus maturation in the secretory pathway**. *Virology* 2007, **368**(2):296-308.

75. Thuy NT, Huy TQ, Nga PT, Morita K, Dunia I, Benedetti L: **A new nidovirus (NamDinh virus NDiV): Its ultrastructural characterization in the C6/36 mosquito cell line**. *Virology* 2013, **444**(1-2):337-342.

76. Knoops K, Kikkert M, Worm SH, Zevenhoven-Dobbe JC, van der Meer Y, Koster AJ, Mommaas AM, Snijder EJ: **SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum**. *PLoS Biol* 2008, **6**(9):e226.

77. Maier HJ, Hawes PC, Cottam EM, Mantell J, Verkade P, Monaghan P, Wileman T, Britton P: **Infectious bronchitis virus generates spherules from zippered endoplasmic reticulum membranes**. *MBio* 2013, **4**(5):e00801-00813.

78. Boursnell ME, Brown TD, Foulds IJ, Green PF, Tomley FM, Binns MM: **Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus**. *J Gen Virol* 1987, **68 (Pt 1)**:57-77.

79. Shi M, Lin XD, Vasilakis N, Tian JH, Li CX, Chen LJ, Eastwood G, Diao XN, Chen MH, Chen X *et al*: **Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses**. *J Virol* 2016, **90**(2):659-669.

80. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM: **Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis**. *Nucleic Acids Res* 1989, **17**(12):4847-4861.

81. Gulyaeva A, Dunowska M, Hoogendoorn E, Giles J, Samborskiy D, Gorbalenya AE: **Domain organization and evolution of the highly divergent 5' coding region of genomes of arteriviruses including the novel possum nidovirus**. *J Virol* 2017, **91**(6).

82. Neuman BW: **Bioinformatics and functional analyses of coronavirus nonstructural proteins involved in the formation of replicative organelles**. *Antiviral Res* 2016, **135**:97-107.

83. Bosch BJ, Rottier PJM: **Nidovirus Entry into Cells**. In: *Nidoviruses.* Edited by Perlman S, Gallagher T, Snijder EJ. Washington, DC: ASM Press; 2008: 157-178.

84. Hogue BG, Machamer CE: **Coronavirus Structural Proteins and Virus Assembly**. In: *Nidoviruses.* Edited by Perlman S, Gallagher T, Snijder EJ. Washington, DC: ASM Press; 2008: 179-200.

85. Faaberg KS: **Arterivirus Structural Proteins and Assembly**. In: *Nidoviruses.* Edited by Perlman S, Gallagher T, Snijder EJ. Washington, DC: ASM Press; 2008: 211-234.

86. Barrette-Ng IH, Ng KK, Mark BL, Van Aken D, Cherney MM, Garen C, Kolodenko Y, Gorbalenya AE, Snijder EJ, James MN: **Structure of arterivirus nsp4. The smallest chymotrypsin-like proteinase with an alpha/beta C-terminal extension and alternate conformations of the oxyanion hole**. *J Biol Chem* 2002, **277**(42):39960-39966.

87. Anand K, Palm GJ, Mesters JR, Siddell SG, Ziebuhr J, Hilgenfeld R: **Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain**. *EMBO J* 2002, **21**(13):3213-3224.

88. Ziebuhr J, Bayer S, Cowley JA, Gorbalenya AE: **The 3C-like proteinase of an invertebrate nidovirus links coronavirus and potyvirus homologs**. *J Virol* 2003, **77**(2):1415-1426.

89. Blanck S, Stinn A, Tsiklauri L, Zirkel F, Junglen S, Ziebuhr J: **Characterization of an alphamesonivirus 3C-like protease defines a special group of nidovirus main proteases**. *J Virol* 2014, **88**(23):13747-13758.

90. Smits SL, Snijder EJ, de Groot RJ: **Characterization of a torovirus main proteinase**. *J Virol* 2006, **80**(8):4157-4167.

91. Ulferts R, Mettenleiter TC, Ziebuhr J: **Characterization of Bafinivirus main protease autoprocessing activities**. *J Virol* 2011, **85**(3):1348-1359.

92. 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**(30):9436-9441.

93. 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**(24):9372-9377.

94. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR: **Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics**. *PLoS Pathog* 2013, **9**(8):e1003565.

95. Chen Y, Cai H, Pan J, Xiang N, Tien P, Ahola T, Guo D: **Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase**. *Proc Natl Acad Sci U S A* 2009, **106**(9):3484-3489.

96. von Grotthuss M, Wyrwicz LS, Rychlewski L: **mRNA cap-1 methyltransferase in the SARS genome**. *Cell* 2003, **113**(6):701-702.

97. 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**(16):8071-8084.

98. 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**(10):e1002294.

99. Bouvet M, Debarnot C, Imbert I, Selisko B, Snijder EJ, Canard B, Decroly E: **In vitro reconstitution of SARS-coronavirus mRNA cap methylation**. *PLoS Pathog* 2010, **6**(4):e1000863.

100. 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**(15):6675-6685.

101. Irie M: **Structure-function relationships of acid ribonucleases: lysosomal, vacuolar, and periplasmic enzymes**. *Pharmacol Ther* 1999, **81**(2):77-89.

102. Robb SM, Gotting K, Ross E, Sanchez AA: **SmedGD 2.0: The Schmidtea mediterranea genome database**. *Genesis* 2015, **53**(8):535-546.

103. Egger B, Lapraz F, Tomiczek B, Muller S, Dessimoz C, Girstmair J, Skunca N, Rawlinson KA, Cameron CB, Beli E *et al*: **A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms**. *Curr Biol* 2015, **25**(10):1347-1353.

104. Soowannayan C, Cowley JA, Michalski WP, Walker PJ: **RNA-binding domain in the nucleocapsid protein of gill-associated nidovirus of penaeid shrimp**. *PLoS One* 2011, **6**(8):e22156.

105. Rahaman J, Siltberg-Liberles J: **Avoiding Regions Symptomatic of Conformational and Functional Flexibility to Identify Antiviral Targets in Current and Future Coronaviruses**. *Genome Biol Evol* 2016, **8**(11):3471-3484.

106. Cowley JA, Walker PJ: **The complete genome sequence of gill-associated virus of Penaeus monodon prawns indicates a gene organisation unique among nidoviruses**. *Arch Virol* 2002, **147**(10):1977-1987.

107. Yu IM, Oldham ML, Zhang J, Chen J: **Crystal structure of the severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein dimerization domain reveals evolutionary linkage between corona- and arteriviridae**. *J Biol Chem* 2006, **281**(25):17134-17139.

108. Krupovic M, Koonin EV: **Multiple origins of viral capsid proteins from cellular ancestors**. *Proc Natl Acad Sci U S A* 2017, **114**(12):E2401-E2410.

109. Zirkel F, Kurth A, Quan PL, Briese T, Ellerbrok H, Pauli G, Leendertz FH, Lipkin WI, Ziebuhr J, Drosten C *et al*: **An insect nidovirus emerging from a primary tropical rainforest**. *MBio* 2011, **2**(3):e00077-00011.

110. Thomas G: **Furin at the cutting edge: from protein traffic to embryogenesis and disease**. *Nat Rev Mol Cell Biol* 2002, **3**(10):753-766.

111. Hijikata M, Kato N, Ootsuyama Y, Nakagawa M, Shimotohno K: **Gene mapping of the putative structural region of the hepatitis C virus genome by in vitro processing analysis**. *Proc Natl Acad Sci U S A* 1991, **88**(13):5547-5551.

112. de Haan CA, Rottier PJ: **Molecular interactions in the assembly of coronaviruses**. *Adv Virus Res* 2005, **64**:165-230.

113. Snijder EJ, Den Boon JA, Spaan WJ, Weiss M, Horzinek MC: **Primary structure and post-translational processing of the Berne virus peplomer protein**. *Virology* 1990, **178**(2):355-363.

114. Jitrapakdee S, Unajak S, Sittidilokratna N, Hodgson RA, Cowley JA, Walker PJ, Panyim S, Boonsaeng V: **Identification and analysis of gp116 and gp64 structural glycoproteins of yellow head nidovirus of Penaeus monodon shrimp**. *J Gen Virol* 2003, **84**(Pt 4):863-873.

115. Méndez EA, Arias CF: **Astroviruses**. In: *Fields Virology.* Edited by Knipe DM, Howley PM, Cohen JI, Griffin DE, Lamb RA, Martin MA, Racaniello VR, Roizman B, vol. 1, 6 edn. Philadelphia, PA: Lippincott Williams & Wilkins; 2013.

116. Zirkel F, Roth H, Kurth A, Drosten C, Ziebuhr J, Junglen S: **Identification and characterization of genetically divergent members of the newly established family Mesoniviridae**. *J Virol* 2013, **87**(11):6346-6358.

117. Pagel M, Meade A, Barker D: **Bayesian estimation of ancestral character states on phylogenies**. *Syst Biol* 2004, **53**(5):673-684.

118. Plant EP, Perez-Alvarado GC, Jacobs JL, Mukhopadhyay B, Hennig M, Dinman JD: **A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal**. *PLoS Biol* 2005, **3**(6):e172.

119. Brandl H, Moon H, Vila-Farre M, Liu SY, Henry I, Rink JC: **PlanMine--a mineable resource of planarian biology and biodiversity**. *Nucleic Acids Res* 2016, **44**(D1):D764-773.

120. Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, Qin XC, Xu J, Holmes EC, Zhang YZ: **Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses**. *Elife* 2015, **4**.

121. Lachnit T, Thomas T, Steinberg P: **Expanding our Understanding of the Seaweed Holobiont: RNA Viruses of the Red Alga Delisea pulchra**. *Front Microbiol* 2015, **6**:1489.

122. Webster CL, Longdon B, Lewis SH, Obbard DJ: **Twenty-Five New Viruses Associated with the Drosophilidae (Diptera)**. *Evol Bioinform Online* 2016, **12**(Suppl 2):13-25.

123. Marzano SY, Nelson BD, Ajayi-Oyetunde O, Bradley CA, Hughes TJ, Hartman GL, Eastburn DM, Domier LL: **Identification of Diverse Mycoviruses through Metatranscriptomics Characterization of the Viromes of Five Major Fungal Plant Pathogens**. *J Virol* 2016, **90**(15):6846-6863.

124. Fauver JR, Grubaugh ND, Krajacich BJ, Weger-Lucarelli J, Lakin SM, Fakoli LS, 3rd, Bolay FK, Diclaro JW, 2nd, Dabire KR, Foy BD *et al*: **West African Anopheles gambiae mosquitoes harbor a taxonomically diverse virome including new insect-specific flaviviruses, mononegaviruses, and totiviruses**. *Virology* 2016, **498**:288-299.

125. Shi M, Neville P, Nicholson J, Eden JS, Imrie A, Holmes EC: **High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia**. *J Virol* 2017, **91**(17).

126. Remnant EJ, Shi M, Buchmann G, Blacquiere T, Holmes EC, Beekman M, Ashe A: **A Diverse Range of Novel RNA Viruses in Geographically Distinct Honey Bee Populations**. *J Virol* 2017, **91**(16).

127. Dolja VV, Koonin EV: **Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer**. *Virus Res* 2018, **244**:36-52.

128. Zhang YZ, Shi M, Holmes EC: **Using Metagenomics to Characterize an Expanding Virosphere**. *Cell* 2018, **172**(6):1168-1172.

129. Tokarz R, Sameroff S, Hesse RA, Hause BM, Desai A, Jain K, Lipkin WI: **Discovery of a novel nidovirus in cattle with respiratory disease**. *J Gen Virol* 2015, **96**(8):2188-2193.

130. Gorbalenya AE, Brinton MA, Cowley J, de Groot R, Gulyaeva A, Lauber C, Neuman B, Ziebuhr J: **ICTV taxonomic proposal 2017.015S Reorganization and expansion of the order Nidovirales at the family and sub-order ranks**. 2017.

131. Gorbalenya AE, Pringle FM, Zeddam JL, Luke BT, Cameron CE, Kalmakoff J, Hanzlik TN, Gordon KH, Ward VK: **The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage**. *J Mol Biol* 2002, **324**(1):47-62.

132. Olendraite I, Lukhovitskaya NI, Porter SD, Valles SM, Firth AE: **Polycipiviridae: a proposed new family of polycistronic picorna-like RNA viruses**. *J Gen Virol* 2017, **98**(9):2368-2378.

133. Le Gall O, Christian P, Fauquet CM, King AM, Knowles NJ, Nakashima N, Stanway G, Gorbalenya AE: **Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T = 3 virion architecture**. *Arch Virol* 2008, **153**(4):715-727.

134. Napthine S, Ling R, Finch LK, Jones JD, Bell S, Brierley I, Firth AE: **Protein-directed ribosomal frameshifting temporally regulates gene expression**. *Nat Commun* 2017, **8**:15582.

135. Enjuanes L, Almazan F, Sola I, Zuniga S: **Biochemical aspects of coronavirus replication and virus-host interaction**. *Annu Rev Microbiol* 2006, **60**:211-230.

136. Belshaw R, Pybus OG, Rambaut A: **The evolution of genome compression and genomic novelty in RNA viruses**. *Genome Res* 2007, **17**(10):1496-1504.

137. Gorbalenya AE: **Big nidovirus genome. When count and order of domains matter**. *Adv Exp Med Biol* 2001, **494**:1-17.

138. Luhtala N, Parker R: **T2 Family ribonucleases: ancient enzymes with diverse roles**. *Trends Biochem Sci* 2010, **35**(5):253-259.

139. Krey T, Bontems F, Vonrhein C, Vaney MC, Bricogne G, Rumenapf T, Rey FA: **Crystal structure of the pestivirus envelope glycoprotein E(rns) and mechanistic analysis of its ribonuclease activity**. *Structure* 2012, **20**(5):862-873.

140. Park B KY: **Immunosuppression induced by expression of a viral RNase enhances susceptibility of Plutella xylostella to microbial pesticides.** *Insect Science* 2012, **19**(1):47-54.

141. Wang Z, Nie Y, Wang P, Ding M, Deng H: **Characterization of classical swine fever virus entry by using pseudotyped viruses: E1 and E2 are sufficient to mediate viral entry**. *Virology* 2004, **330**(1):332-341.

142. Ozhogina OA, Trexler M, Banyai L, Llinas M, Patthy L: **Origin of fibronectin type II (FN2) modules: structural analyses of distantly-related members of the kringle family idey the kringle domain of neurotrypsin as a potential link between FN2 domains and kringles**. *Protein Sci* 2001, **10**(10):2114-2122.

143. Chalmers IW, Hoffmann KF: **Platyhelminth Venom Allergen-Like (VAL) proteins: revealing structural diversity, class-specific features and biological associations across the phylum**. *Parasitology* 2012, **139**(10):1231-1245.

144. Napper CE, Drickamer K, Taylor ME: **Collagen binding by the mannose receptor mediated through the fibronectin type II domain**. *Biochem J* 2006, **395**(3):579-586.

145. Tam EM, Moore TR, Butler GS, Overall CM: **Characterization of the distinct collagen binding, helicase and cleavage mechanisms of matrix metalloproteinase 2 and 14 (gelatinase A and MT1-MMP): the differential roles of the MMP hemopexin c domains and the MMP-2 fibronectin type II modules in collagen triple helicase activities**. *J Biol Chem* 2004, **279**(41):43336-43344.

146. Rauceo JM, De Armond R, Otoo H, Kahn PC, Klotz SA, Gaur NK, Lipke PN: **Threonine-rich repeats increase fibronectin binding in the Candida albicans adhesin Als5p**. *Eukaryot Cell* 2006, **5**(10):1664-1673.

147. Hevia A, Martinez N, Ladero V, Alvarez MA, Margolles A, Sanchez B: **An extracellular Serine/Threonine-rich protein from Lactobacillus plantarum NCIMB 8826 is a novel aggregation-promoting factor with affinity to mucin**. *Appl Environ Microbiol* 2013, **79**(19):6059-6066.

148. Al-Khodor S, Price CT, Kalia A, Abu KY: **Functional diversity of ankyrin repeats in microbial proteins**. *Trends Microbiol* 2010, **18**(3):132-139.

149. Mosavi LK, Cammett TJ, Desrosiers DC, Peng ZY: **The ankyrin repeat as molecular architecture for protein recognition**. *Protein Sci* 2004, **13**(6):1435-1448.

150. Chen DY, Fabrizio JA, Wilkins SE, Dave KA, Gorman JJ, Gleadle JM, Fleming SB, Peet DJ, Mercer AA: **Ankyrin Repeat Proteins of Orf Virus Influence the Cellular Hypoxia Response Pathway**. *J Virol* 2017, **91**(1).

151. Rahman MM, McFadden G: **Modulation of NF-kappaB signalling by microbial pathogens**. *Nat Rev Microbiol* 2011, **9**(4):291-306.

152. Camus-Bouclainville C, Fiette L, Bouchiha S, Pignolet B, Counor D, Filipe C, Gelfi J, Messud-Petit F: **A virulence factor of myxoma virus colocalizes with NF-kappaB in the nucleus and interferes with inflammation**. *J Virol* 2004, **78**(5):2510-2516.

153. Gilmore TD, Wolenski FS: **NF-kappaB: where did it come from and why?** *Immunol Rev* 2012, **246**(1):14-35.

154. Falabella P, Varricchio P, Provost B, Espagne E, Ferrarese R, Grimaldi A, de EM, Fimiani G, Ursini MV, Malva C *et al*: **Characterization of the IkappaB-like gene family in polydnaviruses associated with wasps belonging to different Braconid subfamilies**. *J Gen Virol* 2007, **88**(Pt 1):92-104.

155. Tait SW, Reid EB, Greaves DR, Wileman TE, Powell PP: **Mechanism of inactivation of NF-kappa B by a viral homologue of I kappa b alpha. Signal-induced release of i kappa b alpha results in binding of the viral homologue to NF-kappa B**. *J Biol Chem* 2000, **275**(44):34656-34664.

156. Canton J, Fehr AR, Fernandez-Delgado R, Gutierrez-Alvarez FJ, Sanchez-Aparicio MT, Garcia-Sastre A, Perlman S, Enjuanes L, Sola I: **MERS-CoV 4b protein interferes with the NF-kappaB-dependent innate immune response during infection**. *PLoS Pathog* 2018, **14**(1):e1006838.

157. Shackelton LA, Holmes EC: **The evolution of large DNA viruses: combining genomic information of viruses and their hosts**. *Trends Microbiol* 2004, **12**(10):458-465.

158. Smith EC, Sexton NR, Denison MR: **Thinking Outside the Triangle: Replication Fidelity of the Largest RNA Viruses**. *Annu Rev Virol* 2014, **1**(1):111-132.

159. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R: **Viral mutation rates**. *J Virol* 2010, **84**(19):9733-9748.

160. Sniegowski PD, Gerrish PJ, Johnson T, Shaver A: **The evolution of mutation rates: separating causes from consequences**. *Bioessays* 2000, **22**(12):1057-1066.

161. Lynch M: **Evolution of the mutation rate**. *Trends Genet* 2010, **26**(8):345-352.

162. Beese LS, Steitz TA: **Structural basis for the 3'-5' exonuclease activity of Escherichia coli DNA polymerase I: a two metal ion mechanism**. *EMBO J* 1991, **10**(1):25-33.

163. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

164. Wang Y, Stary JM, Wilhelm JE, Newmark PA: **A functional genomic screen in planarians identifies novel regulators of germ cell development**. *Genes Dev* 2010, **24**(18):2081-2092.

165. Silvester N, Alako B, Amid C, Cerdeno-Tarraga A, Clarke L, Cleland I, Harrison PW, Jayathilaka S, Kay S, Keane T *et al*: **The European Nucleotide Archive in 2017**. *Nucleic Acids Res* 2017.

166. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.

167. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

168. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data**. *Bioinformatics* 2011, **27**(21):2987-2993.

169. King RS, Newmark PA: **In situ hybridization protocol for enhanced detection of gene expression in the planarian Schmidtea mediterranea**. *BMC Dev Biol* 2013, **13**:8.

170. Brubacher JL, Vieira AP, Newmark PA: **Preparation of the planarian Schmidtea mediterranea for high-resolution histology and transmission electron microscopy**. *Nat Protoc* 2014, **9**(3):661-673.

171. Venable JH, Coggeshall R: **A Simplified Lead Citrate Stain for Use in Electron Microscopy**. *J Cell Biol* 1965, **25**:407-408.

172. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **Data growth and its impact on the SCOP database: new developments**. *Nucleic Acids Res* 2008, **36**(Database issue):D419-425.

173. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

174. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic Acids Res* 2014, **42**(Database issue):D222-D230.

175. UniProt Consortium: **UniProt: a hub for protein information**. *Nucleic Acids Res* 2015, **43**(Database issue):D204-D212.

176. Lauber C, Gorbalenya AE: **Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses**. *J Virol* 2012, **86**(7):3890-3904.

177. Adams MJ, Lefkowitz EJ, King AM, Harrach B, Harrison RL, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Mushegian AR *et al*: **Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2016)**. *Arch Virol* 2016, **161**:2921-2949.

178. Brister JR, Ako-Adjei D, Bao Y, Blinkova O: **NCBI viral genomes resource**. *Nucleic Acids Res* 2015, **43**(Database issue):D571-577.

179. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers EW: **GenBank**. *Nucleic Acids Res* 2017, **45**(D1):D37–D42.

180. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Res* 2016, **44**(D1):D733-745.

181. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction**. *Nucleic Acids Res* 2003, **31**(13):3406-3415.

182. Janssen S, Giegerich R: **The RNA shapes studio**. *Bioinformatics* 2015, **31**(3):423-425.

183. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB: **Protein disorder prediction: implications for structural proteomics**. *Structure* 2003, **11**(11):1453-1459.

184. Drozdetskiy A, Cole C, Procter J, Barton GJ: **JPred4: a protein secondary structure prediction server**. *Nucleic Acids Res* 2015, **43**(W1):W389-394.

185. Petersen TN, Brunak S, von HG, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat Methods* 2011, **8**(10):785-786.

186. Duckert P, Brunak S, Blom N: **Prediction of proprotein convertase cleavage sites**. *Protein Eng Des Sel* 2004, **17**(1):107-112.

187. Gorbalenya AE, Lieutaud P, Harris MR, Coutard B, Canard B, Kleywegt GJ, Kravchenko AA, Samborskiy DV, Sidorov IA, Leontovich AM *et al*: **Practical application of bioinformatics by the multidisciplinary VIZIER consortium**. *Antiviral Res* 2010, **87**(2):95-110.

188. Eddy SR: **A new generation of homology search tools based on probabilistic inference**. *Genome Inform* 2009, **23**(1):205-211.

189. Söding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**(7):951-960.

190. Drummond AJ, Suchard MA, Xie D, Rambaut A: **Bayesian phylogenetics with BEAUti and the BEAST 1.7**. *Mol Biol Evol* 2012, **29**(8):1969-1973.

191. Darriba D, Taboada GL, Doallo R, Posada D: **ProtTest 3: fast selection of best-fit models of protein evolution**. *Bioinformatics* 2011, **27**(8):1164-1165.

192. Kass RE, Raftery AE: **Bayes Factors**. *J Am Stat Assoc* 1995, **90**(430):773-795.

193. 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**(13):3320-3323.

194. Heled J, Bouckaert RR: **Looking for trees in the forest: summary tree from posterior samples**. *BMC Evol Biol* 2013, **13**:221.

195. R Core Team: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing; 2013.

196. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276-277.

197. Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL: **Domain enhanced lookup time accelerated BLAST**. *Biol Direct* 2012, **7**:12.

198. Gibney G, Baxevanis AD: **Searching NCBI databases using Entrez**. *Curr Protoc Bioinformatics* 2011, **Chapter 1**:Unit 1 3.

199. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

200. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.

201. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Mol Biol Evol* 2013, **30**(4):772-780.

202. Touw WG, Baakman C, Black J, te Beek TA, Krieger E, Joosten RP, Vriend G: **A series of PDB-related databanks for everyday needs**. *Nucleic Acids Res* 2015, **43**(Database issue):D364-368.

203. Hekkelman ML, Vriend G: **MRS: a fast and compact retrieval system for biological data**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W766-769.

204. Thorn A, Steinfeld R, Ziegenbein M, Grapp M, Hsiao HH, Urlaub H, Sheldrick GM, Gartner J, Kratzner R: **Structure and activity of the only human RNase T2**. *Nucleic Acids Res* 2012, **40**(17):8733-8742.

205. Briknarova K, Gehrmann M, Banyai L, Tordai H, Patthy L, Llinas M: **Gelatin-binding region of human matrix metalloproteinase-2: solution structure, dynamics, and function of the COL-23 two-domain construct**. *J Biol Chem* 2001, **276**(29):27613-27621.

206. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M *et al*: **InterPro in 2017-beyond protein family and domain annotations**. *Nucleic Acids Res* 2017, **45**(D1):D190-D199.

207. Adams MJ, Carstens EB: **Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2012)**. *Arch Virol* 2012, **157**(7):1411-1422.

208. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language**. *Bioinformatics* 2004, **20**(2):289-290.

209. Risler JL, Delorme MO, Delacroix H, Henaut A: **Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix**. *J Mol Biol* 1988, **204**(4):1019-1029.

210. Pedersen KJ: **Slime-secreting cells of planarians**. *Ann N Y Acad Sci* 1963, **106**:424-443.

CHAPTER 5

LAMPA, LArge Multidomain Protein Annotator, and its application to RNA virus polyproteins

Anastasia A. Gulyaeva
Andrey I. Sigorskih[#]
Elena S. Ocheredko[#]
Dmitry V. Samborskiy
Alexander E. Gorbalenya

[#]equal contribution

Bioinformatics (2020)
DOI: 10.1093/bioinformatics/btaa065

ABSTRACT

Motivation: To facilitate accurate estimation of statistical significance of sequence similarity in profile-profile searches, queries should ideally correspond to protein domains. For multidomain proteins, using domains as queries depends on delineation of domain borders, which may be unknown. Thus, proteins are commonly used as queries that complicates establishing homology for similarities close to cut-off levels of statistical significance.

Results: In this report we describe an iterative approach, called LAMPA, LArge Multidomain Protein Annotator, that resolves the above conundrum by gradual expansion of hit coverage of multidomain proteins through re-evaluating statistical significance of hit similarity using ever smaller queries defined at each iteration. LAMPA employs TMHMM and HHsearch for recognition of transmembrane regions and homology, respectively. We used Pfam database for annotating 2985 multidomain proteins (polyproteins) composed of more than 1000 amino acid residues, which dominate proteomes of RNA viruses. Under strict cut-offs, LAMPA outperformed HHsearch-mediated runs using intact polyproteins as queries by three measures: number of and coverage by identified homologous regions, and number of hit Pfam profiles. Compared to HHsearch, LAMPA identified 507 extra homologous regions in 14.4% of polyproteins. This Pfam-based annotation of RNA virus polyproteins by LAMPA was also superior to RefSeq expert annotation by two measures, region number and annotated length, for 69.3% of RNA virus polyprotein entries. We rationalized the obtained results based on dependencies of HHsearch hit statistical significance for local alignment similarity score from lengths and diversities of query-target pairs in computational experiments.

