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

Gulyaeva, A.

CHAPTER 6

General Discussion

PREFACE

Historically, viruses were discovered and characterized experimentally. The advent of nucleic acid sequencing opened up a new way of studying viruses – comparative genomics – allowing to characterize viruses based on their genome sequences, often using results of experimental research on related viruses. In this thesis, we used comparative genomics techniques to characterize various aspects of nidovirus biology and evolution (**chapters 2-4**), while we also developed a method facilitating homology detection and annotation of large and highly divergent polyproteins (**chapter 5**).

Since 2014, when the project described in this thesis started, the importance of comparative genomics in nidovirus characterization has greatly increased together with the rate of nidovirus discovery. The number of nidovirus species recognized by ICTV has almost quadrupled, increasing from 31 in 2014 to 88 in 2017 [1, 2] and to 109 proposed in 2019 (Fig. 1) [3-6]. The reasons behind this explosive growth, observed for other groups of RNA viruses as well, are advancement of NGS technologies, and consequent transformation of metatranscriptome sequencing into a widely-used laboratory technique. Metatranscriptomics allows to screen a broad range of hosts for the presence of RNA viruses with high efficiency. For example, a single metatranscriptomics study conducted by Shi *et al.* in 2016 identified 1,445 novel phylogenetically distinct RNA virus genomes [7] that almost doubled the number of RNA virus species known at the time. Six genomes discovered in that study were later recognized by ICTV as prototypes of novel, divergent nidovirus species [2, 8].

The number of nidoviruses discovered in the past five years far exceeds the number of nidoviruses that were ever propagated in cell culture and characterized experimentally. As the rate with which new nidoviruses are discovered is increasing, so does the cost and complexity of the experiments required to characterize all of them experimentally. It makes laboratory research on all the newly discovered nidoviruses infeasible. Instead, comparative genomics is used to provide a connection between numerous nidoviruses discovered based solely on NGS data, and a few nidoviruses that are subject to comprehensive experimental characterization. Comparative genomics identifies homologous regions of genomes and proteins, allowing to transfer functional annotation from experimentally characterized viruses and hosts to newly discovered virus genomes.

Characterization by comparative genomics can be facilitated by reliable classification of viruses, as taxonomic assignment itself may offer clues about the biology of a newly discovered virus, as well as help to design comparative genomics experiments. Accommodating the known diversity of RNA viruses requires building a multilevel hierarchical classification while dealing with large evolutionary distances. The DEmARC software package that was developed in our group [9, 10] allows to build such classification. Following its publication in 2012, the package was advanced by our group to include a greater choice (1) of methods used to calculate genetic distances between

viruses, and (2) of linkage types used in hierarchical clustering; (3) to implement sequence weights reducing the impact of virus sampling bias; (4) to propose an alternative approach of identifying classification levels; (5) to visualize various aspects of the obtained classification; (6) to provide the user with an in-package tutorial on DEmARC usage. Over the past few years, our group was involved in producing DEmARC-based multilevel classifications of the newly described nidoviruses that ICTV Study Groups on nidoviruses used as the basis for taxonomy proposals [3-6, 11-17]. ICTV approved the first seven of these proposals, including the one delineating four new invertebrate nidovirus families (*Medioni-*, *Euroni-*, *Abysso-* and *Mononiviridae*) [1, 2], and is currently considering the four latest proposals, including a proposal delineating five new vertebrate nidovirus families (*Olifo-*, *Gresna-*, *Cremega-*, *Nanhypo-* and *Nanghoshaviridae*) [3]. On a phylogenetic tree, all four novel invertebrate families cluster with invertebrate nidoviruses that were known before, while novel vertebrate families are basal to *Arteriviridae* (Fig. 1). Also the subfamilies *Coronavirinae* and *Torovirinae* were elevated to the rank of families, *Coronaviridae* and *Tobaniviridae*, respectively, and include now many more species [2].

Comparative genomics analysis of nidoviruses is challenging since, due to the high mutation rate, many novel nidoviruses are highly divergent. One manifestation of nidovirus divergence is the ever-expanding nidovirus host range (Fig. 1) that now includes many exotic animals: from a marsupial mammal [18] to mollusks [7, 19, 20], an urochordate [21] and a flatworm (**chapter 4**). The complexity of the analysis is further compounded by the organization of the nidovirus proteome, which is dominated by large multidomain polyproteins. When a polyprotein sequence is compared to a database of known protein sequences or profiles, distant homologous relationships may remain undetected due to the underestimation of statistical significance of hits by standard tools, designed to annotate smaller proteins with few domains. To address this complication, our group developed a tool, called LAMPA, that gradually splits the polyprotein sequence into smaller queries in a biologically reasonable manner, improving estimation of hit statistical significance (**chapter 5**). For comparison of the delineated queries with databases, LAMPA employs HH-suite, one of the most sensitive software packages for protein homology detection [22, 23].

Usage of state-of-the-art comparative genomics tools was essential for the characterization of three different aspects of the growing nidovirus diversity, described in this thesis: identification of a novel replicative domain universally conserved in all nidoviruses (**chapter 3**), analysis of domain organization and evolution of arterivirus non-structural polyprotein N-terminus that encompassed multiple newly recognized species, including the most divergent WPDV, the genome sequencing of which was completed as part of the study (**chapter 2**), functional description of a novel, highly divergent nidovirus genome identified by mining transcriptome of its host, planarian *S. mediterranea* (**chapter 4**). These and other studies, conducted in recent years, contributed to the evolution of understanding of nidovirus key properties.

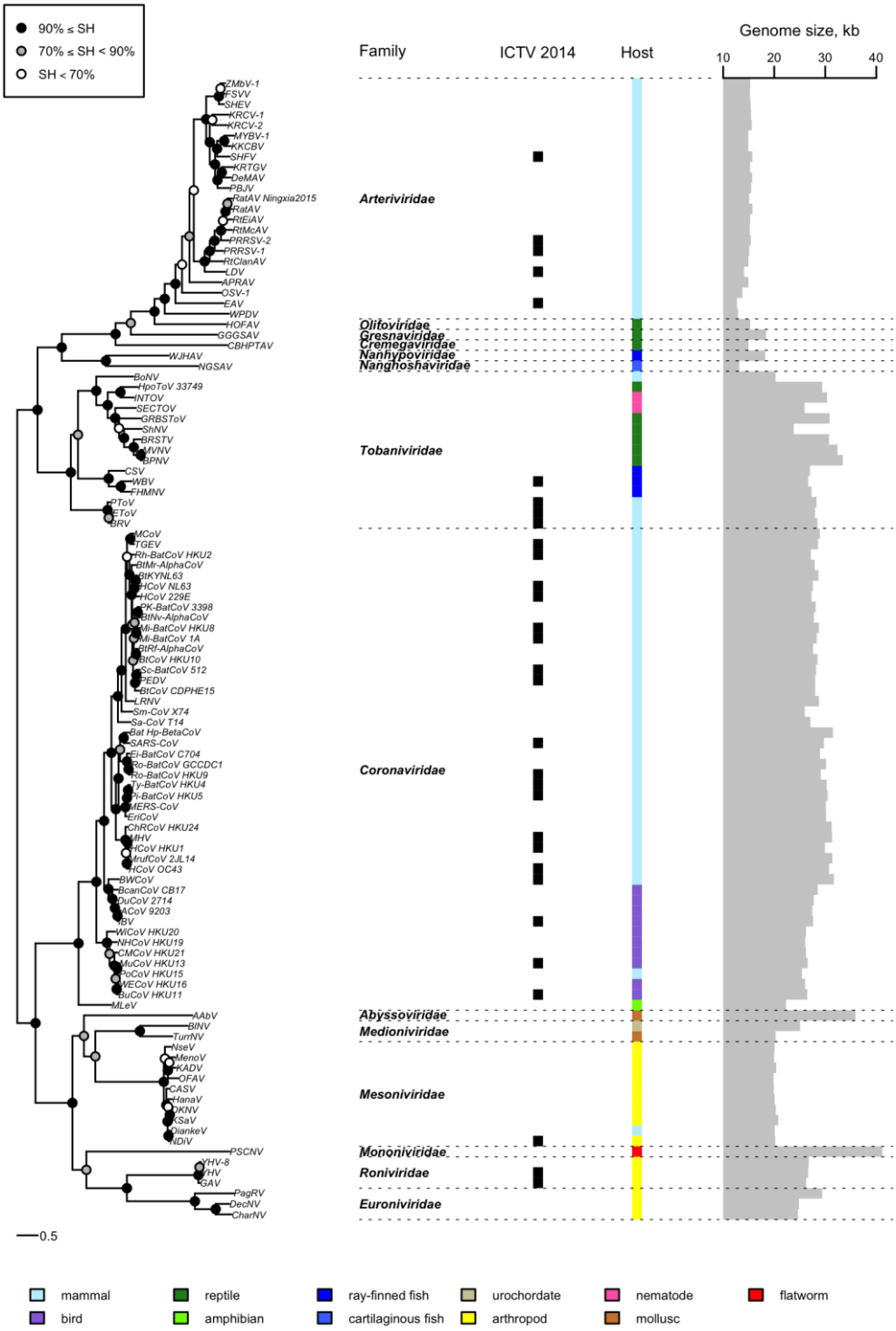
90% ≤ SH
70% ≤ SH < 90%
SH < 70%
Family
ICTV 2014
Host
Genome size, kb
10
20
30
40
ZMbV-1
FSVV
SHEV
KRCV-1
KRCV-2
MYBV-1
KKCBV
SHFV
KRTGV
DeMAV
PBJV
RatAV Ningxia2015
RatAV
RtEiAV
RtMcAV
PRRSV-2
PRRSV-1
RtClanAV
LDV
APRAV
OSV-1
EAV
WPDV
HOFAV
GGGSAV
CBHPTAV
WJHAV
NGSAV
BoNV
HpoToV 33749
INTOV
SECTOV
GRBSToV
ShNV
BRSTV
MVNV
BPNV
CSV
WBV
FHMNV
PToV
EToV
BRV
MCoV
TGEV
Rh-BatCoV HKU2
BtMr-AlphaCoV
BtKYNL63
HCoV NL63
HCoV 229E
PK-BatCoV 3398
BtNv-AlphaCoV
Mi-BatCoV HKU8
Mi-BatCoV 1A
BtRf-AlphaCoV
BtCoV HKU10
Sc-BatCoV 512
PEDV
BtCoV CDPHE15
LRNV
Sm-CoV X74
Sa-CoV T14
Bat Hp-BetaCoV
SARS-CoV
Ei-BatCoV C704
Ro-BatCoV GCCDC1
Ro-BatCoV HKU9
Ty-BatCoV HKU4
Pi-BatCoV HKU5
MERS-CoV
EriCoV
ChRCoV HKU24
MHV
HCoV HKU1
MrufCoV 2JL14
HCoV OC43
BWCoV
BcanCoV CB17
DuCoV 2714
MCoV 9203
IBV
WiCoV HKU20
NHCoV HKU19
CMCoV HKU21
MuCoV HKU13
RoCoV HKU15
WECoV HKU16
BuCoV HKU11
MLeV
AAbV
BINV
TurnNV
NseV
MenoV
HADV
OFAV
CASV
HanaV
OKNV
KSaV
DiankeV
NDiV
PSCNV
YHV-8
HV
GAV
PagRV
DecNV
CharNV
0.5
Arteriviridae
Olifoviridae
Gresnaviridae
Cremegaviridae
Nanhypoviridae
Nanghoshaviridae
Tobaniviridae
Coronaviridae
Abyssoviridae
Medioniviridae
Mesoniviridae
Mononiviridae
Roniviridae
Euroniviridae
mammal
bird
reptile
amphibian
ray-finned fish
cartilaginous fish
urochordate
arthropod
nematode
mollusc
flatworm

