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

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].

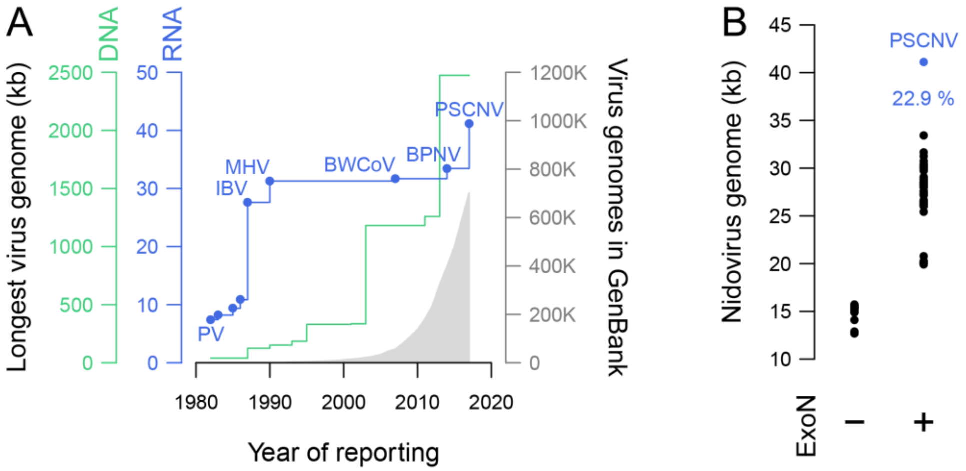

Figure 1 | Genome sizes of nidoviruses. (**A**) Timeline of discovery of largest RNA and DNA virus genomes versus accumulation of virus genome sequences in GenBank (1982–2017). PV, poliovirus; and nidoviruses: IBV, avian bronchitis virus, MHV, mouse hepatitis virus, BWCoV, beluga whale coronavirus SW1, 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

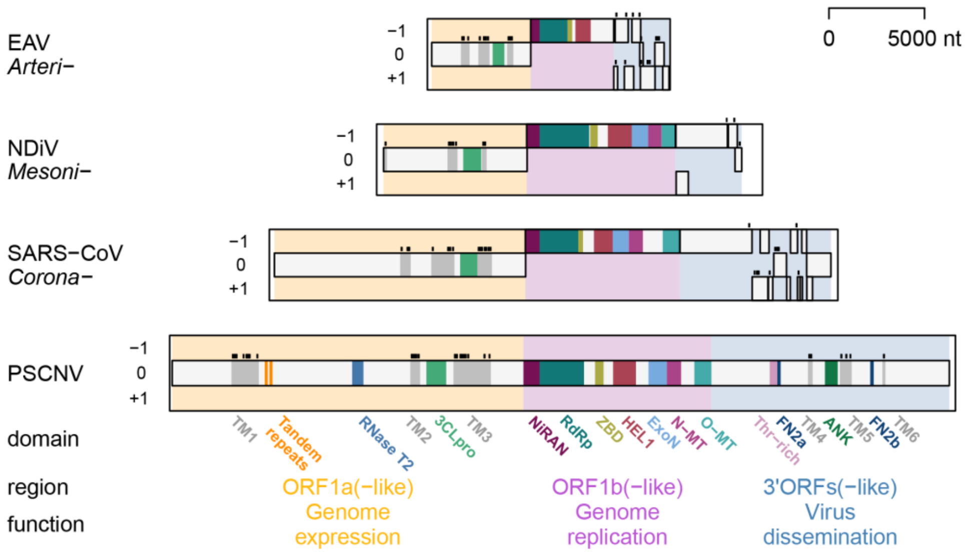

Figure 2 | Genomes and proteomes of nidoviruses. ORFs and encoded protein domains in genomes of viruses representing three nidovirus families and PSCNV. The protein-encoded part of the genomes is split in three adjacent regions, which are colored and labelled accordingly. EAV, equine arteritis virus; 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

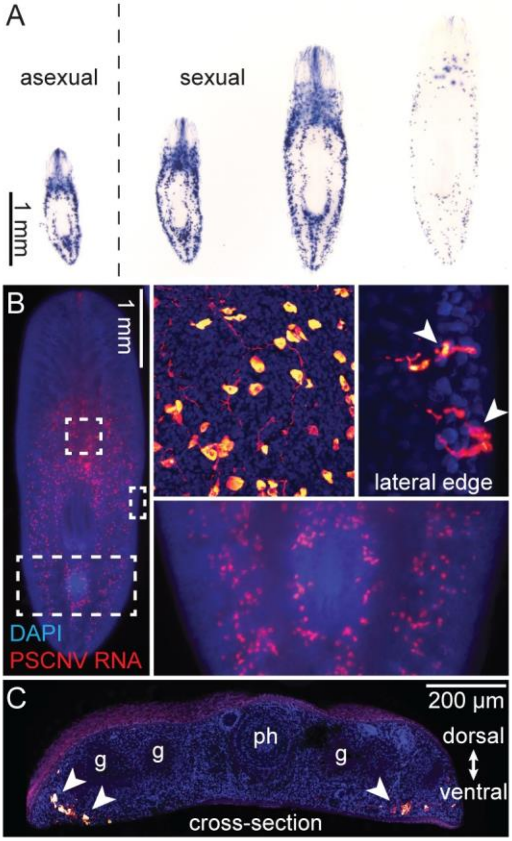

Figure 3 | Expression of PSCNV RNA in planarians. (**A**) PSCNV RNA (blue) detected in asexual (left) and sexual *S. mediterranea* by whole-mount ISH. (**B**) Fluorescent ISH showing PSCNV expression in a sexual planarian. Insets show higher magnification of areas indicated by boxes. Top two insets are 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

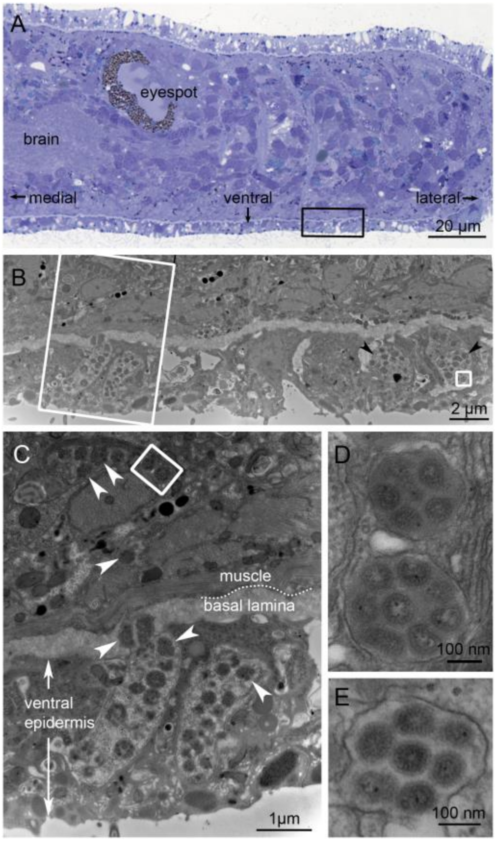

Figure 4 | Putative PSCNV particles revealed by electron microscopy. (**A**) Adjacent histological transverse section, to orient EM images. Black rectangle corresponds to location of (**B**), a low magnification EM view to provide context. White rectangle corresponds to location of (**C**), in which 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.

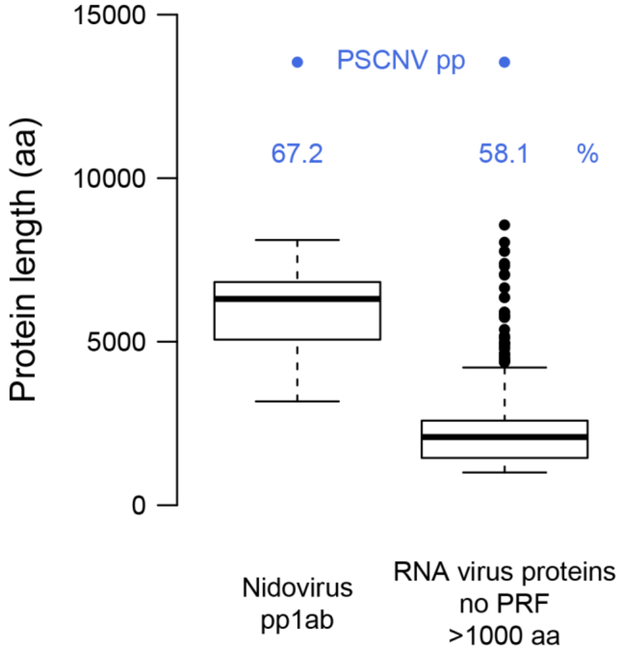

Figure 5 | Largest proteins of nidoviruses and other RNA viruses in comparison with PSCNV polyprotein. Percentage indicates the difference between sizes of the PSCNV polyprotein (pp) and that of the next-largest entity. For details, see SI Materials and Methods.

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.