Availability: LAMPA 1.0.0 R package is placed on GitHub (https://github.com/Gorbalenya-Lab/LAMPA).

1 INTRODUCTION

Due to high-throughput next-generation sequencing, genomics is outpacing functional and structural characterization of proteins [1]. This gap is especially pronounced and fast growing for viruses, whose discovery and characterization in diverse habitats has been driven by metagenomics over the last ten years [2, 3].

In genomics projects, conceptually translated open reading frames (ORFs) are functionally characterized by bioinformatics tools which use homology recognition for annotation. To improve accuracy of protein annotation, bioinformatics tools use iterative searches of databases of individual sequences (e.g. PSI-BLAST [4] vs GenBank [5]), search profile databases (e.g. HMMER [6] or HHsearch [7, 8] vs Pfam [9], or HHblits [8] vs Uniclust30 [10]), and may involve comparison of query and target secondary structure (e.g. HHsearch vs SCOP [11]). Annotation pipelines favor selectivity over sensitivity by imposing stringent cut-offs on similarity between query and database entries. Scores of similarity are interpreted in statistical frameworks using either expectation values (default cut-off E=0.001, BLAST, HMMER, HHsearch) or homology Probability (default cut-off P=95%, HHsearch).

To recognize distant homologs, popular HHsearch was fine-tuned based on a subset of SCOP 1.63 database with less than 20% pairwise sequence identity of structural domains [7], where mean sequence length is equal 178 aa [11] (Fig. 1), typical of functional and structural domain [12]. Its hit statistical significance increases with score of similarity between query and target, and it depends on sizes and diversities of query and target [13]. Specifically, large size increases likelihood of a hit score emerging by chance, while the opposite is true for small size. Notwithstanding HHsearch training on protein domains, it has been routinely used in analysis of proteins of unknown domain organization. For a single-domain protein, statistical significance of hit similarity must be applicable to its domain, since sizes of both are similar. On the other hand, for multidomain queries, statistical support of a hit associated with individual domain may be underestimated due to inflated search space that encompasses other domains of the query protein [4, 14].

The query size issue could be of little practical consequence for proteins having closely related homologs in sequence databases. However for identification of distant relationships, accurate estimation of statistical significance could be impactful. The above problem may be particularly acute for RNA viruses [15], which typically encode large multidomain proteins (>1000 aa) [16]. (Hereafter and for sake of simplicity, we'll use polyprotein to refer to virus multidomain proteins). They are much larger than most proteins of cellular organisms, whose length distributions resemble lognormal, with a mean below 500 aa [17]. Human immunodeficiency virus, Ebola virus, severe acute

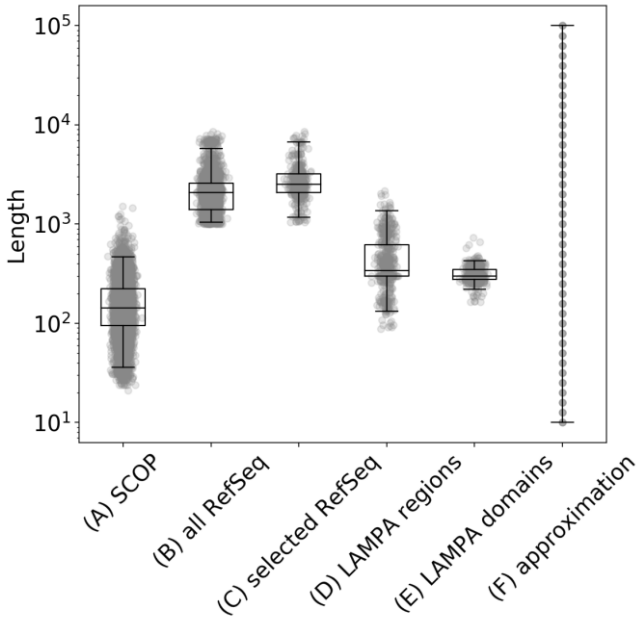


Figure 1 | Length distribution of proteins in datasets relevant to comparison of HHsearch and LAMPA. This plot depicts sizes of six protein datasets labelled from A to F and used or cited in this study. (**A**) 6271 SCOP domains used for HHsearch training (range: 21-1504 aa); (**B**) 2985 RefSeq virus polyproteins (range: 1001-8572 aa); (**C**) 431 RefSeq virus polyproteins which include 507 regions exclusively annotated by LAMPA (range: 1039-8572 aa); (**D**) 507 hit regions generated by LAMPA from 431 RefSeq polyproteins (range: 88-2172 aa); (**E**) 507 domains tentatively demarcated around LAMPA hits (range: 164-732 aa); (**F**) 41 designed sizes of each of three proteins, 123 in total, tested in computational experiments (range: 10 – 100,000 aa).

respiratory syndrome coronavirus, and poliovirus, and very many other eukaryotic viruses encode polyproteins [18, 19]. These polyproteins mediate replication/transcription and promote virus particle formation in either the synthesized form or after being proteolytically processed. Furthermore, the already known proteomes of RNA viruses are exceptionally diverse due to high mutation rate of RNA viruses [20], with many relationships in twilight and midnight zones of homology [21, 22].

In our recent HH-suit-mediated analysis of the largest known polyprotein of RNA virus (PSCNV, 13,556 aa) [23], we initially annotated only three regions by homology (polyprotein 7.1%). To check whether this result could be partially attributed to an underestimation of genuine statistical significance of the similarity between polyprotein domains and target protein profiles, we split the polyprotein using comparative genomics and, indeed, identified three other homologs with high confidence [23].

The above positive experience led us to formalize this approach in R package, called LAMPA, LArge Multidomain Protein Annotator, that we describe in this report. Also we

present proof-of-the principle for LAMPA in study of homology between RNA virus polyproteins and pfamA_31.0 database. It was further supported and expanded by evaluation of dependences of HHsearch statistics for fixed similarity score from lengths and diversities of query and target in computational experiments.

2 METHODS

2.1 Databases and virus protein dataset

We used pfamA_31.0 database [9], accompanying HH-suite [8], as *target* database to identify homology by profile searches and transfer annotation. We were interested in annotating virus proteins and selected a subset of NCBI Viral Genomes Resource database (RefSeq) [1] to serve as *queries* in homology searches and the source of expert annotation (Text S1.1). Only proteins of true RNA viruses that use RNA-dependent RNA polymerase (RdRp), positive and negative single-stranded RNA viruses, (+)ssRNA and (-)ssRNA, respectively, and double-stranded RNA viruses, dsRNA, were included in the query protein dataset (Fig. S1). Protein sequences were obtained from “translation” qualifiers of “CDS” features in RefSeq genome entries. The query database included all 2985 protein sequences of RNA virus genomes listed in “Viral genome browser” table on 2018.07.26 (Table S1), that were 1000 aa or longer (protein length ranged from 1001 to 8572 aa, median=2081 aa; Fig. 1). It was further grouped into 884 clusters using MMseqs2 [24], following the authors recommendations for multidomain proteins and defining sequence identity rate (--cluster-mode 1 --min-seq-id 0.3 --alignment-mode 3) and local alignment coverage (--cov-mode 0 -c 0.8) (see Text S1.2 and Table S1). Most of these proteins are encoded in a single ORF [25]. We parsed RefSeq entries corresponding to the analyzed proteins to extract region annotations from “Region” features [26]. Other annotation features, such as “CDS”, “Protein”, and “Site”, which were not taken into analysis, may overlap with the “Region” or include extra information. For further details about polyprotein query dataset see Text S1.1.

2.2 Comparative sequence analysis

Transmembrane (TM) helices in protein sequences were predicted by TMHMM 2.0c [27]. Secondary structures (SS) of query sequences, regardless of their length, were derived from the predictions made for the respective entire polyproteins by script addss.pl from HH-suite 3.0.0 (2015.03.15) [28], which used PSIPRED 3.5 tool [29]. Query profiles were built and compared to a database by programs HHmake and HHsearch from HH-suite 2.0.16, respectively [7]. In all analyses, parameters of HH-suite programs were left at default values, with the exception of HHmake parameter “-M first”, indicating that columns with residue in the first sequence of the FASTA file are considered match states,

and HHsearch three parameters: “-p 0”, allowing hits with Probability as low as zero; “-norealign”, blocking realignment of reported hits using maximum accuracy (MAC) algorithm; “-alt 10”, enabling reporting up to 10 significant alternative alignments between a query and a target profile [14] (Text S1.3). To identify statistically significant hits and homologous regions, HHsearch hits were subjected to post-processing under three cut-offs: Probability >95%, E-value <10, and hit length of >50 aa of the query sequence. Hits satisfying these thresholds and overlapping on query were combined into a cluster, extreme N- and C-terminal residues of which defined boundaries of region in the query that was homologous to target(s). Statistics of the top-scoring hit in the cluster defined the entire cluster, and name of the top-scoring target profile in the cluster annotated the query region. Unless stated otherwise, all reported analyses used the hits post-processing. Also we used HHblits v.3 [8] for analysis of selected polyproteins as detailed in Text S1.4. Analysis and visualization were performed using R 3.3.0 [30].

2.3 Statistics

P value of Wilcoxon signed rank test (P_W) was calculated using function “wilcox.test” from R package “stats”, with arguments “paired” and “alternative” set to values “TRUE” and “greater”, respectively [30].

2.4 Calculation of HHsearch P-value and Probability dependence from lengths and diversities of query-target pair for fixed hit score

HHsearch uses extreme value distribution (EVD) model for estimating hit's P-value, E-value, and Probability from query-target local alignment similarity score. P-value for a given score is defined as:

$$P_{value}(score) = 1 - \exp(-\exp(-\lambda * (score - \mu))) \qquad (1)$$

where λ and μ are the EVD parameters that optimally approximate the score distribution of false positives for a given pair of query and target profiles. E-value is defined as $P_{value}(score)*N_{DB}$, where N_{DB} is the number of searched target profiles in the database. For calculations of λ and μ, HHsearch uses ‘profile auto-calibration’ that employs two simple artificial neural networks [13]. This default procedure makes use of dependence of λ and μ on four characteristics: profile lengths and sequence diversities of both query and target. The parameters of the neural networks were derived by training on a set of profiles based on 6271 sequences of SCOP20 v1.73 database (minimal, median and maximal protein lengths = 21, 142 and 1504 aa, respectively; 5-to-95% range = 48-to-392 aa) (Fig. 1). Estimation for Probability of detecting homologous relationship (true positives) is also based on the EVD distribution but involves correction by the SS alignment score.

To learn how HHsearch performs on queries of our study with sizes close to or exceeding the largest protein in the training SCOP database, we conducted computational experiments using the HHsearch procedure that generates EVD parameters by adapting corresponding C++ source code into a Python Jupyter notebook (https://github.com/Gorbalenya-Lab/hh-suite-notebooks/tree/LAMPA). We approximated P-value and Probability of hit for fixed local alignment similarity score (including also SS alignment score for Probability) in relation to lengths and/or diversities of the corresponding query and target profiles, one of which may have been set to vary in large range of values (see Text S1.5).

3 RESULTS

3.1 LAMPA, iterative approach for homology recognition and functional annotation of multidomain proteins

LAMPA approach is aimed at improving detection of remote homology in large multidomain proteins (queries). Its multistage iterative procedure includes prediction of TM regions in query by TMHMM at the pre-iteration stage #0 and comparisons of query and its regions with HH-suite profile database(s) (targets) using HHsearch for iterations at stages #1-#3 (Fig. 2). As query, intact protein is used for stages #0 and #1, and various protein regions are used for stages #2 and #3. Iteration is a single execution of a procedure involving protein regions demarcation and submission of regions to HHsearch-mediated homology searches to identify statistically significant hits (values of post-processing cut-offs, specified in 2.2, are default). The approach stages are detailed below:

Stage #0. *Detection of TM regions in original query.* TM region (domain) may include either single or few helices predicted by TMHMM. By default, more than one helix is included in a region if each helix is separated from its neighbor by less than 100 aa. Region boundaries are defined by either helix boundaries (single-helix region) or opposite boundaries of two respective terminal helices (multiple-helix region). TM regions are used to split original query into smaller regions (see stage #2).

Stage #1. *Detection of homology regions in original query.* This is the first iteration of the annotation procedure that uses HHsearch-mediated homology search. Its input and output are the original query and hit annotated regions, respectively.

Stage #2. *Detection of homology regions in split query: query-protein-specific (QP-specific) iterations.* To initiate this stage, the procedure selects regions of the original query that are flanked by either of the following: N- or C-terminus of the original query, TM regions

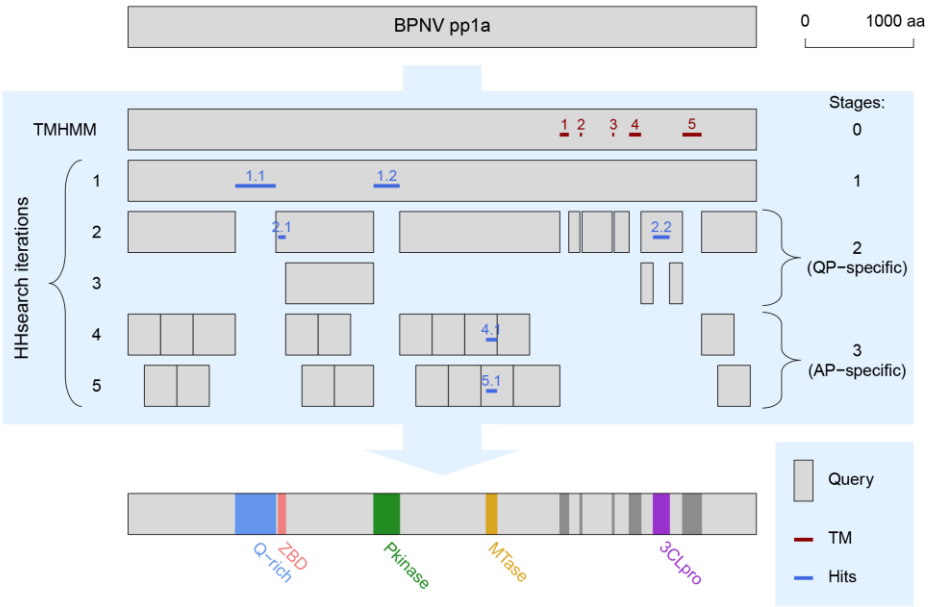


Figure 2 | LAMPA workflow and its application to RNA virus polyprotein. Presented is outline of the LAMPA approach (blue background) applied to polyprotein 1a (pp1a) of ball python nidovirus (BPNV). Grey bars, regions of BPNV pp1a that served as TMHMM or HHsearch queries. Iterations of the procedure and programs used are depicted on the left; stages are indicated on the right. Clusters of TM helices are depicted in dark red, clusters of hits – in dark blue. Hit double digits refer to iteration and hit position on polyprotein from left to right, respectively, except for hits at stage #0 which are labelled with the position only. Hits and annotations obtained on stage #1 represent output of conventional HHsearch. Q-rich, region rich in glutamine residue; ZBD, zinc-binding domain; Pkinase, protein kinase; MTase, methyltransferase; 3CLpro, 3C-like protease. For other details see text.

and hits clusters identified at the stages #0 and #1, respectively. These regions are used as input to HHsearch-mediated homology searches. Obtained hits are used for annotation and to demarcate flanking smaller non-annotated regions. The latter are used to initiate a new iteration in the manner described above. The iterations are repeated until no hits satisfying the cut-offs are identified.

Stage #3. *Detection of homology regions in split query: average-protein-size-specific (AP-specific) iterations.* Non-annotated regions after the stage #2 are split into two overlapping sets of 300 aa queries (default). The most C-terminal queries of both sets are extended to include the remaining part of the respective region, if the remaining part is shorter than 300/2=150 aa (default) and if the extended query does not cover the entire region. The default 300 aa size is close to that of an average protein (AP), hence respective iterations are called AP-specific. Queries are defined starting from either the N-terminus (first AP-specific iteration) or 300/2=150 aa (default) downstream the N-terminus (second

AP-specific iteration) of the non-annotated regions of stage #2. They are run independently. During this stage one and the same region of polyprotein may be found to have homolog and be annotated on both AP-specific iterations, since two sets overlap.

3.2 LAMPA implementation

The above approach was realized as LAMPA 1.0.0 R package (see also Text S1.6) that includes a single command 'LAMPA' with 15 arguments that allow user to specify a single protein query sequence, target database(s), information required to run HH-suit and TMHMM, and parameters of the LAMPA procedure, which are detailed in the package manual. LAMPA package employs two external R packages: seqinr [31] and IRanges [32]. Output of the command is a directory, name of which is identical to the name of the file with query sequence by default. This directory contains a plot (similar to Fig. 2) and two tables summarizing TM predictions and homology annotations made for the query sequence (overlapping with Table S2), as well as files with detailed information about hits constituting each cluster, and a folder with raw data (see package manual for details). Analysis of 2985 virus polyproteins against pfamA_31.0, detailed below, required 2000 min on 16 CPUs for LAMPA to complete (with 0.3 - 2.5 min per query, and approximately extra 1000 min compared to HHsearch). A separate script, not included in the LAMPA package, was used to automate analysis of multiple queries in this study.

3.3 Evaluation of LAMPA performance relative to HHsearch in analysis of RNA virus polyproteins

We evaluated LAMPA performance under default parameter values by querying pfamA_31.0 with 2985 RNA virus polyproteins (see 2.1; Fig. 1). This analysis documents dependence of HHsearch statistics on query size: split protein fragments or regions ('LAMPA') relative to intact proteins ('HHsearch'). Only the most N-terminal cluster of hits was considered in 26 cases of overlapping clusters from the LAMPA AP-specific stage. For annotation-related statistics, we did not consider TM domains (LAMPA stage #0, Fig. 2). The output of the LAMPA stage #1 represented also output of the HHsearch run on intact proteins.

Additionally, HHsearch was also used for further statistical analyses of the difference between outputs of two tools. For these analyses, HHsearch output was not subject to post-processing (see 2.2) that allowed to analyse hits with Probability ≤ 95%, E-value ≥ 10 and size on query ≤ 50 aa (see below). This use of HHsearch was outside the LAMPA framework and required matching of hits obtained by LAMPA and HHsearch for evaluation. We restricted this matching to the top-scoring hits of LAMPA hit clusters and HHsearch that overlapped on query and targeted the same Pfam profile.

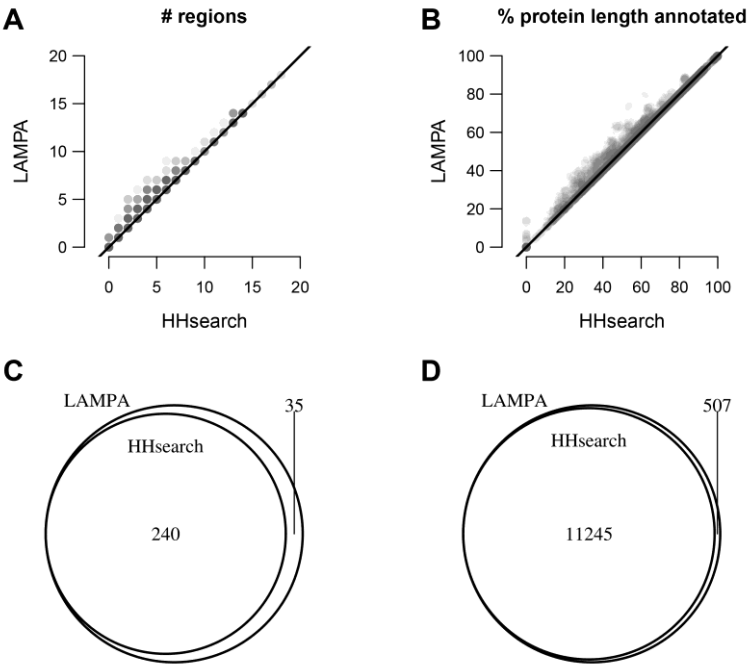


Figure 3 | Gain of homology recognition by LAMPA compared to HHsearch. Presented are four depictions of results of querying pfamA_31.0 with 2985 RNA virus proteins using LAMPA and HHsearch. (**A**) Number of regions (hit clusters) per query protein annotated by the two tools. Each protein is depicted by a transparent grey dot. Since multiple proteins may have the same or similar number of regions annotated by the two tools (X and Y dot coordinates), dots may overlap. Grey density is proportional to the number of overlapping dots. Black line, diagonal. (**B**) Share of protein length (%) annotated by the two tools. For other details see panel A. (**C**) Overlap between Pfam profiles that were linked to RNA virus proteins by the two tools. (**D**) Overlap between RNA virus polypro-tein regions annotated by the two tools.

3.4 LAMPA outperforms HHsearch in recognizing homology and facilitating annotation of RNA virus polyproteins

Neither LAMPA or HHsearch found homology between 163 proteins (5.5% of the dataset) and pfamA_31.0. For 2391 proteins (80.1%), LAMPA and HHsearch hit the same homologous regions, from 1 to 18. For 420 proteins (14.1%), LAMPA annotated from 1 to 3 extra regions on top of 1 to 15 found also by HHsearch (Fig. 3A). For each of the remaining 11 proteins (0.4%), a single region was hit by LAMPA only. Increase in number of annotated regions per protein by LAMPA was statistically significant (P_W=9.5e-86). By design of the procedure, HHsearch outperformed LAMPA for none of the polyproteins. For the three virus genome classes (2273 proteins in total), share of proteins, for which gain in number of annotated regions by LAMPA was observed, varied five-fold: (-)ssRNA viruses (3.1%), dsRNA viruses (10.2%), and (+)ssRNA viruses (15.9%). Among the 712 proteins with unknown virus genome class, LAMPA outperformed HHsearch for 22.2% of polyproteins.

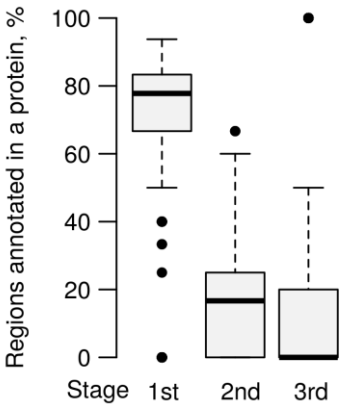


Figure 4 | Contribution of different stages of LAMPA procedure to protein annotation. Contribution of three LAMPA stages to annotation of 431 proteins, including regions exclusively annotated by LAMPA, was measured by percentage of regions annotated in each protein. Total number of regions annotated in each protein was considered 100%, regardless of their actual number and share in the protein. The box-plots, lower and upper limits of the box delimit the first (25%) and third (75%) quartiles, midline limit of the box – median, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box, data beyond that distance are represented by points.

Increase in the number of annotated regions (Fig. 3D) was accompanied by the increase in the polyprotein coverage by annotations, which ranged from 1.0% to 25.5% of polyprotein length (Fig. 3B; P_W=1.18e-72).

Also we compared lists of Pfam profiles hit by LAMPA and HHsearch, and were used for region annotation (Fig. 3C, Table S2). Both tools selected 173 profiles to annotate 5737 virus regions, and extra 67 profiles were used to annotate 5508 and 5947 virus regions by HHsearch and LAMPA, respectively. Also, additional 35 profiles were solely used by LAMPA to annotate 68 virus regions. Key enzymes of RNA viruses (RdRp, helicases, proteases, methyltransferases) dominated the shared part of the LAMPA and HHsearch Pfam profile lists (Fig. S2A). In contrast, the LAMPA-restricted profiles did not include RdRp but included types of enzymes and non-enzymatic proteins not found in the shared list, e.g. seven kinase profiles (Fig. S2B, Table S2). Many protein regions exclusively annotated by LAMPA were from most divergent RNA viruses [33].

3.5 Both QP- and AP-specific stages of LAMPA procedure contributed to gain of annotation

Gain of annotation by LAMPA compared to HHsearch is fully attributed to QP- and AP-specific stages. The gain was observed for 431 polyproteins, with the share of regions exclusively annotated by LAMPA varying from 6.2% to 100.0% (mean = 27.2%) of all recognised regions. Mean percentage of regions annotated in these proteins during the stages #1-#3 were 72.8%, 17.1% and 10.2%, respectively (Fig. 4). During QP- and AP-

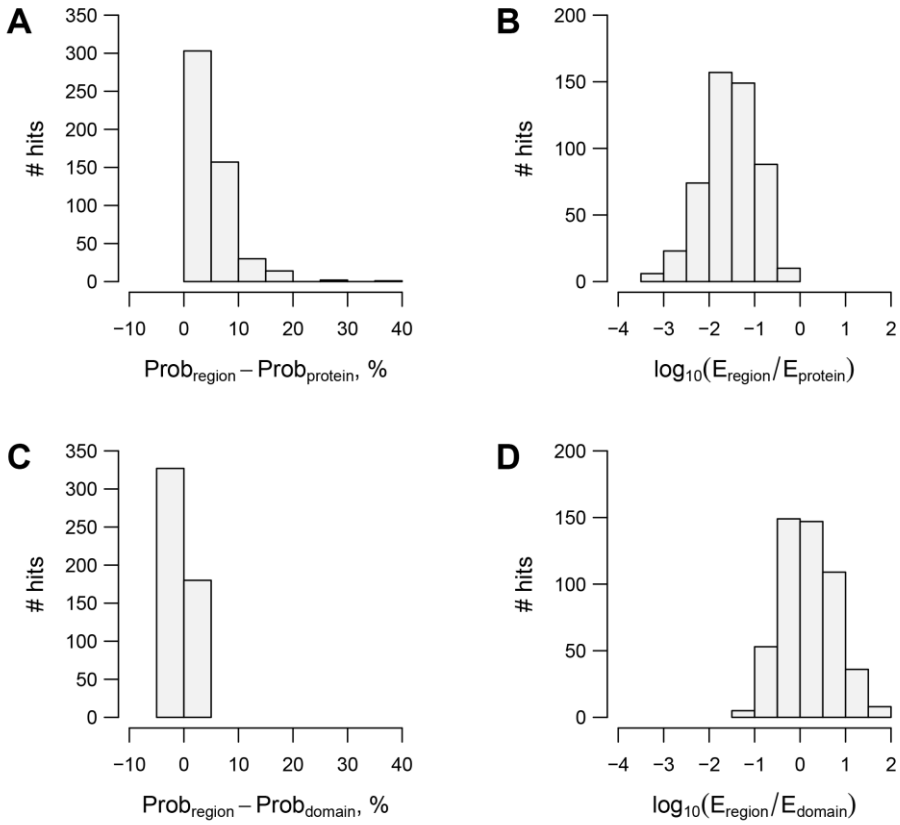


Figure 5 | Gain of hit statistical significance by LAMPA compared to HHsearch. LAMPA hits to region queries, obtained during the QP-specific and AP-specific stages of LAMPA procedure, are compared with matching HHsearch hits to polyprotein queries, in respect to hit Probability (**A**) and E-value (**B**); and with matching HHsearch hits to putative domain queries (operational definition, see text for details), in respect to hit Probability (**C**) and E-value (**D**). Analysed HHsearch hits were not subject to post-processing.

specific stages, regions were identified in 322 proteins (10.8% of the whole dataset) and 126 proteins (4.2%), respectively.