Figure 1 | Nidovirus diversity: phylogeny, taxonomy, host range, and genome size. Shown is the midpoint-rooted phylogeny, family structure of 2019, species recognized by 2014, host group and genome size of nidovirus species under consideration by ICTV in 2019 [3-6]. The phylogeny was reconstructed based on Viralis MSAs [24] of 3CLpro, NiRAN, RdRp, ZBD, HEL1 conserved cores, using IQ-Tree 1.5.5 [25], with an evolutionary model selected for each domain independently. To estimate branch support, SH-like approximate likelihood ratio test with 1000 replicates was conducted. Viruses representing species and species acronyms are identical to those used in 2019 ICTV proposals (except acronym AcCoV was substituted by AAbV). Information about nidovirus hosts and genome lengths was obtained from GenBank entries. Note that for nidoviruses discovered by metatranscriptomics, host misassignment due to contamination remains a persistent concern. For example, it can be hypothesized that two toroviruses (species *Infratovirus 1* [INTOV] and *Sectovirus 1* [SECTOV], represented by the Xinzhou toro-like virus [KX883638.1] and the Xinzhou nematode virus 6 [KX883637.1], respectively) discovered in metatranscriptomes of snake-associated nematodes [7], which were ascribed as their potential hosts, may infect reptiles instead, as the two viruses cluster with reptile viruses on the phylogenetic tree, and their ascribed hosts are likely to be contaminated with reptile materials. Likewise, the actual host of the *Dianke virus* (DiankeV; KY056254.1) may be an arthropod, because while the virus genome was reported to be isolated from the brain of a rodent, it is closely related to arthropod viruses not known to infect vertebrate hosts.

NIDOVIRUSES RE-DEFINED: UNCOUPLING MOST CONSERVED CHARACTERISTICS

Originally nidoviruses were recognized as a distinct virus group sharing a unique mechanism of discontinuous subgenomic RNA synthesis (transcription), which provided the basis for the group name. Nidoviruses are also distinguished by a conserved multi-ORF organization of the genome, utilization of PRF to express the second ORF (ORF1b), four-domain synteny of replicative domains (3CLpro-RdRp-ZBD-HEL1) encoded in the middle part of the genome, monophyletic clustering of the conserved replicative domains, and a genome size ranging from 12 to 34 kb and including the largest known RNA genomes [26-29]. This recurrent association of many characteristics is indicative of genetic coupling that may have originated early in the nidovirus evolution. However, with expanding experimental characterization of a few distantly related nidoviruses and through comparative genomics of the ever-increasing diversity of nidoviruses, the notion of nidovirus-specific conservation has been steadily adjusted.

Since 1993, when the order *Nidovirales* was established [30], some nidoviruses, namely toro- and roniviruses, were shown to deviate from others in using discontinuous subgenomic RNA synthesis [31, 32]. Discoveries made after 2014, including the ones described in this thesis, identified deviations in other characteristics, limiting further the number of those that are universally shared by all nidoviruses.

Variation in genome architecture and expression mechanisms of nidoviruses

All nidovirus genomes discovered up to 2016 followed the common genome plan: overlapping ORF1a and ORF1b occupying the 5'-terminal two-thirds of the genome and encoding non-structural proteins responsible for virus expression and replication, respectively, and multiple smaller 3'ORFs that encode structural and accessory proteins.

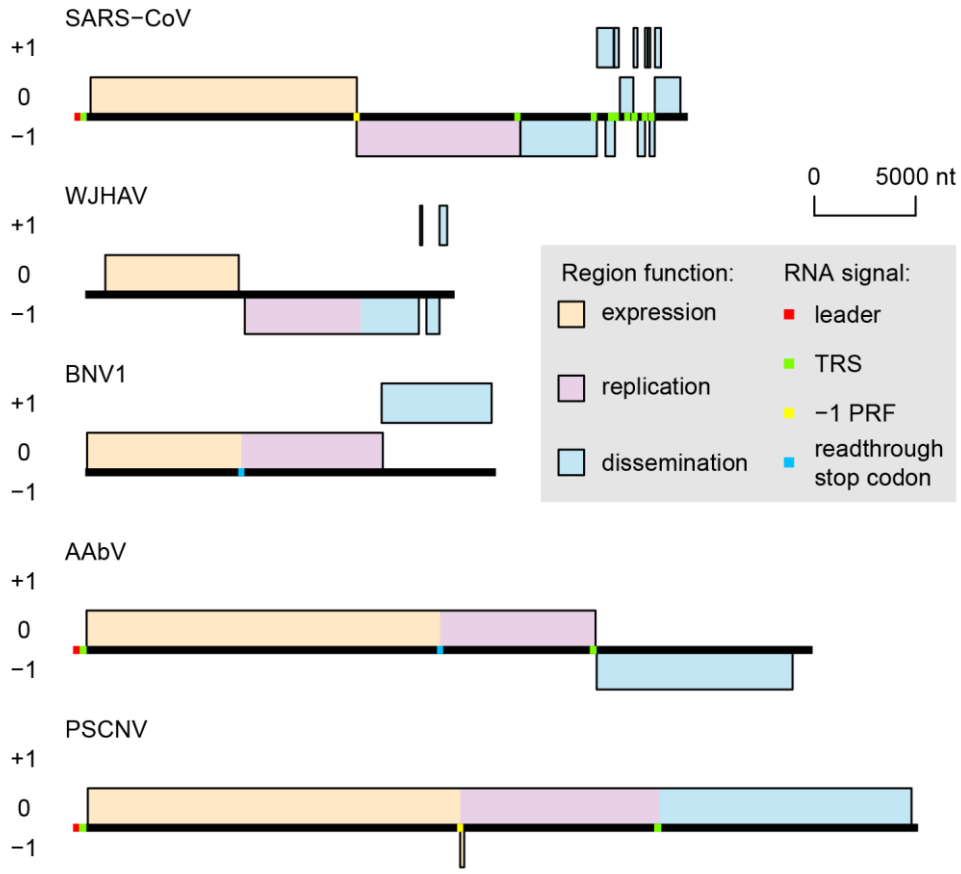

Figure 2 | Nidoviruses with canonical (SARS-CoV) and non-canonical genome ORFs organization. WJHAV, Wuhan Japanese halfbeak arterivirus (MG600020.1); BNV1, Beihai nido-like virus 1 (KX883629.1) classified as species *Turrinivirus 1* (TurrNV) by ICTV; AAbV, Aplysia abyssovirus 1 (GBBW01007738.1); PSCNV, Planarian secretory cell nidovirus (MH933735.1). ORFs are positioned according to their frame, with the most 5'-terminal depicted ORF set as zero. ORF regions are colored according to their predicted function (see inset). Genome signals, described by the discoverers of each virus, are indicated by color (see inset).