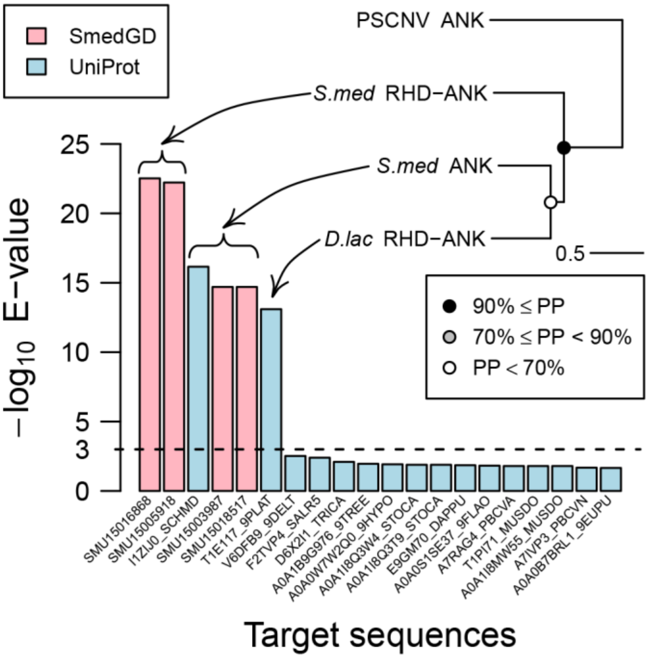

Figure 6 | ANK domain of PSCNV and its homologs. The closest cellular homologs of PSCNV ANK are ranked by similarity (left, above the broken baseline) and depicted through phylogeny (right; reconstructed and rooted by BEAST, summarized as maximum clade credibility tree; PP, posterior probability of clades) 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

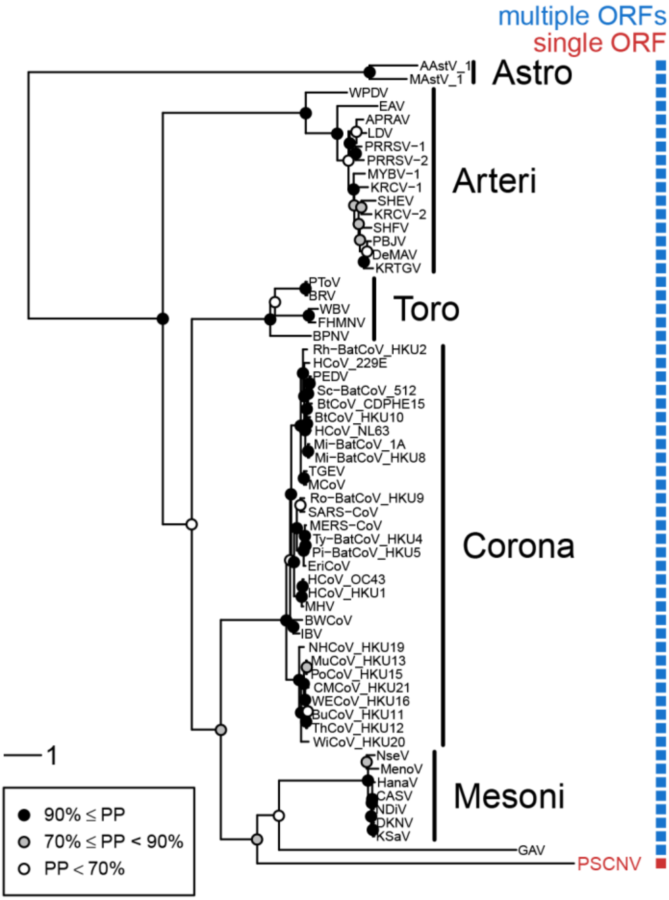

Figure 7 | Phylogeny of PSCNV. RdRp-based Bayesian maximum clade credibility tree and the genomic ORF organization (character state) for PSCNV, a representative set of nidoviruses, and astroviruses (outgroup). PP, posterior probability of clades. For virus names, see Table S1.

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).

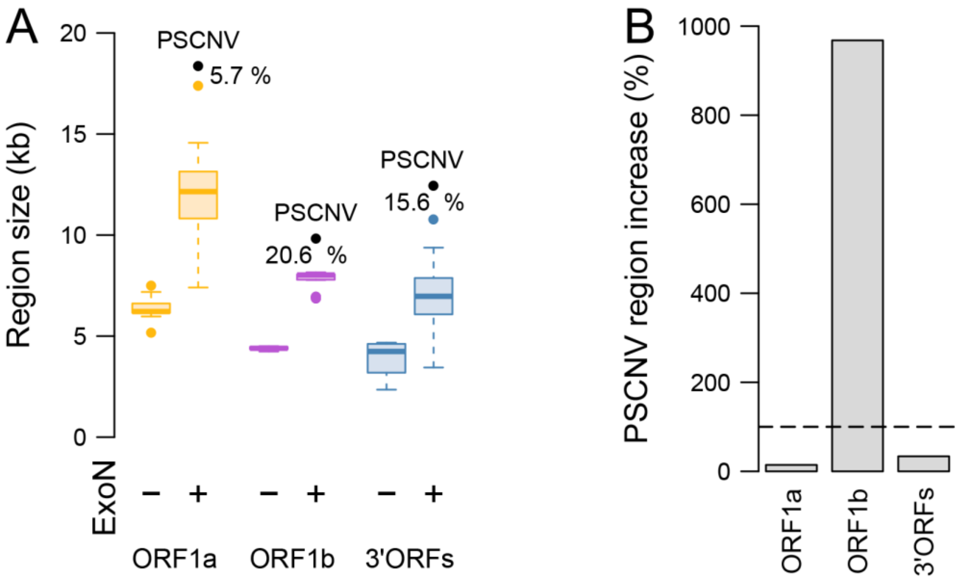

Figure 8 | Nidovirus genome and region size differences. (**A**) Sizes of three nidovirus ORF regions. Percentage indicates the difference between a genome region's size in PSCNV, and that of the next-largest entity. Color scheme as in Fig. 2. (**B**) Size increase of the three genome regions in PSCNV (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

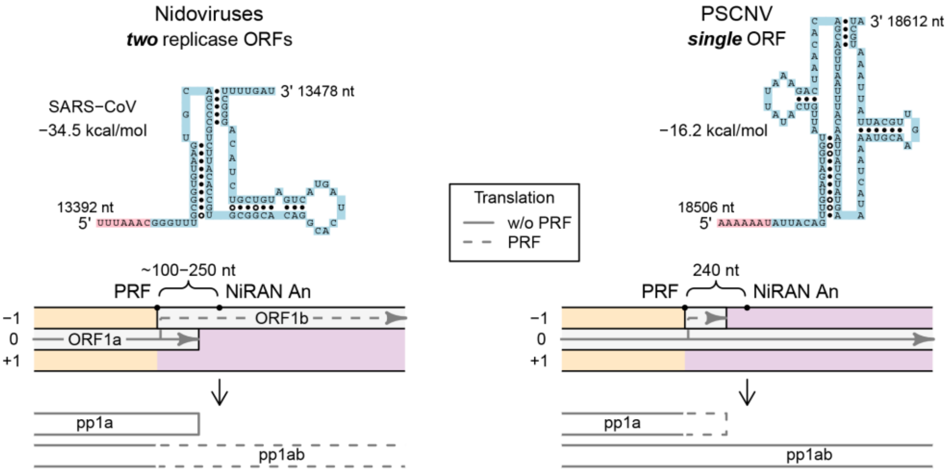

Figure 9 | Genome translation. Comparison of mechanisms by which ORFs 1a and 1b are translated in previously described nidoviruses (left) and PSCNV (right, hypothetical). On the top, RNA structure of the PRF sites, predicted by KnotInFrame, is presented: slippery sequence, pink; pseudoknot, blue.

(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.

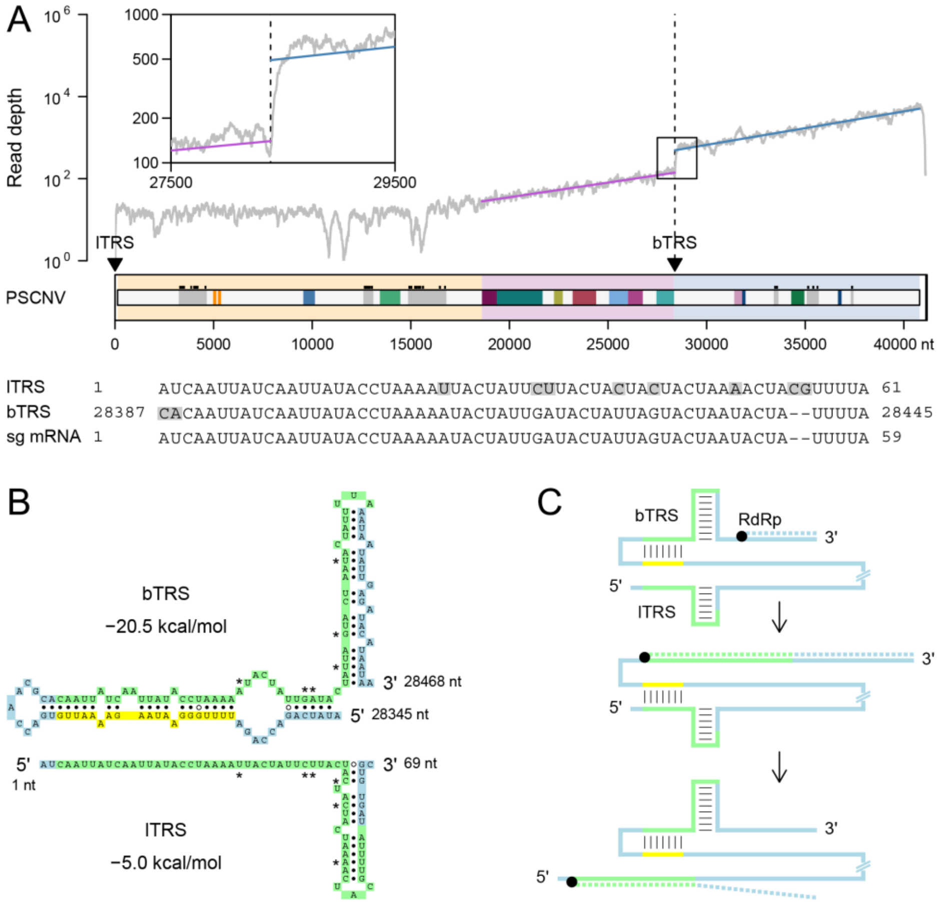
A
Read depth
10^6
10^4
10^2
10^0
1000
500
200
100
27500
29500
lTRS
bTRS
PSCNV
0
5000
10000
15000
20000
25000
30000
35000
40000 nt
lTRS 1 AUCAAUUAUCAAUUAUACCUAAAAUUACUAUUCUUACUACUACUACUAAAACUACGUUUUA 61
bTRS 28387 CACAAUUAUCAAUUAUACCUAAAAAUACUAUUGAUACUAUUAGUACUAAUACUA--UUUUA 28445
sg mRNA 1 AUCAAUUAUCAAUUAUACCUAAAAAUACUAUUGAUACUAUUAGUACUAAUACUA--UUUUA 59
B
bTRS
−20.5 kcal/mol
3' 28468 nt
5' 28345 nt
5'
1 nt
3' 69 nt
lTRS
−5.0 kcal/mol
C
bTRS
RdRp
3'
5'
lTRS
3'
5'
3'
5'

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, $D_1(\text{region})=(p-M)/M*100\%$; 2) the region size variation, $D_2(\text{region})=(p-M)/(M-m)*100\%$; and 3) the region size variation and genome size increase, $D_3(\text{region})=D_2(\text{region})/D_2(\text{genome})*100\%$, 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].