3.6 Increase of hit statistical significance by LAMPA com-pared to HHsearch is modest but common

LAMPA identified 507 clusters of hits on 431 proteins, HHsearch counterparts of which were removed by post-processing under the used thresholds (see 2.2; Fig. 3D). We used the top-scoring hits in these clusters to estimate the gain of statistical significance (Probability and E-value) by LAMPA compared to HHsearch and represent clusters in all analyses described below. We identified matching HHsearch hits for all 507 LAMPA hits (Table S2), with 437 hits (86.2%) having identical coordinates on query. In each pair of hits,

LAMPA hit was characterised by higher Probability and lower E-value (Fig. 5A and 5B). Probability increase by LAMPA compared to HHsearch was in the range from 0.5% to 37.6%, with mean 5.3% (Fig. 5A). Decimal logarithm of LAMPA to HHsearch E-values ratio ranged from -3.4 to -0.2 with mean -1.5 (Fig. 5B). Positive correlation between Probability and −logE-value was accompanied by E-value variation around two orders of magnitude for most Probabilities before and after they were elevated above the cut-off by LAMPA (Fig. S3). Likewise, for E-values around 10^{-1}, Probability varied approximately ±5%, illustrating that choice of statistic in addition to significance cut-off may affect output.

3.7 LAMPA-demarcated regions may approximate authentic domains for purpose of homology detection

The LAMPA region queries may still be (much) larger than the actual domains, natural borders of which remain unknown. Because of this uncertainty, we reasoned that the gain of statistical significance by LAMPA compared to HHsearch might provide only a lower estimate for the actual difference between Probabilities and E-values of the respective hits obtained for the polyprotein and expected for its domains. To improve understanding about how close the obtained LAMPA Probabilities and E-values for protein regions may be to those of the actual domains, we adopted an operational definition of polyprotein domain in relation to homology hit and used it to approximate borders of the actual domains; in total 507 hits on 431 polyproteins (see above) were considered for this purpose. Operational domain was demarcated as LAMPA hit that was extended by 100 aa to the N- and C-terminus; if distance to the polyprotein terminus was less than 100 aa, extension was adjusted accordingly (which was used in 48 of 507 cases). The demarcated domain sizes ranged from 164 to 732 aa (mean=315 aa) that was close to dominant domain size in public databases and narrower compared to the range of 88 to 2172 aa (mean=479 aa) of region queries that produced the original LAMPA hits (Fig. 1). For each of 507 hits, we then compared Probability and E-value values, assigned by LAMPA, to those obtained by HHsearch for a matching hit in a separate analysis that used demarcated domains as queries and involved no hits post-processing (see 2.2; Table S2).

We obtained data for all 507 hits, with 457 hits (90.1 %) having identical coordinates on query in LAMPA and HHsearch analyses. The difference between the two Probability values ranged from -1.8% to 4.6% with mean and median close to zero (both were equal -0.2%); absolute value of the difference didn't exceed 2% in 99.8% of cases (Fig. 5C). Decimal logarithm of the E-values ratio ranged from -1.3 to 1.8, mean 0.2 (Fig. 5D). These differences were evenly distributed and much smaller than those observed in comparison of LAMPA hits to region queries and HHsearch hits to polyprotein queries (Fig. 5A and 5B). Based on these results we concluded that sizes of queries used by LAMPA during iterative

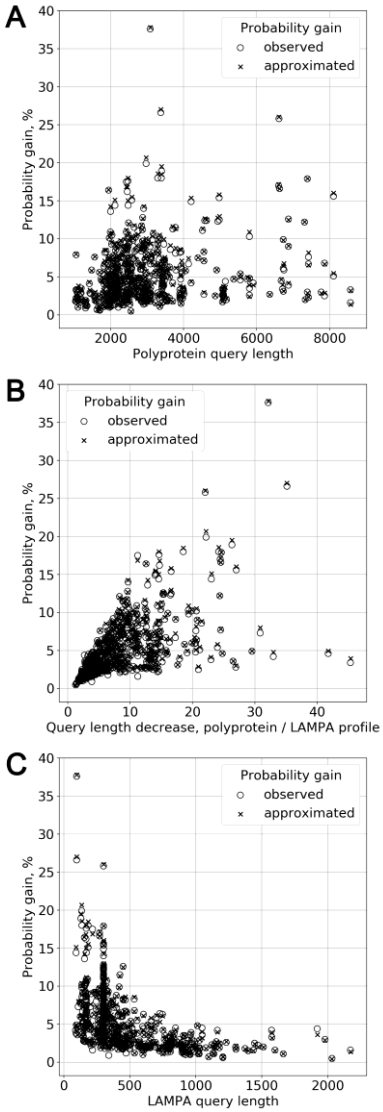


Figure 6 | Relationship between Probability gain by LAMPA and query lengths. Difference between Probabilities of hit to region query (LAMPA stages #2 or #3) vs polyprotein query (HHsearch without hits post-processing) (empty circle), is compared with difference between the respective approximated Probabilities for the matching hit in computational experiments (cross) at the Y axis, for 507 hits in total. These values are plotted against values of three characteristics of respective queries at the X axis: (**A**) polyprotein length (stage #1), (**B**) ratio of polyprotein to query region length (stage #1 vs stage #2/3), and (**C**) query region length (stage #2/3).

stages may be close to those of the respective authentic domains for the purpose of statistical evaluation of homology and annotation transfer under the employed cut-off.

3.8 Increase of statistical significance of hits by LAMPA compared to HHsearch is proportional to respective decrease of query length

We then asked how LAMPA-based increase of statistical significance in 507 hits of 431 proteins in 504 pairs of polyprotein and Pfam profile depended on lengths of polyprotein (original query, varied between 1039 and 8572 aa) and its fragments (queries varied between 88 to 2172 aa at LAMPA stages #2 and #3) (Fig. 1). We observed steady but highly uneven increase of Probability gain for polyproteins in the size range between 1001 and approximately 3000 aa which then levelled (Fig. 6A). That positive dependence was stronger and more common when Probability gain was plotted against relative length decrease in queries of LAMPA compared to HHsearch, which varied in the range from 1x to 45.3x, with 68.2% of the decreases of query length being in the 1-10x range (Fig. 6B). Accordingly, Probability gain fall steeply with increase of the LAMPA query length up to 2172 aa; it was below 10% and 5% for LAMPA queries including more than 448 aa and 747 aa, respectively (Fig. 6C).

3.9 Estimation of hits Probability by LAMPA may be approximated in computational experiment

Non-uniform dependence of Probability gain from query length (Fig. 6A, C) implied other characteristics be involved. Indeed, besides query length, target length and diversities of query and target are used by HHsearch for the calculation of λ and μ that affect hit score P-value (see 2.4). Accordingly, we analysed the relationship between estimates of hit statistical significance and possible lengths of the corresponding query and target profiles systematically using computational experiments. They used local alignment similarity score of HHsearch hit of *full-length* query-target pair for approximating hit Probability on queries of *other observed and computationally generated sizes*, assuming that hit score may not change with query size. This assumption proved to be accurate within a margin of error (see below).

We used the HHsearch neural networks to generate EVD parameters, followed by calculation of Probability, as well as P-value, of hit to polyprotein region from local alignment similarity score of this hit in every full-length query-target pair for which hit Probability gain was observed (in total 507 hits; Figs. 3D and 6; for details see https://github.com/Gorbalenya-Lab/hh-suite-notebooks/tree/LAMPA). First we noted good agreement between gains of Probabilities obtained in computational experiments and LAMPA runs (Fig. 6). They are within of +0.7%/-0.4% deviation of Probability gain estimation by LAMPA for the 95 percentile of hit scores in the dataset (Fig. S4A). The modest difference between the two values is explained by respective deviation of the underlying similarity score of the pairwise HHsearch hit alignment for polyprotein, which was fixed in computational experiments, from region-specific score that is calculated for

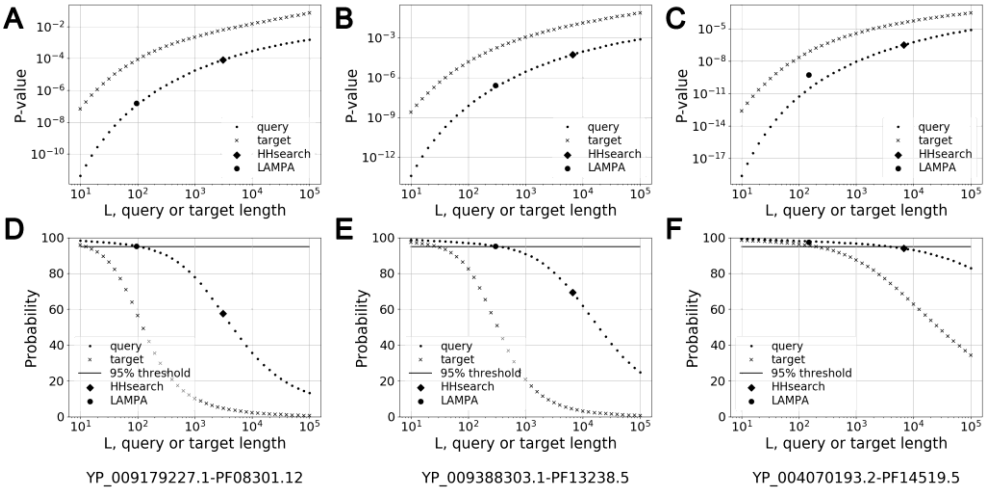


Figure 7 | Relationship between hit statistical significance and profile lengths in computational experiments. HHsearch hit P-value (**A-C**) and Probability (**D-F**) were estimated for 41 designed lengths of *query* or *target*, each of which was equidistant from its immediate neighbour on base 10 logarithmic scale (see Text S1). The 41 pairs of values were plotted to reveal relationship between two characteristics. These plots used hit score values of three query-target pairs, which are specified at the bottom of the figure and whose respective hit statistics values at the #1 stage (HHsearch), and #2 or #3 stages (LAMPA) are also depicted.

actual query and target profiles by LAMPA. Thus, by default, the same hit alignment involving polyprotein and its part as queries might have slightly different scores and also coordinates, further contributing to difference between the respective Probabilities (and P-values, Fig. S4B) in computational experiments.

3.10 P-value and Probability of HHsearch hits depend non-linearly on the lengths and diversities of query and tar-get profiles in computational experiments

The increase of the hit Probability during QP- and AP-specific iterations (Fig. 6) is likely explained by the use of query length in the auto-calibration procedure of HHsearch (see 2.4). We then conducted four computational experiments for three selected query-target pairs (Text S1.5) that were characterized by the largest Probability gain of LAMPA hit at stages #2 (37.6%) and #3 (25.8%), respectively, and associated with the largest decrease of query size (47 fold) (Fig. 7, Fig. S5 and Table S3). They also represent considerable ranges of hit scores (40.2, 41.1, and 67.2 for three pairs) and target diversities (6.7, 11.5, and 7.7). Forty one computationally designed lengths of each of three queries were tested (Fig. 1; Text S1.5).

In the three query-target pairs, both P-value and Probability showed strong non-linear dependence on designed sizes of query and target (Fig. 7) (hereafter we use “designed” to distinguish computational experiment from LAMPA). Specifically, P-value changed steeply,

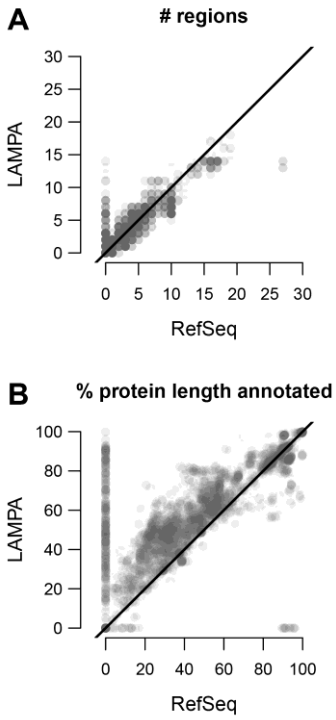


Figure 8 | Summary statistic of annotation coverage by LAMPA and RefSeq experts. Comparison of the number of regions per protein (**A**) or percentage of protein length (protein coverage) (**B**) annotated by LAMPA (stages #1-3) and RefSeq experts, based on analysis 2985 RNA virus proteins. Each protein is represented by a transparent grey dot; dot density is proportional to the number of proteins with identical characteristics. Black line, diagonal.

with curves of designed queries and targets running in parallel relative to each other (Fig. 7A-C). In the designed length range from 100 to 10000 aa, which encompasses most queries and targets of this study, P-value increased by approximately four orders of magnitude for queries of three pairs. This increase was limited to two orders of magnitude for the three selected queries illustrating LAMPA gain versus HHsearch. In contrast, dependence of Probability on length of designed queries and targets followed inverted logistic curve and differed between target and query as well as between the three pairs (Fig. 7D-F). Dependence of Probability on designed query size was most no-ticeable only below the 95% threshold, where it followed growth phase of logistic. The selected LAMPA and HHsearch queries were at different places of this growth phase in two query-target pairs (Fig. 7D,E) and outside the growth phase in third pair (Fig. 7F) which explained different Probability gains of LAMPA hit in these pairs. Hit score and target diversity contributed to variable Probability gain in three pairs (Text S1.5).

3.11 LAMPA can significantly expand RefSeq expert annotation of RNA virus polyproteins

Finally, we compared annotations of the RNA virus polyproteins by LAMPA and HHsearch versus RefSeq experts (Fig. 8, Fig. S6). Concerning the *number* of annotated regions per polyprotein, LAMPA and HHsearch were as good as RefSeq for 38.8 and 41.4% of polyproteins, respectively, while RefSeq expert or LAMPA/HHsearch outperformed the other for 23.3/27.0% and 37.9/31.6% of polyproteins, respectively (Fig. 8A, Fig. S6A). Notably, LAMPA and HHsearch annotated regions in 298 and 291 out of 426 polyproteins with no RefSeq annotation and increased the number of annotated region(s) for further 833 and 652 polyproteins. Increase in the number of annotated regions per protein by LAMPA but not HHsearch was statistically significant (P_W=3.11e-08 and 0.752, respectively). LAMPA and HHsearch annotations covered larger share of polyprotein (mean region length was 312, 321 and 265 aa for LAMPA, HHsearch and RefSeq annotation, respectively). This coverage increase was observed for 78.7 and 77.5% proteins, respectively, (Fig. 8B, Fig. S6B) and was statistically significant (P_W=1.07e-291 and 3.81e-273). We note that the above numbers apply to annotation in the “Region” fields of RefSeq entries. Other fields may record non-redundant annotation which is particularly likely for RefSeq entries with zero regions annotated in the “Region” field. These entries are in minority in the dataset. In summary, LAMPA expands further HHsearch annotation that may already improve RefSeq annotation of RNA virus polyproteins.

4 DISCUSSION

In this report we present an iterative LAMPA pipeline for advanced homology detection in large multidomain proteins and proof-of-the-principle for LAMPA in its application to RNA virus polyproteins. Statistical apparatus of HHsearch, used in LAMPA, was trained on a dataset of structurally defined domains with the median size of 142 aa to ascertain high sensitivity and selectivity, although HHsearch is used for annotation of proteins, regardless of their domain composition and size. This expanded application of HHsearch is due to two factors: 1) in contrast to sequence diversity of query (profile) (see HHblits), domain composition of query received relatively little attention in relation to HHsearch sensitivity; 2) considerable complexity and uncertainty of domain delineation in protein sequences. We have addressed both aspects in this study and offer a practical solution to the detection of distant homology in multidomain proteins using conventional profile-based tools in the LAMPA pipeline, which could be particularly useful in the on-going exploration of the Virosphere [2, 3, 23].

Length along with diversity are the two characteristics of query and target that determine hits Probability and P-value in HHsearch profiles' auto-calibration procedure [13]. We employed this procedure in computational experiments of high accuracy to plot the dependence of hits Probability and P-value from designed query/target lengths of several query-target pairs over a large size range that was beyond those used for tuning the auto-calibration procedure (12 to 1504 aa) and this study (1001 to 8572 aa) (Fig. 1). The produced plots revealed constrained statistic-specific shape of considerable variation for the two statistics characterizing a hit score in relation to query size (Fig. 7). Due to training of the auto-calibration procedure on the *domain* dataset, this variation informs about hit score statistics in application to *single-domain* proteins. When applied to *multidomain* proteins, like those used in this study, it illustrates how statistical significance of hit scores may be underappreciated depending on difference of sizes of the intact protein and its domains. This underappreciation is realized regardless of multidomain protein size, although it may be consistently considerable only for large proteins.

In line with the formula 1 (see 2.4), the computational experiments revealed also complex dependencies of statistical significance of HHsearch hits on designed target length and profile diversities of query and target (Fig. 7, Fig. S5). These dependencies explained variable gains of hit statistical significance by LAMPA compared to HHsearch in different query-target pairs. They also provide theoretical foundation for further efforts of improving the homology recognition by LAMPA through enriching queries using HHblits and targeting several databases, as is discussed below.

For queries including single domain or larger, false positive rate of LAMPA may not be different from that of HHsearch [7, 8], which is used for calculation of hit statistical significance. Our results were obtained with Probability cut-off of 95%, which was chosen to ascertain homology detection and suppress false positives [14]. The user may use E-value instead of Probability or lower the cut-off that will trade confidence in homology detection for increasing polyprotein coverage. We expect LAMPA to outperform HHsearch at these lower cut-offs as well. Due to logistic dependence between Probability and query length (Fig. 7D-F), Probability gains with under 95% cut-offs could be bigger than reported here.

We used TMHMM and HHsearch to functionally annotate polyproteins on structural grounds and by homology, respectively; they were used by LAMPA to delimit uncharacterized polyprotein regions that queried Pfam 31.0 further. (As discussed in Text S1.3, the use of HHsearch in the LAMPA framework was adjusted for analysis of RNA virus polyproteins). Once this iterative query-specific characterization at the QP-stage was exhausted, we used average protein domain size to delimit the remaining non-annotated regions during further database searches. This AP-stage has elements of arbitrariness

which were partially addressed *ad hoc* by using two alternative starting points for query delimitation.

This aspect and the entire pipeline may be advanced further. At the stage #0, other programs in addition to TMHMM may assist with functional annotation, e.g. mapping disordered regions, or regions anomalously enriched with certain amino acid residues, or cleavage sites for particular proteases like it was demonstrated in our recent study [23]. In that study, HHsearch was used to scan several databases, and this provision is also available in the LAMPA 1.0.0 package. Also, iterative profile programs, e.g. PSI-BLAST or HHblits, could be incorporated in the LAMPA to enrich query and improve homology recognition by targeting proteins that are not part of curated profile databases. These improvements could increase relative share of the QP-stage in homology detection and region annotation. In theory, the LAMPA may identify all domains at the #1 and QP-stage, with the AP-stage generating no hits, either due to the lack of queries or homology. Notwithstanding future advances, the current LAMPA version may already complement HHblits, the current top homology search tool. Indeed, under the 95% Probability cut-off HHblits failed to annotate 195 of 507 regions that LAMPA but not HHsearch annotated in 431 polyproteins of this study (Table S2, Text S1.4).

The reported gain of hit statistical significance by LAMPA compared to HHsearch was modest but sufficient to elevate many hits above the Probability 95% cut-off. It improved homology detection and hit coverage in 14.4% of polyproteins which were enriched with sequences that share not more than 30% identity with others in the dataset. Thus, gain of hit statistical significance by LAMPA compared to HHsearch could be larger for viruses that prototype genera or higher rank taxa rather than species dominating our dataset (see Text S1.2).

LAMPA annotation was most frequent for (+)ssRNA viruses, which correlates with their abundance and expanded diversity relative to dsRNA and (-)ssRNA viruses. Most newly detected homologs may already be known in other related viruses, which is evident from names and descriptions of hit Pfam profiles that often refer to viruses and their proteins (Table S2). However, they also include those not reported in literature, e.g. ZBD and MTase domains in pp1a (YP_009052476.1) of BPNV, python tobanivirus (Fig. 2; Table S2). The detection of the MTase domain, which is apparently conserved in the distantly related fish WBV (YP_803214.1) in this genome location, is particularly intriguing. These viruses and other nidoviruses with genomes > 20 kb are known to encode one or two MTases far downstream in the pp1b part of the pp1ab polyprotein [23, 34, 35] that were implicated in the 5’-end mRNA cap formation [36]. These and other functional assignments (Table S2) could be used to direct experimental research and in reconstruction of evolution of RNA viruses.

LAMPA facilitates homology detection and may be used to improve annotation coverage by other tools and experts in genomic projects, as well as in curated databases, including RefSeq. However, other factors besides detection of homology may affect quality of annotation [37, 38] and they were outside the scope of this study.

ACKNOWLEDGEMENTS

We thank Andrey M. Leontovich and Igor A. Sidorov for discussions and assistance.

FUNDING

This work has been supported by the EU Horizon2020 EVAg 653316 project and the LUMC MoBiLe program; AEG was a Leiden University Fund (LUF) Professor.

Conflict of Interest: none declared.

SUPPLEMENTARY INFORMATION

Text S1.1 Virus protein dataset

The RefSeq database was chosen to compile the query virus database for three reasons. First, it is one of the best representations of the known RNA virus genome diversity that is publicly available. Second, RefSeq maintains proper taxonomic representation of viruses that alleviates considerable biases of genome sequencing toward selected viruses of societal significance. Third, RefSeq curates annotation of genome records, which could be used as a standard to compare to [1].

Most viruses are represented by a single polyprotein in our query dataset, but large RNA viruses may encode several, either overlapping or not. Non-overlapping polyproteins are encoded in separate ORFs on single or multiple genome segments (see Table S1). In contrast, polyproteins of some viruses, notably those of nidoviruses and alphaviruses, are expressed from two ORFs using either ribosomal frame-shifting signal or read-through terminal codon [25]. Often, a RefSeq genome entry contains a “CDS” feature attributed to the combination of the two such ORFs, alongside a “CDS” feature attributed to the first ORF. A “CDS” feature attributed to the second ORF may also be included, even though it may not be expressed independently of the first ORF. These extra “CDS” features constitute a source of redundancy, as our query dataset was created by extracting protein

sequences ≥1000 aa from “translation” qualifiers of all “CDS” features of the selected RefSeq genome entries.

Proteins of (+)ssRNA viruses accounted for 47.1% of the query dataset, length of the proteins ranged from 1001 to 8572 aa (polyprotein of a flavi-like Gamboa mosquito virus [39]), median length was 2168 aa. Proteins of (-)ssRNA viruses accounted for 18.2% of the dataset, length of the proteins ranged from 1003 to 4403 aa (L protein of Shayang Spider Virus 1 from the order *Bunyavirales* [40]), median length was 2122 aa. Proteins of dsRNA viruses accounted for 10.9% of the dataset, length of the proteins ranged from 1002 to 7391 aa. Two dsRNA viruses with largest protein sizes, 6359 and 7391 aa, and possibly others with similar large sizes may in fact be (+)ssRNA viruses (polyproteins of Gentian Kobu-sho-associated virus [41, 42] and Ceratobasidium endornavirus D [43, 44]). Median length of the dsRNA virus proteins, included in the dataset, was 1274 aa. For the remaining 23.9% proteins of the dataset, genome type was not specified in the corresponding genome entries, while their lengths ranged from 1001 to 7421 aa, median=1963 aa.

We used RefSeq annotation of the virus sequences as a standard in our study. Although it is useful, the RefSeq remains a project in progress, and its annotation is subject to frequent update and revision. Much of its annotation is based on profile analysis involving Pfam, CDD or other databases. In this respect, our findings using strict significance cut-offs are equally reliable and can be considered true to the extent we could transfer Pfam profiles descriptions to the identified homologous regions of query proteins.

Text S1.2 Redundancy of the virus protein dataset in relation to comparison of LAMPA and HHsearch

Majority of the 2985 polyproteins of the query dataset are encoded by viruses that prototype virus species, which is a main criterion for their selection by RefSeq team to address redundancy problem and ensure their relevance for research and applications. However, known species are distributed highly unevenly among virus families that creates a bias. To evaluate how similar polyproteins of these species are in the protein distance space, we have clustered 2985 sequences using MMseqs2 software (0.8 coverage and 30% identity; single-linkage clustering mode) in analysis that delineated 884 clusters (with number of sequences per cluster varying from 1 to 124; average and median number of sequences per cluster 3.4 and 11, respectively) (Table S1). Inspection of virus taxonomy of these clusters indicates that they correspond loosely to taxa or a subset of taxa of classified viruses at genus/subfamily rank, depending on virus family. We found that 431 polyproteins, for which LAMPA outperformed HHsearch (Fig. 1, C dataset), represent a disproportionally large share of the total number of clusters (14.4% sequences found in 26.1% clusters) and were enriched with polyproteins representing less populated clusters

(231 clusters, average/median: 1.9/4.0 sequences per cluster). Thus, LAMPA outperformed HHsearch for annotation of a larger share of sequences in the clustered dataset than in the original dataset. This observation implies that the main observations and conclusions of our study were not undermined by selection of the RefSeq virus polyproteins as queries, without prior clustering.

Text S1.3 The use of HHsearch in the LAMPA framework for analysis of RNA virus polyproteins

The application of HHsearch to analysis of virus polyproteins in the LAMPA framework required non-default values for two parameters. The first parameter, "-norelaign", was used to switch off maximum accuracy (MAC) realignment algorithm, the postprocessing step at which the hit alignment is improved and the hit's span can be also adjusted, while hit scores (E-value/Probability) remain intact [14]. Although this postprocessing may improve alignment, we observed hit degradation and even its complete loss due to MAC use. A solution to this problem was suggested (https://github.com/soedinglab/hh-suite/issues/153). Second parameter, "-alt 10", increased the maximal number of reported alternative alignments between query and the same target profile to ten. The default maximum of two alternative alignments was found to be problematic, as RNA virus polyproteins may include more than two paralogs.

Also HHsearch may be prone to the overestimation of statistical significance of hits (false positives), if query size is at the low extreme of the size range of the training dataset. In the LAMPA framework, short queries smaller than domain may indeed be used at stage #2, if the query is flanked, from one or both sides, by hits that cover only a portion of the respective domain. These considerations prompted a limit on hit length (>50 aa by default) that also defined minimal length of query at stage #2.