Recent discoveries, including some in our study (**chapter 4**), demonstrated that variation of this conserved genome plan is possible in nidoviruses (Fig. 2). In the Wuhan Japanese halfbeak arterivirus (WJHAV) genome, ORF1b is fused with a gene encoding putative glycoprotein [33], presumably a structural protein. Both Beihai nido-like virus 1 (BNV1) and Aplysia abyssovirus 1 (AAbV) contain two ORFs, a 5'-terminal ORF combining ORF1a- and ORF1b-like regions, and a single 3'-terminal ORF encoding structural protein domains [7, 19, 20]. The PSCNV genome has a single large ORF, which is an equivalent of ORF1a, ORF1b and 3'ORFs fused together (**chapter 4**). Notably, PSCNV ORF encodes a 13,556 aa

polyprotein that is 58-67% larger than the largest single- or multi-ORF polyproteins of other viruses.

Conserved multi-ORF genome organization of nidoviruses is coupled with conserved expression mechanisms of transcription and translation, controlling the relative quantities of functionally different proteins in infected cell. The discovery of nidoviruses with an unusual ORF organization raises the question whether they maintain the canonical stoichiometry of viral proteins, and if so, by what expression mechanisms.

In canonical nidoviruses, ORF1a-encoded proteins are expressed in higher quantities than ORF1b-encoded proteins, due to -1 PRF directing a fraction of ribosomes from ORF1a to ORF1b translation. A similar non-structural proteins ratio may be achieved through different mechanisms in non-canonical nidoviruses BNV1, AAbV and PSCNV. Both BNV1 and AAbV have ORF1a-like and ORF1b-like regions residing in the same reading frame and separated by a stop codon (Fig. 2) [7, 19, 20]. If a readthrough of this stop codon only occurs in a fraction of translation events, proteins encoded in the ORF1a-like region would be expressed in a higher quantity compared to proteins encoded in the ORF1b-like region. The PSCNV genome includes a predicted -1 PRF site with a potential to divert translation from ORF1b-like region of the main ORF to a tiny 39 nt ORF (Fig. 2). If efficiency of frameshifting at the predicted site is limited, ORF1a-like compared to ORF1b-like region of the PSCNV genome would be expressed more frequently (**chapter 4**). The main difference between the -1 PRF-directed mechanisms in canonical nidoviruses and in PSCNV is that -1 PRF directs translation into ORF1b in the former but diverts it from ORF1b-like region in the latter.

Another important feature of protein synthesis in canonical nidoviruses is that structural proteins are expressed starting at a later point in time and in higher quantities than non-structural ones. This is achieved through TRS-guided production of sg mRNAs that are 3'-coterminal with the genome and encompass ORFs encoding structural proteins. Notably, this mechanism may be used also for expression of non-canonical nidoviruses: the single 3'ORF of AAbV [19] and the 3'ORFs-like region of the PSCNV genome (**chapter 4**); based on similarity with canonical nidoviruses, this hypothesis may be extended to the single 3'ORF of BNV1 and the three small 3'ORFs of WJHAV, although these viruses were not studied in this respect (Fig. 2). In both the PSCNV and AAbV, potential leader and body TRSs were identified by comparative genomics as large repeats in the 5'UTR and upstream of the genome region predicted to encode structural proteins, respectively (Fig. 2). A sharp increase in coverage of the genome by RNA-seq reads was observed at the body TRS of both PSCNV and AAbV, consistent with the downstream region being a subject of transcription. Existence of PSCNV sg mRNA species, expected to be expressed when the identified TRSs are employed, was confirmed in a 5'-RACE experiment. Importantly, if

translation of the PSCNV sg mRNA species is initiated at its most 5'-terminal start-codon, it would result in production of a polyprotein identical to the C-terminus of the giant polyprotein expressed from the PSCNV genome (**chapter 4**). Thus, predicted structural proteins of PSCNV may be expressed from both genome and sg mRNA.

Interestingly, the PSCNV may not be the only non-canonical nidovirus, structural proteins of which are synthesized from both genome and sg mRNA. The WJHAV potentially encodes a glycoprotein in the unusually long 3'-terminus of its ORF1b that is located downstream of the otherwise terminal O-MT locus [3, 33]. While the WJHAV was not analyzed in this respect, this genome organization is compatible with both genome and sg mRNA directing synthesis of the glycoprotein (Fig. 2). Production of certain structural proteins from both genome and sg mRNA can also be envisioned for some of the canonical nidoviruses, such that stop-codon of ORF1b and start-codon of the downstream structural ORF are in-frame and separated by few nucleotides in their genomes. If a readthrough of the ORF1b stop-codon would occur, with the stop-codon being decoded by a suppressor tRNA [34], it would lead to a continuation of translation, resulting in the production of pp1ab fused with a structural protein (unpublished observation). For example, SARS-CoV ORF1b and ORF2, encoding S protein, belong to the same reading frame and are separated by 6 nt (Fig. 1 from chapter 1).

Unlike the expression of structural proteins from individual ORFs, observed in most known nidoviruses, expression of multiple structural proteins from a single ORF, predicted for non-canonical BNV1, AAbV and PSCNV (Fig. 2), would require processing of the structural polyprotein by host and/or viral proteases, unless their structural domains function in the context of a single polyprotein. Accordingly, a chymotrypsin-like serine protease domain was detected in the structural polyprotein sequences of BNV1 and AAbV [19, 20], and potential cleavage sites of cellular proteases furin and signal peptidase were identified in the C-terminal region of the PSCNV polyprotein (**chapter 4**).

In addition to the deviations from the canonical nidovirus genome architecture in newly discovered viruses discussed above, a number of small functional ORFs, preceding or overlapping with the nidovirus replicase, were revealed in relatively-well characterized nidoviruses in recent years.

The presence of a short ORF, encoded upstream of ORF1a, but likely to be bypassed by ribosomes due to the poor initiation context of its start-codon, was previously reported for multiple arteri- and coronaviruses [35-39]. A recent study utilizing ribosome profiling confirmed that such ORF is translated in MHV [40]. In addition, two CUG-initiated ORFs, one upstream of ORF1a, and another overlapping with ORF1a 5'-terminus, were

demonstrated to be translated in equine torovirus (EToV); the ORFs were shown to be conserved within the genus *Torovirus* [41].

Another pair of small functional ORFs was discovered in arterivirus PRRSV in 2012. Two ORFs are located in the nsp2-encoding genome region, and can be expressed via -1 and -2 PRF, respectively. PRFs occur during ORF1a translation, and lead to the production of two shortened pp1a proteins with alternative C-termini. Both -1 and -2 PRF are transactivated by a complex of arterivirus nsp1β protein and cellular poly(C) binding protein, and employ the same PRF site, consisting of a slippery sequence and a downstream cytidine-rich genome element. The mechanism was predicted to be conserved in all arteriviruses known at the time of the study, with the exception of EAV, based on conservation of the PRF site and the size of the small ORFs [42-44]. This prediction was recently extended to a number of newly discovered arteriviruses, and a distantly related arterivirus WPDV (**chapter 2**). Notably, the distance between the putative WPDV PRF site elements was predicted to be larger than in other arteriviruses. Its utilization may be coupled with extensive evolution of the WPDV nsp1β protein, which diverged considerably from its orthologs and/or other changes to the transactivating protein complex. In addition, the WPDV protein domain predicted to be expressed as a result of -2 PRF is very short and doesn't include potential transmembrane helices, unlike the analogous domains of other arteriviruses.

NiRAN: a fifth universally conserved domain of nidoviruses

In 1988-1989, four domains, 3CLpro, RdRp, ZBD and HEL1, were delineated in the first sequenced coronavirus [26, 45-47]. Subsequently, they were found to be universally conserved in all nidoviruses [30, 48-50]. Many more protein domains were delineated in further studies but none were conserved across the entire order until our discovery of the NiRAN domain, which remained undetected for almost twenty years, was published in 2015 [51]. Its discovery (**chapter 3**) extended the synteny of universally conserved domains associated with the order *Nidovirales* to 3CLpro-NiRAN-RdRp-ZBD-HEL1. Like all the other domains of the synteny, NiRAN was discovered by bioinformatics analysis of extremely distant homology. It is the only enzymatic domain among these five conserved domains that has no apparent homolog in other RNA viruses (however see below).

The NiRAN domain is encoded upstream of RdRp within the same cleavage product. It was shown to be conserved in all nidoviruses by comparative genomics that included profile-profile, predicted secondary structure, and conserved sequence motifs analyses. Comparative genomics helped formulate initial hypothesis about function of the domain. Based on the domain position in ORF1b-encoded part of pp1ab, harboring replicative enzymes, and its profile of invariant residues, the domain was proposed to be an NTP-dependent RNA ligase, a partner of endoribonuclease NendoU, which was believed to

Figure 3 | Primary and secondary structure similarity between NiRAN and protein kinases. Shown is HHsearch alignment of CoToMeRoAr NiRAN and PFAM Pkinase profiles in HH-suite format. The most conserved residues of NiRAN are highlighted in green, NiRAN motifs are designated on top of the alignment. Key functional residues of Pkinase are highlighted in cyan, selected Pkinase sub-domains are designed below the alignment.

be universally conserved in nidoviruses at the time [52]. Consequently, the ability of the domain to covalently bind nucleotides (nucleotidylation), which constitutes the first stage of the ligase reaction, was probed experimentally using recombinant EAV nsp9; nucleotidylation activity with high substrate specificity for UTP, and a lesser substrate specificity for GTP, was demonstrated. Accordingly, the domain was named nidovirus RdRp-associated nucleotidyltransferase, or NiRAN (**chapter 3**).