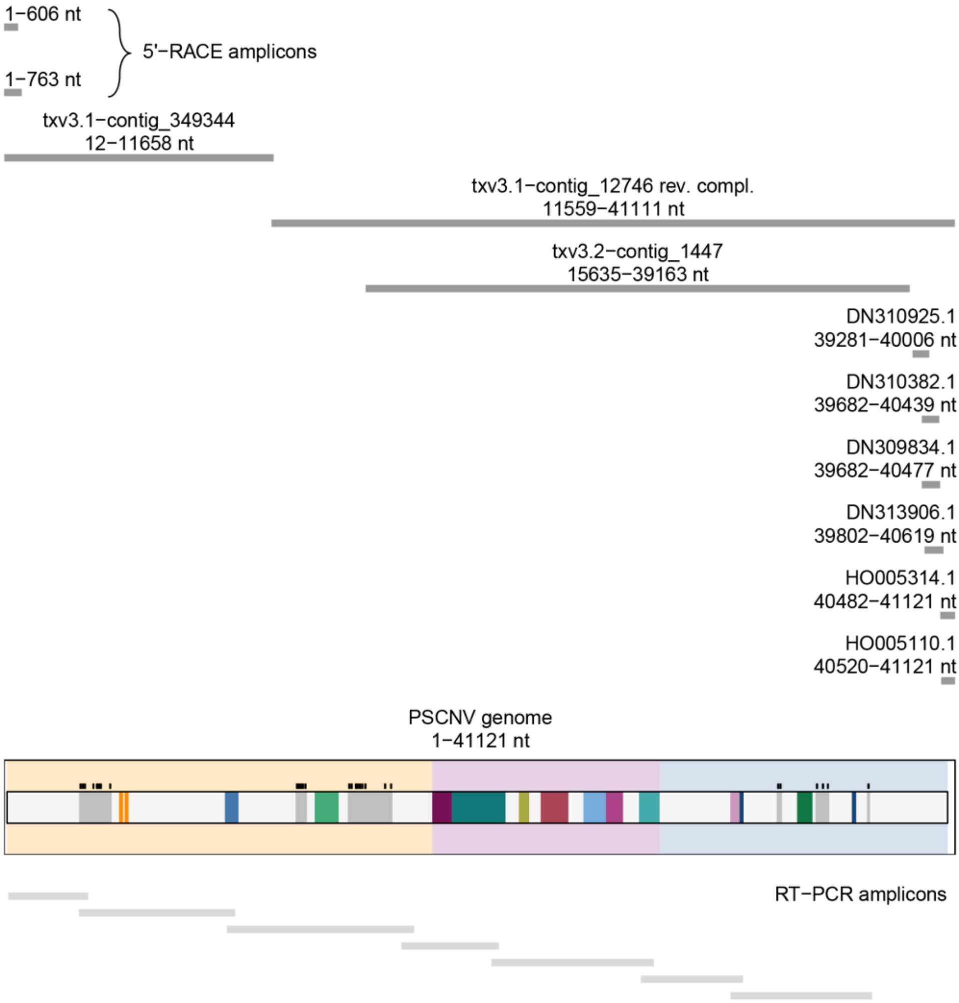

Figure S1 | PSCNV genome assembly and its verification. Contigs and 5’-RACE amplicons, used to assemble the PSCNV genome sequence are shown above the PSCNV genome map (see Fig. 2 for designations) by dark grey lines, with coordinates of the corresponding PSCNV genome regions specified on top of each 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).

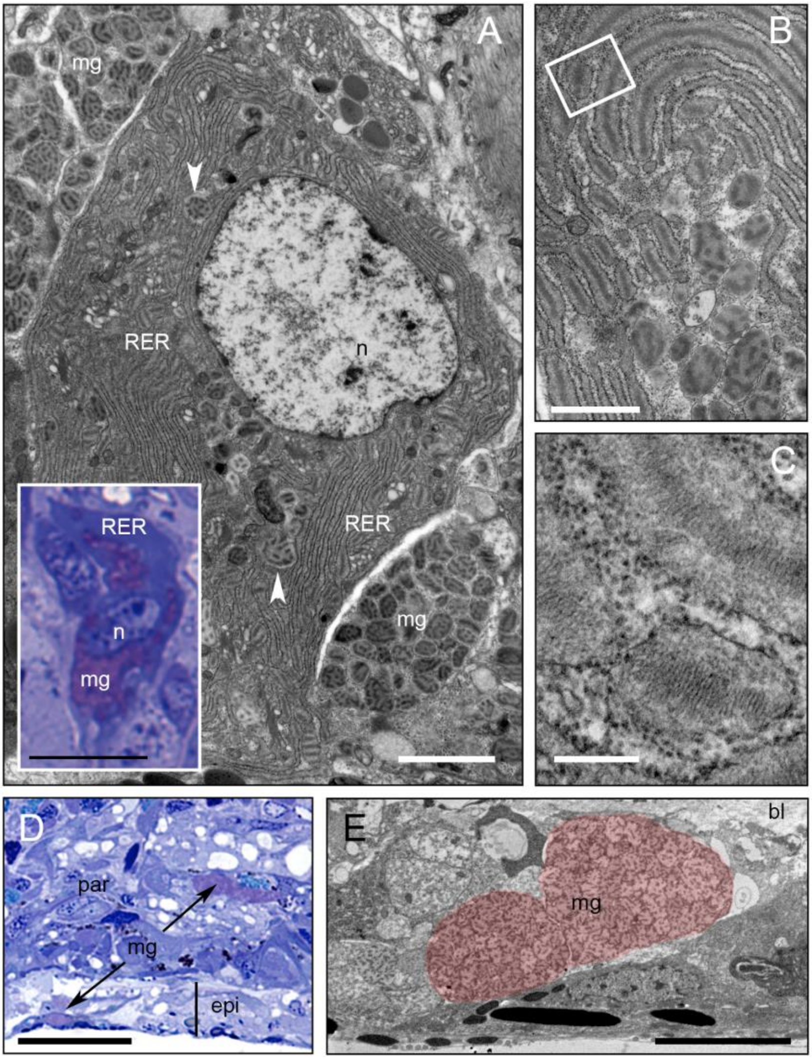

Figure S2 | Characteristics of mucus cells in *S. mediterranea*. (**A**) Transmission electron micrograph of typical mucus cell [210]; n = nucleus. Cell bodies of such cells are filled with rough endoplasmic reticulum (RER). Distinctive mottled structures indicated by arrowheads are mucus granules. Extensions 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.

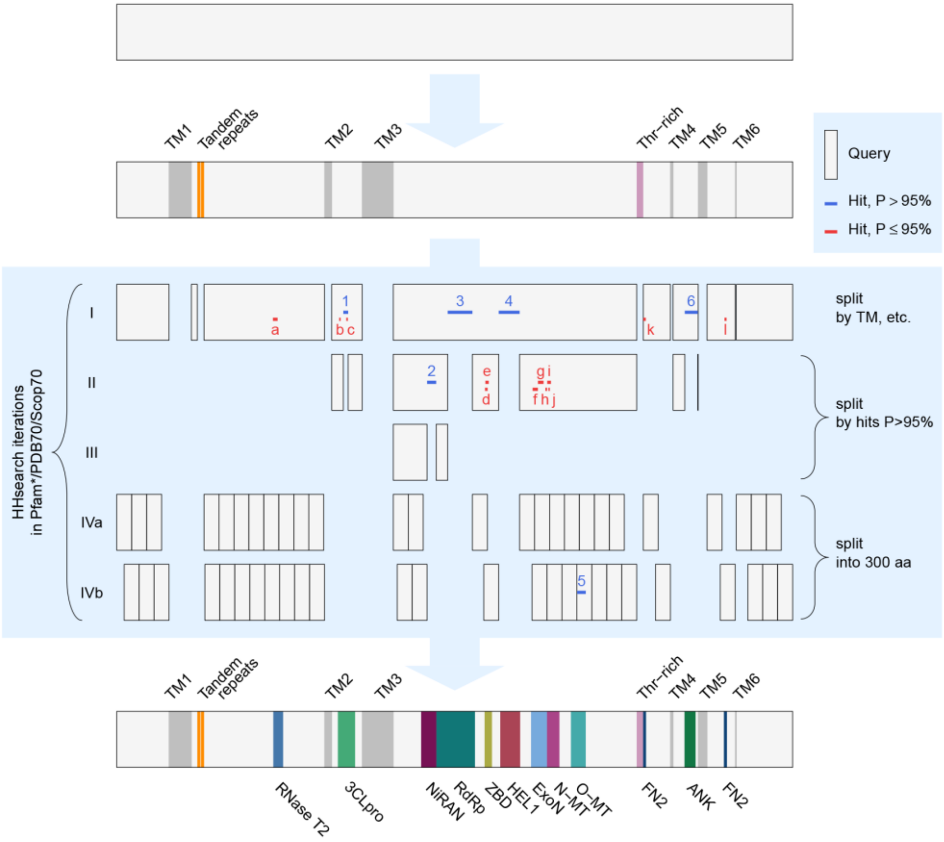

Figure S3 | Outline of iterative HHsearch-based procedure to annotate PSCNV polyprotein. Grey bars on the top and bottom represent PSCNV polyprotein with annotation available prior to the procedure and obtained as a result of the procedure (see Table S5), respectively. Outline of the procedure (see SI 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.