Text S1.4 The use of HHblits to evaluate LAMPA gain of RNA virus polyproteins annotation

We used 431 polyproteins, which include 507 regions annotated by LAMPA but not HHsearch, as queries for HHblits to see whether this tool could annotate these regions. The polyprotein queries were initially enriched with homologs by running HHblits v.3 [8] against Uniclust30_2018_08 database [10] with 1, 2, or 3 search iterations and default other options (i.e. 0.001 for E-value cutoff for alignment extension and max. diversity threshold Neff=20, which stops further iterations). The enriched queries were then used for HHblits search in PfamA database with default options and only one search iteration (subsequent search iterations showed no significant improvement of hit score and coverage). Finally the obtained HHblits hits on PfamA profiles were mapped on corresponding LAMPA hits to 507 regions (Table S2). HHblits hit was considered as

matching, if it had Probability value above 95% and covered more than 70% of query region of respective LAMPA hit alignment. We observed that 195 of 507 regions were either not reported by HHblits at all (37) or were attributed with Probability value under the 95% cut-off (155) or had low query coverage (3).

Text S1.5 Dependence of P-value and Probability of fixed HHsearch hit score from size and diversity in query-target pairs of LAMPA analysis

We conducted several computational experiments using HHsearch neural networks. First, we assessed dependence of the Probability gain on query length, using different measures, in 507 query-target pairs from the hit list of LAMPA analysis of RNA virus polyproteins (Table S2). The obtained results were compared with those obtained in the LAMPA analysis and presented on Fig.6. Then, we selected three query-target pairs from the above list (Table S3) and conducted four in-depth computational experiments (for details see https://github.com/Gorbalenya-Lab/hh-suite-notebooks/tree/LAMPA). In first three experiments, diversities of query and target profiles were fixed at their respective real values (hereafter, the 'real' refers to characteristics of the full-length query or target profile). In the first experiment, we estimated P-value and Probability for computationally generated 41 different lengths of *query*, each of which was equidistant from its immediate neighbour on base 10 logarithmic scale in the query length space that ranged from 10^1 to 10^5 aa, with the target length fixed at its real value. In complementary second experiment, we estimated values of two statistics for the 41 length variants of the *target*, as specified above, and with the query length fixed at its real value. Results of these two experiments for three selected query-target pairs (Table S3) were combined separately for P-value and Probability, respectively (Fig. 7). In the third experiment, we estimated Probability for all combinations of the 41 length variants of the *query* and *target*. Results of this experiment were visualised using contour plots that depict change of Probability in the query length vs target length space (Fig. S5A-C). In the fourth experiment, lengths of query and target profiles were fixed at their respective real values. Then, we estimated Probability for all combinations of computationally generated 43 diversities of *query* and *target*, each of which was equidistant from its immediate neighbour on linear scale in the diversity space that ranged from 1 to 15. Results of this experiment were visualised using contour plots that depict change of Probability in the query diversity vs target diversity space (Fig. S5D-F).

Several factors contributed to variable Probability gain by LAMPA in the three query-target pairs (Table S3). In the YP_004070193.2-PF14519.5 pair, it was limited to 3.4% because of high HHsearch hit score = 67.2 that defined Probability = 94% which was close to the LAMPA 95% cut-off (Figs. 7F and S5C). Likewise relatively low scores, 40.2 and 41.1, defined high Probability gains in pairs YP_009179227.1-PF08301.12 and YP_009388303.1-

PF13238.5 (Figs. 7D,E and S5A,B). This gain was smaller in the second pair because of higher diversity of its target profile (PF13238.5 vs PF08301.12 – 11.5 and 6.1, respectively) (Fig. S5D,E).

The dependence of Probability on lengths and diversities of the query and target profiles is complex and remarkably symmetrical (Fig. S5). The actual Probability values strongly depend on the external parameters (hit score, query and target lengths for Fig. S5D-F plots). Notably, it can show non-monotonous changes for a fixed query or target diversity over most of the range values. In the present study, query profiles were based on a single sequence (diversity = 1), with Probability estimation only increasing with further increase of the observed diversity in three target profiles (Fig. S5D-F).

Text S1.6 Instructions regarding the usage of LAMPA R package

The package is provided on GitHub: https://github.com/Gorbalenya-Lab/LAMPA. It can be installed using R commands *library(devtools); install_github('Gorbalenya-Lab/LAMPA')* and loaded using R command *library(LAMPA)*. The package contains a single user-level function, that is called also LAMPA. To display detailed information about the usage of this function, use R command *help(LAMPA)*.

While we run the analysis of RNA virus polyproteins using HHmake and HHsearch programs from HH-suite 2.0.16 and script addss.pl from HH-suite 3.0.0 against pfamA_31.0, the package is expected to work with other versions of these HH-suite programs and scripts as well, provided that they have the same input and output data formats. Other databases compatible with the HH-suite programs can also be used. Running LAMPA based solely on HH-suite v.3.x is technically possible but may be affected by HHsearch v.3.x issue which leads to overuse of random access memory (RAM) during searches of large databases and could cause job crushed (https://github.com/soedinglab/hh-suite/issues/124).

Single run of the LAMPA function conducts the annotation procedure for a single query sequence. To apply the function to multiple query sequences, user can employ R loop *for* iterating over query sequences and running the LAMPA function for each query sequence in succession [30]; the number of central processing units (CPUs) utilized in HHsearch searches can be regulated via the LAMPA argument *cpu*. Alternatively, user can employ R package doParallel to run the LAMPA function for multiple query sequences in parallel [45]; it is recommended to set value of the LAMPA argument *cpu* to 1 in this case.

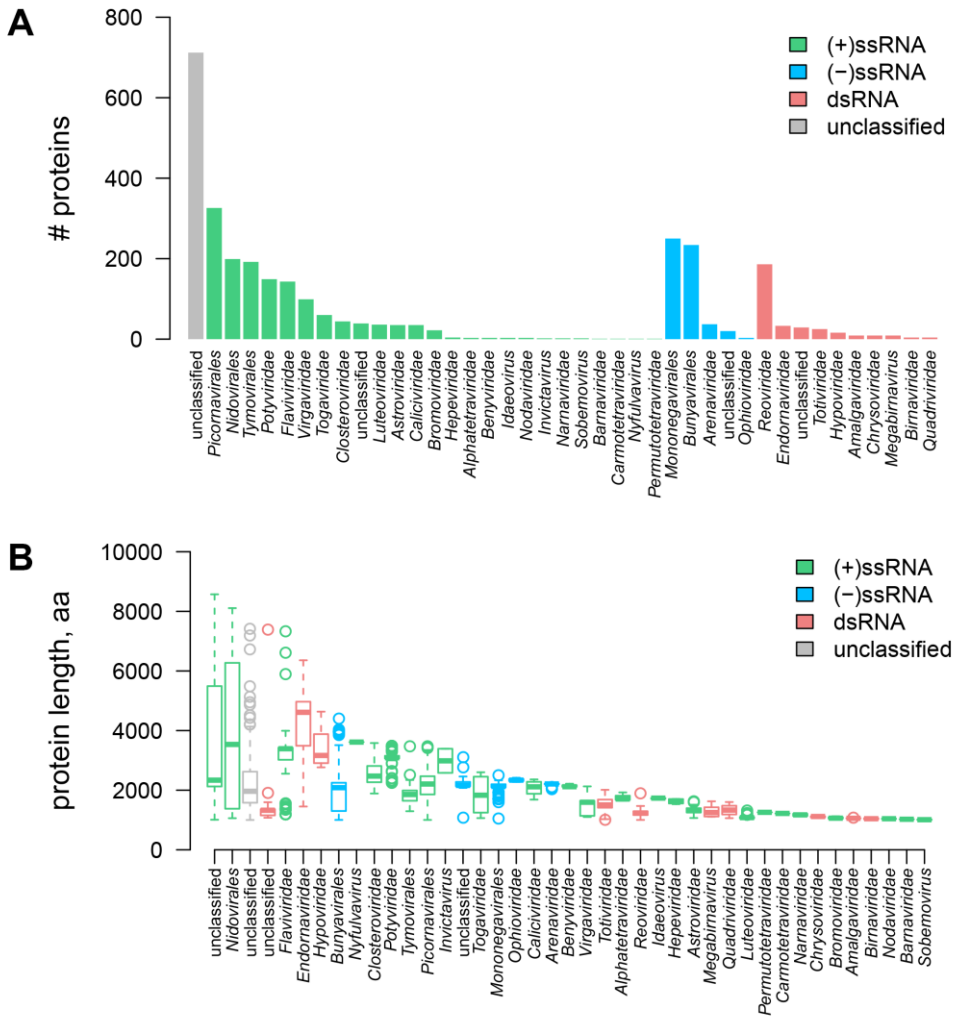


Figure S1 | Composition of the analysed RNA virus polyprotein dataset. (**A**) Number of proteins belonging to different taxonomic groups. (**B**) Length of proteins from different taxonomic groups. Virus taxonomy for each protein were derived from the corresponding genome RefSeq entry; only the most senior taxonomic rank specified in the entry is shown for each protein.

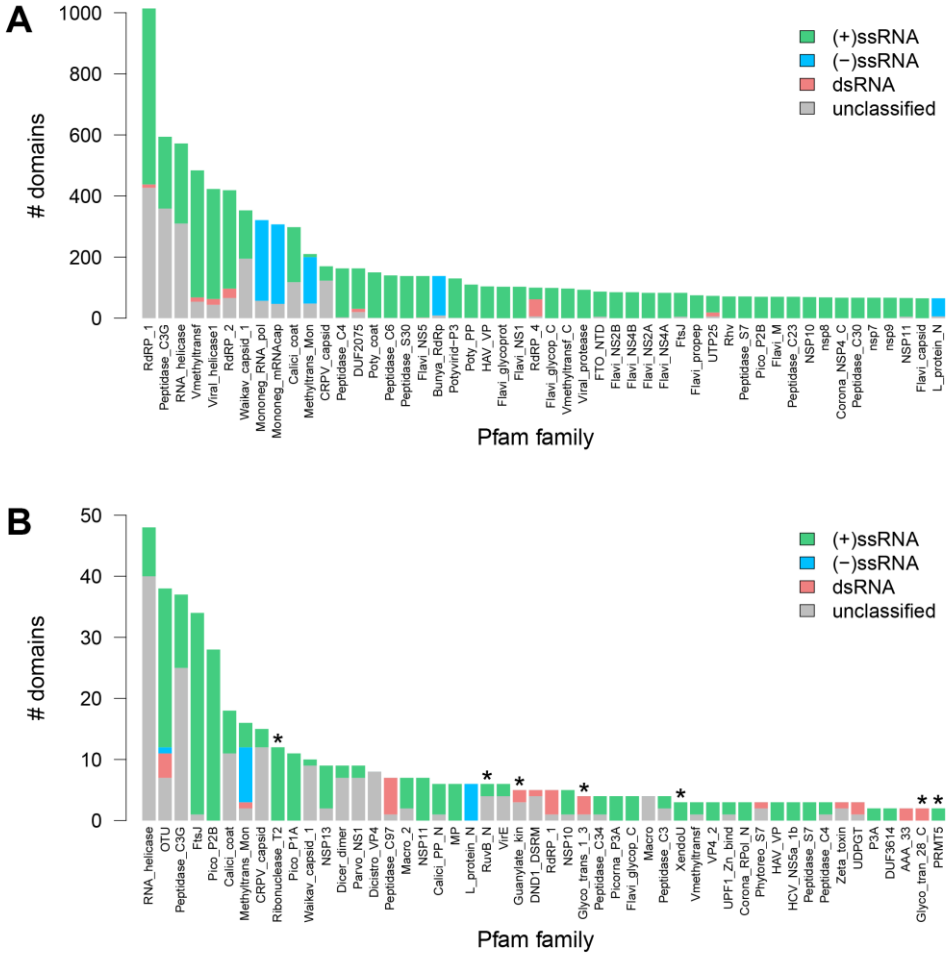


Figure S2 | Target profiles that dominated LAMPA hit lists of RNA virus polyproteins. Fifty Pfam profiles that were most frequently hit by RNA virus polyproteins during (**A**) stage #1 and (**B**) stages #2-#3 of the LAMPA procedure. Pfam profiles, not hit at stage #1 (unique to LAMPA compared to conventional HHsearch), are highlighted with asterisks.

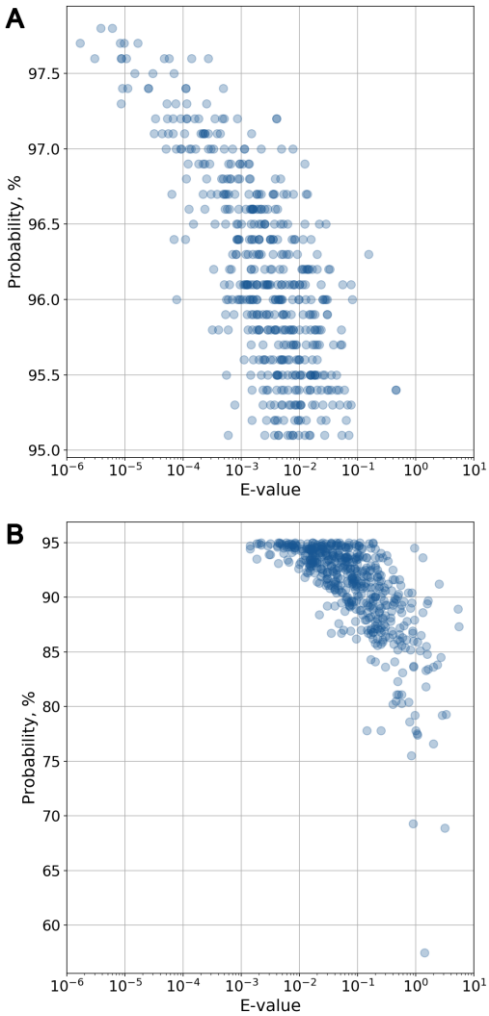


Figure S3 | Relationship between Probability and E-value for HHsearch hits. The plots show relationship between Probability and E-value for 507 hits that were elevated above 95% Probability cut-off by LAMPA at stages #2 and #3 (**A**) compared to stage #1 that is equivalent to HHsearch output (**B**). Probabilities and E-values of hits are inversely related, and this relationship is modulated by hits' secondary structure scores that are distributed in a wide range (from -3.6 to 18.8) and affect Probability but not E-value. Variation of Probability values decreases and E-values in logarithmic scale increases after hits were elevated above 95% Probability cut-off. Both these trends are determined by the properties of hit score auto-calibration procedure; in particular by the observed dependence of Probability and P-,E-value on query profile length, see Figure 7.

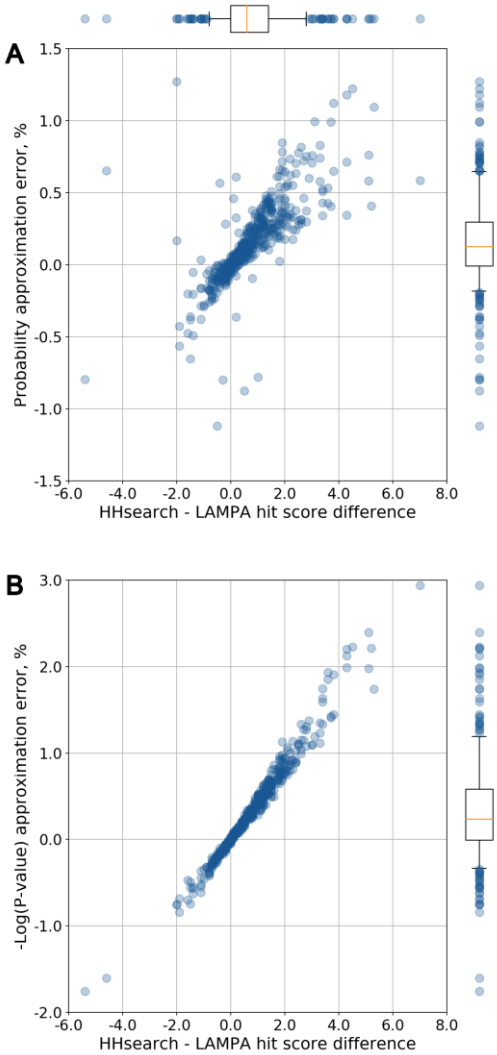


Figure S4 | Statistic approximation error and its dependence on hit score accuracy of query in computational experiments. In computational experiments, hit statistics were calculated for each query, regardless of its length, using fixed hit score(s) obtained for respective intact polyprotein. The depicted plots show relationship between deltas of hit statistic (Y axis) and its score (X axis) calculated for polyprotein and its region, which were used as queries at stages #1 vs #2 and #3 of LAMPA. The delta of hit statistic, Probability (panel **A**) and P-value (panel **B**), is equal to error of statistic approximated in computational experiments. Hit score used to calculate Probability but not P-value is composite and includes secondary structure score. Box-and-whisker summary statistic for two variables: box, 25%-75% range, whiskers 2.5%-97.5% range.

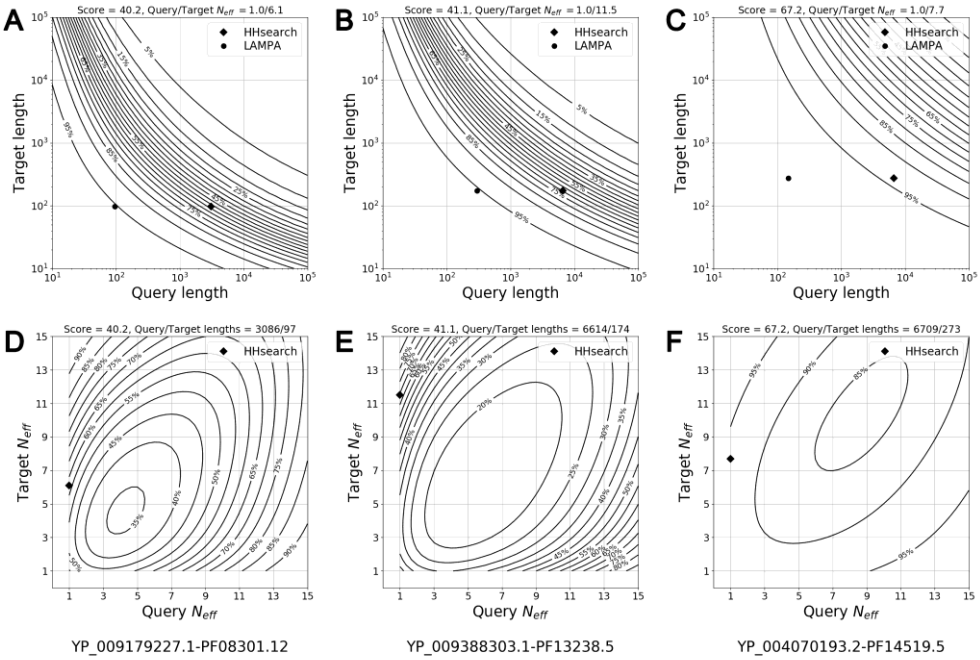


Figure S5 | Relationship of hit Probability to query and target lengths and diversities in computational experiments. Presented are results of estimation of HHsearch hit Probability for different combinations of either query and target lengths (**A-C**) or query and target profile diversities (**D-F**), which were computationally generated. Diamond and circle labels in A-C panels indicate lengths of profiles used to detect the hit by HHsearch (without hits post-processing) and LAMPA (stage #2 or #3), respectively. Diamond label in D-F panels indicates real values of target and query diversities. Three query-target pairs used for panels A and D, B and E, and C and F are indicated at the bottom.

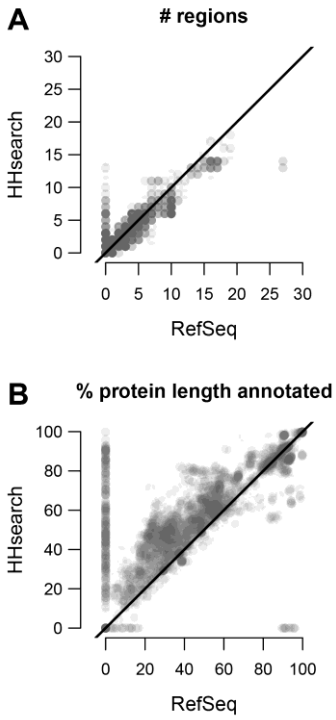


Figure S6 | Summary statistic of annotation coverage by HHsearch and RefSeq experts. Comparison of the number of regions per protein (**A**) or percentage of protein length (protein coverage) (**B**) annotated by HHsearch (LAMPA stage #1) and RefSeq experts, based on analysis 2985 RNA virus proteins. Each protein is represented by a transparent grey dot; dot density is proportional to the number of proteins with identical characteristics. Black line, diagonal.

Table S1 | RNA virus polyproteins used for testing LAMPA.

Table is available from https://doi.org/10.1093/bioinformatics/btaa065

Table S2 | Hits between RNA virus polyproteins and PfamA profiles identified during QP-specific and AP-specific stages of LAMPA.

Table is available from https://doi.org/10.1093/bioinformatics/btaa065

Table S3 | Characteristics affecting estimation of statistical significance of similarity in three query-target pairs.

Table is available from https://doi.org/10.1093/bioinformatics/btaa065

REFERENCES

1. Brister JR, Ako-Adjei D, Bao Y, Blinkova O: **NCBI viral genomes resource**. *Nucleic Acids Res* 2015, **43**(Database issue):D571-577.
2. Suttle CA: **Marine viruses--major players in the global ecosystem**. *Nat Rev Microbiol* 2007, **5**(10):801-812.
3. Zhang YZ, Chen YM, Wang W, Qin XC, Holmes EC: **Expanding the RNA Virosphere by Unbiased Metagenomics**. *Annu Rev Virol* 2019.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
5. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I: **GenBank**. *Nucleic Acids Res* 2019, **47**(D1):D94-D99.
6. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W29-37.
7. Söding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**(7):951-960.
8. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment**. *Nat Methods* 2012, **9**(2):173-175.
9. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A *et al*: **The Pfam protein families database in 2019**. *Nucleic Acids Res* 2019, **47**(D1):D427-D432.
10. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M: **Uniclust databases of clustered and deeply annotated protein sequences and alignments**. *Nucleic Acids Res* 2017, **45**(D1):D170-D176.
11. Fox NK, Brenner SE, Chandonia JM: **SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures**. *Nucleic Acids Res* 2014, **42**(Database issue):D304-309.
12. Wheelan SJ, Marchler-Bauer A, Bryant SH: **Domain size distributions can predict domain boundaries**. *Bioinformatics* 2000, **16**(7):613-618.
13. Remmert M: **Fast, sensitive protein sequence searches using iterative pairwise comparison of hidden Markov models**. *Doctoral dissertation.* Munich: Ludwig Maximilian University; 2011.
14. Söding J, Remmert M, Hauser A: **User Guide 2.0.15: HH-suite for sensitive protein sequence searching based on HMM-HMM alignment**. In: *HH-suite.* 2012.
15. Baltimore D: **Expression of animal virus genomes**. *Bacteriol Rev* 1971, **35**(3):235-241.

16. Das K, Arnold E: **Negative-Strand RNA Virus L Proteins: One Machine, Many Activities**. *Cell* 2015, **162**(2):239-241.

17. Zhang J: **Protein-length distributions for the three domains of life**. *Trends Genet* 2000, **16**(3):107-109.

18. Dougherty WG, Semler BL: **Expression of virus-encoded proteinases: functional and structural similarities with cellular enzymes**. *Microbiol Rev* 1993, **57**(4):781-822.

19. Gorbalenya AE, Snijder EJ: **Viral cysteine proteinases**. *Perspectives in Drug Discovery and Design* 1996, **6**(1):64-86.

20. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R: **Viral mutation rates**. *J Virol* 2010, **84**(19):9733-9748.

21. Kuchibhatla DB, Sherman WA, Chung BY, Cook S, Schneider G, Eisenhaber B, Karlin DG: **Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently "orphan" viral proteins**. *J Virol* 2014, **88**(1):10-20.

22. Habermann BH: **Oh Brother, Where Art Thou? Finding Orthologs in the Twilight and Midnight Zones of Sequence Similarity**. In: *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods.* Edited by Pontarotti P. Cham: Springer International Publishing; 2016: 393-419.

23. Saberi A, Gulyaeva AA, Brubacher JL, Newmark PA, Gorbalenya AE: **A planarian nidovirus expands the limits of RNA genome size**. *PLoS Pathog* 2018, **14**(11):e1007314.

24. Steinegger M, Soding J: **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets**. *Nat Biotechnol* 2017, **35**(11):1026-1028.

25. Firth AE, Brierley I: **Non-canonical translation in RNA viruses**. *J Gen Virol* 2012, **93**(Pt 7):1385-1409.

26. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Res* 2016, **44**(D1):D733-745.

27. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences**. *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:175-182.

28. Steinegger M, Meier M, Mirdita M, Vohringer H, Haunsberger SJ, Soding J: **HH-suite3 for fast remote homology detection and deep protein annotation**. *BMC Bioinformatics* 2019, **20**(1):473.

29. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices**. *J Mol Biol* 1999, **292**(2):195-202.

30. R Core Team: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing; 2018.

31. Charif D, Lobry JR: **SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis**. In: *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations.* Edited by Bastolla U, Porto M, Roman HE, Vendruscolo M. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007: 207-232.

32. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges**. *PLoS Comput Biol* 2013, **9**(8):e1003118.

33. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS *et al*: **Redefining the invertebrate RNA virosphere**. *Nature* 2016, **540**:539-543.

34. Schutze H, Ulferts R, Schelle B, Bayer S, Granzow H, Hoffmann B, Mettenleiter TC, Ziebuhr J: **Characterization of White bream virus reveals a novel genetic cluster of nidoviruses**. *J Virol* 2006, **80**(23):11598-11609.

35. Stenglein MD, Jacobson ER, Wozniak EJ, Wellehan JF, Kincaid A, Gordon M, Porter BF, Baumgartner W, Stahl S, Kelley K *et al*: **Ball python nidovirus: a candidate etiologic agent for severe respiratory disease in Python regius**. *MBio* 2014, **5**(5):e01484-01414.

36. Decroly E, Ferron F, Lescar J, Canard B: **Conventional and unconventional mechanisms for capping viral mRNA**. *Nat Rev Microbiol* 2011, **10**(1):51-65.

37. Punta M, Ofran Y: **The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function**. *PLoS Comput Biol* 2008, **4**(10):e1000160.

38. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A *et al*: **A large-scale evaluation of computational protein function prediction**. *Nat Methods* 2013, **10**(3):221-227.

39. Shi M, Lin XD, Vasilakis N, Tian JH, Li CX, Chen LJ, Eastwood G, Diao XN, Chen MH, Chen X *et al*: **Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses**. *J Virol* 2016, **90**(2):659-669.

40. Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, Qin XC, Xu J, Holmes EC, Zhang YZ: **Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses**. *Elife* 2015, **4**.