Three possible functional roles were proposed for NiRAN, with each being compatible with some but not all available experimental data (**chapter 3**). First, as suggested initially, NiRAN might be an NTP-dependent RNA ligase. However, the fact that conservation of the potential counterpart of the ligase, NendoU, was shown to be restricted to vertebrate nidoviruses [50], as well as EAV NiRAN preference for UTP and GTP, uncharacteristic for ligases, make this hypothesis less plausible. Second, NiRAN might be a guanylyltransferase (GTase) catalyzing the second reaction of the mRNA capping pathway (Fig. 3 from chapter 1), though preference of EAV NiRAN for UTP over GTP substrate would be difficult to explain in this context. Third, NiRAN might be involved in protein-primed RNA synthesis,

either in the capacity of a protein covalently linked to the 5'-terminal nucleotide of the nascent RNA strand, or by transferring a nucleotide to such protein. Notably, a protein-priming role would be in perfect agreement with the EAV NiRAN substrate specificity: 5'-terminal nucleotide of EAV genome and sg mRNAs is G, hence GTP is required to initiate their synthesis; 3'-end of the EAV genome is polyadenylated, hence UTP is required to initiate synthesis of minus-strand templates. Also, it would be consistent with RdRp-based phylogenetic clustering of nidoviruses with RNA viruses that employ protein-primed RNA synthesis and 3C(L) proteases [26, 53, 54]. However, it would be difficult to reconcile this role with the presence of a cap structure at the 5'-end of nidovirus mRNAs [32, 55, 56], as the priming protein would have to be removed prior to capping. Besides, primase-dependent and *de novo* mechanisms of RNA synthesis initiation in nidoviruses were proposed [57, 58], although they were challenged most recently, when coronavirus nsp8 was shown to possess oligo(U)-templated 3'-terminal adenylyltransferase, rather than template-dependent RNA polymerase activity [59].

The functional role of NiRAN in the virus life cycle could be informed by the function of its homologs. However, we were unable to identify NiRAN homologs in a carefully controlled large-scale analysis that detected homologs for all other domains universally conserved in nidoviruses. This result also left the evolutionary origin of NiRAN uncertain. Nidovirus-wide conservation of NiRAN, as well as the apparent absence of NiRAN homologs in other viruses, indicated that it is a genetic marker of the order *Nidovirales*, only the second after the previously discovered ZBD (**chapter 3**). NiRAN is the only major replicative enzyme known to be exclusively associated with a large monophyletic group of (+)ssRNA viruses [60].

Four years after our study reporting the NiRAN domain discovery (**chapter 3**) was completed, the structure of the SARS-CoV nsp12, first of this protein, shed new light upon the NiRAN domain [61]. The N-terminus of SARS-CoV nsp12 up to NiRAN motif B_N is not visible in the structure and is likely to be highly flexible, while the structure of the NiRAN C-terminus, as well as the downstream Interface and RdRp domains were resolved. Comparison of the partial SARS-CoV NiRAN structure with a database of available protein structures identified significant although limited structural similarity between the NiRAN domain and the protein kinases (Z-score 7.9, RMSD 3.3 Å for the top hit with tyrosine kinase JAK1 structure 6C7Y, chain A). The strongest similarity was observed in the kinase nucleotide-binding site. The authors of the nsp12 structure noted that several of the kinase active site residues aligned with identical, highly conserved residues of NiRAN [61], while others did not have such counterparts in NiRAN either due to the respective residues variation or their structure not being solved. The authors concluded that the reported NiRAN nucleotidylation activity is compatible with the kinase-like fold of the domain, as observed for human pseudokinase SelO [61, 62].

In light of these new findings, we revisited our sequence-based analysis that led to the delineation of the NiRAN domain described in **chapter 3.** Inspection of the original hit list showed that the third best HHsearch hit of the CoToMeRoAr NiRAN profile (combined NiRAN sequences of all nidovirus subfamilies known at the time) in the Pfam database was with Pkinase profile (PF00069), representing protein kinases. Statistical support of the hit was far weaker than common thresholds of significance (Probability=21% vs 95%, E-value=160 vs 0.001), indicating that if NiRAN and protein kinases are indeed homologs, as the SARS-CoV nsp12 structural study suggests, they may have diverged beyond reliable recognition by most advanced sequence-based methods. Such pronounced divergence would be most compatible with conservation of the structural fold but emergence of a new function, often associated with replacement of otherwise key conserved residues.

Unlike the structural comparison, which was limited to the resolved structure of the NiRAN C-terminus (downstream of motif B_N) [61], the HHsearch hit covered all three major motifs of NiRAN (Fig. 3). This hit length, 109 match columns, was outstanding among the other (poorly supported) hits of this profile-database comparison, the mean length of which was 27 match columns. Compared to protein kinases, NiRAN domains of ExoN-encoding large nidoviruses included a large insertion that separated motifs B_N and C_N. Absolutely conserved Lys of NiRAN A_N motif, identified as the most likely target of nucleotidylation in **chapter 3**, aligned with signature Lys of protein kinase sub-domain II. This residue helps to anchor and orient nucleotide, used by kinase as a phosphate donor, by forming ionic bonds with its α- and β- phosphates. Absolutely conserved Glu of NiRAN A_N motif aligned with signature Glu of protein kinase sub-domain III, which helps to stabilize interactions between signature Lys of sub-domain II and nucleotide [63-65]. NiRAN motif B_N including invariant Arg residue and highly conserved Ser/Thr and Asp residues (**chapters 3, 4**) mapped to a protein kinase region of relatively low conservation, indicative of NiRAN-specific conservation and function.

N-terminal half of NiRAN C_N motif aligned with catalytic protein kinase sub-domain VIB, although its signature DxxxxN motif was not conserved in NiRAN: Asp of the signature, believed to catalyze the phosphotransfer reaction, mapped to a variable NiRAN residue, while Asn of the signature mapped to a highly conserved Asn of NiRAN, that is nevertheless substituted by Asp in several arteriviruses (Fig. 3, Fig. S1 from chapter 3) [63-65]. C-terminal half of NiRAN C_N motif aligned with protein kinase sub-domain VII. Conserved signature of sub-domain VII, tripeptide Asp-Phe-Gly, whose Asp residue is believed to chelate metal ion necessary to orient the γ-phosphate of the nucleotide for transfer, aligned to an absolutely conserved dipeptide of NiRAN, Asp-Phe, and a third residue that is a strictly conserved Gly in corona- and toroviruses, and a strictly conserved Glu in arteri-, mesoni- and roniviruses (Fig. 3, Fig. S1 from chapter 3) [63-65].

The similarity between NiRAN and protein kinases, observed in the HHsearch hit, doesn't extend to include upstream sub-domain I of protein kinases, characterized by a signature GxGxxG motif and responsible for covering and anchoring nontransferable phosphates of nucleotide cofactor, or downstream sub-domains VIII – XI, which include subdomain VIII, characterized by signature Ala-Pro-Glu tripeptide and responsible for recognition of peptide substrate [63-65]. These subdomains may either be absent or diverged beyond recognition in nidovirus nsp9/nsp12.

Thus, NiRAN and protein kinases may share elements of the kinase fold and some of its functionally important residues, which are associated with nucleotide binding, but not the Asp residue key for the phosphotransferase activity of kinases. Unless NiRAN adopted other residues to compensate for the lack of the catalytic Asp, the domain is unlikely to be a bona fide protein kinase. On the other hand, its nucleotidyltransferase activity seems to be compatible with the conservation of the nucleotide-binding site in the NiRAN and kinase domains.

Could identification of protein kinases as plausible NiRAN homologs prompt revision of the NiRAN assignment as a marker of the order *Nidovirales* (**chapter 3**)? Indeed, some may argue that protein kinases are encoded by other viruses: large DNA viruses [66], as well as some toroviruses (Fig. 2 from chapter 1) [29, 67]. On the other hand, we believe that unique properties of NiRAN, including its extremely divergent sequence, unparalleled association with RdRp structurally and functionally, and a place within the nidovirus domain synteny, clearly separate NiRAN from other homologs.

Synteny of key replicative domains remains the most conserved marker of nidoviruses

With the identification of the NiRAN domain (**chapter 3**), one hallmark of nidoviruses – universally conserved replicative domains encoded in a certain order (synteny) – expanded to include five domains: 3CLpro-NiRAN-RdRp-ZBD-HEL1. They remain one of the few characteristics that readily distinguish the ever-growing diversity of nidoviruses from other viruses.

