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

Gulyaeva, A.

CHAPTER 1

General Introduction

PREFACE

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

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

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

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

NIDOVIRUS DIVERSITY AND TAXONOMY

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

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

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

NIDOVIRUS GENOME ORGANIZATION

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

NIDOVIRUS LIFE CYCLE

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

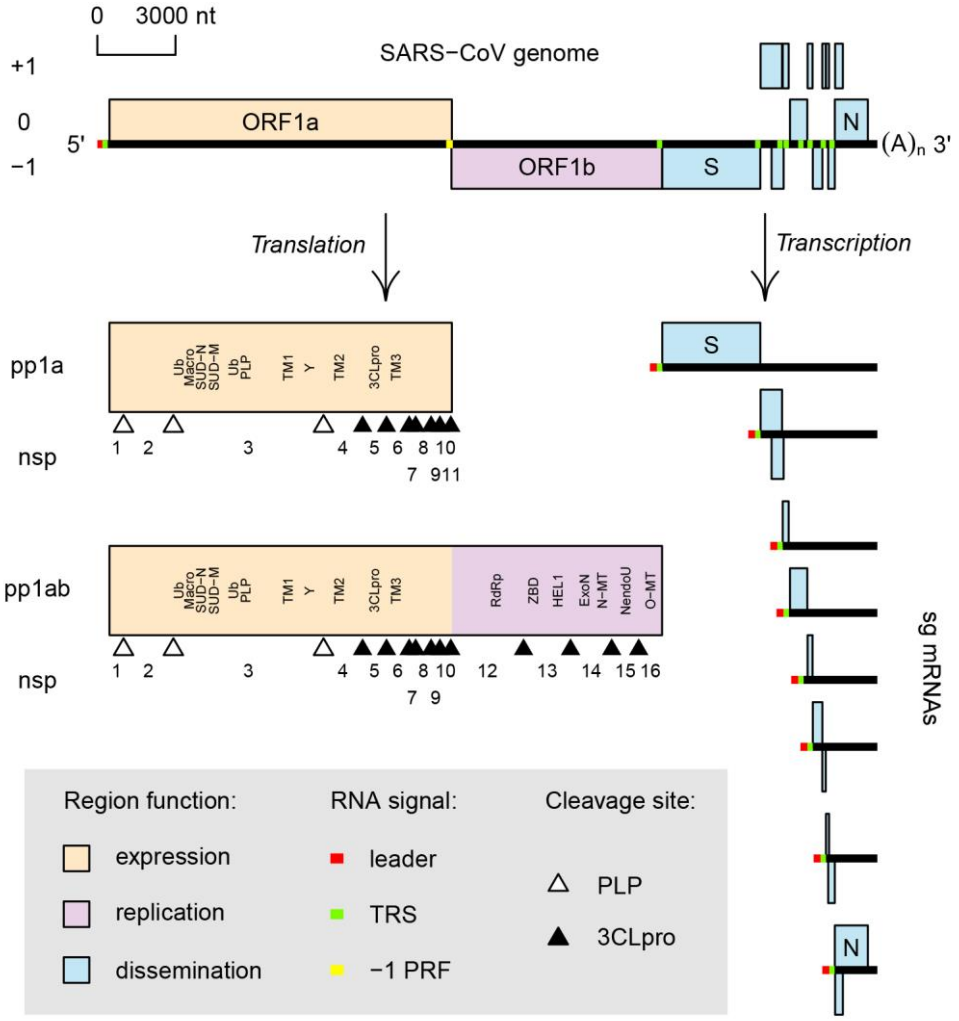

Figure 1 | SARS-CoV genome organization and expression. Genome (top), products of genome translation (left) and transcription (right) are shown. ORFs and polyprotein regions are colored according to their predominant function (see inset). Genome ORFs are depicted in their frame, with ORF1a frame set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