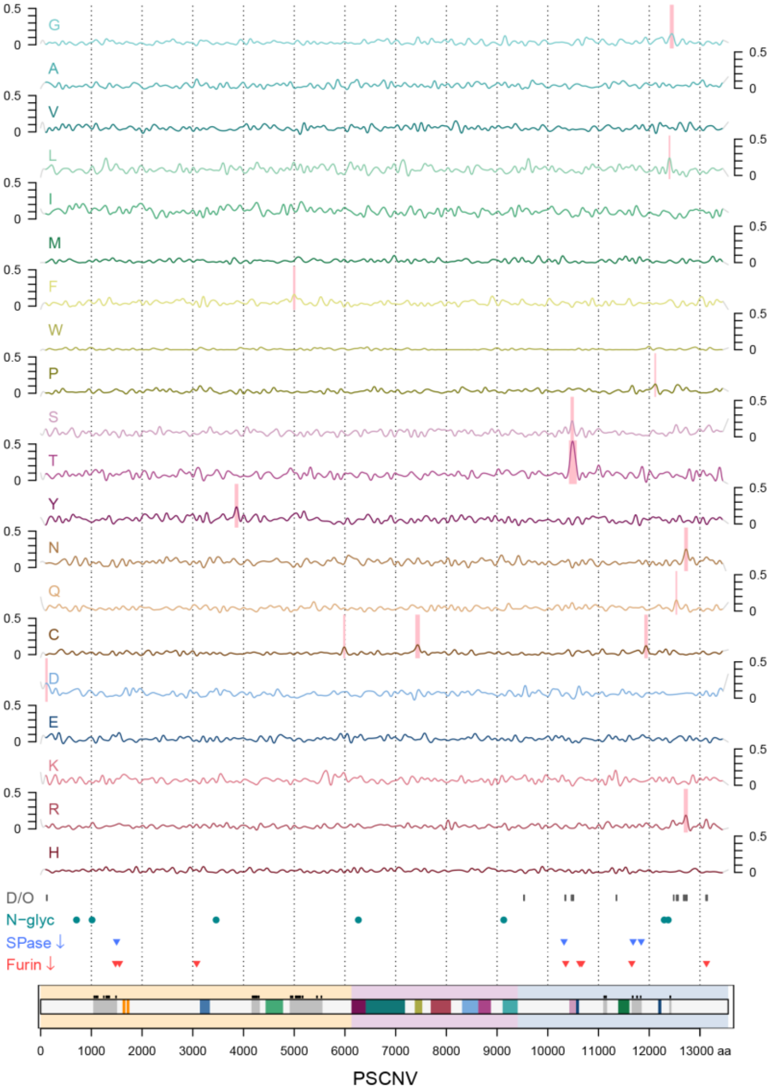

Figure S4 | Density distribution of twenty amino acid residues and predicted functional sites of PSCNV polyprotein. Top: first derivative of cumulative amino acid residue content is plotted for each of the 20 residues with residue-specific colors; values corresponding to the N- and C- terminal 100 residues 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.

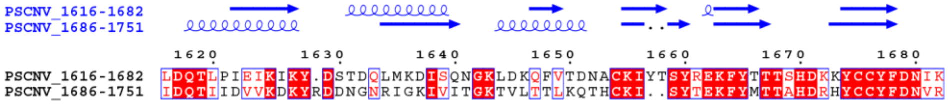
PSCNV_1616-1682
PSCNV_1686-1751
1620
1630
1640
1650
1660
1670
1680
PSCNV_1616-1682 LDQTLPIEIKIKY.DSTDQLMKDISQNGKLDKQFVTDNACKIYTSYREKFYTTTSHDKKYCCYFDNIK
PSCNV_1686-1751 IDQTIIDVVKDKYRDDNGNRIGKIVITGKTVLTTLKQTHCKI..SYTEKFYMTTAHDRHYCCYFDNVR

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.

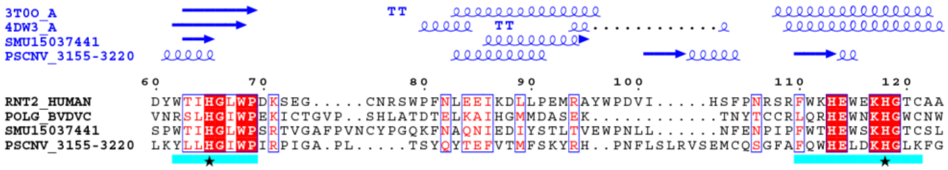
3T0O_A
4DW3_A
SMU15037441
PSCNV_3155-3220
TT
TT
60
70
80
90
100
110
120
RNT2_HUMAN DYWTIHGLWPDKSEG.....CNRSWPFNLEEIKDLLPEMRAYWPDVI.....HSFPNRSRFWKHEWEKHGTCAA
POLG_BVDVC VNRSLHGIWPEKICTGVP..SHLATDTELKAIHGMMDASEK............TNYTCCRLQRHEWNKHGWCNW
SMU15037441 SPWTIHGLWPSRTVGAFPVNCYPGQKFNAQNIEDIYSTLTVEWPNLL......NFENPIPFWTHEWSKHGTCSL
PSCNV_3155-3220 LKYLLHGIWPIRPIGA.PL.....TSYQYTEFVTMFSKYRH..PNFLSLRVSEMCQSGFAFQWHELDKHGLKFG

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.

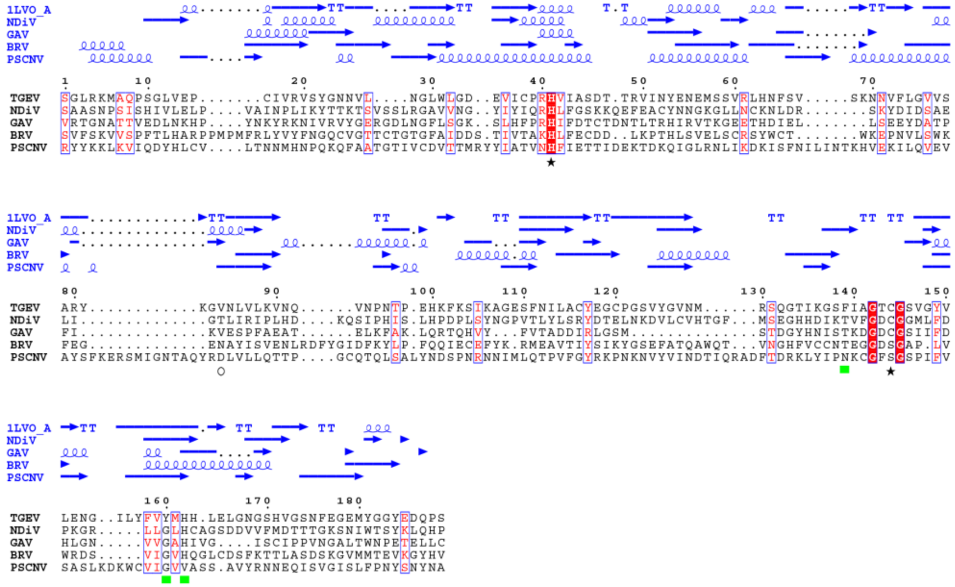
1LVO_A
NDiV
GAV
BRV
PSCNV
T.T
1
10
20
30
40
50
60
70
TGEV SGLRKMAQPSGLVEP.......CIVRVSYGNNVL....NGLWLGD..EVICPRHVIASDT.TRVINYENEMSSVRLHNFSV.....SKNNVFLGVVS
NDiV SAASNPSISHIVLELP....VAINPLIKYTTKTSVSSLRGAVVNG..YIYIQRHLFGSKKQEFEACYNNGKGLLNCKNLDR.......SKYDIDSAE
GAV VRTGNATTVEDLNKHP....YNKYRKNIVRVYGERGDLNGFLSGK..SLHFPRHIFDTCTDNTLTRHIRVTKGEETHDIEL.......LSEEYDATP
BRV SVFSKVVSPFTLHARPPMPMFRLYVYFNGQCVGTTCTGTGFAIDDS.TIVTAKHLFECDD..LKPTHLSVELSCRSYWCT.......WKEPNVLSWK
PSCNV RYYKKLKVIQDYHLCV....LTNNMHNPQKQFAATGTIVCDVTTMRYYIATVNHFIETTIDEKTDKQIGLRNLIKDKISFNILINTKHVEKILQVEV
★
TT
80
90
100
110
120
130
140
150
TGEV ARY............KGVNLVLKVNQ......VNPNTP.EHKFKSIKAGESFNILACYEGCPGSVYGVNM......RSQGTIKGSFIAGTCGSVGYV
NDiV LI..............GTLIRIPLHD....KQSIPHIS.LHPDPLSYNGPVTLYLSRYDTELNKDVLCVHTGF...MSEGHHDIKTVFGDCGGMLFD
GAV FI..............KVESPFAEAT......ELKFAK.LQRTQHVY...FVTADDIRLGSM..............STDGYHNISTKDGDCGSIIFD
BRV FEG.............ENAYISVENLRDFYGIDFKYLP.FQQIECEFYK.RMEAVTIYSIKYGSEFATQAWQT...VNGHFVCCNTEGGDSGAP.LV
PSCNV AYSFKERSMIGNTAQYRDLVLLQTTP....GCQTQLSALYNDSPNRNNIMLQTPVFGYRKPNKNVYVINDTIQRADFTDRKLYIPNKCGFSGSPIFV
○
🟩
★
160
170
180
TGEV LENG..ILYFVYMHH.LELGNGSHVGSNFEGEMYGGYEDQPS
NDiV PKGR.....LLGLHCAGSDDVVFMDTTTGKSNIWTSYKLQHP
GAV HLGN.....VVGAHIVG....ISCIPPVNGALTWNPETELLC
BRV WRDS.....VIGVHQGLCDSFKTTLASDSKGVMMTEVKGYHV
PSCNV SASLKDKWCVIGVVASS.AVYRNNEQISVGISLFPNYSNYNA
🟩 🟩

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.