While remaining under a strong purifying selection, these domains may have accepted rare or unique substitutions of key residues, revealing adaptation of the associated functions in most divergent nidoviruses. Namely, PSCNV, prototyping a nidovirus suborder [2], contains a number of remarkable substitutions in three domains of the synteny, 3CLpro, NiRAN and RdRp (**chapter 4**). A substrate pocket of PSCNV 3CLpro contains Val residue in place of His residue absolutely conserved in other nidoviruses; the substitution was predicted to confer an unusual substrate specificity to the enzyme (**chapter 4**) [68]. PSCNV NiRAN has a substitution in one out of the seven residues absolutely conserved in

other nidoviruses (**chapters 3, 4**). PSCNV RdRp has a Gly-Asp-Asp signature in its catalytic motif C, instead of Ser-Asp-Asp signature characteristic for nidoviruses (**chapter 4**) [28].

In agreement with their nidovirus-wide conservation, all domains of the synteny, when tested in experiments, proved to be essential for nidovirus replication (**chapter 3**) [69-72]. Nidovirus synteny of invariably encoded replicative domains is one of the very few conserved replicative domain architectures that accommodate the enormous diversity of (+)ssRNA viruses [60].

Genomes of nidoviruses can be far larger than previously believed

Following the sequencing of the entire 31.4 kb MHV genome in 1991 [73], the largest RNA virus genome known at the time, the apparent upper genome size limit increased only slightly over the years of nidovirus discovery, reaching a plateau (Fig. 1A from chapter 4). By the time this study started (end of 2014), the 33.5 kb genome of torovirus BPNV was the largest RNA virus genome known [29], and it seemed that the genome size of ~35 kb may represent the natural limit of the RNA virus genome size [7]. However, recent discoveries of two novel nidoviruses, PSCNV (**chapter 4**) and AAbV [19, 20], challenged this notion. While the AAbV genome size is just above 35 kb: 35.9 kb (7.4% larger than BPNV), PSCNV has a giant genome size by RNA virus standards: 41.1 kb (22.9% larger than BPNV, 14.5% larger than AAbV).

Several factors are believed to restrict the RNA virus genome size, including the fragility of RNA molecules, the selective advantage provided by the fast replication of small genomes, as well as error-prone replication of RNA viruses, that is believed to have a potential to cause error catastrophe in longer genomes [74]. Nidoviruses with large genomes uniquely possess a proofreading enzyme, exoribonuclease, that reduces the error rate of replication [27, 50, 75, 76]. It can be hypothesized that newly discovered nidoviruses with extremely large genomes have acquired properties that allowed them to decrease the error rate of replication even further, permitting them to maintain longer genomes, and increasing their capacity for adaptation. Interestingly, nidoviruses with the two longest known genomes, AAbV and PSCNV, infect exotic hosts, a flatworm and a mollusk, respectively (Fig. 1). It is tempting to suggest that unknown host factors may also play a role in the viability of these extraordinary long genomes.

The genome expansion of nidoviruses was described by a theoretical model [77]. According to the model, the nidovirus genome expansion was dominated by a consecutive expansion of genome regions responsible for genome replication (ORF1b), genome expression (ORF1a) and genome dissemination (3'ORFs region) in the course of the nidovirus evolution. The model, which considers extant nidoviruses as "frozen" at different points along the evolutionary trajectory, predicted that further expansion of the

nidovirus genome would be initiated by the expansion of ORF1b [77]. Notably, the ORF1b-like genome region of PSCNV is 9.8 kb, which is considerably larger than the largest ORF1b region of a previously known nidovirus – 8.2 kb ORF1b of human coronavirus OC43 [78]. This region size increase is especially remarkable because, unlike sizes of ORF1a and 3'ORFs, the size of nidovirus ORF1b is tightly constrained (Fig. 8A from chapter 4). When accounting for genome and region size variation, the increase of ORF1b-like region size in PSCNV was shown to be almost ten times greater than what would be expected if sizes of all genome regions in PSCNV increased uniformly (Fig. 8B from chapter 4). This increase of the ORF1b-like region size corroborates the nidovirus genome expansion model, and thus may indicate that nidoviruses with even larger genomes can be discovered in the future.

INNOVATION IN NIDOVIRUS GENOMES: DUPLICATION AND GENE ACQUISITION

Most large-scale evolutionary changes in nidoviruses can be attributed to aberrant homologous and non-homologous recombination, the mechanisms behind deletions, duplications and gene acquisitions [79, 80]. These evolutionary events are most frequently observed in the two regions of nidovirus genome controlling nidovirus-host interactions: pre-TM2 region of ORF1a and 3'ORFs. Several notable examples of deletions, duplications and gene acquisitions, mapping to these genome regions, were described in recent years, including some in our studies (**chapters 2, 4**).

May tandem repeats be common in ORF1a of nidoviruses?

One of the most common mechanisms of genome and protein innovation is the generation of tandem repeats. Possibly due to fast evolution, adjacent and highly similar tandem repeats were rarely observed in the genomes of RNA viruses, and reported only in a single nidovirus, betacoronavirus HCoV-HKU1, prior to 2014 [81, 82].

A recent study uncovered tandem repeats in nsp3 of a gammacoronavirus called duck-dominant coronavirus (DdCoV; classified as species *Duck coronavirus 2714* [DuCoV_2714] by ICTV), in a position similar to that of HCoV-HKU1 repeats. Four analyzed isolates of DdCoV all harbored five almost-identical copies of a 23 aa charged residue-rich repeat in nsp3 [83].

As shown in this thesis, the arterivirus ORF1a pre-TM2 region also contains repeats positioned in close proximity to each other: three copies of the PxPxPR motif, separated by ~10 aa, were identified within the HVR domain of EAV and WPDV (**chapter 2**). At least one copy of this motif was also found within the Hinge or HVR domain of almost all other arteriviruses (Fig. 3B from chapter 2). PxPxPR motifs may be recognized by cellular Src homology 3 (SH3) protein domains, implicated in signal transduction [84]. The same

function was previously suggested for the canonical SH3-binding motifs PxxP detected in the nsp2 sequence of PRRSV-1 [85]. Given the small size of PxPxPR motifs and their scattered position within the fast-evolving Hinge and HVR domains of arteriviruses, they might have emerged by either point mutation fixed by selection, or duplication followed by diversification.

Also, we described two types of tandem repeats in the newly discovered invertebrate nidovirus PSCNV (**chapter 4**). Two tandem repeats of 67 and 66 aa, separated by 3 aa and sharing 41.1% identity, were found in the pre-TM2 genome region. No homologs of these repeats, which could have pointed to their function, were identified. Further, PSCNV encodes an array of at least three tandem ankyrin repeats in the 3'ORFs-like region of its genome; their origin and function is discussed below.

Complex diversification of PLP paralogs in arteriviruses

Gene duplication is commonly followed by diversification of repeats, which could be advantageous for the virus [86]. Multiple PLP domains are present in the majority of known vertebrate nidoviruses and were one of the first recognized duplications in RNA viruses [87]. Their analysis in arteriviruses (**chapter 2**) offers an insight into the process of repeats diversification and its implications. Our study involved 14 arterivirus species recognized by ICTV as of 2016 [1], which included the most divergent arterivirus, WPDV, important for understanding limits of divergence within this family. We relied on previous research regarding the organization and function of the pp1a/1ab N-terminus in five arterivirus species (for review see [88-91]), as well as on resolved tertiary structures of PRRSV-2 PLP1a and PLP1b, and EAV PLP2 [92-94].

The number of PLP domains encoded in the pp1a/1ab N-terminus of 14 arterivirus species was found to vary from three to four (Fig. 3 from chapter 2), which were referred to as PLP1a, PLP1b, PLP1c and PLP2, based on the order of encoding.

In line with expectations, we found very limited sequence similarity between predicted PLP1 domains of arteriviruses and the WPDV pp1a that was restricted to the immediate vicinity of their catalytic residues and led to tentative identification of three PLP domains. The PLP1a domain of WPDV was predicted to be proteolytically inactive, like its EAV counterpart [95], as both enzymes lack catalytic Cys residue. Since WPDV and EAV do not form a monophyletic lineage, it remains uncertain whether this loss of the catalytic residue was independent or not in two viruses. Regardless of being enzymatically active or not, PLP1a of all arteriviruses retain catalytic His residue that is part of a unique motif (HxxxxxF). This unusual pattern of conservation, involving active and defective enzymes, suggests a distinct noncatalytic function for this residue.

In contrast to WPDV, PLP1 domains of non-WPDV arteriviruses exhibited a considerable degree of sequence and/or structural similarity. Their PLP1a domain is associated with the N-terminal zinc-finger domain, not found in WPDV. Sequence similarity between profiles of PLP1a and PLP1b or PLP1c was low, as was structural similarity between PLP1a and PLP1b of PRRSV-2. PLP1b and PLP1c were found to share a significant sequence similarity; surprisingly, EAV PLP1b exhibited a stronger similarity towards PLP1c, rather than PLP1b sequences. Notably, most residues conserved in PLP1a and PLP1b/c of non-WPDV arteriviruses mapped to the C- and N-terminal subdomains of the PLP structure, respectively (Fig. 11 from chapter 2), which is consistent with different functional specializations of these domains.