41. Kobayashi K, Atsumi G, Iwadate Y, Tomita R, Chiba K-i, Akasaka S, Nishihara M, Takahashi H, Yamaoka N, Nishiguchi M *et al*: **Gentian Kobu-sho-associated virus: a tentative, novel double-stranded RNA virus that is relevant to gentian Kobu-sho syndrome**. *Journal of General Plant Pathology* 2013, **79**(1):56-63.

42. Bekal S, Domier LL, Gonfa B, McCoppin NK, Lambert KN, Bhalerao K: **A novel flavivirus in the soybean cyst nematode**. *J Gen Virol* 2014, **95**(Pt 6):1272-1280.

43. Ong JWL, Li H, Sivasithamparam K, Dixon KW, Jones MGK, Wylie SJ: **Novel Endorna-like viruses, including three with two open reading frames, challenge the membership criteria and taxonomy of the Endornaviridae**. *Virology* 2016, **499**:203-211.

44. Valverde RA, Khalifa ME, Okada R, Fukuhara T, Sabanadzovic S, Ictv Report C: **ICTV Virus Taxonomy Profile: Endornaviridae**. *J Gen Virol* 2019.

45. Microsoft Corporation, Weston S: **doParallel: Foreach Parallel Adaptor for the 'parallel' Package**. In.; 2018.

CHAPTER 6

General Discussion

PREFACE

Historically, viruses were discovered and characterized experimentally. The advent of nucleic acid sequencing opened up a new way of studying viruses – comparative genomics – allowing to characterize viruses based on their genome sequences, often using results of experimental research on related viruses. In this thesis, we used comparative genomics techniques to characterize various aspects of nidovirus biology and evolution (**chapters 2-4**), while we also developed a method facilitating homology detection and annotation of large and highly divergent polyproteins (**chapter 5**).

Since 2014, when the project described in this thesis started, the importance of comparative genomics in nidovirus characterization has greatly increased together with the rate of nidovirus discovery. The number of nidovirus species recognized by ICTV has almost quadrupled, increasing from 31 in 2014 to 88 in 2017 [1, 2] and to 109 proposed in 2019 (Fig. 1) [3-6]. The reasons behind this explosive growth, observed for other groups of RNA viruses as well, are advancement of NGS technologies, and consequent transformation of metatranscriptome sequencing into a widely-used laboratory technique. Metatranscriptomics allows to screen a broad range of hosts for the presence of RNA viruses with high efficiency. For example, a single metatranscriptomics study conducted by Shi *et al.* in 2016 identified 1,445 novel phylogenetically distinct RNA virus genomes [7] that almost doubled the number of RNA virus species known at the time. Six genomes discovered in that study were later recognized by ICTV as prototypes of novel, divergent nidovirus species [2, 8].

The number of nidoviruses discovered in the past five years far exceeds the number of nidoviruses that were ever propagated in cell culture and characterized experimentally. As the rate with which new nidoviruses are discovered is increasing, so does the cost and complexity of the experiments required to characterize all of them experimentally. It makes laboratory research on all the newly discovered nidoviruses infeasible. Instead, comparative genomics is used to provide a connection between numerous nidoviruses discovered based solely on NGS data, and a few nidoviruses that are subject to comprehensive experimental characterization. Comparative genomics identifies homologous regions of genomes and proteins, allowing to transfer functional annotation from experimentally characterized viruses and hosts to newly discovered virus genomes.

Characterization by comparative genomics can be facilitated by reliable classification of viruses, as taxonomic assignment itself may offer clues about the biology of a newly discovered virus, as well as help to design comparative genomics experiments. Accommodating the known diversity of RNA viruses requires building a multilevel hierarchical classification while dealing with large evolutionary distances. The DEmARC software package that was developed in our group [9, 10] allows to build such classification. Following its publication in 2012, the package was advanced by our group to include a greater choice (1) of methods used to calculate genetic distances between

viruses, and (2) of linkage types used in hierarchical clustering; (3) to implement sequence weights reducing the impact of virus sampling bias; (4) to propose an alternative approach of identifying classification levels; (5) to visualize various aspects of the obtained classification; (6) to provide the user with an in-package tutorial on DEmARC usage. Over the past few years, our group was involved in producing DEmARC-based multilevel classifications of the newly described nidoviruses that ICTV Study Groups on nidoviruses used as the basis for taxonomy proposals [3-6, 11-17]. ICTV approved the first seven of these proposals, including the one delineating four new invertebrate nidovirus families (*Medioni-*, *Euroni-*, *Abysso-* and *Mononiviridae*) [1, 2], and is currently considering the four latest proposals, including a proposal delineating five new vertebrate nidovirus families (*Olifo-*, *Gresna-*, *Cremega-*, *Nanhypo-* and *Nanghoshaviridae*) [3]. On a phylogenetic tree, all four novel invertebrate families cluster with invertebrate nidoviruses that were known before, while novel vertebrate families are basal to *Arteriviridae* (Fig. 1). Also the subfamilies *Coronavirinae* and *Torovirinae* were elevated to the rank of families, *Coronaviridae* and *Tobaniviridae*, respectively, and include now many more species [2].

Comparative genomics analysis of nidoviruses is challenging since, due to the high mutation rate, many novel nidoviruses are highly divergent. One manifestation of nidovirus divergence is the ever-expanding nidovirus host range (Fig. 1) that now includes many exotic animals: from a marsupial mammal [18] to mollusks [7, 19, 20], an urochordate [21] and a flatworm (**chapter 4**). The complexity of the analysis is further compounded by the organization of the nidovirus proteome, which is dominated by large multidomain polyproteins. When a polyprotein sequence is compared to a database of known protein sequences or profiles, distant homologous relationships may remain undetected due to the underestimation of statistical significance of hits by standard tools, designed to annotate smaller proteins with few domains. To address this complication, our group developed a tool, called LAMPA, that gradually splits the polyprotein sequence into smaller queries in a biologically reasonable manner, improving estimation of hit statistical significance (**chapter 5**). For comparison of the delineated queries with databases, LAMPA employs HH-suite, one of the most sensitive software packages for protein homology detection [22, 23].

Usage of state-of-the-art comparative genomics tools was essential for the characterization of three different aspects of the growing nidovirus diversity, described in this thesis: identification of a novel replicative domain universally conserved in all nidoviruses (**chapter 3**), analysis of domain organization and evolution of arterivirus non-structural polyprotein N-terminus that encompassed multiple newly recognized species, including the most divergent WPDV, the genome sequencing of which was completed as part of the study (**chapter 2**), functional description of a novel, highly divergent nidovirus genome identified by mining transcriptome of its host, planarian *S. mediterranea* (**chapter 4**). These and other studies, conducted in recent years, contributed to the evolution of understanding of nidovirus key properties.

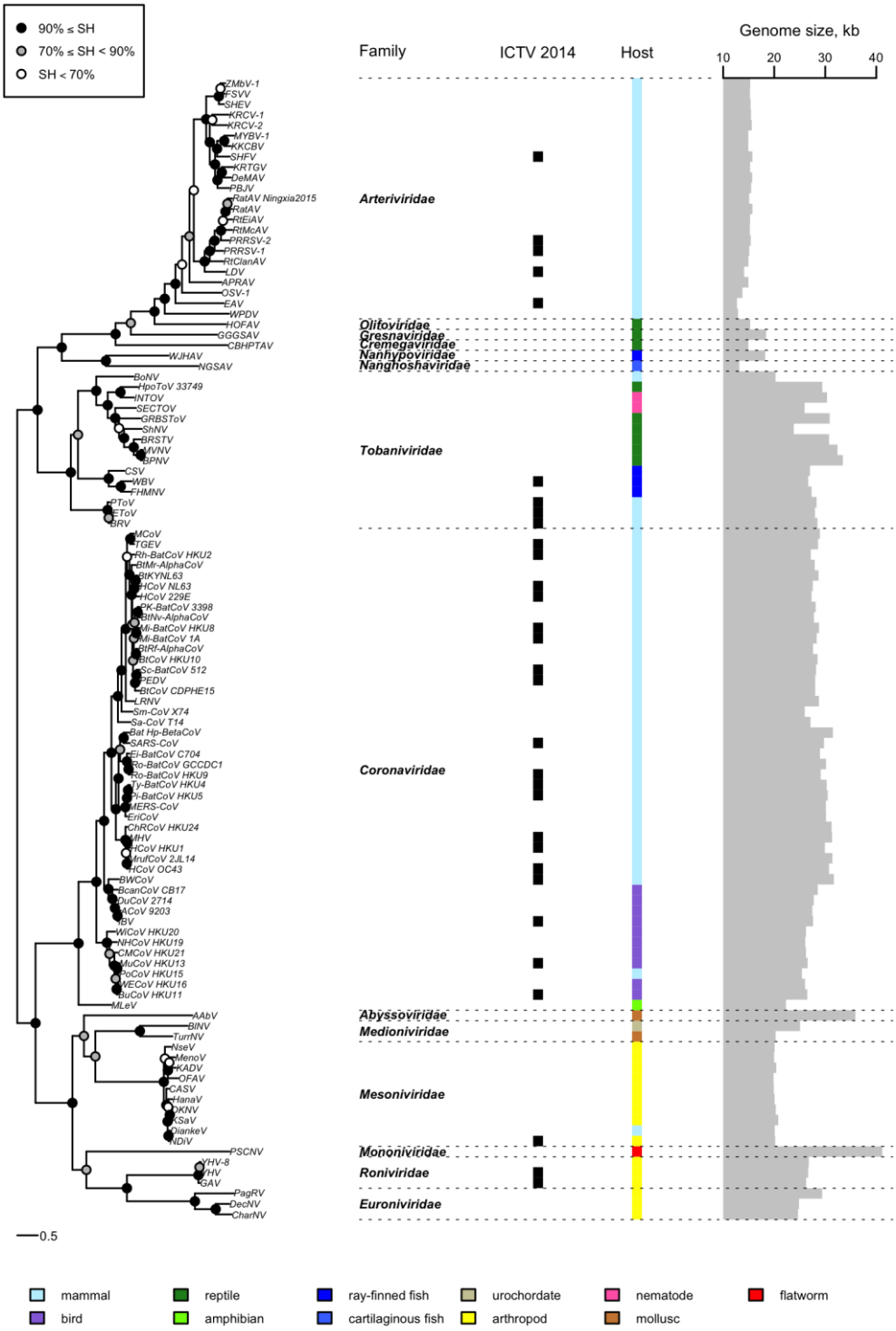

90% ≤ SH
70% ≤ SH < 90%
SH < 70%
Family
ICTV 2014
Host
Genome size, kb
10
20
30
40
ZMbV-1
FSVV
SHEV
KRCV-1
KRCV-2
MYBV-1
KKCBV
SHFV
KRTGV
DeMAV
PBJV
RatAV Ningxia2015
RatAV
RtEiAV
RtMcAV
PRRSV-2
PRRSV-1
RtClanAV
LDV
APRAV
OSV-1
EAV
WPDV
HOFAV
GGGSAV
CBHPTAV
WJHAV
NGSAV
BoNV
HpoToV 33749
INTOV
SECTOV
GRBSToV
ShNV
BRSTV
MVNV
BPNV
CSV
WBV
FHMNV
PToV
EToV
BRV
MCoV
TGEV
Rh-BatCoV HKU2
BtMr-AlphaCoV
BtKYNL63
HCoV NL63
HCoV 229E
PK-BatCoV 3398
BtNv-AlphaCoV
Mi-BatCoV HKU8
Mi-BatCoV 1A
BtRf-AlphaCoV
BtCoV HKU10
Sc-BatCoV 512
PEDV
BtCoV CDPHE15
LRNV
Sm-CoV X74
Sa-CoV T14
Bat Hp-BetaCoV
SARS-CoV
Ei-BatCoV C704
Ro-BatCoV GCCDC1
Ro-BatCoV HKU9
Ty-BatCoV HKU4
Tylo-BatCoV HKU5
MERS-CoV
EriCoV
ChRCoV HKU24
MHV
HCoV HKU1
MrufCoV 2JL14
HCoV OC43
BWCoV
BcanCoV CB17
DuCoV 2714
ACoV 9203
BV
WiCoV HKU20
NHCoV HKU19
CMCoV HKU21
MuCoV HKU13
RoCoV HKU15
MECoV HKU16
BuCoV HKU11
MLeV
AAbV
BINV
TurrNV
NseV
MenoV
KADV
OFAV
CASV
HanaV
OKNV
KSaV
DiankeV
NDiV
PSCNV
YHV-8
HV
GAV
PagRV
DecNV
CharNV
0.5
Arteriviridae
Olifoviridae
Gresnaviridae
Cremegaviridae
Nanhypoviridae
Nanghoshaviridae
Tobaniviridae
Coronaviridae
Abyssoviridae
Medioniviridae
Mesoniviridae
Mononiviridae
Roniviridae
Euroniviridae
mammal
bird
reptile
amphibian
ray-finned fish
cartilaginous fish
urochordate
arthropod
nematode
mollusc
flatworm

Figure 1 | Nidovirus diversity: phylogeny, taxonomy, host range, and genome size. Shown is the midpoint-rooted phylogeny, family structure of 2019, species recognized by 2014, host group and genome size of nidovirus species under consideration by ICTV in 2019 [3-6]. The phylogeny was reconstructed based on Viralis MSAs [24] of 3CLpro, NiRAN, RdRp, ZBD, HEL1 conserved cores, using IQ-Tree 1.5.5 [25], with an evolutionary model selected for each domain independently. To estimate branch support, SH-like approximate likelihood ratio test with 1000 replicates was conducted. Viruses representing species and species acronyms are identical to those used in 2019 ICTV proposals (except acronym AcCoV was substituted by AAbV). Information about nidovirus hosts and genome lengths was obtained from GenBank entries. Note that for nidoviruses discovered by metatranscriptomics, host misassignment due to contamination remains a persistent concern. For example, it can be hypothesized that two toroviruses (species *Infratovirus 1* [INTOV] and *Sectovirus 1* [SECTOV], represented by the Xinzhou toro-like virus [KX883638.1] and the Xinzhou nematode virus 6 [KX883637.1], respectively) discovered in metatranscriptomes of snake-associated nematodes [7], which were ascribed as their potential hosts, may infect reptiles instead, as the two viruses cluster with reptile viruses on the phylogenetic tree, and their ascribed hosts are likely to be contaminated with reptile materials. Likewise, the actual host of the *Dianke virus* (DiankeV; KY056254.1) may be an arthropod, because while the virus genome was reported to be isolated from the brain of a rodent, it is closely related to arthropod viruses not known to infect vertebrate hosts.

NIDOVIRUSES RE-DEFINED: UNCOUPLING MOST CONSERVED CHARACTERISTICS

Originally nidoviruses were recognized as a distinct virus group sharing a unique mechanism of discontinuous subgenomic RNA synthesis (transcription), which provided the basis for the group name. Nidoviruses are also distinguished by a conserved multi-ORF organization of the genome, utilization of PRF to express the second ORF (ORF1b), four-domain synteny of replicative domains (3CLpro-RdRp-ZBD-HEL1) encoded in the middle part of the genome, monophyletic clustering of the conserved replicative domains, and a genome size ranging from 12 to 34 kb and including the largest known RNA genomes [26-29]. This recurrent association of many characteristics is indicative of genetic coupling that may have originated early in the nidovirus evolution. However, with expanding experimental characterization of a few distantly related nidoviruses and through comparative genomics of the ever-increasing diversity of nidoviruses, the notion of nidovirus-specific conservation has been steadily adjusted.

Since 1993, when the order *Nidovirales* was established [30], some nidoviruses, namely toro- and roniviruses, were shown to deviate from others in using discontinuous subgenomic RNA synthesis [31, 32]. Discoveries made after 2014, including the ones described in this thesis, identified deviations in other characteristics, limiting further the number of those that are universally shared by all nidoviruses.

Variation in genome architecture and expression mechanisms of nidoviruses

All nidovirus genomes discovered up to 2016 followed the common genome plan: overlapping ORF1a and ORF1b occupying the 5’-terminal two-thirds of the genome and encoding non-structural proteins responsible for virus expression and replication, respectively, and multiple smaller 3’ORFs that encode structural and accessory proteins.

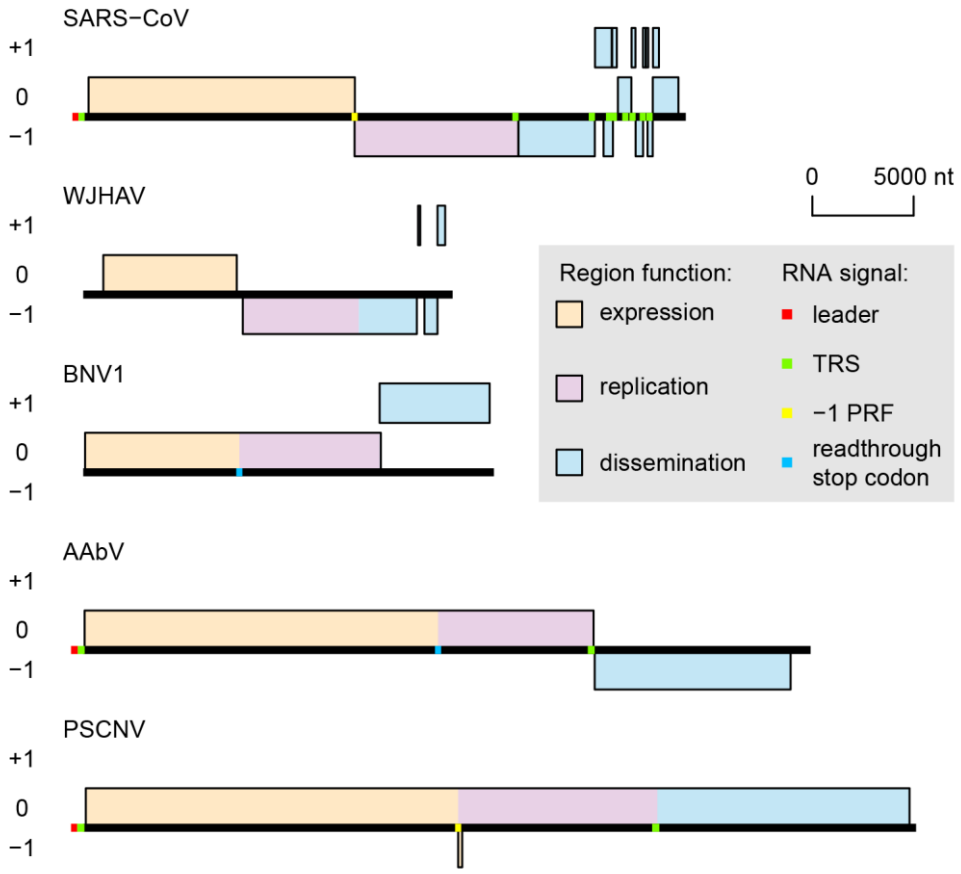


Figure 2 | Nidoviruses with canonical (SARS-CoV) and non-canonical genome ORFs organization. WJHAV, Wuhan Japanese halfbeak arterivirus (MG600020.1); BNV1, Beihai nido-like virus 1 (KX883629.1) classified as species *Turrinivirus 1* (TurrNV) by ICTV; AAbV, Aplysia abyssovirus 1 (GBBW01007738.1); PSCNV, Planarian secretory cell nidovirus (MH933735.1). ORFs are positioned according to their frame, with the most 5'-terminal depicted ORF set as zero. ORF regions are colored according to their predicted function (see inset). Genome signals, described by the discoverers of each virus, are indicated by color (see inset).

Recent discoveries, including some in our study (**chapter 4**), demonstrated that variation of this conserved genome plan is possible in nidoviruses (Fig. 2). In the Wuhan Japanese halfbeak arterivirus (WJHAV) genome, ORF1b is fused with a gene encoding putative glycoprotein [33], presumably a structural protein. Both Beihai nido-like virus 1 (BNV1) and Aplysia abyssovirus 1 (AAbV) contain two ORFs, a 5'-terminal ORF combining ORF1a- and ORF1b-like regions, and a single 3'-terminal ORF encoding structural protein domains [7, 19, 20]. The PSCNV genome has a single large ORF, which is an equivalent of ORF1a, ORF1b and 3'ORFs fused together (**chapter 4**). Notably, PSCNV ORF encodes a 13,556 aa

polyprotein that is 58-67% larger than the largest single- or multi-ORF polyproteins of other viruses.

Conserved multi-ORF genome organization of nidoviruses is coupled with conserved expression mechanisms of transcription and translation, controlling the relative quantities of functionally different proteins in infected cell. The discovery of nidoviruses with an unusual ORF organization raises the question whether they maintain the canonical stoichiometry of viral proteins, and if so, by what expression mechanisms.

In canonical nidoviruses, ORF1a-encoded proteins are expressed in higher quantities than ORF1b-encoded proteins, due to -1 PRF directing a fraction of ribosomes from ORF1a to ORF1b translation. A similar non-structural proteins ratio may be achieved through different mechanisms in non-canonical nidoviruses BNV1, AAbV and PSCNV. Both BNV1 and AAbV have ORF1a-like and ORF1b-like regions residing in the same reading frame and separated by a stop codon (Fig. 2) [7, 19, 20]. If a readthrough of this stop codon only occurs in a fraction of translation events, proteins encoded in the ORF1a-like region would be expressed in a higher quantity compared to proteins encoded in the ORF1b-like region. The PSCNV genome includes a predicted -1 PRF site with a potential to divert translation from ORF1b-like region of the main ORF to a tiny 39 nt ORF (Fig. 2). If efficiency of frameshifting at the predicted site is limited, ORF1a-like compared to ORF1b-like region of the PSCNV genome would be expressed more frequently (**chapter 4**). The main difference between the -1 PRF-directed mechanisms in canonical nidoviruses and in PSCNV is that -1 PRF directs translation into ORF1b in the former but diverts it from ORF1b-like region in the latter.

Another important feature of protein synthesis in canonical nidoviruses is that structural proteins are expressed starting at a later point in time and in higher quantities than non-structural ones. This is achieved through TRS-guided production of sg mRNAs that are 3'-coterminal with the genome and encompass ORFs encoding structural proteins. Notably, this mechanism may be used also for expression of non-canonical nidoviruses: the single 3'ORF of AAbV [19] and the 3'ORFs-like region of the PSCNV genome (**chapter 4**); based on similarity with canonical nidoviruses, this hypothesis may be extended to the single 3'ORF of BNV1 and the three small 3'ORFs of WJHAV, although these viruses were not studied in this respect (Fig. 2). In both the PSCNV and AAbV, potential leader and body TRSs were identified by comparative genomics as large repeats in the 5'UTR and upstream of the genome region predicted to encode structural proteins, respectively (Fig. 2). A sharp increase in coverage of the genome by RNA-seq reads was observed at the body TRS of both PSCNV and AAbV, consistent with the downstream region being a subject of transcription. Existence of PSCNV sg mRNA species, expected to be expressed when the identified TRSs are employed, was confirmed in a 5'-RACE experiment. Importantly, if

translation of the PSCNV sg mRNA species is initiated at its most 5'-terminal start-codon, it would result in production of a polyprotein identical to the C-terminus of the giant polyprotein expressed from the PSCNV genome (**chapter 4**). Thus, predicted structural proteins of PSCNV may be expressed from both genome and sg mRNA.

Interestingly, the PSCNV may not be the only non-canonical nidovirus, structural proteins of which are synthesized from both genome and sg mRNA. The WJHAV potentially encodes a glycoprotein in the unusually long 3'-terminus of its ORF1b that is located downstream of the otherwise terminal O-MT locus [3, 33]. While the WJHAV was not analyzed in this respect, this genome organization is compatible with both genome and sg mRNA directing synthesis of the glycoprotein (Fig. 2). Production of certain structural proteins from both genome and sg mRNA can also be envisioned for some of the canonical nidoviruses, such that stop-codon of ORF1b and start-codon of the downstream structural ORF are in-frame and separated by few nucleotides in their genomes. If a readthrough of the ORF1b stop-codon would occur, with the stop-codon being decoded by a suppressor tRNA [34], it would lead to a continuation of translation, resulting in the production of pp1ab fused with a structural protein (unpublished observation). For example, SARS-CoV ORF1b and ORF2, encoding S protein, belong to the same reading frame and are separated by 6 nt (Fig. 1 from chapter 1).

Unlike the expression of structural proteins from individual ORFs, observed in most known nidoviruses, expression of multiple structural proteins from a single ORF, predicted for non-canonical BNV1, AAbV and PSCNV (Fig. 2), would require processing of the structural polyprotein by host and/or viral proteases, unless their structural domains function in the context of a single polyprotein. Accordingly, a chymotrypsin-like serine protease domain was detected in the structural polyprotein sequences of BNV1 and AAbV [19, 20], and potential cleavage sites of cellular proteases furin and signal peptidase were identified in the C-terminal region of the PSCNV polyprotein (**chapter 4**).

In addition to the deviations from the canonical nidovirus genome architecture in newly discovered viruses discussed above, a number of small functional ORFs, preceding or overlapping with the nidovirus replicase, were revealed in relatively-well characterized nidoviruses in recent years.

The presence of a short ORF, encoded upstream of ORF1a, but likely to be bypassed by ribosomes due to the poor initiation context of its start-codon, was previously reported for multiple arteri- and coronaviruses [35-39]. A recent study utilizing ribosome profiling confirmed that such ORF is translated in MHV [40]. In addition, two CUG-initiated ORFs, one upstream of ORF1a, and another overlapping with ORF1a 5'-terminus, were

demonstrated to be translated in equine torovirus (EToV); the ORFs were shown to be conserved within the genus *Torovirus* [41].

Another pair of small functional ORFs was discovered in arterivirus PRRSV in 2012. Two ORFs are located in the nsp2-encoding genome region, and can be expressed via -1 and -2 PRF, respectively. PRFs occur during ORF1a translation, and lead to the production of two shortened pp1a proteins with alternative C-termini. Both -1 and -2 PRF are transactivated by a complex of arterivirus nsp1β protein and cellular poly(C) binding protein, and employ the same PRF site, consisting of a slippery sequence and a downstream cytidine-rich genome element. The mechanism was predicted to be conserved in all arteriviruses known at the time of the study, with the exception of EAV, based on conservation of the PRF site and the size of the small ORFs [42-44]. This prediction was recently extended to a number of newly discovered arteriviruses, and a distantly related arterivirus WPDV (**chapter 2**). Notably, the distance between the putative WPDV PRF site elements was predicted to be larger than in other arteriviruses. Its utilization may be coupled with extensive evolution of the WPDV nsp1β protein, which diverged considerably from its orthologs and/or other changes to the transactivating protein complex. In addition, the WPDV protein domain predicted to be expressed as a result of -2 PRF is very short and doesn't include potential transmembrane helices, unlike the analogous domains of other arteriviruses.