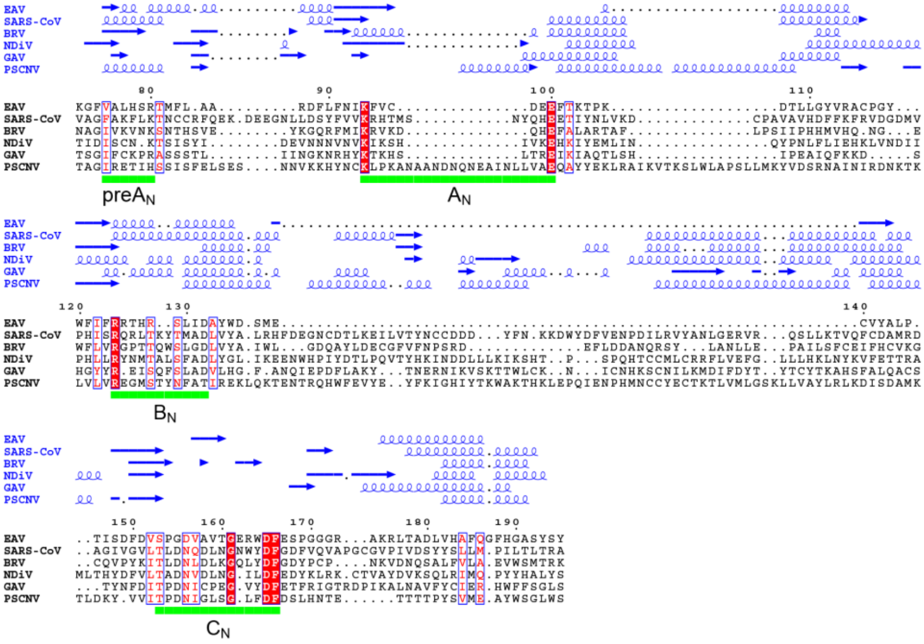
EAV
SARS-CoV
BRV
NDiV
GAV
PSCNV
80
90
100
110
EAV KGFVALHSRTMFL.AA.........RDFLFNIKFVC...............DEEFTKTPK..................DTLLGYVRACPGY...
SARS-CoV VAGFAKFLKTNCCRFQEK.DEEGNLLDSYFVVKRHTMS...........NYQHEETIYNLVKD.............CPAVAVHDFFKFRVDGDMV
BRV NAGIVKVNKSNTHSVE........YKGQRFMIKRVKD..............QHEFALARTAF..............LPSIIPHHMVHQ.NG...E
NDiV TIDISCN.KTSISYI........DEVNNNVNVKIKSH.............IVKEHKIYEMLIN...............QYPNLFLIEHKLVNDII
GAV TSGIFCKPRASSSTL........IINGKNRHYKTKHS.............LTREIKIAQTLSH................IPEAIQFKKD.....S
PSCNV TAGIRETIHSSISFELSES....NNVKKHYNCKLPKANAANDNQNEAINLLVAEQAYYEKLRAIKVTKSLWLAPSLLMKYVDSRNAINIRDNKTK
preA$_N$
A$_N$
120
130
140
EAV WFIFRRTHR..SLIDAYWD.SME..CVYALP.
SARS-CoV PHISRQRLTKYTMADLVYA.LRHFDEGNCDTLKEILVTYNCCDDD...YFN.KKDWYDFVENPDILRVYANLGERVR...QSLLKTVQFCDAMRD
BRV WFLVRGPTTQWSLGDLVYA.IWL...GDQAYLDECGFVFNPSRD.............EFLDDANQRSY...LANLLE...PAILSFCEIFHCVKG
NDiV PHLLRYNMTALSFADLYGL.IKEENWHPIYDTLPQVTYHKINDDLLLKIKSHT..P...SPQHTCCMLCRRFLVEFG..LLLHKLNYKVFETTRA
GAV HGYYR.EISQFSLADVLHG.F.ANQIEPDFLAKY...TNERNIKVSKTTWLCK..N...ICNHKSCNILKMDIFDYT..YTCYTKAHSFALQACS
PSCNV LVLVREGMSTYNFATIREKLQKTENTRQHWFEVYE..YFKIGHIYTKWAKTHKLEPQIENPHMNCCYECTKTLVMLGSKLLVAYLRLKDISDAMK
B$_N$
150
160
170
180
190
EAV ..TISDFDVSPGDVAVTGERWDFESPGGGR...AKRLTADLVHAFQGFHGASYSY
SARS-CoV ..AGIVGVLTLDNQDLNGNWYDFGDFVQVAPGCGVPIVDSYYSLLM.PILTLTRA
BRV ..CQVPYKITLDNLDLKGQLYDFGDYPCP....NKVDNQSALFVLA.EVWSMTRK
NDiV MLTHYDFVLTADNVDLNG.ILDFEDYKLRK.CTVAYDVKSQLRIMQ.PYYHALYS
GAV ...TYNFDITPDNICPEG.VYDFETFRIGTRDPIKALNAVFYCIER.HWFFSGLS
PSCNV TLDKY.VVITPDNIGLSG.LFDFDSLHNTE......TTTTPYSVME.AYWSGLWS
C$_N$

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.

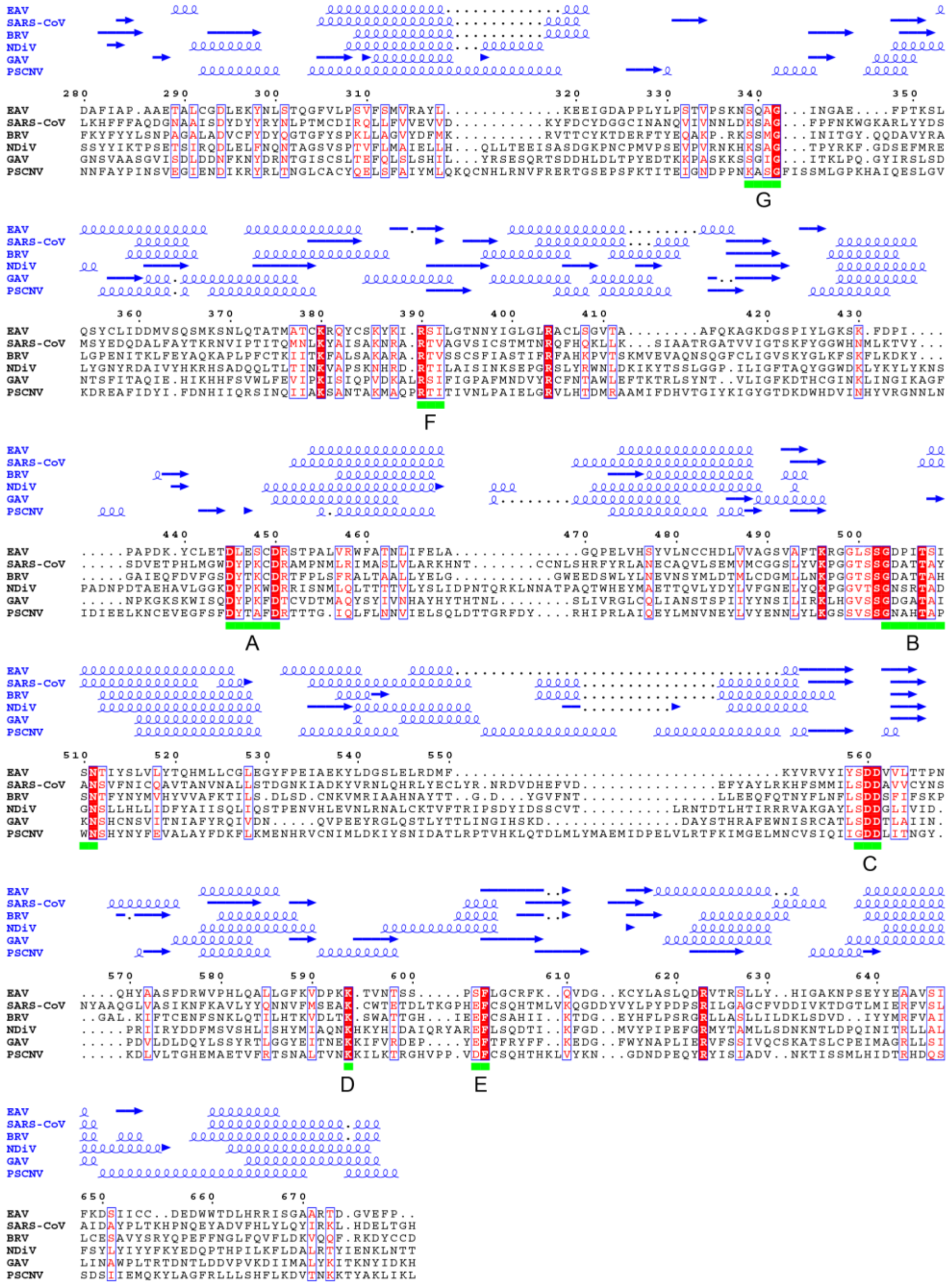

Figure S9 | MSA of RdRp domains from five distantly related nidoviruses and PSCNV (6632–7125 aa). Conserved motifs are underlined in green. Absolutely conserved residues are shown on red background and partially conserved residues in red font. Secondary structure is shown in blue. Residue numbering on top of the alignment refers to EAV nsp9.