The PLP2 domain was shown to be universally conserved in all 14 arterivirus species, while sequence similarity between PLP2 and PLP1 domains was below commonly accepted thresholds of statistical significance for profile-profile analysis. Likewise, structural similarity between PRRSV PLP1a/1b and PLP2 was not statistically significant.

The established pattern of inter-domains similarity between different PLPs and arteriviruses allows to suggest an evolutionary scenario that might have led to the emergence of PLP arrays in arteriviruses. Barring a minor possibility that PLP1 domains were acquired by ancestors of the WPDV and non-WPDV arterivirus lineages independently, the MRCA of these viruses may have already encoded one, two or three PLP1 domains and a PLP2 domain. Presence of one ancestral PLP1 domain in the MRCA of known arteriviruses would imply that arrays of PLP1 domains were generated in WPDV and non-WPDV arteriviruses independently, and the weak similarity observed between the respective PLP1 domains of the two lineages is a result of parallel evolution (variant of convergence). An alternative scenario would be the presence of two or three PLP1 domains in the arterivirus MRCA, with ancestral PLP1a already bearing its distinguishing features, lack of catalytic Cys and a unique HxxxxxF motif involving catalytic His. Subsequent evolution of PLP1 domains in different lineages might have involved duplications followed by diversification, deletions, and convergence. Ancestors of non-WPDV PLP1a and PLP1b/c likely emerged as a result of a duplication that occurred prior to the existence of MRCA of non-WPDV arteriviruses. The MRCA of non-WPDV arteriviruses might have possessed a single ancestral PLP1b/c domain, which would imply a subsequent duplication of the domain in the non-EAV lineage. Alternatively, the MRCA of non-WPDV arteriviruses might have possessed ancestral PLP1b and PLP1c, with ortholog of PLP1b being lost in the EAV lineage.

Thus, an array of orthologous PLP domains of arteriviruses emerged as a result of a broad range of evolutionary events, where multiple PLP duplications were followed by

diversification, loss and possibly convergence of orthologous domains, involved loss of catalytic activity and different functional specialization of orthologs.

Newly described nidovirus includes repeats of uncertain origins

While the evolutionary origin of the tandem repeats described above can be tentatively ascribed to duplication, the PSCNV genome encodes at least two pairs of repeats separated by considerable distances (**chapter 4**), whose evolutionary origin remains uncertain.

The potential leader and body TRSs of PSCNV (see above) are ~60 nt long, share 86% sequence identity, and are separated by 28,327 nt. They might have emerged as a result of duplication, but incremental extension of the similar regions by point mutations fixed by selection (convergence), associated with genome expansion, seems to be more likely. A shorter ancestral genome might have already had TRSs at the respective positions in the genome, as is typical for nidoviruses. Expansion of the genome would have necessitated TRSs extension: short motifs identical to TRSs can be encountered in a long genome just by chance, jeopardizing genome expression. Consequently, gradual expansion of the genome could have created evolutionary pressure, causing identical sequences of TRSs to gradually extend along with the genome through convergence mechanism. Expansion of virus sampling in the PSCNV clade could help in resolving evolutionary history of this intriguing similarity involving exceptionally long TRS-like elements.

In addition, PSCNV encodes two highly divergent fibronectin type 2 (FN2) domains separated by 1,572 aa in the polyprotein. Besides duplication in ancestral virus, FN2 domains might have been acquired independently from unknown sources, since they are only remotely similar and located far apart. The possible function of the two FN2 domains is discussed below.

Newly described nidovirus acquired domains rarely or never observed in viruses

Another major source of genome and protein innovation is domain acquisition from other species. Many domains found in subsets of nidoviruses may have been acquired through this mechanism (see Introduction). Our study of PSCNV was particularly insightful in this respect since it expanded the previously known proteome repertoire of nidoviruses or even larger groups of viruses (**chapter 4**). Besides encoding orthologs of canonical replicative domains of nidoviruses, PSCNV was found to also encode such domains as ribonuclease T2 (RNase T2), two fibronectin type 2 domains, and an ankyrin repeats domain (ANK).

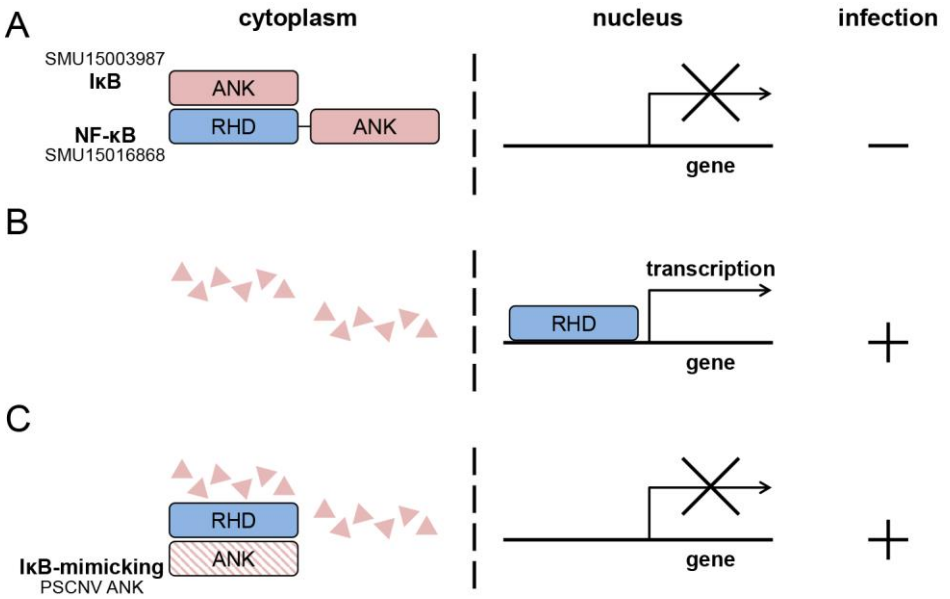

Figure 4 | Proposed roles of PSCNV ANK and its host homologs in modulation of antiviral immune response. (**A**) In the absence of an inducing signal, NF-κB protein (SMU15016868) resides in the cytoplasm, bound by inhibitors: its own ANK domain and protein IκB (SMU15003987). (**B**) In response to viral infection, inhibitors are degraded, allowing the NF-κB transcription factor to enter the nucleus and modulate gene expression to promote antiviral immune response. (**C**) IκB-mimicking viral protein (PSCNV ANK) may retain the NF-κB transcription factor in the cytoplasm after its inhibitors were degraded, thus downregulating the immune response.

The RNase T2 homolog is encoded in the pre-TM2 region of the giant PSCNV ORF. RNases T2 cleave ssRNA in an acidic environment, and are encoded by a broad range of cellular organisms, as well as two groups of viruses, (+)ssRNA pestiviruses and dsDNA polydnaviruses [96]. Viral RNases T2 were implicated in modulating host immune response [97, 98], and we proposed a similar function for PSCNV RNase T2.

Also, we identified two FN2 domains in the PSCNV polyprotein, which are encoded far downstream from RNase T2 in a 3'ORFs-like region. Never before found in viruses, FN2 domains are common in vertebrates and invertebrates as modules of multidomain proteins involved in diverse processes [99, 100]; when studied they were found responsible for protein-protein interactions: binding of gelatin and collagen [101, 102]. We speculated that FN2 domains of PSCNV could also bind collagen. Since PSCNV was found to infect mucus-producing cells, PSCNV FN2 domains might help to adhere PSCNV virus particles to the collagen-containing mucus excreted by the host [103], facilitating spread of the virus among host population.

The two FN2 domains flank the PSCNV polyprotein region that includes the ANK domain, another mediator of protein-protein interactions. The ANK domain is ubiquitous in proteins of diverse cellular organisms, and dsDNA viruses with large genomes, but was not detected in proteins of RNA viruses before [104].

The PSCNV RNase T2, FN2 domains and the ANK domain were likely acquired from other viruses or hosts via non-homologous recombination. Due to the lack of close homologs, evolutionary origins of RNase T2 and FN2 domains in PSCNV remain unknown. In contrast, the PSCNV ANK domain clusters confidently with ANK domains of a pair of host proteins, SMU15016868 and SMU15003987, indicating that the ANK domain might have been acquired from an ancestor of the host, flatworm *Schmidtea mediterranea*.

Further analysis of the domain architecture of these host homologs led us to a hypothesis about the possible functional role of PSCNV ANK (**chapter 4**) in a striking parallel with a process documented for several dsDNA viruses [105-107]. Namely, the host ANK-containing proteins have domain architectures suggestive of their interaction: SMU15016868 is characteristic for NF-κB protein, N-RHD-ANK-C (RHD is a Rel homology domain), while SMU15003987 is characteristic for its inhibitor IκB, N-ANK-C [108]. Based on studies of several viruses [105, 106], the NF-κB protein is expected to reside in the cytoplasm, bound by inhibitors, its own ANK domain and protein IκB, in the absence of a viral infection (Fig. 4A). A viral infection would trigger degradation of NF-κB inhibitors, allowing NF-κB transcription factor to enter the nucleus and modulate gene expression to promote an antiviral immune response (Fig. 4B). Thus, we proposed PSCNV ANK to act as a IκB-mimicking protein, retaining a NF-κB transcription factor in the cytoplasm after the degradation of its inhibitors, and thus downregulating the immune response (Fig. 4C).