NiRAN: a fifth universally conserved domain of nidoviruses

In 1988-1989, four domains, 3CLpro, RdRp, ZBD and HEL1, were delineated in the first sequenced coronavirus [26, 45-47]. Subsequently, they were found to be universally conserved in all nidoviruses [30, 48-50]. Many more protein domains were delineated in further studies but none were conserved across the entire order until our discovery of the NiRAN domain, which remained undetected for almost twenty years, was published in 2015 [51]. Its discovery (**chapter 3**) extended the synteny of universally conserved domains associated with the order *Nidovirales* to 3CLpro-NiRAN-RdRp-ZBD-HEL1. Like all the other domains of the synteny, NiRAN was discovered by bioinformatics analysis of extremely distant homology. It is the only enzymatic domain among these five conserved domains that has no apparent homolog in other RNA viruses (however see below).

The NiRAN domain is encoded upstream of RdRp within the same cleavage product. It was shown to be conserved in all nidoviruses by comparative genomics that included profile-profile, predicted secondary structure, and conserved sequence motifs analyses. Comparative genomics helped formulate initial hypothesis about function of the domain. Based on the domain position in ORF1b-encoded part of pp1ab, harboring replicative enzymes, and its profile of invariant residues, the domain was proposed to be an NTP-dependent RNA ligase, a partner of endoribonuclease NendoU, which was believed to

```
Probab=20.84  E-value=1.6e+02  Score=21.55  Aligned_cols=109  Identities=21%  Similarity=0.167  Sum_probs=60.7

                                 AN                                                  BN
Q ss_pred                ccCceEEEEEccc--h--------HHHHHHHHHhc-CccccceEEEeccCccceeccCCCcccHHHHHHHhcCCCcc.
Q AF227196m         84   INGKNRHYRTKHS--L--------TRRIKIAQTLS---HIPEAIQFKKD--SHGYYR-EISQFSLAIVLHGFANQIEPDf  147 (252)
Q Consensus         69   ~~d~~~~vK~~~~~~~~~~~~~~~E~~~y~~L~~~~~~~h~~f~~~~g~~~l~R~~lT~ySL~DL~ya~r~~~~~~.  136 (249)
                         ..+..+.+|+...  .          ..|......+.. ..+.-.+++.. .....++-.....-||.+++-....++
T Consensus         22   ~~~~~~~~K~~~~~~~~~~~~~~~~~@~~~~~~l~~~~i~~~~~~~~~~~~~~~v~~~~~~~~L~~~l~~~~~~~~~.   96 (260)
T PF00069_consen    22   KNGQKVAVKIIDKSEIEEEEEREENIREIKILRRLRHPNIVQILDVFQD-DNYLYIVMEYCPGGSLQDYLQKNKPLS---.   96 (260)
T ss_dssp                TTTEEEEEEEEESTTHHHHHHHHHHHHHHHHHHHTBTTBCHEEEEEEE-SSEEEEEEEEETTEBHHHHHHHHSSBB---.
T ss_pred                CCCeEEEEEEccccccccccccchhhhhhhcccccccccccccccccccc-cccccccccccccccccccccccccc---.
Confidence               3345688888776  1        125556666644 44444555555 3334455556665688888873333 11
                                 II                 III

                                                                                                CN
Q ss_pred                HHHHHHhhhhccCccc...CCCCChhHHHHHHHHHHHHHHHHhHHHHHHHHHcCcee-eecCCCccCC----Cceeec
Q AF227196m        148   LAKYTNERNIKVSKTTwlcKNICNHKSCNILKMDIFDYTYTCYTKAHSFALQACSTYNF-DITPDNICPE----G-VYDF  221 (252)
Q Consensus        137   ~~l~e~l~~~~~~~~~...d~~en~~~~~~~~~lg~~~~~a~l~~~~~~~~~~~g~~~-vlTlDNqDLn----G~~YDF  208 (249)
                                                   ..          .+...+..+++....+-..|++- .|+|+|+=++    ..+-||
T Consensus         97   ----------------...-----------------~~~~~~qi~~~L~~Lh~~~i~H~dikp~NIl~~~~~~~~l~Df  142 (260)
T PF00069_consen    97   ----------------...----EE-----------EILKIAYQILEALAYLHSKGIVHRDIKPENILLDENGEVKLIDF  142 (260)
T ss_dssp                ----------------...----HH-----------HHHHHHHHHHHHHHHHHTTEEESSBSGGGEEESTTSEEEESSG
T ss_pred                ----------------...----cc-----------cccccccccccccccccccccccccccccccccccccccccccc
Confidence                                          11          12233344445555655556664 4999998655    455688
                                                                                   VIB          VII

Q ss_pred                cccc
Q AF227196m        222   ETFR  225 (252)
Q Consensus        209   gDf~  212 (249)
                         |-..
T Consensus        143   g~~~  146 (260)
T PF00069_consen   143   GSSV  146 (260)
T ss_dssp                TTTE
T ss_pred                cccc
Confidence               8654
```

Figure 3 | Primary and secondary structure similarity between NiRAN and protein kinases. Shown is HHsearch alignment of CoToMeRoAr NiRAN and PFAM Pkinase profiles in HH-suite format. The most conserved residues of NiRAN are highlighted in green, NiRAN motifs are designated on top of the alignment. Key functional residues of Pkinase are highlighted in cyan, selected Pkinase sub-domains are designed below the alignment.

be universally conserved in nidoviruses at the time [52]. Consequently, the ability of the domain to covalently bind nucleotides (nucleotidylation), which constitutes the first stage of the ligase reaction, was probed experimentally using recombinant EAV nsp9; nucleotidylation activity with high substrate specificity for UTP, and a lesser substrate specificity for GTP, was demonstrated. Accordingly, the domain was named nidovirus RdRp-associated nucleotidyltransferase, or NiRAN (**chapter 3**).

Three possible functional roles were proposed for NiRAN, with each being compatible with some but not all available experimental data (**chapter 3**). First, as suggested initially, NiRAN might be an NTP-dependent RNA ligase. However, the fact that conservation of the potential counterpart of the ligase, NendoU, was shown to be restricted to vertebrate nidoviruses [50], as well as EAV NiRAN preference for UTP and GTP, uncharacteristic for ligases, make this hypothesis less plausible. Second, NiRAN might be a guanylyltransferase (GTase) catalyzing the second reaction of the mRNA capping pathway (Fig. 3 from chapter 1), though preference of EAV NiRAN for UTP over GTP substrate would be difficult to explain in this context. Third, NiRAN might be involved in protein-primed RNA synthesis,

either in the capacity of a protein covalently linked to the 5'-terminal nucleotide of the nascent RNA strand, or by transferring a nucleotide to such protein. Notably, a protein-priming role would be in perfect agreement with the EAV NiRAN substrate specificity: 5'-terminal nucleotide of EAV genome and sg mRNAs is G, hence GTP is required to initiate their synthesis; 3'-end of the EAV genome is polyadenylated, hence UTP is required to initiate synthesis of minus-strand templates. Also, it would be consistent with RdRp-based phylogenetic clustering of nidoviruses with RNA viruses that employ protein-primed RNA synthesis and 3C(L) proteases [26, 53, 54]. However, it would be difficult to reconcile this role with the presence of a cap structure at the 5'-end of nidovirus mRNAs [32, 55, 56], as the priming protein would have to be removed prior to capping. Besides, primase-dependent and *de novo* mechanisms of RNA synthesis initiation in nidoviruses were proposed [57, 58], although they were challenged most recently, when coronavirus nsp8 was shown to possess oligo(U)-templated 3'-terminal adenylyltransferase, rather than template-dependent RNA polymerase activity [59].

The functional role of NiRAN in the virus life cycle could be informed by the function of its homologs. However, we were unable to identify NiRAN homologs in a carefully controlled large-scale analysis that detected homologs for all other domains universally conserved in nidoviruses. This result also left the evolutionary origin of NiRAN uncertain. Nidovirus-wide conservation of NiRAN, as well as the apparent absence of NiRAN homologs in other viruses, indicated that it is a genetic marker of the order *Nidovirales*, only the second after the previously discovered ZBD (**chapter 3**). NiRAN is the only major replicative enzyme known to be exclusively associated with a large monophyletic group of (+)ssRNA viruses [60].

Four years after our study reporting the NiRAN domain discovery (**chapter 3**) was completed, the structure of the SARS-CoV nsp12, first of this protein, shed new light upon the NiRAN domain [61]. The N-terminus of SARS-CoV nsp12 up to NiRAN motif B_N is not visible in the structure and is likely to be highly flexible, while the structure of the NiRAN C-terminus, as well as the downstream Interface and RdRp domains were resolved. Comparison of the partial SARS-CoV NiRAN structure with a database of available protein structures identified significant although limited structural similarity between the NiRAN domain and the protein kinases (Z-score 7.9, RMSD 3.3 Å for the top hit with tyrosine kinase JAK1 structure 6C7Y, chain A). The strongest similarity was observed in the kinase nucleotide-binding site. The authors of the nsp12 structure noted that several of the kinase active site residues aligned with identical, highly conserved residues of NiRAN [61], while others did not have such counterparts in NiRAN either due to the respective residues variation or their structure not being solved. The authors concluded that the reported NiRAN nucleotidylation activity is compatible with the kinase-like fold of the domain, as observed for human pseudokinase SelO [61, 62].

In light of these new findings, we revisited our sequence-based analysis that led to the delineation of the NiRAN domain described in **chapter 3.** Inspection of the original hit list showed that the third best HHsearch hit of the CoToMeRoAr NiRAN profile (combined NiRAN sequences of all nidovirus subfamilies known at the time) in the Pfam database was with Pkinase profile (PF00069), representing protein kinases. Statistical support of the hit was far weaker than common thresholds of significance (Probability=21% vs 95%, E-value=160 vs 0.001), indicating that if NiRAN and protein kinases are indeed homologs, as the SARS-CoV nsp12 structural study suggests, they may have diverged beyond reliable recognition by most advanced sequence-based methods. Such pronounced divergence would be most compatible with conservation of the structural fold but emergence of a new function, often associated with replacement of otherwise key conserved residues.

Unlike the structural comparison, which was limited to the resolved structure of the NiRAN C-terminus (downstream of motif B_N) [61], the HHsearch hit covered all three major motifs of NiRAN (Fig. 3). This hit length, 109 match columns, was outstanding among the other (poorly supported) hits of this profile-database comparison, the mean length of which was 27 match columns. Compared to protein kinases, NiRAN domains of ExoN-encoding large nidoviruses included a large insertion that separated motifs B_N and C_N. Absolutely conserved Lys of NiRAN A_N motif, identified as the most likely target of nucleotidylation in **chapter 3**, aligned with signature Lys of protein kinase sub-domain II. This residue helps to anchor and orient nucleotide, used by kinase as a phosphate donor, by forming ionic bonds with its α- and β- phosphates. Absolutely conserved Glu of NiRAN A_N motif aligned with signature Glu of protein kinase sub-domain III, which helps to stabilize interactions between signature Lys of sub-domain II and nucleotide [63-65]. NiRAN motif B_N including invariant Arg residue and highly conserved Ser/Thr and Asp residues (**chapters 3, 4**) mapped to a protein kinase region of relatively low conservation, indicative of NiRAN-specific conservation and function.

N-terminal half of NiRAN C_N motif aligned with catalytic protein kinase sub-domain VIB, although its signature DxxxxN motif was not conserved in NiRAN: Asp of the signature, believed to catalyze the phosphotransfer reaction, mapped to a variable NiRAN residue, while Asn of the signature mapped to a highly conserved Asn of NiRAN, that is nevertheless substituted by Asp in several arteriviruses (Fig. 3, Fig. S1 from chapter 3) [63-65]. C-terminal half of NiRAN C_N motif aligned with protein kinase sub-domain VII. Conserved signature of sub-domain VII, tripeptide Asp-Phe-Gly, whose Asp residue is believed to chelate metal ion necessary to orient the γ-phosphate of the nucleotide for transfer, aligned to an absolutely conserved dipeptide of NiRAN, Asp-Phe, and a third residue that is a strictly conserved Gly in corona- and toroviruses, and a strictly conserved Glu in arteri-, mesoni- and roniviruses (Fig. 3, Fig. S1 from chapter 3) [63-65].

The similarity between NiRAN and protein kinases, observed in the HHsearch hit, doesn't extend to include upstream sub-domain I of protein kinases, characterized by a signature GxGxxG motif and responsible for covering and anchoring nontransferable phosphates of nucleotide cofactor, or downstream sub-domains VIII – XI, which include subdomain VIII, characterized by signature Ala-Pro-Glu tripeptide and responsible for recognition of peptide substrate [63-65]. These subdomains may either be absent or diverged beyond recognition in nidovirus nsp9/nsp12.

Thus, NiRAN and protein kinases may share elements of the kinase fold and some of its functionally important residues, which are associated with nucleotide binding, but not the Asp residue key for the phosphotransferase activity of kinases. Unless NiRAN adopted other residues to compensate for the lack of the catalytic Asp, the domain is unlikely to be a bona fide protein kinase. On the other hand, its nucleotidyltransferase activity seems to be compatible with the conservation of the nucleotide-binding site in the NiRAN and kinase domains.

Could identification of protein kinases as plausible NiRAN homologs prompt revision of the NiRAN assignment as a marker of the order *Nidovirales* (**chapter 3**)? Indeed, some may argue that protein kinases are encoded by other viruses: large DNA viruses [66], as well as some toroviruses (Fig. 2 from chapter 1) [29, 67]. On the other hand, we believe that unique properties of NiRAN, including its extremely divergent sequence, unparalleled association with RdRp structurally and functionally, and a place within the nidovirus domain synteny, clearly separate NiRAN from other homologs.

Synteny of key replicative domains remains the most conserved marker of nidoviruses

With the identification of the NiRAN domain (**chapter 3**), one hallmark of nidoviruses – universally conserved replicative domains encoded in a certain order (synteny) – expanded to include five domains: 3CLpro-NiRAN-RdRp-ZBD-HEL1. They remain one of the few characteristics that readily distinguish the ever-growing diversity of nidoviruses from other viruses.

While remaining under a strong purifying selection, these domains may have accepted rare or unique substitutions of key residues, revealing adaptation of the associated functions in most divergent nidoviruses. Namely, PSCNV, prototyping a nidovirus suborder [2], contains a number of remarkable substitutions in three domains of the synteny, 3CLpro, NiRAN and RdRp (**chapter 4**). A substrate pocket of PSCNV 3CLpro contains Val residue in place of His residue absolutely conserved in other nidoviruses; the substitution was predicted to confer an unusual substrate specificity to the enzyme (**chapter 4**) [68]. PSCNV NiRAN has a substitution in one out of the seven residues absolutely conserved in

other nidoviruses (**chapters 3, 4**). PSCNV RdRp has a Gly-Asp-Asp signature in its catalytic motif C, instead of Ser-Asp-Asp signature characteristic for nidoviruses (**chapter 4**) [28].

In agreement with their nidovirus-wide conservation, all domains of the synteny, when tested in experiments, proved to be essential for nidovirus replication (**chapter 3**) [69-72]. Nidovirus synteny of invariably encoded replicative domains is one of the very few conserved replicative domain architectures that accommodate the enormous diversity of (+)ssRNA viruses [60].

Genomes of nidoviruses can be far larger than previously believed

Following the sequencing of the entire 31.4 kb MHV genome in 1991 [73], the largest RNA virus genome known at the time, the apparent upper genome size limit increased only slightly over the years of nidovirus discovery, reaching a plateau (Fig. 1A from chapter 4). By the time this study started (end of 2014), the 33.5 kb genome of torovirus BPNV was the largest RNA virus genome known [29], and it seemed that the genome size of ~35 kb may represent the natural limit of the RNA virus genome size [7]. However, recent discoveries of two novel nidoviruses, PSCNV (**chapter 4**) and AAbV [19, 20], challenged this notion. While the AAbV genome size is just above 35 kb: 35.9 kb (7.4% larger than BPNV), PSCNV has a giant genome size by RNA virus standards: 41.1 kb (22.9% larger than BPNV, 14.5% larger than AAbV).

Several factors are believed to restrict the RNA virus genome size, including the fragility of RNA molecules, the selective advantage provided by the fast replication of small genomes, as well as error-prone replication of RNA viruses, that is believed to have a potential to cause error catastrophe in longer genomes [74]. Nidoviruses with large genomes uniquely possess a proofreading enzyme, exoribonuclease, that reduces the error rate of replication [27, 50, 75, 76]. It can be hypothesized that newly discovered nidoviruses with extremely large genomes have acquired properties that allowed them to decrease the error rate of replication even further, permitting them to maintain longer genomes, and increasing their capacity for adaptation. Interestingly, nidoviruses with the two longest known genomes, AAbV and PSCNV, infect exotic hosts, a flatworm and a mollusk, respectively (Fig. 1). It is tempting to suggest that unknown host factors may also play a role in the viability of these extraordinary long genomes.

The genome expansion of nidoviruses was described by a theoretical model [77]. According to the model, the nidovirus genome expansion was dominated by a consecutive expansion of genome regions responsible for genome replication (ORF1b), genome expression (ORF1a) and genome dissemination (3'ORFs region) in the course of the nidovirus evolution. The model, which considers extant nidoviruses as "frozen" at different points along the evolutionary trajectory, predicted that further expansion of the

nidovirus genome would be initiated by the expansion of ORF1b [77]. Notably, the ORF1b-like genome region of PSCNV is 9.8 kb, which is considerably larger than the largest ORF1b region of a previously known nidovirus – 8.2 kb ORF1b of human coronavirus OC43 [78]. This region size increase is especially remarkable because, unlike sizes of ORF1a and 3'ORFs, the size of nidovirus ORF1b is tightly constrained (Fig. 8A from chapter 4). When accounting for genome and region size variation, the increase of ORF1b-like region size in PSCNV was shown to be almost ten times greater than what would be expected if sizes of all genome regions in PSCNV increased uniformly (Fig. 8B from chapter 4). This increase of the ORF1b-like region size corroborates the nidovirus genome expansion model, and thus may indicate that nidoviruses with even larger genomes can be discovered in the future.

INNOVATION IN NIDOVIRUS GENOMES: DUPLICATION AND GENE ACQUISITION

Most large-scale evolutionary changes in nidoviruses can be attributed to aberrant homologous and non-homologous recombination, the mechanisms behind deletions, duplications and gene acquisitions [79, 80]. These evolutionary events are most frequently observed in the two regions of nidovirus genome controlling nidovirus-host interactions: pre-TM2 region of ORF1a and 3'ORFs. Several notable examples of deletions, duplications and gene acquisitions, mapping to these genome regions, were described in recent years, including some in our studies (**chapters 2, 4**).

May tandem repeats be common in ORF1a of nidoviruses?

One of the most common mechanisms of genome and protein innovation is the generation of tandem repeats. Possibly due to fast evolution, adjacent and highly similar tandem repeats were rarely observed in the genomes of RNA viruses, and reported only in a single nidovirus, betacoronavirus HCoV-HKU1, prior to 2014 [81, 82].

A recent study uncovered tandem repeats in nsp3 of a gammacoronavirus called duck-dominant coronavirus (DdCoV; classified as species *Duck coronavirus 2714* [DuCoV_2714] by ICTV), in a position similar to that of HCoV-HKU1 repeats. Four analyzed isolates of DdCoV all harbored five almost-identical copies of a 23 aa charged residue-rich repeat in nsp3 [83].

As shown in this thesis, the arterivirus ORF1a pre-TM2 region also contains repeats positioned in close proximity to each other: three copies of the PxPxPR motif, separated by ~10 aa, were identified within the HVR domain of EAV and WPDV (**chapter 2**). At least one copy of this motif was also found within the Hinge or HVR domain of almost all other arteriviruses (Fig. 3B from chapter 2). PxPxPR motifs may be recognized by cellular Src homology 3 (SH3) protein domains, implicated in signal transduction [84]. The same

function was previously suggested for the canonical SH3-binding motifs PxxP detected in the nsp2 sequence of PRRSV-1 [85]. Given the small size of PxPxPR motifs and their scattered position within the fast-evolving Hinge and HVR domains of arteriviruses, they might have emerged by either point mutation fixed by selection, or duplication followed by diversification.

Also, we described two types of tandem repeats in the newly discovered invertebrate nidovirus PSCNV (**chapter 4**). Two tandem repeats of 67 and 66 aa, separated by 3 aa and sharing 41.1% identity, were found in the pre-TM2 genome region. No homologs of these repeats, which could have pointed to their function, were identified. Further, PSCNV encodes an array of at least three tandem ankyrin repeats in the 3'ORFs-like region of its genome; their origin and function is discussed below.

Complex diversification of PLP paralogs in arteriviruses

Gene duplication is commonly followed by diversification of repeats, which could be advantageous for the virus [86]. Multiple PLP domains are present in the majority of known vertebrate nidoviruses and were one of the first recognized duplications in RNA viruses [87]. Their analysis in arteriviruses (**chapter 2**) offers an insight into the process of repeats diversification and its implications. Our study involved 14 arterivirus species recognized by ICTV as of 2016 [1], which included the most divergent arterivirus, WPDV, important for understanding limits of divergence within this family. We relied on previous research regarding the organization and function of the pp1a/1ab N-terminus in five arterivirus species (for review see [88-91]), as well as on resolved tertiary structures of PRRSV-2 PLP1a and PLP1b, and EAV PLP2 [92-94].

The number of PLP domains encoded in the pp1a/1ab N-terminus of 14 arterivirus species was found to vary from three to four (Fig. 3 from chapter 2), which were referred to as PLP1a, PLP1b, PLP1c and PLP2, based on the order of encoding.

In line with expectations, we found very limited sequence similarity between predicted PLP1 domains of arteriviruses and the WPDV pp1a that was restricted to the immediate vicinity of their catalytic residues and led to tentative identification of three PLP domains. The PLP1a domain of WPDV was predicted to be proteolytically inactive, like its EAV counterpart [95], as both enzymes lack catalytic Cys residue. Since WPDV and EAV do not form a monophyletic lineage, it remains uncertain whether this loss of the catalytic residue was independent or not in two viruses. Regardless of being enzymatically active or not, PLP1a of all arteriviruses retain catalytic His residue that is part of a unique motif (HxxxxxF). This unusual pattern of conservation, involving active and defective enzymes, suggests a distinct noncatalytic function for this residue.

In contrast to WPDV, PLP1 domains of non-WPDV arteriviruses exhibited a considerable degree of sequence and/or structural similarity. Their PLP1a domain is associated with the N-terminal zinc-finger domain, not found in WPDV. Sequence similarity between profiles of PLP1a and PLP1b or PLP1c was low, as was structural similarity between PLP1a and PLP1b of PRRSV-2. PLP1b and PLP1c were found to share a significant sequence similarity; surprisingly, EAV PLP1b exhibited a stronger similarity towards PLP1c, rather than PLP1b sequences. Notably, most residues conserved in PLP1a and PLP1b/c of non-WPDV arteriviruses mapped to the C- and N-terminal subdomains of the PLP structure, respectively (Fig. 11 from chapter 2), which is consistent with different functional specializations of these domains.

The PLP2 domain was shown to be universally conserved in all 14 arterivirus species, while sequence similarity between PLP2 and PLP1 domains was below commonly accepted thresholds of statistical significance for profile-profile analysis. Likewise, structural similarity between PRRSV PLP1a/1b and PLP2 was not statistically significant.

The established pattern of inter-domains similarity between different PLPs and arteriviruses allows to suggest an evolutionary scenario that might have led to the emergence of PLP arrays in arteriviruses. Barring a minor possibility that PLP1 domains were acquired by ancestors of the WPDV and non-WPDV arterivirus lineages independently, the MRCA of these viruses may have already encoded one, two or three PLP1 domains and a PLP2 domain. Presence of one ancestral PLP1 domain in the MRCA of known arteriviruses would imply that arrays of PLP1 domains were generated in WPDV and non-WPDV arteriviruses independently, and the weak similarity observed between the respective PLP1 domains of the two lineages is a result of parallel evolution (variant of convergence). An alternative scenario would be the presence of two or three PLP1 domains in the arterivirus MRCA, with ancestral PLP1a already bearing its distinguishing features, lack of catalytic Cys and a unique HxxxxxF motif involving catalytic His. Subsequent evolution of PLP1 domains in different lineages might have involved duplications followed by diversification, deletions, and convergence. Ancestors of non-WPDV PLP1a and PLP1b/c likely emerged as a result of a duplication that occurred prior to the existence of MRCA of non-WPDV arteriviruses. The MRCA of non-WPDV arteriviruses might have possessed a single ancestral PLP1b/c domain, which would imply a subsequent duplication of the domain in the non-EAV lineage. Alternatively, the MRCA of non-WPDV arteriviruses might have possessed ancestral PLP1b and PLP1c, with ortholog of PLP1b being lost in the EAV lineage.

Thus, an array of orthologous PLP domains of arteriviruses emerged as a result of a broad range of evolutionary events, where multiple PLP duplications were followed by

diversification, loss and possibly convergence of orthologous domains, involved loss of catalytic activity and different functional specialization of orthologs.

Newly described nidovirus includes repeats of uncertain origins

While the evolutionary origin of the tandem repeats described above can be tentatively ascribed to duplication, the PSCNV genome encodes at least two pairs of repeats separated by considerable distances (**chapter 4**), whose evolutionary origin remains uncertain.

The potential leader and body TRSs of PSCNV (see above) are ~60 nt long, share 86% sequence identity, and are separated by 28,327 nt. They might have emerged as a result of duplication, but incremental extension of the similar regions by point mutations fixed by selection (convergence), associated with genome expansion, seems to be more likely. A shorter ancestral genome might have already had TRSs at the respective positions in the genome, as is typical for nidoviruses. Expansion of the genome would have necessitated TRSs extension: short motifs identical to TRSs can be encountered in a long genome just by chance, jeopardizing genome expression. Consequently, gradual expansion of the genome could have created evolutionary pressure, causing identical sequences of TRSs to gradually extend along with the genome through convergence mechanism. Expansion of virus sampling in the PSCNV clade could help in resolving evolutionary history of this intriguing similarity involving exceptionally long TRS-like elements.

In addition, PSCNV encodes two highly divergent fibronectin type 2 (FN2) domains separated by 1,572 aa in the polyprotein. Besides duplication in ancestral virus, FN2 domains might have been acquired independently from unknown sources, since they are only remotely similar and located far apart. The possible function of the two FN2 domains is discussed below.

Newly described nidovirus acquired domains rarely or never observed in viruses

Another major source of genome and protein innovation is domain acquisition from other species. Many domains found in subsets of nidoviruses may have been acquired through this mechanism (see Introduction). Our study of PSCNV was particularly insightful in this respect since it expanded the previously known proteome repertoire of nidoviruses or even larger groups of viruses (**chapter 4**). Besides encoding orthologs of canonical replicative domains of nidoviruses, PSCNV was found to also encode such domains as ribonuclease T2 (RNase T2), two fibronectin type 2 domains, and an ankyrin repeats domain (ANK).