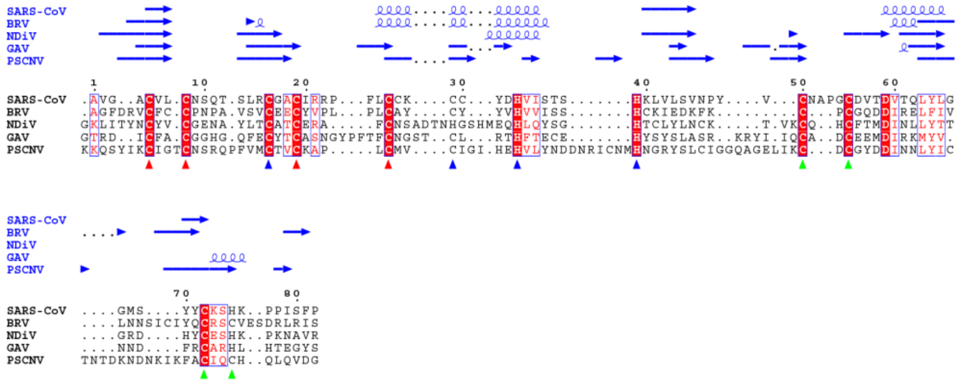
SARS-CoV
BRV
NDiV
GAV
PSCNV
1
10
20
30
40
50
60
SARS-CoV .AVG..ACVL.CNSQT.SLRCGACIRRP...FLCCK....CC...YDHVISTS.......HKLVLSVNPY....V...CNAPGCDVTDVTQLYLG
BRV .AGFDRVCFC.CPNPA.VSVCEECYVPL...PLCAY....CY...YVHVVISS.......HCKIEDKFK.........C...PCGQDDIRELFIV
NDiV GKLITYNCYV.CGENA.YLTCATCERA.....FCNSADTNHGSHMEQHLQYSG.......HTCLYLNCK.....T.VKCQ..HCFTMDINLLYTT
GAV GTRD..ICFA.CGGHG.QFECYTCASNGYPFTFCNGST..CL...RTHFTECE.......HYSYSLASR..KRYI.IQCA..DCEEMDIRKMYV.
PSCNV KKQSYIKCIGTCNSRQPFVMCTVCKAP.....LCMV....CIGI.HEHVLYNDDNRICNMHNGRYSLCIGGQAGELIKC...DCGYDDINNLYIC
SARS-CoV
BRV
NDiV
GAV
PSCNV
70
80
SARS-CoV ....GMS....YYCKSHK..PPISFP
BRV ....LNNSICIYQCRSCVESDRLRIS
NDiV ....GRD....HYCESHK..PKNAVR
GAV ....NND....FRCARHL..HTEGYS
PSCNV TNTDKNDNKIKFACIQCH..QLQVDG

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.

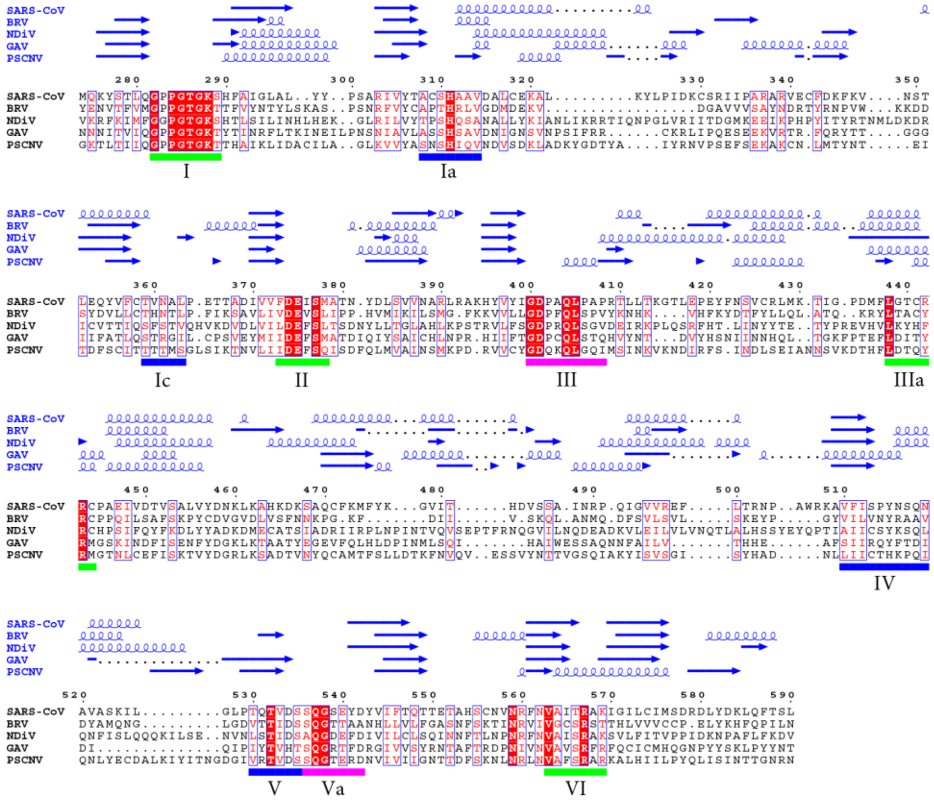

Figure S11 | MSA of HEL1 domains from four distantly related nidoviruses and PSCNV (7718–8056 aa). Conserved motifs are highlighted by color indicating their predominant function [47]: NTP binding and hydrolysis, green; nucleic acid binding, blue; coupling of NTP and nucleic acid binding, purple. Absolutely 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.

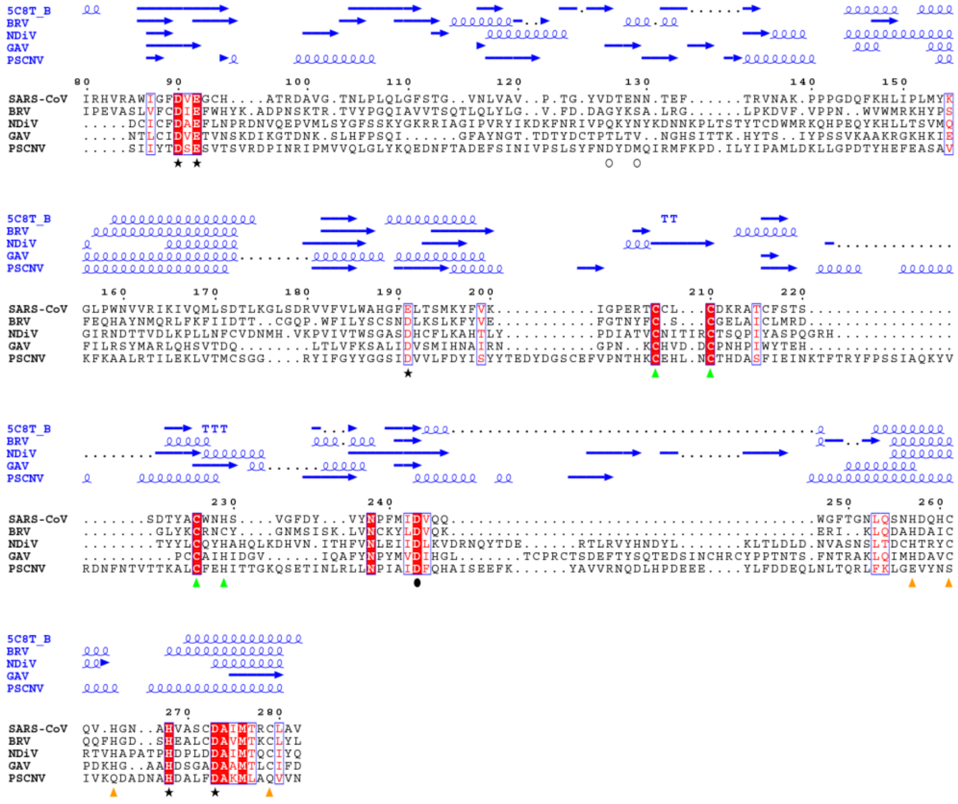
5C8T_B
BRV
NDiV
GAV
PSCNV
80 90 100 110 120 130 140 150
SARS-CoV IRHVRAWIGFDVEGCH....ATRDAVG.TNLPLQLGFSTG..VNLVAV..P.TG.YVDTENN.TEF......TRVNAK.PPPGDQFKHLIPLMYK
BRV IPEVASLVFCDIEFWHYK.ADPNSKTR.TVYPGQIAVVTSQTLQLYLG..V.FD.DAGYKSA.LRG......LPKDVF.VPPN..WVWMRKHYPS
NDiV .....DCICFDAEFLNPRDNVQEPVMLSYGFSSKYGKRRIAGIPVRYIKDKFNRIVPQKYNYKDNNKPLTSTYTCDWMRKQHPEQYKHLLTSVMQ
GAV .....NTLCIDVETVNSKDIKGTDNK.SLHFPSQI......GFAYNGT.TDTYDCTPTLTV..NGHSITTK.HYTS..IYPSSVKAAKRGKHKIE
PSCNV .....SIIYTDSESVTSVRDPINRIPMVVQLGLYKQEDNFTADEFSINIVPSLSYFNDYDMQIRMFKPD.ILYIPAMLDKLLGPDTYHEFEASAV
TT
160 170 180 190 200 210 220
SARS-CoV GLPWNVVRIKIVQMLSDTLKGLSDRVVFVLWAHGFELTSMKYFVK...........IGPERTCCL...CDKRATCFSTS................
BRV FEQHAYNMQRLFKFIIDTT..CGQP.WFILYSCSNDLKSLKFYVE...........FGTNYFC.S...CGELAICLMRD................
NDiV GIRNDTTVDLKPLLNFCVDNMH.VKPVIVTWSGASDHCFLKAHTLY..........PDIATVCNITIRCTSQPIYASPQGRH.............
GAV FILRSYMARLQHSVTDQ........LTLVFKSALIDVSMIHNAIRN..........GPN..KCHVD.DCPNHPIWYTEH................
PSCNV KFKAALRTILEKLVTMCSGG....RYIFGYYGGSIDVVLFDYISYYTEDYDGSCEFVPNTHKCEHL.NCTHDASFIEINKTFTRYFPSSIAQKYV
TTT
230 240 250 260
SARS-CoV .......SDTYACWNHS....VGFDY...VYNPFMIDVQQ.....................................WGFTGNLQSNHDQHC
BRV ........GLYKCRNCY....GNMSISK.LVNCKYLDVQK.....................................ERI..KLQDAHDAIC
NDiV ........TYYLCQYHAHQLKDHVN.ITHFVNLEIIDLKVDRNQYTDE......RTLRVYHNDYL.......KLTLDLD.NVASNSLTDCHTRYC
GAV ..........PCCAIHIDGV......IQAFYNPYMVDIHGL.......TCPRCTSDEFTYSQTEDSINCHRCYPPTNTS.FNTRAKLQIMHDAVC
PSCNV RDNFNTVTTKALCFEHITTGKQSETINLRLLNPIAIDFQHAISEEFK......YAVVRNQDLHPDEEE....YLFDDEQLNLTQRLFKLGEVYNS
270 280
SARS-CoV QV.HGN..AHVASCDAIMTRCLAV
BRV QQFHGD..SHEALCDAVMTKCLYL
NDiV RTVHAPATPHDPLDDAIMTQCIYQ
GAV PDKHG..AAHDSGADAAMTLCIFD
PSCNV IVKQDADNAHDALFDAKMLAQVVN