FUTURE PERSPECTIVES

Comparative genomics reveals patterns of natural variation, forming the basis for evolutionary hypotheses that inform experimental research. This thesis contains multiple examples of the connection between comparative genomics and bench. Experimental characterization of the novel, universally conserved domain of nidoviruses, NiRAN, was inspired by a hypothesis about its RNA ligase activity, formulated based solely on comparative genomics data (**chapter 3**), while hypotheses and models suggested in **chapters 2 and 4** create a basis for future experimental research. In **chapter 2**, methods of comparative genomics allowed to make several predictions about the N-termini of arteriviral polyproteins: the position of enzymatically active PLP domains of fourteen arteriviruses, few of which were previously characterized in this respect, as well as potential nsp2 PRF and SH3-binding sites were identified. In **chapter 4**, an extremely divergent and unusual nidovirus, PSCNV, was extensively analyzed by methods of comparative genomics, leading to hypotheses about various aspects of its biology, including the nature of replicative domains and structural proteins that it encodes,

mechanisms of differential protein expression that it employs, and ways of evading host immune defenses that it uses. Besides, existence of a related virus infecting a flatworm *Planaria torva* was also predicted in **chapter 4**. All these predictions could benefit from experimental verification.

The explosive growth of the number and diversity of newly discovered nidoviruses, which has been observed in recent years, is likely to continue and accelerate even more in the future. These developments will bring new insights about the biology and evolution of nidoviruses, but may also present a challenge. The unprecedented influx of new divergent nidovirus genome sequences calls for the development of tools allowing to reliably classify them [9, 10], and to detect homology despite their enormous genetic divergence (**chapter 5**). Importantly, nidovirus genomes that will be discovered in the future can also be instrumental in verifying and advancing hypotheses formulated in this thesis. For example, the discovery of a sister virus for WPDV (**chapter 2**) or PSCNV (**chapter 4**), which are separated from the currently known nidoviruses by long genetic distances, would make it possible to test hypotheses about these viruses by analyzing the conservation of their predicted functional genome and proteome elements.

REFERENCES

1. Adams MJ, Lefkowitz EJ, King AM, Harrach B, Harrison RL, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Mushegian AR *et al*: **Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2016)**. *Arch Virol* 2016, **161**:2921-2949.
2. Siddell SG, Walker PJ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, Gorbalenya AE, Harrach B, Harrison RL, Junglen S *et al*: **Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018)**. *Arch Virol* 2019.
3. Gorbalenya AE, Brinton MA, de Groot RJ, Gulyaeva AA, Lauber C, Neuman BW, Ziebuhr J: **Pending ICTV taxonomic proposal 2019.023S Create five new families and a new suborder of vertebrate viruses in the order Nidovirales.** 2019.
4. Brinton MA, Gulyaeva AA, Balasuriya UBR, Dunowska M, Faaberg KS, Goldberg T, Leung F-C, Nauwynck HJ, Snijder EJ, Stadejek T *et al*: **Pending ICTV taxonomic proposal 2019.020S Create one new genus (Nuarterivirus); move the existing subgenus Pedartevirus to the genus Iotaarterivirus; rename one species from the subgenus Pedartevirus; create one new species in the new genus Nuarterivirus; create one new subgenus and two new species in the existing genus Betaarterivirus.** 2019.
5. Ziebuhr J, Baker S, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Neuman BW, Perlman S, Poon LLM *et al*: **Pending ICTV taxonomic proposal 2019.021S Create ten new species and a new genus in the subfamily Orthocoronavirinae of the family Coronaviridae and five new species and a new genus in the subfamily Serpentovirinae of the family Tobaniviridae.** 2019.
6. Gorbalenya AE, Gulyaeva AA, Hobson-Peters J, Junglen S, Morita K, Sawabe K, Vasilakis N, Ziebuhr J: **Pending ICTV taxonomic proposal 2019.022S Create one new species in the genus Alphamesonivirus of the family Mesoniviridae and one new species in the genus Okavirus of the family Roniviridae.** 2019.
7. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS *et al*: **Redefining the invertebrate RNA virosphere**. *Nature* 2016, **540**:539-543.
8. Adams MJ, Lefkowitz EJ, King AM, Harrach B, Harrison RL, Knowles NJ, Kropinski AM, Krupovic M, Kuhn JH, Mushegian AR *et al*: **50 years of the International Committee on Taxonomy of Viruses: progress and prospects**. *Arch Virol* 2017, **162**(5):1441-1446.
9. Lauber C, Gorbalenya AE: **Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses**. *J Virol* 2012, **86**(7):3890-3904.
10. Lauber C, Gorbalenya AE: **Toward genetics-based virus taxonomy: comparative analysis of a genetics-based classification and the taxonomy of picornaviruses**. *J Virol* 2012, **86**(7):3905-3915.

11. Brinton MA, Gulyaeva AA, Balasuriya UBR, Dunowska M, Faaberg KS, Leung FC, Nauwynck HJ, Snijder EJ, Stadejek T, Gorbalenya AE: **ICTV taxonomic proposal 2015.014a-cS In the family Arteriviridae create 10 species (1 unassigned, 9 in the genus Arterivirus) and rename one species**. 2015.

12. Ziebuhr J, Baric RS, Baker S, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Neuman BW, Perlman S *et al*: **ICTV taxonomic proposal 2015.003a-eS Create 12 species in the family Coronaviridae**. 2015.

13. Gorbalenya AE, Gulyaeva AA, Hobson-Peters J, Junglen S, Morita K, Sawabe K, Vasilakis N, Ziebuhr J: **ICTV taxonomic proposal 2015.004a,bS In the family Mesoniviridae, create four species in genus Alphamesonivirus and two unassigned in the family**. 2015.

14. Gorbalenya AE, Brinton MA, Cowley J, de Groot R, Gulyaeva A, Lauber C, Neuman B, Ziebuhr J: **ICTV taxonomic proposal 2017.015S Reorganization and expansion of the order Nidovirales at the family and sub-order ranks**. 2017.

15. Brinton MA, Gulyaeva A, Balasuriya UBR, Dunowska M, Faaberg KS, Goldberg T, Leung FC-C, Nauwynck HJ, Snijder EJ, Stadejek T *et al*: **ICTV taxonomic proposal 2017.012S Expansion of the rank structure of the family Arteriviridae and renaming its taxa**. 2017.

16. Ziebuhr J, Baric RS, Baker S, de Groot RJ, Drosten C, Gulyaeva A, Haagmans BL, Neuman BW, Perlman S, Poon LLM *et al*: **ICTV taxonomic proposal 2017.013S Reorganization of the family Coronaviridae into two families, Coronaviridae (including the current subfamily Coronavirinae and the new subfamily Letovirinae) and the new family Tobaniviridae (accommodating the current subfamily Torovirinae and three other subfamilies), revision of the genus rank structure and introduction of a new subgenus rank**. 2017.

17. Gorbalenya AE, Brinton MA, Cowley J, de Groot R, Gulyaeva A, Lauber C, Neuman B, Ziebuhr J: **ICTV taxonomic proposal 2017.014S Establishing taxa at the ranks of subfamily, genus, sub-genus and species in six families of invertebrate nidoviruses**. 2017.

18. Dunowska M, Biggs PJ, Zheng T, Perrott MR: **Identification of a novel nidovirus associated with a neurological disease of the Australian brushtail possum (Trichosurus vulpecula)**. *Vet Microbiol* 2012, **156**(3-4):418-424.

19. Bukhari K, Mulley G, Gulyaeva AA, Zhao L, Shu G, Jiang J, Neuman BW: **Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family Abyssoviridae, and from a sister group to the Coronavirinae, the proposed genus Alphaletovirus**. *Virology* 2018, **524**:160-171.

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

21. Zondag LE, Rutherford K, Gemmell NJ, Wilson MJ: **Uncovering the pathways underlying whole body regeneration in a chordate model, Botrylloides leachi using de novo transcriptome analysis**. *BMC Genomics* 2016, **17**:114.

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

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

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

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

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

27. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE: **Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage**. *J Mol Biol* 2003, **331**(5):991-1004.

28. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ: **Nidovirales: evolving the largest RNA virus genome**. *Virus Res* 2006, **117**(1):17-37.

29. Stenglein MD, Jacobson ER, Wozniak EJ, Wellehan JF, Kincaid A, Gordon M, Porter BF, Baumgartner W, Stahl S, Kelley K *et al*: **Ball python nidovirus: a candidate etiologic agent for severe respiratory disease in Python regius**. *MBio* 2014, **5**(5):e01484-01414.

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

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

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

33. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L *et al*: **The evolutionary history of vertebrate RNA viruses**. *Nature* 2018, **556**:197-202.

34. Firth AE, Brierley I: **Non-canonical translation in RNA viruses**. *J Gen Virol* 2012, **93**(Pt 7):1385-1409.

35. Snijder EJ, Meulenberg JJ: **The molecular biology of arteriviruses**. *J Gen Virol* 1998, **79 (Pt 5)**:961-979.

36. Molenkamp R, van Tol H, Rozier BC, van der Meer Y, Spaan WJ, Snijder EJ: **The arterivirus replicase is the only viral protein required for genome replication and subgenomic mRNA transcription**. *J Gen Virol* 2000, **81**(Pt 10):2491-2496.