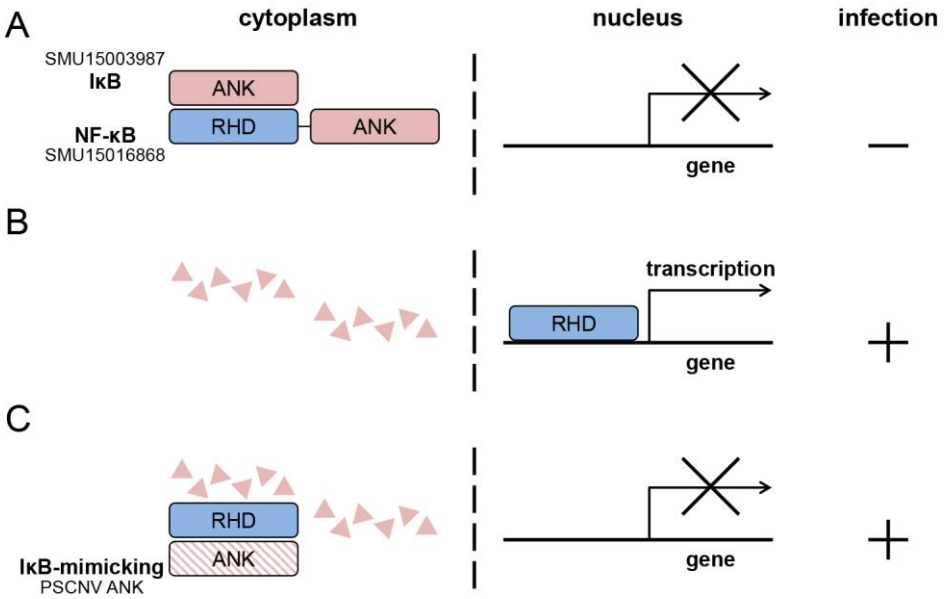


Figure 4 | Proposed roles of PSCNV ANK and its host homologs in modulation of antiviral immune response. (**A**) In the absence of an inducing signal, NF-κB protein (SMU15016868) resides in the cytoplasm, bound by inhibitors: its own ANK domain and protein IκB (SMU15003987). (**B**) In response to viral infection, inhibitors are degraded, allowing the NF-κB transcription factor to enter the nucleus and modulate gene expression to promote antiviral immune response. (**C**) IκB-mimicking viral protein (PSCNV ANK) may retain the NF-κB transcription factor in the cytoplasm after its inhibitors were degraded, thus downregulating the immune response.

The RNase T2 homolog is encoded in the pre-TM2 region of the giant PSCNV ORF. RNases T2 cleave ssRNA in an acidic environment, and are encoded by a broad range of cellular organisms, as well as two groups of viruses, (+)ssRNA pestiviruses and dsDNA polydnaviruses [96]. Viral RNases T2 were implicated in modulating host immune response [97, 98], and we proposed a similar function for PSCNV RNase T2.

Also, we identified two FN2 domains in the PSCNV polyprotein, which are encoded far downstream from RNase T2 in a 3'ORFs-like region. Never before found in viruses, FN2 domains are common in vertebrates and invertebrates as modules of multidomain proteins involved in diverse processes [99, 100]; when studied they were found responsible for protein-protein interactions: binding of gelatin and collagen [101, 102]. We speculated that FN2 domains of PSCNV could also bind collagen. Since PSCNV was found to infect mucus-producing cells, PSCNV FN2 domains might help to adhere PSCNV virus particles to the collagen-containing mucus excreted by the host [103], facilitating spread of the virus among host population.

The two FN2 domains flank the PSCNV polyprotein region that includes the ANK domain, another mediator of protein-protein interactions. The ANK domain is ubiquitous in proteins of diverse cellular organisms, and dsDNA viruses with large genomes, but was not detected in proteins of RNA viruses before [104].

The PSCNV RNase T2, FN2 domains and the ANK domain were likely acquired from other viruses or hosts via non-homologous recombination. Due to the lack of close homologs, evolutionary origins of RNase T2 and FN2 domains in PSCNV remain unknown. In contrast, the PSCNV ANK domain clusters confidently with ANK domains of a pair of host proteins, SMU15016868 and SMU15003987, indicating that the ANK domain might have been acquired from an ancestor of the host, flatworm *Schmidtea mediterranea*.

Further analysis of the domain architecture of these host homologs led us to a hypothesis about the possible functional role of PSCNV ANK (**chapter 4**) in a striking parallel with a process documented for several dsDNA viruses [105-107]. Namely, the host ANK-containing proteins have domain architectures suggestive of their interaction: SMU15016868 is characteristic for NF-κB protein, N-RHD-ANK-C (RHD is a Rel homology domain), while SMU15003987 is characteristic for its inhibitor IκB, N-ANK-C [108]. Based on studies of several viruses [105, 106], the NF-κB protein is expected to reside in the cytoplasm, bound by inhibitors, its own ANK domain and protein IκB, in the absence of a viral infection (Fig. 4A). A viral infection would trigger degradation of NF-κB inhibitors, allowing NF-κB transcription factor to enter the nucleus and modulate gene expression to promote an antiviral immune response (Fig. 4B). Thus, we proposed PSCNV ANK to act as a IκB-mimicking protein, retaining a NF-κB transcription factor in the cytoplasm after the degradation of its inhibitors, and thus downregulating the immune response (Fig. 4C).

FUTURE PERSPECTIVES

Comparative genomics reveals patterns of natural variation, forming the basis for evolutionary hypotheses that inform experimental research. This thesis contains multiple examples of the connection between comparative genomics and bench. Experimental characterization of the novel, universally conserved domain of nidoviruses, NiRAN, was inspired by a hypothesis about its RNA ligase activity, formulated based solely on comparative genomics data (**chapter 3**), while hypotheses and models suggested in **chapters 2 and 4** create a basis for future experimental research. In **chapter 2**, methods of comparative genomics allowed to make several predictions about the N-termini of arteriviral polyproteins: the position of enzymatically active PLP domains of fourteen arteriviruses, few of which were previously characterized in this respect, as well as potential nsp2 PRF and SH3-binding sites were identified. In **chapter 4**, an extremely divergent and unusual nidovirus, PSCNV, was extensively analyzed by methods of comparative genomics, leading to hypotheses about various aspects of its biology, including the nature of replicative domains and structural proteins that it encodes,

mechanisms of differential protein expression that it employs, and ways of evading host immune defenses that it uses. Besides, existence of a related virus infecting a flatworm *Planaria torva* was also predicted in **chapter 4**. All these predictions could benefit from experimental verification.

The explosive growth of the number and diversity of newly discovered nidoviruses, which has been observed in recent years, is likely to continue and accelerate even more in the future. These developments will bring new insights about the biology and evolution of nidoviruses, but may also present a challenge. The unprecedented influx of new divergent nidovirus genome sequences calls for the development of tools allowing to reliably classify them [9, 10], and to detect homology despite their enormous genetic divergence (**chapter 5**). Importantly, nidovirus genomes that will be discovered in the future can also be instrumental in verifying and advancing hypotheses formulated in this thesis. For example, the discovery of a sister virus for WPDV (**chapter 2**) or PSCNV (**chapter 4**), which are separated from the currently known nidoviruses by long genetic distances, would make it possible to test hypotheses about these viruses by analyzing the conservation of their predicted functional genome and proteome elements.

REFERENCES

1. Adams MJ, Lefkowitz EJ, King AM, Harrach B, Harrison RL, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Mushegian AR *et al*: **Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2016)**. *Arch Virol* 2016, **161**:2921-2949.
2. 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.
3. Gorbalenya AE, Brinton MA, de Groot RJ, Gulyaeva AA, Lauber C, Neuman BW, Ziebuhr J: **Pending ICTV taxonomic proposal 2019.023S Create five new families and a new suborder of vertebrate viruses in the order Nidovirales.** 2019.
4. Brinton MA, Gulyaeva AA, Balasuriya UBR, Dunowska M, Faaberg KS, Goldberg T, Leung F-C, Nauwynck HJ, Snijder EJ, Stadejek T *et al*: **Pending ICTV taxonomic proposal 2019.020S Create one new genus (Nuarterivirus); move the existing subgenus Pedartevirus to the genus Iotaarterivirus; rename one species from the subgenus Pedartevirus; create one new species in the new genus Nuarterivirus; create one new subgenus and two new species in the existing genus Betaarterivirus.** 2019.
5. Ziebuhr J, Baker S, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Neuman BW, Perlman S, Poon LLM *et al*: **Pending ICTV taxonomic proposal 2019.021S Create ten new species and a new genus in the subfamily Orthocoronavirinae of the family Coronaviridae and five new species and a new genus in the subfamily Serpentovirinae of the family Tobaniviridae.** 2019.
6. Gorbalenya AE, Gulyaeva AA, Hobson-Peters J, Junglen S, Morita K, Sawabe K, Vasilakis N, Ziebuhr J: **Pending ICTV taxonomic proposal 2019.022S Create one new species in the genus Alphamesonivirus of the family Mesoniviridae and one new species in the genus Okavirus of the family Roniviridae.** 2019.
7. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS *et al*: **Redefining the invertebrate RNA virosphere**. *Nature* 2016, **540**:539-543.
8. Adams MJ, Lefkowitz EJ, King AM, Harrach B, Harrison RL, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Mushegian AR *et al*: **50 years of the International Committee on Taxonomy of Viruses: progress and prospects**. *Arch Virol* 2017, **162**(5):1441-1446.
9. Lauber C, Gorbalenya AE: **Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses**. *J Virol* 2012, **86**(7):3890-3904.
10. Lauber C, Gorbalenya AE: **Toward genetics-based virus taxonomy: comparative analysis of a genetics-based classification and the taxonomy of picornaviruses**. *J Virol* 2012, **86**(7):3905-3915.

11. Brinton MA, Gulyaeva AA, Balasuriya UBR, Dunowska M, Faaberg KS, Leung FC, Nauwynck HJ, Snijder EJ, Stadejek T, Gorbalenya AE: **ICTV taxonomic proposal 2015.014a-cS In the family Arteriviridae create 10 species (1 unassigned, 9 in the genus Arterivirus) and rename one species**. 2015.

12. Ziebuhr J, Baric RS, Baker S, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Neuman BW, Perlman S *et al*: **ICTV taxonomic proposal 2015.003a-eS Create 12 species in the family Coronaviridae**. 2015.

13. Gorbalenya AE, Gulyaeva AA, Hobson-Peters J, Junglen S, Morita K, Sawabe K, Vasilakis N, Ziebuhr J: **ICTV taxonomic proposal 2015.004a,bS In the family Mesoniviridae, create four species in genus Alphamesonivirus and two unassigned in the family**. 2015.

14. Gorbalenya AE, Brinton MA, Cowley J, de Groot R, Gulyaeva A, Lauber C, Neuman B, Ziebuhr J: **ICTV taxonomic proposal 2017.015S Reorganization and expansion of the order Nidovirales at the family and sub-order ranks**. 2017.

15. Brinton MA, Gulyaeva A, Balasuriya UBR, Dunowska M, Faaberg KS, Goldberg T, Leung FC-C, Nauwynck HJ, Snijder EJ, Stadejek T *et al*: **ICTV taxonomic proposal 2017.012S Expansion of the rank structure of the family Arteriviridae and renaming its taxa**. 2017.

16. Ziebuhr J, Baric RS, Baker S, de Groot RJ, Drosten C, Gulyaeva A, Haagmans BL, Neuman BW, Perlman S, Poon LLM *et al*: **ICTV taxonomic 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 subfamily Torovirinae and three other subfamilies), revision of the genus rank structure and introduction of a new subgenus rank**. 2017.

17. Gorbalenya AE, Brinton MA, Cowley J, de Groot R, Gulyaeva A, Lauber C, Neuman B, Ziebuhr J: **ICTV taxonomic proposal 2017.014S Establishing taxa at the ranks of subfamily, genus, sub-genus and species in six families of invertebrate nidoviruses**. 2017.

18. Dunowska M, Biggs PJ, Zheng T, Perrott MR: **Identification of a novel nidovirus associated with a neurological disease of the Australian brushtail possum (Trichosurus vulpecula)**. *Vet Microbiol* 2012, **156**(3-4):418-424.

19. 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.

20. 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.

21. Zondag LE, Rutherford K, Gemmell NJ, Wilson MJ: **Uncovering the pathways underlying whole body regeneration in a chordate model, Botrylloides leachi using de novo transcriptome analysis**. *BMC Genomics* 2016, **17**:114.

22. Söding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**(7):951-960.

23. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment**. *Nat Methods* 2012, **9**(2):173-175.

24. Gorbalenya AE, Lieutaud P, Harris MR, Coutard B, Canard B, Kleywegt GJ, Kravchenko AA, Samborskiy DV, Sidorov IA, Leontovich AM *et al*: **Practical application of bioinformatics by the multidisciplinary VIZIER consortium**. *Antiviral Res* 2010, **87**(2):95-110.

25. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ: I**Q-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. *Mol Biol Evol* 2015, **32**(1):268-274.

26. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM: **Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis**. *Nucleic Acids Res* 1989, **17**(12):4847-4861.

27. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, 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**(5):991-1004.

28. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ: **Nidovirales: evolving the largest RNA virus genome**. *Virus Res* 2006, **117**(1):17-37.

29. Stenglein MD, Jacobson ER, Wozniak EJ, Wellehan JF, Kincaid A, Gordon M, Porter BF, Baumgartner W, Stahl S, Kelley K *et al*: **Ball python nidovirus: a candidate etiologic agent for severe respiratory disease in Python regius**. *MBio* 2014, **5**(5):e01484-01414.

30. den Boon JA, Snijder EJ, Chirnside ED, de Vries AA, Horzinek MC, Spaan WJ: **Equine arteritis virus is not a togavirus but belongs to the coronaviruslike superfamily**. *J Virol* 1991, **65**(6):2910-2920.

31. Cowley JA, Dimmock CM, Walker PJ: **Gill-associated nidovirus of Penaeus monodon prawns transcribes 3'-coterminal subgenomic mRNAs that do not possess 5'-leader sequences**. *J Gen Virol* 2002, **83**(Pt 4):927-935.

32. van Vliet AL, Smits SL, Rottier PJ, de Groot RJ: **Discontinuous and non-discontinuous subgenomic RNA transcription in a nidovirus**. *EMBO J* 2002, **21**(23):6571-6580.

33. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L *et al*: **The evolutionary history of vertebrate RNA viruses**. *Nature* 2018, **556**:197-202.

34. Firth AE, Brierley I: **Non-canonical translation in RNA viruses**. *J Gen Virol* 2012, **93**(Pt 7):1385-1409.

35. Snijder EJ, Meulenberg JJ: **The molecular biology of arteriviruses**. *J Gen Virol* 1998, **79 (Pt 5)**:961-979.

36. Molenkamp R, van Tol H, Rozier BC, van der Meer Y, Spaan WJ, Snijder EJ: **The arterivirus replicase is the only viral protein required for genome replication and subgenomic mRNA transcription**. *J Gen Virol* 2000, **81**(Pt 10):2491-2496.

37. Archambault D, Kheyar A, de Vries AA, Rottier PJ: **The intraleader AUG nucleotide sequence context is important for equine arteritis virus replication**. *Virus Genes* 2006, **33**(1):59-68.

38. Brian DA, Baric RS: **Coronavirus genome structure and replication**. *Curr Top Microbiol Immunol* 2005, **287**:1-30.

39. Wu HY, Guan BJ, Su YP, Fan YH, Brian DA: **Reselection of a genomic upstream open reading frame in mouse hepatitis coronavirus 5'-untranslated-region mutants**. *J Virol* 2014, **88**(2):846-858.

40. Irigoyen N, Firth AE, Jones JD, Chung BY, Siddell SG, Brierley I: **High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling**. *PLoS Pathog* 2016, **12**(2):e1005473.

41. Stewart H, Brown K, Dinan AM, Irigoyen N, Snijder EJ, Firth AE: **Transcriptional and Translational Landscape of Equine Torovirus**. *J Virol* 2018, **92**(17).

42. Fang Y, Treffers EE, Li Y, Tas A, Sun Z, van der Meer Y, de Ru AH, van Veelen PA, Atkins JF, Snijder EJ *et al*: **Efficient -2 frameshifting by mammalian ribosomes to synthesize an additional arterivirus protein**. *Proc Natl Acad Sci U S A* 2012, **109**(43):E2920-E2928.

43. Li Y, Treffers EE, Napthine S, Tas A, Zhu L, Sun Z, Bell S, Mark BL, van Veelen PA, van Hemert MJ *et al*: **Transactivation of programmed ribosomal frameshifting by a viral protein**. *Proc Natl Acad Sci U S A* 2014, **111**(21):E2172-E2181.

44. Napthine S, Treffers EE, Bell S, Goodfellow I, Fang Y, Firth AE, Snijder EJ, Brierley I: **A novel role for poly(C) binding proteins in programmed ribosomal frameshifting**. *Nucleic Acids Res* 2016, **44**(12):5491-5503.

45. Hodgman TC: **A new superfamily of replicative proteins**. *Nature* 1988, **333**(6168):22-23.

46. Gorbalenya AE, Koonin EV: **Birnavirus RNA polymerase is related to polymerases of positive strand RNA viruses**. *Nucleic Acids Res* 1988, **16**(15):7735.

47. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM: **A novel superfamily of nucleoside triphosphate-binding motif containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination**. *FEBS Lett* 1988, **235**(1-2):16-24.

48. Cowley JA, Dimmock CM, Spann KM, Walker PJ: **Gill-associated virus of Penaeus monodon prawns: an invertebrate virus with ORF1a and ORF1b genes related to arteri- and coronaviruses**. *J Gen Virol* 2000, **81**(Pt 6):1473-1484.

49. Zirkel F, Kurth A, Quan PL, Briese T, Ellerbrok H, Pauli G, Leendertz FH, Lipkin WI, Ziebuhr J, Drosten C *et al*: **An insect nidovirus emerging from a primary tropical rainforest**. *MBio* 2011, **2**(3):e00077-00011.

50. Nga PT, Parquet MC, Lauber C, Parida M, Nabeshima T, Yu F, Thuy NT, Inoue S, Ito T, Okamoto K *et al*: **Discovery of the first insect nidovirus, a missing evolutionary link in the emergence of the largest RNA virus genomes**. *PLoS Pathog* 2011, **7**(9):e1002215.

51. Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, Janssen GM, 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**(17):8416-8434.

52. Ivanov KA, Hertzig T, Rozanov M, Bayer S, Thiel V, Gorbalenya AE, Ziebuhr J: **Major genetic marker of nidoviruses encodes a replicative endoribonuclease**. *Proc Natl Acad Sci U S A* 2004, **101**(34):12694-12699.

53. Gorbalenya AE, Pringle FM, Zeddam JL, Luke BT, Cameron CE, Kalmakoff J, Hanzlik TN, Gordon KH, Ward VK: **The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage**. *J Mol Biol* 2002, **324**(1):47-62.

54. Wolf YI, Kazlauskas D, Iranzo J, Lucia-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV: **Origins and Evolution of the Global RNA Virome**. *MBio* 2018, **9**(6).

55. 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**(2):557-565.

56. Sagripanti JL, Zandomeni RO, Weinmann R: **The cap structure of simian hemorrhagic fever virion RNA**. *Virology* 1986, **151**(1):146-150.

57. Imbert I, Guillemot JC, Bourhis JM, Bussetta C, Coutard B, Egloff MP, Ferron F, Gorbalenya AE, Canard B: **A second, non-canonical RNA-dependent RNA polymerase in SARS coronavirus**. *EMBO J* 2006, **25**(20):4933-4942.

58. Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, Snijder EJ, Canard B, Imbert I: **One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities**. *Proc Natl Acad Sci U S A* 2014, **111**(37):E3900-3909.

59. Tvarogova J, Madhugiri R, Bylapudi G, Ferguson LJ, Karl N, Ziebuhr J: **Identification and Characterization of a Human Coronavirus 229E Nonstructural Protein 8-Associated RNA 3'-Terminal Adenylyltransferase Activity**. *J Virol* 2019, **93**(12).

60. Gorbalenya AE, Koonin EV: **Comparative analysis of amino-acid sequences of key enzymes of replication and expression of positive-strand RNA viruses: validity of approach and functional and evolutionary implications**. *Sov Sci Rev D Physicochem Biol* 1993, **11**:1-84.

61. Kirchdoerfer RN, Ward AB: **Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors**. *Nat Commun* 2019, **10**(1):2342.

62. Sreelatha A, Yee SS, Lopez VA, Park BC, Kinch LN, Pilch S, Servage KA, Zhang J, Jiou J, Karasiewicz-Urbanska M *et al*: **Protein AMPylation by an Evolutionarily Conserved Pseudokinase**. *Cell* 2018, **175**(3):809-821 e819.

63. Hanks SK, Hunter T: **Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification**. *FASEB J* 1995, **9**(8):576-596.

64. Hanks SK: **Genomic analysis of the eukaryotic protein kinase superfamily: a perspective**. *Genome Biol* 2003, **4**(5):111.

65. Zhao Z, Jin Q, Xu JR, Liu H: **Identification of a fungi-specific lineage of protein kinases closely related to tyrosine kinases**. *PLoS One* 2014, **9**(2):e89813.

66. Jacob T, Van den Broeke C, Favoreel HW: **Viral serine/threonine protein kinases**. *J Virol* 2011, **85**(3):1158-1173.

67. Tokarz R, Sameroff S, Hesse RA, Hause BM, Desai A, Jain K, Lipkin WI: **Discovery of a novel nidovirus in cattle with respiratory disease**. *J Gen Virol* 2015, **96**(8):2188-2193.

68. Ziebuhr J, Snijder EJ, Gorbalenya AE: **Virus-encoded proteinases and proteolytic processing in the Nidovirales**. *J Gen Virol* 2000, **81**(Pt 4):853-879.

69. Seybert A, Posthuma CC, van Dinten LC, Snijder EJ, Gorbalenya AE, Ziebuhr J: **A complex zinc finger controls the enzymatic activities of nidovirus helicases**. *J Virol* 2005, **79**(2):696-704.

70. te Velthuis AJ, van den Worm SH, Sims AC, Baric RS, Snijder EJ, van Hemert MJ: **Zn(2+) inhibits coronavirus and arterivirus RNA polymerase activity in vitro and zinc ionophores block the replication of these viruses in cell culture**. *PLoS Pathog* 2010, **6**(11):e1001176.

71. van Dinten LC, van Tol H, Gorbalenya AE, Snijder EJ: **The predicted metal-binding region of the arterivirus helicase protein is involved in subgenomic mRNA synthesis, genome replication, and virion biogenesis**. *J Virol* 2000, **74**(11):5213-5223.

72. van Dinten LC, Rensen S, Gorbalenya AE, Snijder EJ: **Proteolytic processing of the open reading frame 1b-encoded part of arterivirus replicase is mediated by nsp4 serine protease and Is essential for virus replication**. *J Virol* 1999, **73**(3):2027-2037.

73. Lee HJ, Shieh CK, Gorbalenya AE, Koonin EV, La Monica N, Tuler J, Bagdzhadzhyan A, Lai MM: **The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative proteases and RNA polymerase**. *Virology* 1991, **180**(2):567-582.

74. Holmes EC: **The Evolution and Emergence of RNA Viruses.** New York: Oxford University Press; 2009.

75. 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**(22):12135-12144.

76. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR: **Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics**. *PLoS Pathog* 2013, **9**(8):e1003565.

77. Lauber C, Goeman JJ, Parquet MC, 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**(7):e1003500.

78. St-Jean JR, Jacomy H, Desforges M, Vabret A, Freymuth F, Talbot PJ: **Human respiratory coronavirus OC43: genetic stability and neuroinvasion**. *J Virol* 2004, **78**(16):8824-8834.

79. Lai MM: **RNA recombination in animal and plant viruses**. *Microbiol Rev* 1992, **56**(1):61-79.

80. Simon-Loriere E, Holmes EC: **Why do RNA viruses recombine?** *Nat Rev Microbiol* 2011, **9**(8):617-626.

81. Woo PC, Lau SK, Chu CM, Chan KH, Tsoi HW, Huang Y, Wong BH, Poon RW, Cai JJ, Luk WK *et al*: **Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia**. *J Virol* 2005, **79**(2):884-895.

82. Woo PC, Lau SK, Yip CC, Huang Y, Tsoi HW, Chan KH, Yuen KY: **Comparative analysis of 22 coronavirus HKU1 genomes reveals a novel genotype and evidence of natural recombination in coronavirus HKU1**. *J Virol* 2006, **80**(14):7136-7145.

83. Zhuang QY, Wang KC, Liu S, Hou GY, Jiang WM, Wang SC, Li JP, Yu JM, Chen JM: **Genomic Analysis and Surveillance of the Coronavirus Dominant in Ducks in China**. *PLoS One* 2015, **10**(6):e0129256.

84. Kowanetz K, Szymkiewicz I, Haglund K, Kowanetz M, Husnjak K, Taylor JD, Soubeyran P, Engstrom U, Ladbury JE, Dikic I: **Identification of a novel proline-arginine motif involved in CIN85-dependent clustering of Cbl and down-regulation of epidermal growth factor receptors**. *J Biol Chem* 2003, **278**(41):39735-39746.

85. Ropp SL, Wees CE, Fang Y, Nelson EA, Rossow KD, Bien M, Arndt B, Preszler S, Steen P, Christopher-Hennings J *et al*: **Characterization of emerging European-**

like porcine reproductive and respiratory syndrome virus isolates in the United States. *J Virol* 2004, **78**(7):3684-3703.

86. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models**. *Nat Rev Genet* 2010, **11**(2):97-108.

87. Gorbalenya AE, Koonin EV, Lai MM: **Putative papain-related thiol proteases of positive-strand RNA viruses. Identification of rubi- and aphthovirus proteases and delineation of a novel conserved domain associated with proteases of rubi-, alpha- and coronaviruses**. *FEBS Lett* 1991, **288**(1-2):201-205.

88. Snijder EJ, Kikkert M, Fang Y: **Arterivirus molecular biology and pathogenesis**. *J Gen Virol* 2013, **94**(Pt 10):2141-2163.

89. Nedialkova DD, Gorbalenya AE, Snijder EJ: **Arterivirus Papain-like Proteinase 1a**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2199-2204.

90. Nedialkova DD, Gorbalenya AE, Snijder EJ: **Arterivirus Papain-like Proteinase 1ß**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2205-2210.

91. Kikkert M, Snijder EJ, Gorbalenya AE: **Arterivirus nsp2 Cysteine Proteinase**. In: *Handbook of Proteolytic Enzymes.* Edited by Rawlings ND, Salvesen GS, vol. 2, 3 edn. London: Academic Press; 2013: 2210-2215.

92. Sun Y, Xue F, Guo Y, Ma M, Hao N, Zhang XC, Lou Z, Li X, Rao Z: **Crystal structure of porcine reproductive and respiratory syndrome virus leader protease Nsp1alpha**. *J Virol* 2009, **83**(21):10931-10940.