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

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

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

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

NIDOVIRUS PROTEOME

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

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

Pre-TM2 region of ORF1a

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

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

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

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

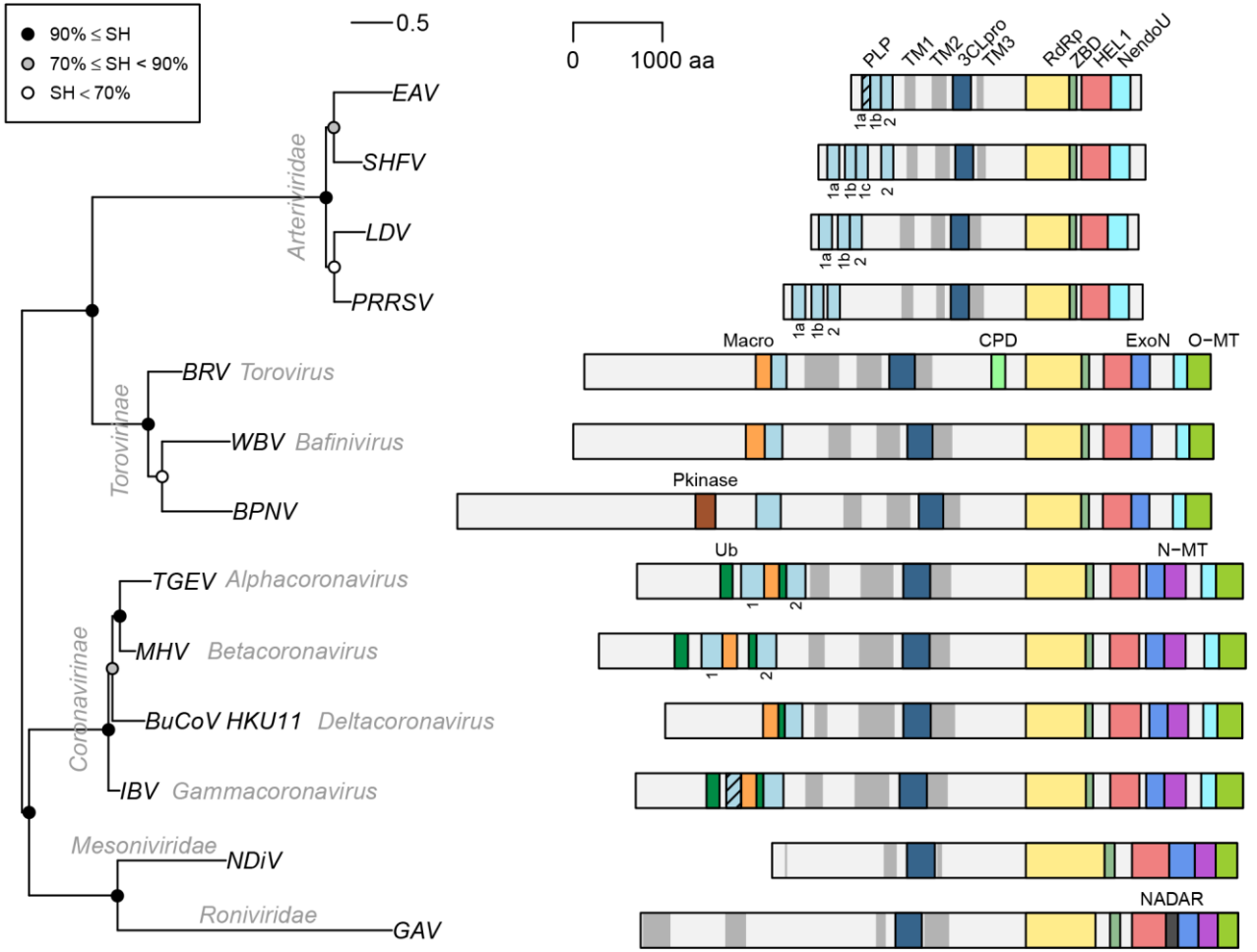
90% ≤ SH
70% ≤ SH < 90%
SH < 70%
0.5
EAV
SHFV
LDV
PRRSV
Arteriviridae
BRV Torovirus
WBV Bafinivirus
BPNV
Torovirinae
TGEV Alphacoronavirus
MHV Betacoronavirus
BuCoV HKU11 Deltacoronavirus
IBV Gammacoronavirus
Coronavirinae
Mesoniviridae
NDiV
Roniviridae
GAV
0
1000 aa
PLP
TM1
TM2
3CLpro
TM3
RdRp
ZBD
HEL1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

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

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

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

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

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

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

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

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

TM2-3CLpro-TM3 region of ORF1a

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

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

Post-TM3 region of ORF1a

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

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

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

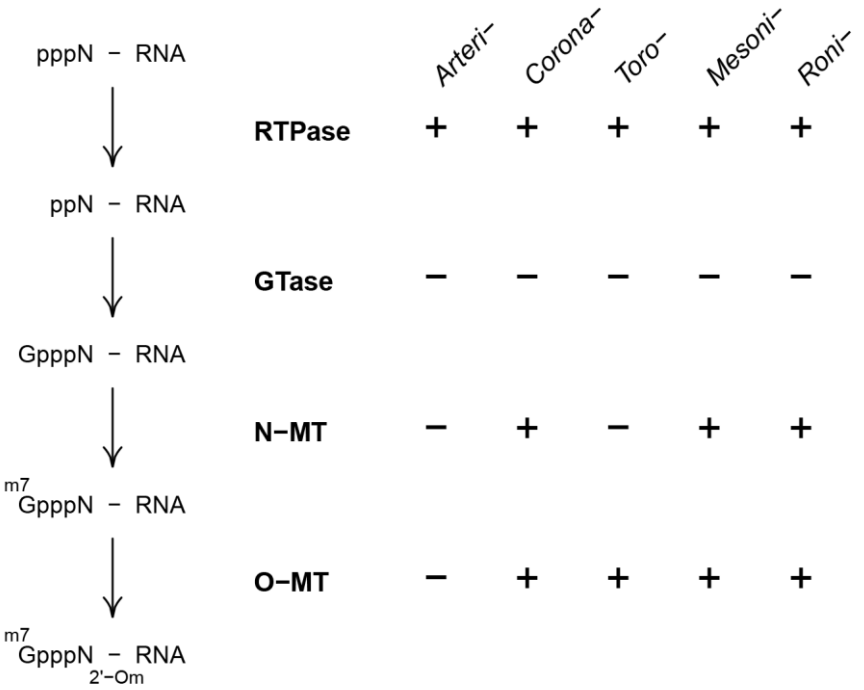

Figure 3 | Capping pathway and enzymes in relation to the proteome of nidoviruses. The conventional mRNA capping pathway is shown on the left, with the enzymes catalyzing the respective four reactions listed in bold. Further to the right, presence of these enzymes in viruses of five nidovirus (sub)families,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