37. Archambault D, Kheyar A, de Vries AA, Rottier PJ: **The intraleader AUG nucleotide sequence context is important for equine arteritis virus replication**. *Virus Genes* 2006, **33**(1):59-68.

38. Brian DA, Baric RS: **Coronavirus genome structure and replication**. *Curr Top Microbiol Immunol* 2005, **287**:1-30.

39. Wu HY, Guan BJ, Su YP, Fan YH, Brian DA: **Reselection of a genomic upstream open reading frame in mouse hepatitis coronavirus 5'-untranslated-region mutants**. *J Virol* 2014, **88**(2):846-858.

40. Irigoyen N, Firth AE, Jones JD, Chung BY, Siddell SG, Brierley I: **High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling**. *PLoS Pathog* 2016, **12**(2):e1005473.

41. Stewart H, Brown K, Dinan AM, Irigoyen N, Snijder EJ, Firth AE: **Transcriptional and Translational Landscape of Equine Torovirus**. *J Virol* 2018, **92**(17).

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

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

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

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

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

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

48. Cowley JA, Dimmock CM, Spann KM, Walker PJ: **Gill-associated virus of Penaeus monodon prawns: an invertebrate virus with ORF1a and ORF1b genes related to arteri- and coronaviruses**. *J Gen Virol* 2000, **81**(Pt 6):1473-1484.

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

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

51. Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, Janssen GM, Ruben M, Overkleeft HS, van Veelen PA, Samborskiy DV, Kravchenko AA, Leontovich AM *et al*: **Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses**. *Nucleic Acids Res* 2015, **43**(17):8416-8434.

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

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

54. Wolf YI, Kazlauskas D, Iranzo J, Lucia-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV: **Origins and Evolution of the Global RNA Virome**. *MBio* 2018, **9**(6).

55. Lai MM, Patton CD, Stohlman SA: **Further characterization of mRNA's of mouse hepatitis virus: presence of common 5'-end nucleotides**. *J Virol* 1982, **41**(2):557-565.

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

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

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

59. Tvarogova J, Madhugiri R, Bylapudi G, Ferguson LJ, Karl N, Ziebuhr J: **Identification and Characterization of a Human Coronavirus 229E Nonstructural Protein 8-Associated RNA 3'-Terminal Adenylyltransferase Activity**. *J Virol* 2019, **93**(12).

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

61. Kirchdoerfer RN, Ward AB: **Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors**. *Nat Commun* 2019, **10**(1):2342.

62. Sreelatha A, Yee SS, Lopez VA, Park BC, Kinch LN, Pilch S, Servage KA, Zhang J, Jiou J, Karasiewicz-Urbanska M *et al*: **Protein AMPylation by an Evolutionarily Conserved Pseudokinase**. *Cell* 2018, **175**(3):809-821 e819.

63. Hanks SK, Hunter T: **Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification**. *FASEB J* 1995, **9**(8):576-596.

64. Hanks SK: **Genomic analysis of the eukaryotic protein kinase superfamily: a perspective**. *Genome Biol* 2003, **4**(5):111.

65. Zhao Z, Jin Q, Xu JR, Liu H: **Identification of a fungi-specific lineage of protein kinases closely related to tyrosine kinases**. *PLoS One* 2014, **9**(2):e89813.

66. Jacob T, Van den Broeke C, Favoreel HW: **Viral serine/threonine protein kinases**. *J Virol* 2011, **85**(3):1158-1173.

67. Tokarz R, Sameroff S, Hesse RA, Hause BM, Desai A, Jain K, Lipkin WI: **Discovery of a novel nidovirus in cattle with respiratory disease**. *J Gen Virol* 2015, **96**(8):2188-2193.

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

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

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

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

72. van Dinten LC, Rensen S, Gorbalenya AE, Snijder EJ: **Proteolytic processing of the open reading frame 1b-encoded part of arterivirus replicase is mediated by nsp4 serine protease and Is essential for virus replication**. *J Virol* 1999, **73**(3):2027-2037.

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

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

75. Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR: **High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants**. *J Virol* 2007, **81**(22):12135-12144.

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

77. Lauber C, Goeman JJ, Parquet MC, Nga PT, Snijder EJ, Morita K, Gorbalenya AE: **The footprint of genome architecture in the largest genome expansion in RNA viruses**. *PLoS Pathog* 2013, **9**(7):e1003500.

78. St-Jean JR, Jacomy H, Desforges M, Vabret A, Freymuth F, Talbot PJ: **Human respiratory coronavirus OC43: genetic stability and neuroinvasion**. *J Virol* 2004, **78**(16):8824-8834.

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

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

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

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

83. Zhuang QY, Wang KC, Liu S, Hou GY, Jiang WM, Wang SC, Li JP, Yu JM, Chen JM: **Genomic Analysis and Surveillance of the Coronavirus Dominant in Ducks in China**. *PLoS One* 2015, **10**(6):e0129256.

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

85. Ropp SL, Wees CE, Fang Y, Nelson EA, Rossow KD, Bien M, Arndt B, Preszler S, Steen P, Christopher-Hennings J *et al*: **Characterization of emerging European-**

like porcine reproductive and respiratory syndrome virus isolates in the United States. *J Virol* 2004, **78**(7):3684-3703.

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

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

88. Snijder EJ, Kikkert M, Fang Y: **Arterivirus molecular biology and pathogenesis**. *J Gen Virol* 2013, **94**(Pt 10):2141-2163.

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

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

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

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

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

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

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

96. Luhtala N, Parker R: **T2 Family ribonucleases: ancient enzymes with diverse roles**. *Trends Biochem Sci* 2010, **35**(5):253-259.

97. Krey T, Bontems F, Vonrhein C, Vaney MC, Bricogne G, Rumenapf T, Rey FA: **Crystal structure of the pestivirus envelope glycoprotein E(rns) and mechanistic analysis of its ribonuclease activity**. *Structure* 2012, **20**(5):862-873.

98. Park B KY: **Immunosuppression induced by expression of a viral RNase enhances susceptibility of Plutella xylostella to microbial pesticides.** *Insect Science* 2012, **19**(1):47-54.

99. Ozhogina OA, Trexler M, Banyai L, Llinas M, Patthy L: **Origin of fibronectin type II (FN2) modules: structural analyses of distantly-related members of the kringle family idey the kringle domain of neurotrypsin as a potential link between FN2 domains and kringles**. *Protein Sci* 2001, **10**(10):2114-2122.

100. Chalmers IW, Hoffmann KF: **Platyhelminth Venom Allergen-Like (VAL) proteins: revealing structural diversity, class-specific features and biological associations across the phylum**. *Parasitology* 2012, **139**(10):1231-1245.

101. Napper CE, Drickamer K, Taylor ME: **Collagen binding by the mannose receptor mediated through the fibronectin type II domain**. *Biochem J* 2006, **395**(3):579-586.

102. Tam EM, Moore TR, Butler GS, Overall CM: **Characterization of the distinct collagen binding, helicase and cleavage mechanisms of matrix metalloproteinase 2 and 14 (gelatinase A and MT1-MMP): the differential roles of the MMP hemopexin c domains and the MMP-2 fibronectin type II modules in collagen triple helicase activities**. *J Biol Chem* 2004, **279**(41):43336-43344.

103. Bocchinfuso DG, Taylor P, Ross E, Ignatchenko A, Ignatchenko V, Kislinger T, Pearson BJ, Moran MF: **Proteomic profiling of the planarian Schmidtea mediterranea and its mucous reveals similarities with human secretions and those predicted for parasitic flatworms**. *Mol Cell Proteomics* 2012, **11**(9):681-691.

104. Al-Khodor S, Price CT, Kalia A, Abu KY: **Functional diversity of ankyrin repeats in microbial proteins**. *Trends Microbiol* 2010, **18**(3):132-139.

105. Falabella P, Varricchio P, Provost B, Espagne E, Ferrarese R, Grimaldi A, de EM, Fimiani G, Ursini MV, Malva C *et al*: **Characterization of the IkappaB-like gene family in polydnaviruses associated with wasps belonging to different Braconid subfamilies**. *J Gen Virol* 2007, **88**(Pt 1):92-104.

106. Tait SW, Reid EB, Greaves DR, Wileman TE, Powell PP: **Mechanism of inactivation of NF-kappa B by a viral homologue of I kappa b alpha. Signal-induced release of i kappa b alpha results in binding of the viral homologue to NF-kappa B**. *J Biol Chem* 2000, **275**(44):34656-34664.

107. Camus-Bouclainville C, Fiette L, Bouchiha S, Pignolet B, Counor D, Filipe C, Gelfi J, Messud-Petit F: **A virulence factor of myxoma virus colocalizes with NF-kappaB in the nucleus and interferes with inflammation**. *J Virol* 2004, **78**(5):2510-2516.

108. Gilmore TD, Wolenski FS: **NF-kappaB: where did it come from and why?** *Immunol Rev* 2012, **246**(1):14-35.