93. Xue F, Sun Y, Yan L, Zhao C, Chen J, Bartlam M, Li X, Lou Z, Rao Z: **The crystal structure of porcine reproductive and respiratory syndrome virus nonstructural protein Nsp1beta reveals a novel metal-dependent nuclease**. *J Virol* 2010, **84**(13):6461-6471.

94. van Kasteren PB, Bailey-Elkin BA, James TW, Ninaber DK, Beugeling C, Khajehpour M, Snijder EJ, Mark BL, Kikkert M: **Deubiquitinase function of arterivirus papain-like protease 2 suppresses the innate immune response in infected host cells**. *Proc Natl Acad Sci U S A* 2013, **110**(9):E838-E847.

95. den Boon JA, Faaberg KS, Meulenberg JJ, Wassenaar AL, Plagemann PG, Gorbalenya AE, Snijder EJ: **Processing and evolution of the N-terminal region of the arterivirus replicase ORF1a protein: identification of two papainlike cysteine proteases**. *J Virol* 1995, **69**(7):4500-4505.

96. Luhtala N, Parker R: **T2 Family ribonucleases: ancient enzymes with diverse roles**. *Trends Biochem Sci* 2010, **35**(5):253-259.

97. Krey T, Bontems F, Vonrhein C, Vaney MC, Bricogne G, Rumenapf T, Rey FA: **Crystal structure of the pestivirus envelope glycoprotein E(rns) and mechanistic analysis of its ribonuclease activity**. *Structure* 2012, **20**(5):862-873.

98. Park B KY: **Immunosuppression induced by expression of a viral RNase enhances susceptibility of Plutella xylostella to microbial pesticides.** *Insect Science* 2012, **19**(1):47-54.

99. Ozhogina OA, Trexler M, Banyai L, Llinas M, Patthy L: **Origin of fibronectin type II (FN2) modules: structural analyses of distantly-related members of the kringle family idey the kringle domain of neurotrypsin as a potential link between FN2 domains and kringles**. *Protein Sci* 2001, **10**(10):2114-2122.

100. Chalmers IW, Hoffmann KF: **Platyhelminth Venom Allergen-Like (VAL) proteins: revealing structural diversity, class-specific features and biological associations across the phylum**. *Parasitology* 2012, **139**(10):1231-1245.

101. Napper CE, Drickamer K, Taylor ME: **Collagen binding by the mannose receptor mediated through the fibronectin type II domain**. *Biochem J* 2006, **395**(3):579-586.

102. Tam EM, Moore TR, Butler GS, Overall CM: **Characterization of the distinct collagen binding, helicase and cleavage mechanisms of matrix metalloproteinase 2 and 14 (gelatinase A and MT1-MMP): the differential roles of the MMP hemopexin c domains and the MMP-2 fibronectin type II modules in collagen triple helicase activities**. *J Biol Chem* 2004, **279**(41):43336-43344.

103. Bocchinfuso DG, Taylor P, Ross E, Ignatchenko A, Ignatchenko V, Kislinger T, Pearson BJ, Moran MF: **Proteomic profiling of the planarian Schmidtea mediterranea and its mucous reveals similarities with human secretions and those predicted for parasitic flatworms**. *Mol Cell Proteomics* 2012, **11**(9):681-691.

104. Al-Khodor S, Price CT, Kalia A, Abu KY: **Functional diversity of ankyrin repeats in microbial proteins**. *Trends Microbiol* 2010, **18**(3):132-139.

105. Falabella P, Varricchio P, Provost B, Espagne E, Ferrarese R, Grimaldi A, de EM, Fimiani G, Ursini MV, Malva C *et al*: **Characterization of the IkappaB-like gene family in polydnaviruses associated with wasps belonging to different Braconid subfamilies**. *J Gen Virol* 2007, **88**(Pt 1):92-104.

106. Tait SW, Reid EB, Greaves DR, Wileman TE, Powell PP: **Mechanism of inactivation of NF-kappa B by a viral homologue of I kappa b alpha. Signal-induced release of i kappa b alpha results in binding of the viral homologue to NF-kappa B**. *J Biol Chem* 2000, **275**(44):34656-34664.

107. Camus-Bouclainville C, Fiette L, Bouchiha S, Pignolet B, Counor D, Filipe C, Gelfi J, Messud-Petit F: **A virulence factor of myxoma virus colocalizes with NF-kappaB in the nucleus and interferes with inflammation**. *J Virol* 2004, **78**(5):2510-2516.

108. Gilmore TD, Wolenski FS: **NF-kappaB: where did it come from and why?** *Immunol Rev* 2012, **246**(1):14-35.

SUMMARY

The order *Nidovirales* is a monophyletic group of positive-sense single-stranded RNA viruses that infect vertebrate and invertebrate hosts, and include viruses with largest RNA genomes. A set of hallmark characteristics distinguish nidoviruses from other RNA viruses: genome organization, mechanisms of genome expression, a synteny of conserved replicative domains. Only a few selected nidoviruses are subject of comprehensive experimental research. At the same time, the advent of next generation sequencing has greatly accelerated the rate of nidovirus discovery. As a result, genome sequence is the only characteristic available for a large and ever growing share of nidoviruses. These developments determine the key role of comparative genomics in further nidovirus characterization. Comparative genomics identifies homologous regions of genomes and proteins, facilitating evolutionary studies, and functional and structural characterization of newly discovered and already known viruses. Specifically, it promotes transfer of functional annotation from experimentally characterized viruses and hosts to newly discovered virus genomes, and defines constraints of natural variation for all viruses, including experimentally characterized. In this thesis, we used comparative genomics to characterize various aspects of nidovirus biology and evolution. This study was conducted in collaboration with other researchers, who discovered new viruses and sequenced their genomes (Chapters 2 and 4), or characterized virus proteins experimentally following bioinformatics sequence analysis (Chapter 3). **Chapter 1** provides background on nidoviruses and techniques of comparative genomics available by the end of 2014, when the project that resulted in this thesis started. **Chapter 2** describes characterization of arterivirus polyprotein 1ab N-terminus encoding multiple papain-like proteases. The analysis relied on previous research on this region and included 5'-terminus of the divergent wobbly possum disease virus genome, sequencing of which was completed as part of the study. The study offers insight into the role and contribution of gene duplication to nidovirus adaptation. **Chapter 3** presents discovery of the fifth replicative domain universally conserved in all nidoviruses, nidovirus RdRp-associated nucleotidyltransferase or NiRAN. NiRAN conservation in nidoviruses, its evolutionary origin, biochemical activity and potential function were analyzed. **Chapter 4** focuses on discovery and characterization of a highly divergent nidovirus with the largest known RNA genome, planarian secretory cell nidovirus or PSCNV. Both unique and conserved features of its genome, proteome and expression were revealed in this study. Moreover, PSCNV discovery advanced our understanding of RNA genome expansion limits. **Chapter 5** addresses an important technical challenge of nidovirus comparative genomics. Proteomes of RNA viruses, including nidoviruses, are dominated by large multidomain polyproteins, although standard tools for homology detection were trained on single-domain proteins. Consequently, homologous relationships of domains in polyproteins may

remain undetected due to underestimation of hits statistical significance. To mitigate this problem, we introduced a tool, called LArge Multidomain Protein Annotator or LAMPA, that gradually splits polyprotein sequence into smaller queries in a biologically reasonable manner, improving estimation of hits statistical significance and annotation coverage. **Chapter 6** discusses how discoveries of recent years, including the ones described in this thesis, advanced our understanding of two fundamental aspects of nidovirus biology. First, we reexamine nidovirus hallmarks, prompted by discovery of novel and divergent nidoviruses and their bioinformatics analysis. Second, we review new insights into the mechanisms of large scale sequence change in nidovirus genomes, which have the largest RNA genome size range.

SAMENVATTING

De orde *Nidovirales* is een monofyletische groep van positief-sense enkelstrengige RNA-virussen die gewervelde en ongewervelde gastheren infecteren en die virussen met de grootste RNA-genomen omvatten. Een aantal kenmerken onderscheidt nidovirussen van andere RNA-virussen: genoomorganisatie, mechanismen van genoomexpressie en syntenie van geconserveerde replicatie domeinen. Slechts enkele nidovirussen zijn onderwerp van uitgebreid experimenteel onderzoek. Tegelijkertijd heeft de opkomst van ”next generation sequencing” de snelheid waarmee nieuwe nidovirussen worden ontdekt aanzienlijk verhoogd. Als gevolg hiervan is de genoomsequentie de enige beschikbare eigenschap voor de karakterisatie van een steeds groter aantal nidovirussen. Deze ontwikkelingen bepalen de sleutelrol die vergelijkende genomics speelt bij verdere karakterisering van nidovirussen. Vergelijkende genomics identificeert homologe regio's van genomen en eiwitten, waardoor evolutionaire studies en functionele- en structurele karakterisering van nieuw ontdekte en reeds bekende virussen worden vereenvoudigd. In het bijzonder bevordert het de overdracht van functionele annotatie van experimenteel gekarakteriseerde virussen en gastheren naar nieuw ontdekte virus genomen en definieert het beperkingen van natuurlijke variatie voor alle virussen, inclusief experimenteel gekarakteriseerde. In dit proefschrift hebben we vergelijkende genomics gebruikt om verschillende aspecten van de biologie en evolutie van nidovirussen te karakteriseren. Deze studie is uitgevoerd in samenwerking met andere onderzoekers, die nieuwe virussen hebben ontdekt en daarvan genoomsequenties hebben bepaald (hoofdstukken 2 en 4), of viruseiwitten experimenteel hebben gekarakteriseerd op basis van sequentie analyse en bio-informatica voorspellingen (hoofdstuk 3). **Hoofdstuk 1** geeft achtergrondinformatie over nidovirussen en technieken van vergelijkende genomics die eind 2014 beschikbaar waren, toen dit project van start ging. **Hoofdstuk 2** beschrijft de karakterisatie van de N-terminus van arterivirus polyproteïne 1ab, die codeert voor meerdere papaïne-achtige proteasen. Deze analyse is gebaseerd op eerder onderzoek naar deze regio en omvat de 5'-terminus van het enigszins verwante wobbly possum disease virus genoom, waarvan de sequentie werd bepaald als onderdeel van de studie. De studie biedt inzicht in de rol en bijdrage van gen duplicatie aan nidovirus-aanpassing. **Hoofdstuk 3** presenteert de ontdekking van het vijfde replicatieve domein dat geconserveerd is in alle nidovirussen, het nidovirus RdRp-geassocieerde nucleotidyltransferase of NiRAN. De conservering van NiRAN in nidovirussen, de evolutionaire oorsprong ervan, de biochemische activiteit en potentiële functie werden geanalyseerd. **Hoofdstuk 4** richt zich op de ontdekking en karakterisering van een zeer afwijkend nidovirus met het grootste bekende RNA-genoom, het planarian secretory cell nidovirus of PSCNV. Zowel unieke als geconserveerde kenmerken van het genoom, proteoom en expressie zijn in deze studie aangetoond. Bovendien heeft de ontdekking van

PSCNV ons begrip van de grenzen van RNA-genoomuitbreiding verbeterd. **Hoofdstuk 5** gaat in op een belangrijke technische uitdaging van vergelijkende genomics van nidovirussen. Het proteoom van RNA-virussen, waaronder nidovirussen, worden gedomineerd door grote multidomein polyproteïnen, terwijl standaardtools voor homologiedetectie worden getraind op eiwitten met een enkel domein. Derhalve kunnen homologe verhoudingen tussen domeinen in polyproteïnen onopgemerkt blijven vanwege een onderschatting van de statistische significantie van hits. Om dit probleem te verminderen, hebben we een tool geïntroduceerd, LArge Multidomain Protein Annotator, of LAMPA genaamd, die polyproteïnesequenties geleidelijk opsplitst in kleinere zoekopdrachten op een biologisch relevante manier, waardoor de schatting van de statistische significantie van hits en de annotatiedekking worden verbeterd. **Hoofdstuk 6** bespreekt hoe ontdekkingen van de afgelopen jaren, inclusief degene die in dit proefschrift zijn beschreven, ons begrip van twee fundamentele aspecten van de nidovirusbiologie hebben verbeterd. Eerst bekijken we de nidovirus-kenmerken opnieuw, aangespoord door de ontdekking van nieuwe en uiteenlopende nidovirussen en hun bioinformatische-analyse. Ten tweede bespreken we nieuwe inzichten in de mechanismen die ten grondslag liggen aan grote sequentie veranderingen in nidovirussen, waarvan, vergeleken met andere RNA virus ordes, de lengte van het RNA genoom het meest kan variëren.

LIST OF ABBREVIATIONS

(-)ssRNA	negative-sense single-stranded RNA
(+)ssRNA	positive-sense single-stranded RNA
2'-PDE	2',5'-phosphodiesterase
3CLpro (3CL^{pro})	3C-like protease
aa	amino acid
AAbV	aplysia abyssovirus 1
AIC	Akaike information criterion
AMP, ADP, ATP	adenosine mono-, di-, triphosphate
ANK	ankyrin domain
APRAV	African pouched rat arterivirus
AsD	arterivirus-specific domain
BIC	Bayesian information criterion
BNV1	Beihai nido-like virus 1
BPNV	ball python nidovirus
BRV	Breda virus
BSA	bovine serum albumin
CAVV	Cavally virus
CIP	calf intestine alkaline phosphatase
CMP, CDP, CTP	cytidine mono-, di-, triphosphate
CoV	coronavirus
CPD	cyclic phosphodiesterase
CPE	cytopathic effect
CPU	central processing unit
CR domain	cysteine-rich domain
DdCoV	duck-dominant coronavirus
DEmARC	DivErsity pArtitioning by hieRarchical Clustering

DeMAV	De Brazza's monkey arterivirus
DNA	deoxyribonucleic acid
dsRNA	double-stranded RNA
E	nidovirus envelope protein
EAV	equine arteritis virus
EM	electron microscopy
ER	endoplasmic reticulum
EToV	equine torovirus
EVD	extreme value distribution
ExoN	DEDDh subfamily exoribonuclease
FN2	fibronectin type II domain
FSBG	5’-(4-fluorosulfonylbenzoyl)guanosine
GAV	gill-associated virus
GMP, GDP, GTP	guanosine mono-, di-, triphosphate
GTase	guanylyltransferase
HE	hemagglutinin-esterase
HEL1	superfamily 1 helicase
HGT	horizontal gene transfer
HMM	hidden Markov model
HVR	hypervariable region
IBV	infectious bronchitis virus
ICTV	International Committee on Taxonomy of Viruses
InfV	influenza virus
ISH	*in situ* hybridization
kb	kilobase
KRCV	Kibale red colobus virus
KRTGV	Kibale red-tailed guenon virus

LAMPA	LArge Multidomain Protein Annotator
LDV	lactate dehydrogenase-elevating virus
LGT	lateral gene transfer
M	nidovirus matrix protein
MAR	mono-ADP-ribose
MCMC	Markov chain Monte Carlo
MERS	Middle East respiratory syndrome
MHV	mouse hepatitis virus
ML	maximum likelihood
MMP-2	matrix metalloproteinase-2
Mpro (M^{pro})	main protease
MRCA	most recent common ancestor
mRNA	messenger RNA
MSA	multiple sequence alignment
MTase	methyltransferase
N	nidovirus nucleocapsid protein
n.a.	not applicable
n.d.	not done
NAD	nicotinamide adenine dinucleotide
NDiV	Nam Dinh virus
NendoU	uridylate-specific endonuclease
NGS	next generation sequencing
NiRAN	nidovirus RdRp-associated nucleotidyltransferase
NMP, NDP, NTP	nucleoside mono-, di-, triphosphate
N-MT	SAM-dependent N7-methyltransferase
nsp	non-structural protein
nt	nucleotide

O-MT	SAM-dependent 2'-O-methyltransferase
ORF	open reading frame
p.i.	post infection
p.t.	post transfection
PAR	poly-ADP-ribose
PBJV	Pebjah virus
PCBP	poly(C) binding protein
PDB	Protein Data Bank
Pkinase	protein kinase
PLP	papain-like protease
polyA	polyadenylate
pp	polyprotein
PPD	pairwise patristic distance
PRF	programmed ribosomal frameshifting
PRRSV	porcine reproductive and respiratory syndrome virus
PSCNV	planarian secretory cell nidovirus
PSSM	position-specific scoring matrix
PV	poliovirus
RdRp	RNA-dependent RNA polymerase
RHD	Rel homology domain
(RLM) RACE	(RNA ligase-mediated) rapid amplification of cDNA ends
RMSD	root mean square deviation
RNA	ribonucleic acid
RNase T2	ribonuclease T2
RNP	RNA-protein
RsD	ronivirus-specific domain
RTC	replication-transcription complex

RTPase	RNA 5’-triphosphotase
S	nidovirus spike protein
SAM	S-adenosyl methionine
SARS	severe acute respiratory syndrome
SD	standard deviation
sg	subgenomic
SH3 domain	Src homology 3 domain
SHEV	simian hemorrhagic encephalitis virus
SHFV	simian hemorrhagic fever virus
SI	standard inoculum
SPase	signal peptidase
SPR	subtree pruning and regrafting
SUD	“SARS-unique” domain
TAP	tobacco acid pyrophosphatase
TGEV	transmissible gastroenteritis virus
TM	transmembrane
tRNA	transfer RNA
TRS	transcription-regulating sequence
Ub	ubiquitin
UMP, UDP, UTP	uridine mono-, di-, triphosphate
UTR	untranslated region
WBV	white bream virus
WJHAV	Wuhan Japanese halfbeak arterivirus
WPDV	wobbly possum disease virus
wt	wild-type
ZBD	zinc-binding domain
ZnF	zinc finger

CURRICULUM VITAE

Anastasia Gulyaeva was born on November 6, 1991 in Moscow, Russia (USSR at the time). In June 2009 she graduated from the physico-mathematical lyceum № 1580 in Moscow. In September 2009 Anastasia enrolled in studies at the Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia. In the course of her studies she conducted rotation projects in the research groups of Prof. dr. V.I. Muronetz, Prof. dr. A.A. Mironov and Prof. dr. A.V. Alexeevsky. In July 2012 Anastasia participated in the MoBiLe Bioinformatics Summer School, where she was working on a scientific assignment in the research group of Prof. dr. P.A.C. 't Hoen in the Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. In 2014 Anastasia graduated from the Lomonosov Moscow State University after defending her MSc. thesis dedicated to the usage of sequence weights in the hierarchical classification of viral genomes, and supervised by Dr. A.M. Leontovich, Dr. I.A. Sidorov and Prof. dr. A.E. Gorbalenya. In the same year, she started her doctoral research in the Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands under supervision of Dr. I.A. Sidorov and Prof. dr. A.E. Gorbalenya. Her doctoral research resulted in the present thesis entitled “Comparative genomics of nidoviruses: towards understanding the biology and evolution of the largest RNA viruses”.

LIST OF PUBLICATIONS

Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, **Gulyaeva AA**, Haagmans BL, Lauber C, Leontovich AM, Neuman BW, Penzar D, Perlman S, Poon LLM, Samborskiy DV, Sidorov IA, Sola I, Ziebuhr J: The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 2020, 5:536–544.

Gulyaeva AA, Sigorskih AI[#], Ocheredko ES[#], Samborskiy DV, Gorbalenya AE: LAMPA, LArge Multidomain Protein Annotator, and its application to RNA virus polyproteins. *Bioinformatics* 2020.

Kanitz M, Blanck S, Heine A, **Gulyaeva AA**, Gorbalenya AE, Ziebuhr J, Diederich WE: Structural basis for catalysis and substrate specificity of a 3C-like cysteine protease from a mosquito mesonivirus. *Virology* 2019, 533:21-33.

Nijhuis RHT[#], Sidorov IA[#], Chung PK, Wessels E, **Gulyaeva AA**, de Vries JJ, Claas ECJ, Gorbalenya AE: PCR assays for detection of human astroviruses: In silico evaluation and design, and in vitro application to samples collected from patients in the Netherlands. *J Clin Virol* 2018, 108:83-89.

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.

Saberi A[#], **Gulyaeva AA**[#], Brubacher JL, Newmark PA, Gorbalenya AE: A planarian nidovirus expands the limits of RNA genome size. *PLoS Pathog* 2018, 14(11):e1007314.

Gulyaeva AA[#], Dunowska M[#], Hoogendoorn E, Giles J, Samborskiy D, Gorbalenya AE: Domain Organization and Evolution of the Highly Divergent 5' Coding Region of Genomes of Arteriviruses, Including the Novel Possum Nidovirus. *J Virol* 2017, 91(6).

Lehmann KC, **Gulyaeva AA**, Zevenhoven-Dobbe JC, Janssen GM, 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(17):8416-8434.

Lehmann KC, Hooghiemstra L, **Gulyaeva AA**, Samborskiy DV, Zevenhoven-Dobbe JC, Snijder EJ, Gorbalenya AE, Posthuma CC: Arterivirus nsp12 versus the coronavirus nsp16

2'-O-methyltransferase: comparison of the C-terminal cleavage products of two nidovirus pp1ab polyproteins. *J Gen Virol* 2015, 96(9):2643-2655.

[#]equal contribution

ICTV proposals

Gorbalenya AE, Brinton MA, de Groot RJ, **Gulyaeva AA**, Lauber C, Neuman BW, Ziebuhr J: Pending ICTV taxonomic proposal 2019.023S Create five new families and a new suborder of vertebrate viruses in the order *Nidovirales*. 2019.

Brinton MA, **Gulyaeva AA**, Balasuriya UBR, Dunowska M, Faaberg KS, Goldberg T, Leung F-C, Nauwynck HJ, Snijder EJ, Stadejek T *et al*: Pending ICTV taxonomic proposal 2019.020S Create one new genus (*Nuarterivirus*); move the existing subgenus *Pedartevirus* to the genus *Iotaarterivirus*; rename one species from the subgenus *Pedartevirus*; create one new species in the new genus *Nuarterivirus*; create one new subgenus and two new species in the existing genus *Betaarterivirus*. 2019.

Ziebuhr J, Baker S, Baric RS, de Groot RJ, Drosten C, **Gulyaeva AA**, Haagmans BL, Neuman BW, Perlman S, Poon LLM *et al*: Pending ICTV taxonomic proposal 2019.021S Create ten new species and a new genus in the subfamily *Orthocoronavirinae* of the family *Coronaviridae* and five new species and a new genus in the subfamily *Serpentovirinae* of the family *Tobaniviridae*. 2019.

Gorbalenya AE, **Gulyaeva AA**, Hobson-Peters J, Junglen S, Morita K, Sawabe K, Vasilakis N, Ziebuhr J: Pending ICTV taxonomic proposal 2019.022S Create one new species in the genus *Alphamesonivirus* of the family *Mesoniviridae* and one new species in the genus *Okavirus* of the family *Roniviridae*. 2019.

Gorbalenya AE, Brinton MA, Cowley J, de Groot R, **Gulyaeva AA**, Lauber C, Neuman B, Ziebuhr J: ICTV taxonomic proposal 2017.015S Reorganization and expansion of the order *Nidovirales* at the family and sub-order ranks. 2017.

Brinton MA, **Gulyaeva AA**, Balasuriya UBR, Dunowska M, Faaberg KS, Goldberg T, Leung FC-C, Nauwynck HJ, Snijder EJ, Stadejek T *et al*: ICTV taxonomic proposal 2017.012S Expansion of the rank structure of the family *Arteriviridae* and renaming its taxa. 2017.

Ziebuhr J, Baric RS, Baker S, de Groot RJ, Drosten C, **Gulyaeva AA**, Haagmans BL, Neuman BW, Perlman S, Poon LLM *et al*: ICTV taxonomic 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 subfamily *Torovirinae* and three other subfamilies), revision of the genus rank structure and introduction of a new subgenus rank. 2017.

Gorbalenya AE, Brinton MA, Cowley J, de Groot R, **Gulyaeva AA**, Lauber C, Neuman B, Ziebuhr J: ICTV taxonomic proposal 2017.014S Establishing taxa at the ranks of subfamily, genus, sub-genus and species in six families of invertebrate nidoviruses. 2017.

Brinton MA, **Gulyaeva AA**, Balasuriya UBR, Dunowska M, Faaberg KS, Leung FC, Nauwynck HJ, Snijder EJ, Stadejek T, Gorbalenya AE: ICTV taxonomic proposal 2015.014a-cS In the family *Arteriviridae* create 10 species (1 unassigned, 9 in the genus *Arterivirus*) and rename one species. 2015.

Ziebuhr J, Baric RS, Baker S, de Groot RJ, Drosten C, **Gulyaeva AA**, Haagmans BL, Lauber C, Neuman BW, Perlman S *et al*: ICTV taxonomic proposal 2015.003a-eS Create 12 species in the family *Coronaviridae*. 2015.

Gorbalenya AE, **Gulyaeva AA**, Hobson-Peters J, Junglen S, Morita K, Sawabe K, Vasilakis N, Ziebuhr J: ICTV taxonomic proposal 2015.004a,bS In the family *Mesoniviridae*, create four species in genus *Alphamesonivirus* and two unassigned in the family. 2015.

ACKNOWLEDGEMENTS

I would like to thank all the people who supported me on my PhD journey. First of all, I want to express my gratitude to my promotor Sasha Gorbalenya and co-promotor Igor Sidorov. Sasha, thank you for your guidance, advice, sharing your knowledge and ideas with me, and for all the opportunities that you gave me. Igor, thank you for always being there for me when I had difficulties. I am very grateful to my Moscow colleague Dmitry Samborskiy for finding amazing solutions to the most difficult bioinformatics problems. I would like to thank all my co-authors for pleasant and productive collaborations. Over the past five years, I have been fortunate to supervise several students of the MoBiLe Bioinformatics Summer School: Vanya Kuznetsov, Sveta Iarovenko, Andrey Sigorskih, Lena Ocheredko and Dima Penzar, thank you for your excellent work. I loved this experience and learned a lot from it! I would like to thank my former teachers and mentors from the Lomonosov Moscow State University, my supervisors from the MoBiLe Bioinformatics Summer School, and an MSc. thesis advisor Andrey Mikhailovich Leontovich, all of whom inspired me to pursue career in bioinformatics. I am also very grateful to Louis Kroes and Eric Snijder for the support of my work at the Department of Medical Microbiology, and to all my colleagues for wonderful and stimulating research environment. I would like to thank Ien Dobbelaar, Ineke van Ballegooijen-Molijn, Manon Stijnman, Sophie Greve, Esther Quakkelaar, Annemieke Hofman-Jansen and Marianne Parlevliet-de Gelder for helping me with administrative issues, and Hans van der Geest for maintaining computational environment for my work over these years. I am very grateful to Jeroen Corver and Tim Dalebout for translating thesis summary into Dutch. Finally, from the bottom of my heart, I would like to thank my family, and most importantly, my mum. Without your love, encouragement and support, my PhD journey would be impossible.