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.

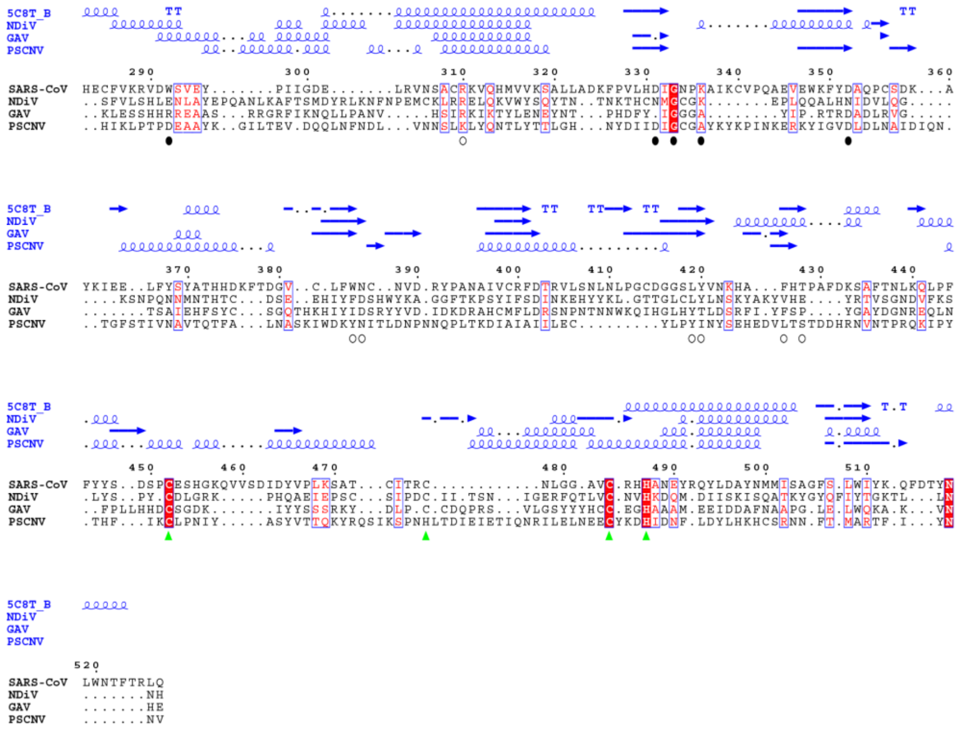
5C8T_B
NDiV
GAV
PSCNV
TT
290
300
310
320
330
340
350
360
SARS-CoV HECFVKRVDWSVEY.......PIIGDE.......LRVNSACRKVQHMVVKSALLADKFPVLHDIGNPKAIKCVPQAEVEWKFYDAQPCSDK...A
NDiV ..SFVLSHLENLAYEPQANLKAFTSMDYRLKNFNPEMCKLRRELQKVWYSQYTN..TNKTHCNMGCGK.......EPLQQALHNIDVLQG.....
GAV ..KLESSHHRREAAS...RRGRFIKNQLLPANV.....HSIRKIKTYLENEYNT...PHDFY.IGGGA........YIP.RTRDADLRVG.....
PSCNV ..HIKLPTPDEAAYK..GILTEV.DQQLNFNDL...VNNSLKLYQNTLYTTLGH...NYDIIDIGCGAYKYKPINKERKYIGVDLDLNAIDIQN.
TT
TT
TT
370
380
390
400
410
420
430
440
SARS-CoV YKIEE..LFYSYATHHDKFTDGV..C.LFWNC..NVD.RYPANAIVCRFDTRVLSNLNLPGCDGGSLYVNKHA...FHTPAFDKSAFTNLKQLPF
NDiV ....KSNPQNNMNTHTC...DSE..EHIYFDSHWYKA.GGFTKPSYIFSDINKEHYYKL.GTTGLCLYLNSKYAKYVHE....YRTVSGNDVFKS
GAV .......TSAIEHFSYC...SGQTHKHIYIDSRYYVD.IDKDRAHCMFLDRSNPNTNNWKQIHGLHYTLDSRFI.YFSP....YGAYDGNREQLN
PSCNV ..TGFSTIVNAVTQTFA...LNASKIWDKYNITLDNPNNQPLTKDIAIAIILEC.........YLPYINYSEHEDVLTSTDDHRNVNTPRQKIPY
T.T
450
460
470
480
490
500
510
SARS-CoV FYYS..DSPCESHGKQVVSDIDYVPLKSAT...CITRC............NLGG.AVC.RHHANEYRQYLDAYNMMISAGFS.LWIYK.QFDTYN
NDiV .LYS..PY.CDLGRK.....PHQAEIEPSC...SIPDC.II.TSN...IGERFQTLVC.NVHKDQM.DIISKISQATKYGYQFIYTGKTL...LN
GAV ..FPLLHHDCSGDK.......IYYSSRKY...DLP.C.CDQPRS...VLGSYYYHCC.EGHAAAM.EEIDDAFNAAPG.LE.LWQKA.K...VN
PSCNV .THF...IKCLPNIY.....ASYVTTQKYRQSIKSPNHLTDIEIETIQNRILELNEECYKDHIDNF.LDYLHKHCSRNN.FT.MARTF.I...YN
520
SARS-CoV LWNTFTRLQ
NDiV ....NH
GAV ....HE
PSCNV ....NV

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.

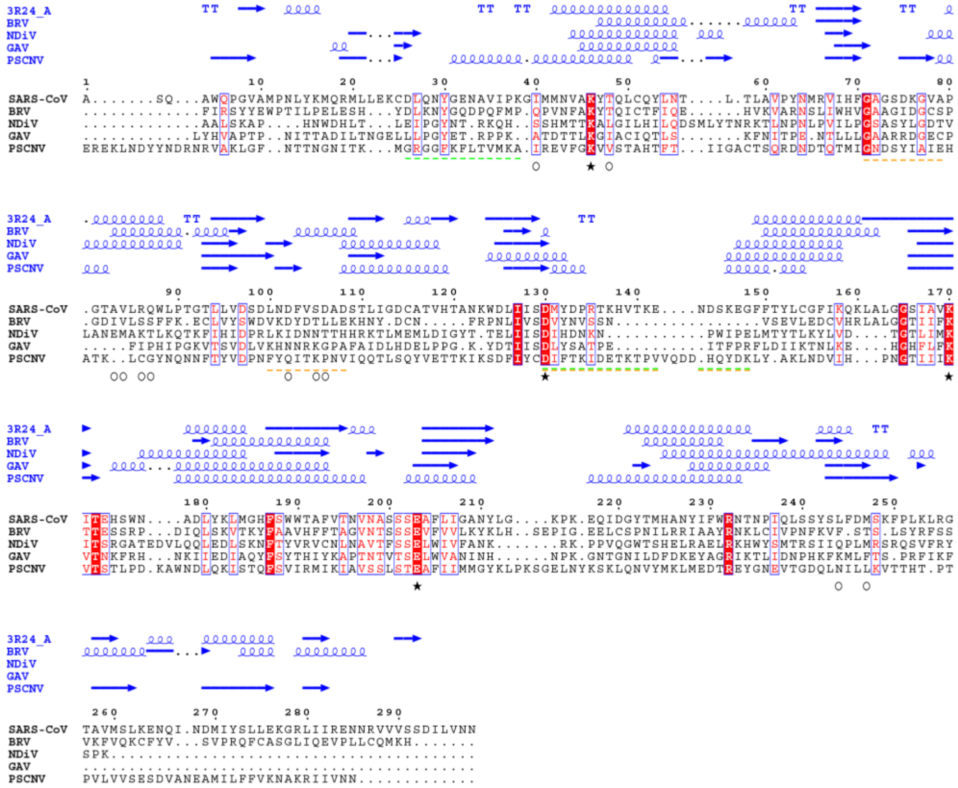

Figure S14 | MSA of O-MT domains from four distantly related nidoviruses and PSCNV (9110–9406 aa). Columns containing SARS-CoV O-MT catalytic tetrad residues are marked by black stars. SARS-CoV O-MT residues involved in interaction with nsp10 are marked by empty circles. Loops constituting SAM-binding 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.