ORF1b region

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

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

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

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

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

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

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

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

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

3'ORFs region

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

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

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

NIDOVIRUS MACROEVOLUTION

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

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

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

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

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

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

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

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

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

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

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

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

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

TOOLS OF NIDOVIRUS COMPARATIVE GENOMICS

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

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

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

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

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

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

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

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

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

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

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

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

SCIENTIFIC QUESTIONS

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

REFERENCES

1. Geoghegan JL, Holmes EC: **Evolutionary Virology at 40**. *Genetics* 2018, **210**(4):1151-1162.
2. Baltimore D: **Expression of animal virus genomes**. *Bacteriol Rev* 1971, **35**(3):235-241.
3. Beijerinck MW: **Concerning a contagium vivum fluidum as cause of the spot disease of tobacco leaves.** *Verh Kon Akad Wetensch* 1898, **6**:3-21.
4. Zhang YZ, Shi M, Holmes EC: **Using Metagenomics to Characterize an Expanding Virosphere**. *Cell* 2018, **172**(6):1168-1172.
5. Fehr AR, Perlman S: **Coronaviruses: an overview of their replication and pathogenesis**. *Methods Mol Biol* 2015, **1282**:1-23.
6. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL *et al*: **A pneumonia outbreak associated with a new coronavirus of probable bat origin**. *Nature* 2020, **579**(7798):270-273.
7. Bailey AL, Lauck M, Sibley SD, Friedrich TC, Kuhn JH, Freimer NB, Jasinska AJ, Phillips-Conroy JE, Jolly CJ, Marx PA *et al*: **Zoonotic Potential of Simian Arteriviruses**. *J Virol* 2015, **90**(2):630-635.
8. Graham RL, Donaldson EF, Baric RS: **A decade after SARS: strategies for controlling emerging coronaviruses**. *Nat Rev Microbiol* 2013, **11**(12):836-848.
9. Joyce GF: **The antiquity of RNA-based evolution**. *Nature* 2002, **418**(6894):214-221.
10. Snijder EJ, Kikkert M, Fang Y: **Arterivirus molecular biology and pathogenesis**. *J Gen Virol* 2013, **94**(Pt 10):2141-2163.
11. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM: **Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis**. *Nucleic Acids Res* 1989, **17**(12):4847-4861.
12. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE: **Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage**. *J Mol Biol* 2003, **331**(5):991-1004.
13. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ: **Nidovirales: evolving the largest RNA virus genome**. *Virus Res* 2006, **117**(1):17-37.
14. Stenglein MD, Jacobson ER, Wozniak EJ, Wellehan JF, Kincaid A, Gordon M, Porter BF, Baumgartner W, Stahl S, Kelley K *et al*: **Ball python nidovirus: a candidate etiologic agent for severe respiratory disease in Python regius**. *MBio* 2014, **5**(5):e01484-01414.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

39. Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, Ziebuhr J: **Discovery of an RNA virus 3'->5' exoribonuclease that is critically involved in coronavirus RNA synthesis**. *Proc Natl Acad Sci U S A* 2006, **103**(13):5108-5113.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

99. Chen Y, Su C, Ke M, Jin X, Xu L, Zhang Z, Wu A, Sun Y, Yang Z, Tien P *et al*: **Biochemical and structural insights into the mechanisms of SARS coronavirus RNA ribose 2'-O-methylation by nsp16/nsp10 protein complex**. *PLoS Pathog* 2011, **7**(10):e1002294.

100. Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, Decroly E: **RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex**. *Proc Natl Acad Sci U S A* 2012, **109**(24):9372-9377.

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

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

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

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

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

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

107. Deng Z, Lehmann KC, Li X, Feng C, Wang G, Zhang Q, Qi X, Yu L, Zhang X, Feng W *et al*: **Structural basis for the regulatory function of a complex zinc-binding domain in a replicative arterivirus helicase resembling a nonsense-mediated mRNA decay helicase**. *Nucleic Acids Res* 2014, **42**(5):3464-3477.

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

109. Ivanov KA, Ziebuhr J: **Human coronavirus 229E nonstructural protein 13: characterization of duplex-unwinding, nucleoside triphosphatase, and RNA 5'-triphosphatase activities**. *J Virol* 2004, **78**(14):7833-7838.

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

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

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

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

114. Decroly E, Imbert I, Coutard B, Bouvet M, Selisko B, Alvarez K, Gorbalenya AE, Snijder EJ, Canard B: **Coronavirus nonstructural protein 16 is a cap-0 binding enzyme possessing (nucleoside-2'O)-methyltransferase activity**. *J Virol* 2008, **82**(16):8071-8084.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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