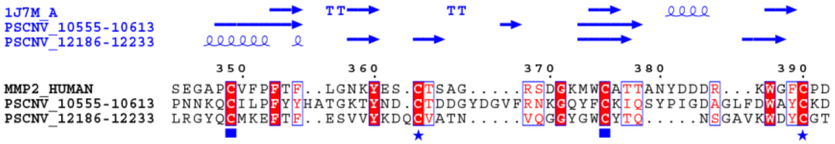
1J7M_A
PSCNV_10555-10613
PSCNV_12186-12233
TT
TT
350
360
370
380
390
MMP2_HUMAN SEGAPCVFPFTF..LGNKYES.CTSAG.....RSDGKMWCATTANYDDDR...KWGFCPD
PSCNV_10555-10613 PNNKQCILPFYYHATGKTYND.CTDDGYDGVFRNKGQYFCKIQSYPIGDAGLFDWAYCKD
PSCNV_12186-12233 LRGYQCMKEFTF..ESVVYKDQCVATN.....VQGGYGWCYTQ.....NSGAVKWDYCGT

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.

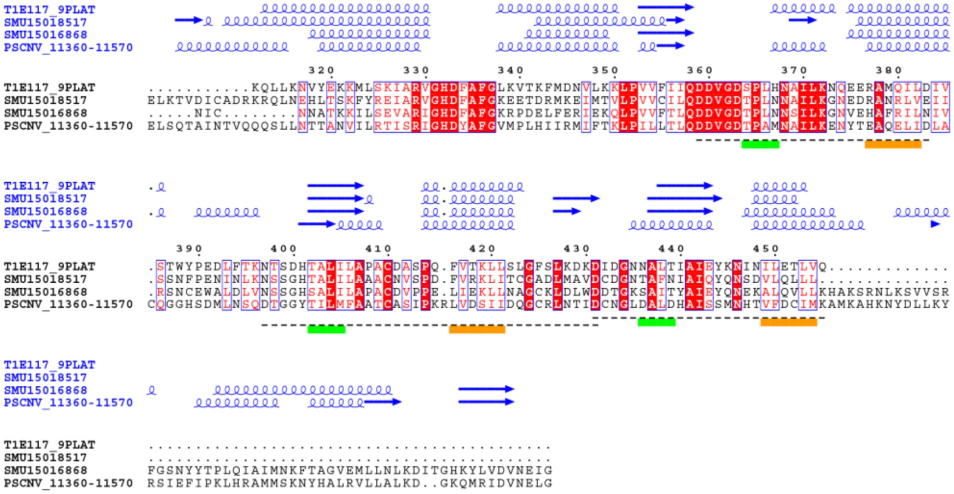
T1E117_9PLAT
SMU15018517
SMU15016868
PSCNV_11360-11570
320
330
340
350
360
370
380
T1E117_9PLAT ...........KQLLKNVYEKKMLSKIARVGHDFAFGLKVTKFMDNVLKKLPVVFIIQDDVGDSPLHNAILKNQEERAMQILDIV
SMU15018517 ELKTVDICADRKRQLNEHLTSKFYREIARVGHDFAFGKEETDRMKEIMTVLPVVCILQDDVGDTPLNNAILKGNEDRANRLVEII
SMU15016868 .....NIC.........NNATKKILSEVARIGHDFAFGKRPDELFERIEKQLPVVFTLQDDVGDTPLNNSILKGNVEHAFRILNIV
PSCNV_11360-11570 ELSQTAINTVQQQSLLNTTANVILRTISRIGHDYAFGVMPLHIIRMIFTKLPILLTLQDDVGDTPAMNAILKENYTEAQELIDLA
T1E117_9PLAT
SMU15018517
SMU15016868
PSCNV_11360-11570
390
400
410
420
430
440
450
T1E117_9PLAT .STWYPEDLFTKNTSDHTALILAPACDASPQ.FVTKLLSLGFSLKDKDIDGNNALTIAIEYKNINILETLVQ.............
SMU15018517 .SSNFPENINLKNSSGHTALILAAACNVSPD.FVRKLITCGADLMAVDCDGNTAFNIAIQYQNSDVLQLLL..............
SMU15016868 .RSNCEWALDLVNSSGHSALILAPACDVPPE.LIEKLLNAGCKLDLWDDTGKSAITYAIEYQNEKALQVLLKHAKSRNLKSVVSR
PSCNV_11360-11570 CQGGHSDMLNSQDTGGYTILMFAATCASIPKRLVDSIIDQGCRLNTIDCNGLDALDHAISSMNHTVFDCIMKAMKAHKNYDLLKY
T1E117_9PLAT
SMU15018517
SMU15016868
PSCNV_11360-11570
T1E117_9PLAT ....
SMU15018517 ....
SMU15016868 FGSNYYTPLQIAIMNKFTAGVEMLLNLKDITGHKYLVDVNEIG
PSCNV_11360-11570 RSIEFIPKLHRAMMSKNYHALRVLLALKD..GKQMRIDVNELG

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.

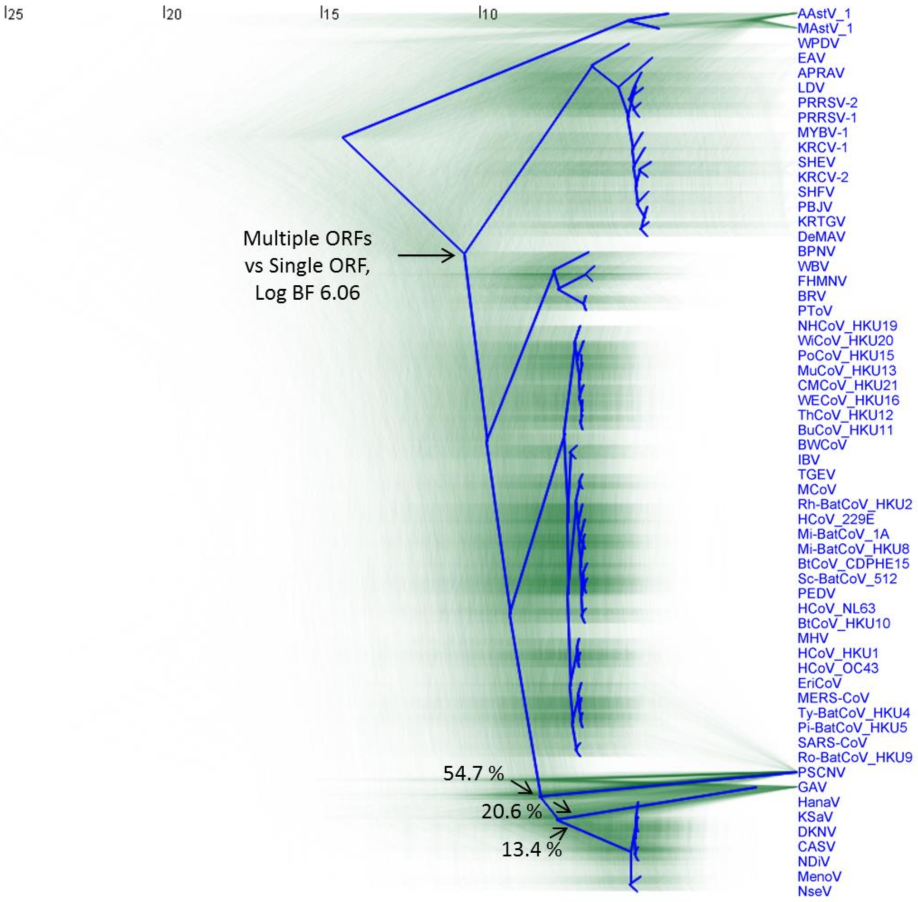

Figure S17 | Phylogeny reconstructed by BEAST based on the alignment of RdRp core of PSCNV, nidoviruses, and astroviruses. Bayesian sample of trees is shown in green, consensus tree with the highest clade support is shown in blue. Support for multiple ORFs vs single ORF in the genome of MRCA of nidoviruses 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).

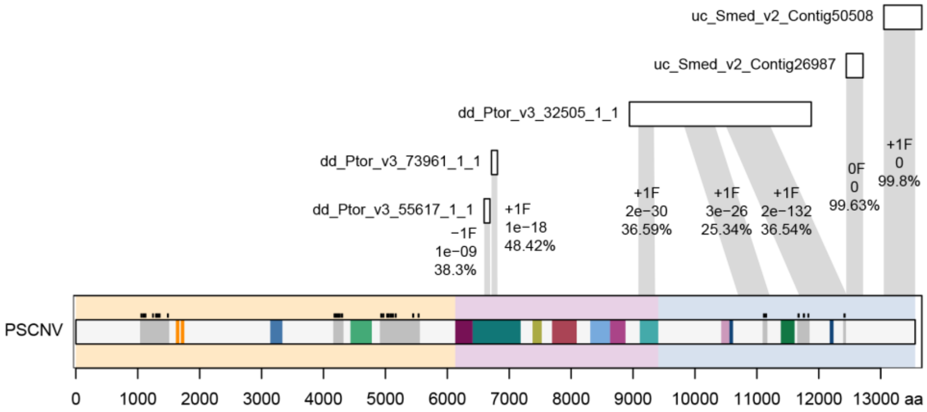

Figure S18 | Statistically significant BLAST hits between translated contigs of PlanMine database and PSCNV polyprotein. Contigs from two assemblies, dd_Ptor_v3 and uc_Smed_v2, are shown as white rectangles. For each hit, depicted as a grey band, a frame in which the contig was translated ("F" stands 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.

					Genome region, nt		
(Sub)family	Species	Acronym	Accession number	Genome, 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)

			Accession		Genome region, nt		
(Sub)family	**Species**	**Acronym**	**number**	**Genome, 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.